UNIVERSITÉ DE STRASBOURG

UNIVERSITY OF STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA TERRE, DE L'UNIVERS ET DE L'ENVIRONNEMENT

PhD Thesis presented by

ANDRÉ STUMPF

Prepared jointly at:
Laboratoire Image, Ville, Environnement – CNRS / University of Strasbourg
Institut de Physique du Globe de Strasbourg – CNRS / University of Strasbourg
Faculty of Geo-information Science and Earth Observation – ITC / University of Twente

For obtaining the degree

Doctor of the University of Strasbourg (PhD)

Landslide recognition and monitoring with remotely sensed data from passive optical sensors

Public defence on December 18th, 2013

Examination committee

Prof. Rémi Michel	Professor, University Pierre et Marie Curie, Paris	Reviewer
Prof. Sébastien Lefèvre	Professor, University South Brittany, Vannes	Reviewer
Prof. Jean-Michel Dischler	Professor, University of Strasbourg, Strasbourg	Reviewer
Prof. Michel Jaboyedoff	Professor, University of Lausanne, Lausanne	Oral examinator
Dr. Jean-Philippe Malet	Senior Researcher, CNRS, University of Strasbourg, Strasbourg	Oral examinator
Dr. Christiane Weber	Director of Research, CNRS, University of Strasbourg, Strasbourg	Promotor
Dr. Norman Kerle	Associate Professor, University of Twente, Enschede	Invited
Dr. Anne Puissant	Associate Professor, University of Strasbourg, Strasbourg	Invited

Image Ville Environnement

Ecole et Observatoire des Sciences de la Terre

ITC

Summary: Landslide inventory mapping and monitoring are indispensable for hazard assessment and disaster management. The enhanced availability of VHR satellites, UAVs and consumer grade digital cameras offers a great potential to support those tasks at regional and local scales, and to complement established techniques such as in situ instrumentation, radar, and laser scanning. A lack of image processing tools for the efficient extraction process-relevant information from different types of optical imagery still complicates the exploitation of optical data and hinders the implementation of operational services. This doctoral thesis is dedicated to the development and application of image processing techniques for the mapping, characterization and monitoring of landslides with optical remote sensing data. A comprehensive review of innovative remote sensing techniques for landslide monitoring shows the potential and limitations of available techniques and guides the selection of the most appropriate combination of sensors – platforms – image analysis methods according to the observed process and end-user needs. For the efficient detection of landslides after major triggering events at the regional scale, a method for rapid mapping combining image segmentation, feature extraction, supervised learning is developed. For detailed landslide investigations at the local scale, this study elaborates image processing chains for detection of surface fissures in time-series of UAV images as geo-indicators of landslide activity, the measurement of horizontal surface displacements from VHR satellite images using stereo-photogrammetric and image correlation methods, and 3D measurements from terrestrial photographs based on multi-view open-source photogrammetry.

Keywords: Landslide, Mapping, Monitoring, Optical remote sensing, Earth observation, Image classification, Machine learning, Photogrammetry, Image correlation

Résumé: La cartographie, l'inventaire et le suivi de glissements de terrain sont indispensables pour l'évaluation de l'aléa glissements de terrain et la gestion des catastrophes. La disponibilité croissante des satellites THR, des drones et des appareils photo numériques grand public offre un grand potentiel pour soutenir ces tâches à l'échelle régionale et locale en complément de techniques établies telles que l'instrumentation in-situ, radar, et les acquisitions par scanner laser. Un manque d'outils de traitement d'image pour l'extraction efficace d'informations pertinentes à partir de différents types d'imagerie optique complique encore l'exploitation des données optiques et entrave la mise en œuvre de services opérationnels. Cette thèse est consacrée à l'élaboration et l'application de techniques de traitement d'image pour la cartographie, la caractérisation et la surveillance des glissements de terrain en exploitant des données d'imagerie optique. Un état de l'art approfondi des techniques de télédétection innovantes pour la surveillance des glissements de terrain est proposé et démontre le potentiel et les limites des techniques et propose des critères pour le choix des capteurs disponibles (plateformes et méthodes d'analyse d'images) selon le processus observé et les besoins des utilisateurs. Pour la cartographie rapide des glissements de terrain lors de catastrophes majeures, une méthode qui combine segmentation d'image et apprentissage supervisé est développée pour l'analyse des images satellitaires THR à travers plusieurs exemples en Chine, au Brésil, à Haïti, en Italie et en France. Pour l'analyse de glissements de terrain à l'échelle locale, la recherche a élaboré des chaînes de traitement d'images pour la détection de fissures à partir de séries temporelles d'images de drones comme possible géo-indicateurs de l'activité des glissements, la mesure des champs de déplacements horizontaux à partir d'images satellitaires THR utilisant en utilisant des méthodes stéréophotogrammétrie et par corrélation d'image, et les mesures 3D à partir de photographies terrestres basées sur des méthodes de photogrammétrie multi-images.

Mots-clés: Glissement de terrain, Cartographie, Surveillance, Observation de la Terre, Télédétection optique, Classification d'images, Apprentissage automatique, Photogrammétrie, Corrélation d'images

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisors Jean-Philippe Malet and Norman Kerle who made this research possible in the first place and guided it with enthusiasm, ambition and scientific acumen. Conducting my studies within different departments with focuses on geography, geophysics, earth observation and computer science was not always an easy task but due to the continuous support and approachable attitude of Jean-Philippe and Norman I benefited greatly from this interdisciplinary environment. In this context I am also very grateful to my promoter Christiane Weber and co-supervisor Anne Puissant who helped a lot to resolve both scientific and administrative issues.

To defend this thesis I have invited Prof. Rémi Michel, Prof. Sébastien Lefèvre, Prof. Jean-Michel Dischler, and Prof. Michel Jaboyedoff and I am indebted to them for kindly accepting this invitation and dedicating time to review the manuscript.

I was fortunate to collaborate and discuss with scientists from many different disciplines who shared with me their insights into image processing, machine learning, photogrammetry, natural hazards management, landslide mechanics and computer programming. I thank Pascal Allemand for hosting me at LGL and initiating me to multi-view photogrammetry, and in particular Marc Pierrot-Deseilligny who created the marvellous MicMac library and was very open to discuss all my related theoretical and practical questions. Thomas A. Lampert and Nicolas Lachiche from ICube helped me to fill gaps in my knowledge on machine learning and computer vision, and inspired me to consider accuracy assessment and ground truth uncertainties with greater rigor. Julien Travelletti who dedicated his PhD research to the investigation of the landslides in the Ubaye valley stimulated my work by generously sharing the insights and datasets he gathered. I thank Grégoire Skupinski, Patrice Ulrich, Alexandre Mathieu, Julien Gance, Alexandre Remaître, and Romy Schlögel for their collaboration during the field campaigns in the Ubaye valley. The many weeks of field work at the Super-Sauze landslide were not only crucial to acquire data and assess obtained results but also always an opportunity for exchange with my colleagues and fellow PhD students outside the four walls of an office. Very fruitful research and project work also resulted from collaborations with Camille Kurtz, Ping Lu, Simon Rougier, Uwe Niethammer, Sabrina Rothmund, Veronica Tofani, Samuel Segoni, Nicolas Casagli, Clément Michoud, Antonio Abellán, and Michel Jaboyedoff.

A large part of the development and data analysis in this thesis was based on the R Project for Statistical Computing which greatly facilitated my work. I'm very grateful to the R Core Development team and the many package authors such as Andy Liaw, Robert Hijmans, Edzer Pebesma, and Adrian Baddeley, who contributed to this research without knowing.

During the course of my thesis I received generous support from several institutions. I was hosted by Laboratoire Image, Ville, Environnement (CNRS / University of Strasbourg, the Institut de Physique du Globe de Strasbourg (CNRS / University of Strasbourg), and the Faculty of Geo-information Science and Earth Observation (ITC / University of Twente) who provided a very open, aspiring and international research environment. Financial and logistic support was provided by the SafeLand project (EC-FP7), the FOSTER project (ANR COSINUS), the Observatoire Multidisciplinaire des Instabilités de Versants (SNO-INSU OMIV), the Centre national d'Etudes Spatiales (CNES), and by the Site d'Etude, d'Observation, de Logement et d'Accueil autour de la Nature et de l'Environnement en Ubaye (Séolane).

Unjustly, in academia teaching is often considered as a necessary duty rather than a virtue in itself. All the more I appreciate the excellent courses on scientific data analysis prepared by Harald van der Werff and David Rossiter as well as the efforts of countless

further scientists who make their course materials available online. I also like to thank Anne Puissant for giving me the chance to gain further teaching experience at the Faculty of Geography (University of Strasbourg) and all Master students who followed my courses with enthusiasm.

I was lucky to share the last four years with many amicable officemates, flatmates, and colleagues from diverse scientific and cultural backgrounds. Some of them eventually became friends and I hope we can keep in touch despite great geographic distances. I will remember especially Byron who shared with me his house and a garden with great biodiversity, and almost sacrificed his tendon for our victory at ITC's soccer cup. My thoughts also go to my friends and fellow PhD students in Enschede Mustafa, Simone, Xuanmei, Sharon, Wiebke, Tolga, Özgün, Mireia, Jahanzeb, Khamarrul, Saibal and Tapas. My colleagues and friends in Strasbourg gave me a very warm welcome and made me feel home despite my lasting struggle with the French language. A big hug goes to Simon and Sarah, Paul, Arnaud, Pierre, Alexandre, Romy and Estelle. I arrived in France at the same time than Tom who became an estimated colleague and a good friend. The hike through the Alps and the all night long conversations about live and science were a great pleasure and inspiration. I hope we will have the chance to follow this up from time to time.

My deepest gratitude goes to my family who always encouraged me to follow my own path and kept the door wide open whenever I decided to come back home for a while. As I remember it was my mother who bought me the first book about natural sciences and encouraged my curiosity on our Sunday afternoon walks. This thesis is dedicated to her.

Last but not least my thoughts are with Mariana whose unconditional love and support have been a great resource ever since we met. I look forward to all the wonders and challenges that lie ahead of us.

RESUME LONG

Au cours du siècle dernier (1903-2004), environ 16.000 personnes ont été tuées par des glissements de terrain en Europe et la perte économique estimée s'élève à 1700 millions US $ (Nadim et al., 2006). Les mécanismes des glissements de terrain ne sont pas encore bien compris (van Asch and Malet, 2009), et les méthodes d'évaluation de l'aléa et du risque se sont pas standardisées (van Westen et al., 2006). La création d'inventaire et de catalogues d'évènements est le fondement d'une évaluation quantitative. Leur élaboration et leur mise à jour est un travail considérable et les résultats incluent toujours un degré élevé de subjectivité. (Wills and McCrink, 2002). Plus précisément, les occurrences spatiales et temporelles, le type de mouvement, l'intensité du phénomène (par exemple exprimé en terme de vitesses de déplacement) et d'autres paramètres descriptifs doivent être intégrés à ces inventaires à l'échelle régionale et à l'échelle locale. Les challenges dans ce domaine concernent surtout la cartographie régionale post-événement et la caractérisation détaillée du type de glissement de terrain, de sa géométrie et de son intensité. Celle-ci inclut le suivi continu de pentes instables afin de collecter des séries quantitatives sur les déplacements et les volumes, et plus généralement afin de prévoir les événements et de mettre en place des systèmes d'alerte.

Les progrès notables (résolution spatiale et spectrale, répétitivité des mesures) et la diversification des technologies dans le domaine de la télédétection passive (imagerie optique issue de plateformes d'acquisition terrestre, aérienne et spatiale) ont considérablement amélioré les capacités de cartographie et de surveillance des versants instables. Dans ce contexte, la télédétection passive contribue à améliorer les connaissances sur la cinématique des glissements de terrain. Les technologies de télédétection active (imagerie radar et LiDAR) sont également couramment utilisées pour l'investigation de glissements de terrain. En particulier, l'interférométrie radar est utilisée dans des services opérationnels de gestion des risques. Cependant, l'interférométrie radar et les données LiDAR comportent plusieurs limitations telles qu'une couverture spatiale et temporelle limitée et des coûts élevés.
L'imagerie optique est une technologie de télédétection rentable et polyvalente dont le potentiel est, à l'heure actuelle, sous-utilisé car elle permettrait de fournir des informations complémentaires et de surmonter plusieurs limitations des techniques actuelles.

Dans ce contexte, l'objectif de ce travail de thèse est de proposer des techniques de traitement des images optiques, robustes et le plus automatisée possible, afin d'exploiter au mieux le volume de données disponibles pour caractériser et suivre les glissements de terrain. Il s'agit également de développer/proposer des chaînes de traitement d'images opérationnelles adaptées aux images optique à Très Haute Résolution (THR) issues des satellites, des drones et des appareils photographiques numériques 'grands publics'. Ces développements sont destinés à la cartographie rapide des glissements de terrain à l'échelle régionale, et à leur caractérisation détaillée et à leur suivi, à l'échelle locale. Les objectifs spécifiques de la recherche sont les suivants :

• le développement d'une chaîne de traitement générique des images satellite à Très Haute Résolution (THR) spatiale, pour la cartographie rapide des glissements de terrain, fondé sur des algorithmes d'apprentissage automatique et sur une approche d'analyse d'images dite 'orientée objet' (OBIA; Object-Based Image Analysis) ;
• le développement et la mise en œuvre d'un algorithme d'apprentissage actif afin de réduire le besoin de données d'entraînement avec une attention particulière accordée aux contraintes spatiales et à l'incertitude liée à la vérité terrain ;

• le développement d'une méthode d'analyse d'images pour détecter des indicateurs géologiques d'instabilités (e.g. fissures de surface) ;
• l'élaboration d'une chaîne de traitement afin de mesurer les déplacements de la surface d'un glissement de terrain à partir d'images satellites THR monoscopiques et stéréoscopiques ;
• l'adaptation et la comparaison de bibliothèques de photogrammétrie open-source pour la reconstruction 3D de la surface d'un glissement de terrain et le suivi des déplacements à partir des photographies terrestres.

Une analyse approfondie des innovations technologiques récentes dans le domaine de la surveillance des glissements de terrain par télédétection est présentée dans le **chapitre 1**. Celui-ci fournit un aperçu détaillé des systèmes et des méthodes de traitement actuellement disponibles. Des critères permettant de comparer les différentes techniques et méthodes sont proposés. Les capacités et les limites de chacune des techniques sont comparées sur la base de critères tels que les résolutions spatiale, spectrale, temporelle, la couverture spatiale, la précision de la mesure, les coûts des données et de traitement et la maturité de la méthode. Au final, selon les types de glissements de terrain, leurs vitesses de déplacement, et leurs échelles d'investigations, des directives sont proposées aux utilisateurs afin de choisir la technologie la plus adaptée. Celles-ci s'adressent à la fois aux scientifiques et aux gestionnaires du territoire en charge de la cartographie des glissements de terrain et de la gestion des risques.

Le **chapitre 2** porte sur le développement, l'implémentation et la validation des chaînes de traitement génériques pour la cartographie des glissements de terrain à partir de données multi-sources (images satellites THR, photographies aériennes et données topographiques). Plusieurs algorithmes pour la segmentation d'image, pour l'extraction et la sélection d'attributs et pour l'apprentissage supervisé sont adaptés et combinés dans des chaînes de traitement. Les techniques développées sont appliquées et validées pour la cartographie des glissements de terrain sur cinq sites d'étude (en Chine, au Brésil, à Haïti, en Italie et en France). Les résultats obtenus sur l'ensemble des sites d'étude sont satisfaisants en termes de précision des classifications.

De nouveaux attributs particulièrement adaptés à la reconnaissance des glissements de terrain sont proposés et ont montré leur généricité. Les améliorations en termes de précision de classification sont significatives sur chaque site d'étude. Une méthode supervisée innovante permet de sélectionner les attributs les plus pertinents. Elle permet de réduire la complexité de l'apprentissage automatique et elle améliore son efficacité et sa précision. Les relations entre la segmentation (paramètre d'échelle), l'importance des différents attributs des objets et la précision des résultats sont étudiées. Les résultats montrent qu'une sur-segmentation fournit généralement des précisions plus élevées. Des programmes de ré-échantillonnages stratifiés sont également proposés et permettent de compenser les biais qui résultent du déséquilibre de la taille des classes.

Afin de réduire le temps d'apprentissage des données d'entrainements dans le processus supervisé, un algorithme d'apprentissage actif original, qui utilise les contraintes spatiales et le coût en terme de temps d'étiquetage, lors du processus de sélection des échantillons, est proposé. Ce nouveau plan d'échantillonnage fondé sur un nombre réduit de régions, permet de diminuer considérablement le temps d'étiquetage de l'utilisateur. Une comparaison, entre les erreurs restantes et les incertitudes calculées sur la base des cartographies expertes, révèle que la précision obtenue est dans la plage de la variance entre plusieurs experts. Cela indique que la chaîne de traitement proposée peut être utilisée pour des applications opérationnelles de cartographie rapide pour lesquelles la qualité finale est comparable à une cartographie experte et manuelle.

Le **chapitre 3** décrit le développement et l'application de techniques de télédétection pour l'acquisition d'information sur les structures de surface, les déplacements et les volumes des glissements de terrain à l'échelle locale à partir d'images satellites THR, de drones et de photographies terrestres.

Une première méthode de traitement d'images combine l'utilisation de filtres gaussiens originaux, d'opérateurs de morphologie mathématique et des méthodes d'analyse d'images orientées-objets pour la cartographie semi-automatique des fissures à partir de séries d'images de drones. Les cartes qui en résultent sont validées avec des cartes expertes (interprétation manuelle) sur l'évolution des fissures par rapport aux vitesses de déplacement et à la géométrie de la surface de glissement.

Une deuxième chaîne de traitement est ensuite proposée pour le calcul des modèles de surface à haute résolution spatiale et des champs de déplacement à partir d'images satellites THR. Elle est appliquée à trois sites différents et elle permet de mesurer des taux de déplacement horizontal avec une précision, pour la première fois, décimétrique. De plus, une telle précision peut être obtenue sans points de contrôle au sol, ce qui est un atout essentiel pour l'automatisation de la surveillance opérationnelle dans un futur proche. L'ensemble des données bi-temporelles analysées couvre une période de deux mois et les champs de déplacement obtenus permettent de mieux comprendre les taux de déplacement.

Enfin, trois chaines de traitement de données multi-stéréoscopiques sont proposées afin de reconstruire des surfaces tridimensionnelles. Les résultats sont comparés en termes de capacité de reconstruction multi-dates et de suivi 3D de glissements de terrain à partir des photographies terrestres. Les comparaisons avec des données LiDAR terrestre et aérien montrent que la stéréophotogrammétrie à partir des bibliothèques open-source permet d'atteindre des précisions semblables mais avec un coût d'équipement et de fonctionnement nettement inférieur. Les volumes et les taux de déplacement 3D sont conformes, respectivement, aux mesures à partir de nuages de points LiDAR et aux observations GNSS. La détection des changements entre nuages de points permet la quantification des processus géomorphologiques dominants (éboulements, ruissellement) et donne un aperçu des champs de déformation pendant les phases d'accélération.

En conclusion, bien que l'utilisation de l'imagerie optique comme outil opérationnel dans les investigations de glissements de terrain nécessite encore des développements, les méthodes proposées dans cette thèse de doctorat sont très prometteurs et pour certains utiles dans un contexte opérationnel de surveillance de versants instables.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

APPENDICES

Introduction

The term landslide is defined as *"... the movement of a mass of rock. earth or debris down a slope*" (c.f. Cruden, 1991) and includes different types such as slides, topples, falls and flows (Cruden and Varnes, 1996). **Landslides** are an important agent in the evolution of mountain ranges (Larsen and Montgomery, 2012) and a major natural hazard with **severe socio-economic impacts worldwide** (Nadim et al., 2006; Kjekstad and Highland, 2009; Petley, 2012). In the context of climate change, there is a "*high confidence*" that changes in the hydro-meteorological pattern (precipitation, air temperature) and melting of the permafrost will affect slope stability in some regions of the world (IPCC, 2012).

The **reduction of landslide hazard** not only depends on spatio-temporal information about the frequency and distribution of landslides, and of their predisposing and triggering factors, but also on the quantification and understanding of landslide kinematics and the underlying mechanical processes. Main tasks in landslide hazard assessment and risk management are the construction of **landslide inventories**, the **fast characterization** of landslide processes in case of disaster, **long-term landslide monitoring**, and landslide **forecast/warning** with statistical or process-based models.

Earth Observation techniques in landslide investigations

Remote sensing techniques can be employed at different spatial scales to gather information about the distribution and kinematics of landslides, and about their conditioning factors (Metternicht et al., 2005; van Westen et al., 2008; Guzzetti et al., 2012; Kirschbaum et al., 2012). Especially in mountainous terrains, **remote sensing provides a valuable tool to obtain spatially distributed information**, and is a complement or alternative to time-consuming and often only point-wise *in-situ* measurements.

Evolution of landslide remote monitoring from 1900 to present day

The first documented cases for the systematic application of remote sensing in landslide investigations can be traced back to use of theodolites and triangulation networks (Figure 0-1a) for the monitoring of deep-seated movements in the Swiss Alps (Ganz, 1914). The possibility to document landslides and their impacts with photographs resulted from the invention of **photography in the early 19th century**. One of the earliest conserved examples are pictures of a rockslide that destroyed the village of Elm (Swiss Canton of Glarus, Figure 0-1b) on 11 September 1881 and was attributed to improper mining practices (Heim, 1882). A remarkable example for those early works is a series of **trigonometric measurements** (Figure 0-1a) carried out at the Monte Arbino (between 1889 and 1928) which documented the movement and acceleration of a large rockslide and led to the evacuation of 14 houses before the final failure on 2 October 1928 (Figure 0-1c). It can be considered as the first successful application of remote sensing for landslide early warning (Zölly, 1948; Bonnard, 2006). At about the same time, continuous trigonometric measurements at the Kilchenstock (Linthal, Swiss Alps) permitted to document the relationship between displacement and meteorological variables for the first time. Based on the observed acceleration an evacuation

of the town of Linthal was issued but the predicted large failure never happened (Heim, 1932).

The first **photo-theodolites** were developed in the 1920s and quickly adopted for topographic surveys. A first example for the use of such instruments on slope processes is given in Finsterwalder (1931) who documented glacier displacement during the late 1920s. This technique also permitted to estimate the volume of the Monte Arbino rockslide as approximately 20 million m^3 (Zölly, 1948).

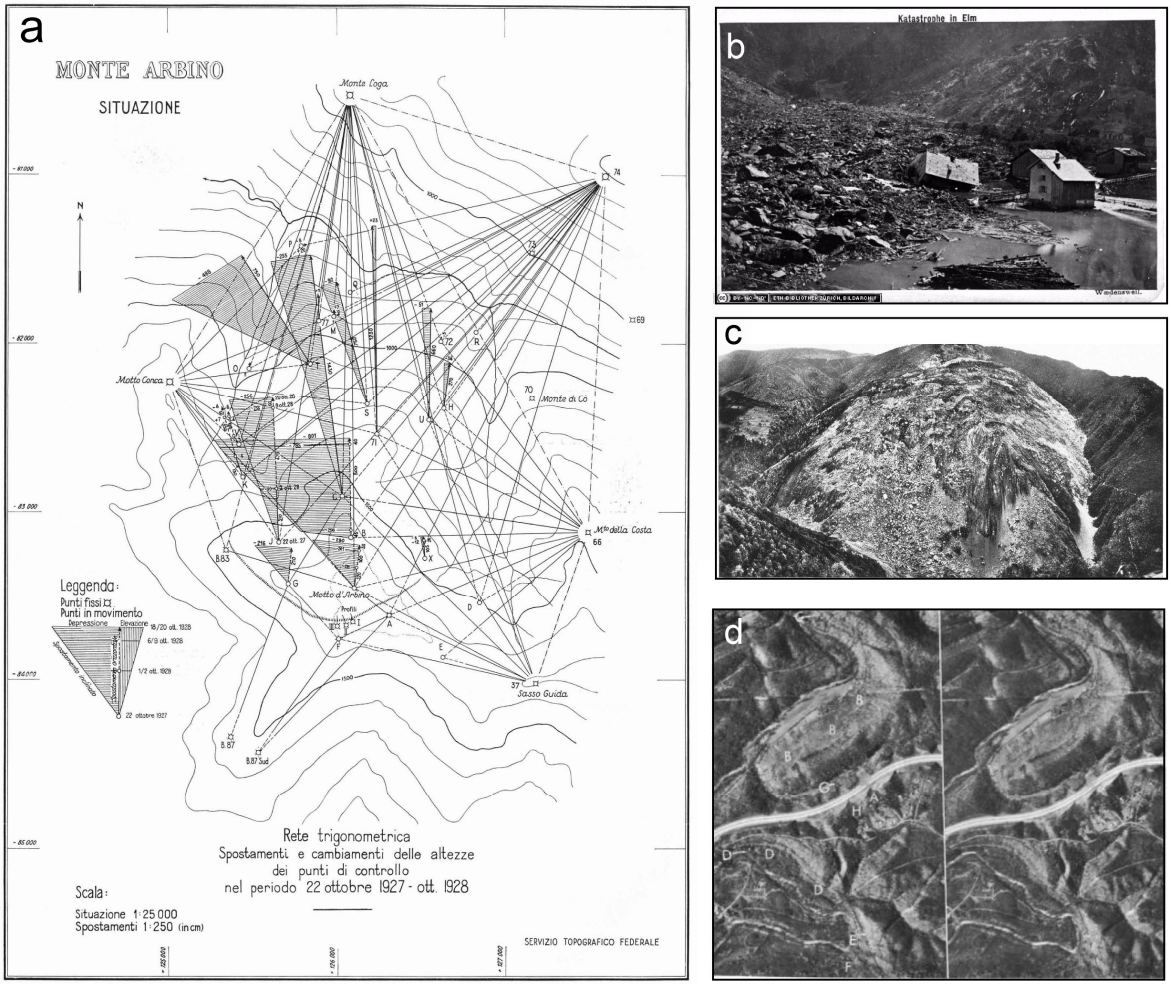


Figure 0-1: Historical examples for the use of remote sensing in landslide investigations. (a) The trigonometric network at the Monte Arbino in the late 1920s (Zölly, 1948). (b) Terrestrial photographs showing the impact of the 11 September 1881 rockslide on the village Elm (taken by Richard, T. Waedensweil in September 1881, archive of the ETH-Zürich). (c) Terrestrial photograph of the Monte Arbino rockslide taken short after a major rockslide on 2 October 1928 (Zölly, 1948). (d) Stereoscopic image-pair recorded on 13 April 1973 showing the Sewickley Water Works (Pennsylvania, USA) landslide (Pomeroy, 1982).

The first **aerial photographs** were recorded in 1858 but only with the enhancement of the stereoautograph (a predecessor of modern stereo-plotters); in the 1910s, precise elevation measurements from stereo-pairs of aerial photographs became possible and greatly facilitated topographic surveys. While the scientific literature before the Second World War reveals no applications of stereoautographs to landslides, the visual interpretation of aerial photographs quickly became a commonly used technique for their recognition. A great interest in black

and white and near-infrared aerial photographs arose from extensive highway construction after Second World War in the U.S. Significant efforts were made to refine methods and training for visual image interpretation with emphasize on stereoscopic image analysis (Figure 0-1 d). As a result, first handbooks and guidelines for the use of aerial photography in landslide investigation were developed (e.g. Johnstone et al., 1953; Cleaves, 1961; Rib and Liang, 1978).

A breakthrough for Earth Observation was the **launch of the first Landsat satellite** (1972) which, for the first time, provided images of the earth surface at regular intervals and with global coverage. An important step toward images with higher resolution was the invention of a charge-coupled device (CCD), in 1969 at AT&T Bell Labs, which soon after led to the launch of the first digital military satellite with a CCD sensor (KH11, USA, 1976). The first civilian satellite that used this technology was the French SPOT satellite launched in 1986. Several studies demonstrated that the visual detection of large landslide features is possible with LANDSAT (30m.pixel^{-1}, Figure 0-2a) and SPOT (10m.pixel^{-1}) images but a clear delineation of the affected area was generally not possible (Mantovani et al., 1996). Consequently, researchers started to develop new methods for change detection and image matching targeting the delineation of affected areas (Nichol and Wong, 2005) and displacement measurements (Scambos et al., 1992; Yamaguchi et al., 2003), respectively.

The **first civilian Synthetic Aperture Radar (SAR)** satellite ERS-1, launched in 1991, was the first system to provide global coverage of SAR images with a revisit time of only 35 days. As a consequence, research dedicated to the enhancement of interferometric methods for the generation of digital elevation models (DEMs) and ground displacement measurements flourished (Bamler and Hartl, 1998) and resulted in the first measurement of a coseismic displacement field from space (Massonnet et al., 1993). A first application of SAR interferometry to the monitoring of landslide was demonstrated for the La Clapière landslide and deviated only a few millimetres from ground measurements (Fruneau et al., 1996). Similar results were reported from the application of Ground-Based InSAR (GB-InSAR) for landslide monitoring providing high temporal and spatial resolution at accuracies better than 2 mm (Tarchi et al., 2003).

Another milestone in Earth Observation was set with the launch of the **IKONOS satellite** in 1999 which became the first civilian satellite with sub-metre resolution and opened the door for a **new generation of very high resolution (VHR) satellites**. VHR satellites provide images with spatial resolution comparable to traditional aerial surveys (Figure 0-2c, d) but with global coverage and short revisit times of up to one day. The **visual interpretation** of VHR satellite images recently became widespread for landslide inventory mapping but it was also recognized early that **more automated image analysis techniques** should be developed to fully exploit the regularly acquired images (Hervás et al., 2003). Due to the development of enhanced methods for image orientation and matching during the last decade, displacement measurements based on aerial stereo-photogrammetry (Casson et al., 2003) and VHR satellite images (Delacourt et al., 2004) came more and more into focus.

Laser technology was invented in the 1960s. The first laser-based electronic distance meters (EDMs) became available in the mid-1960s and the first theodolites with an integrated EDM (also termed electronic tacheometer or total station) was released in the 1970s (Rüeger, 1990). In recent years, **Light Detection and Ranging (LiDAR)** has greatly reshaped the practices in topographic surveying as it allows a very dense spatial sampling of the topography (Figure 0-2e) with centimetre accuracy. Applications of LiDAR in landslide research comprise mapping, monitoring of displacement and deformation, as well as volumetric analysis (Jaboyedoff et al., 2010).

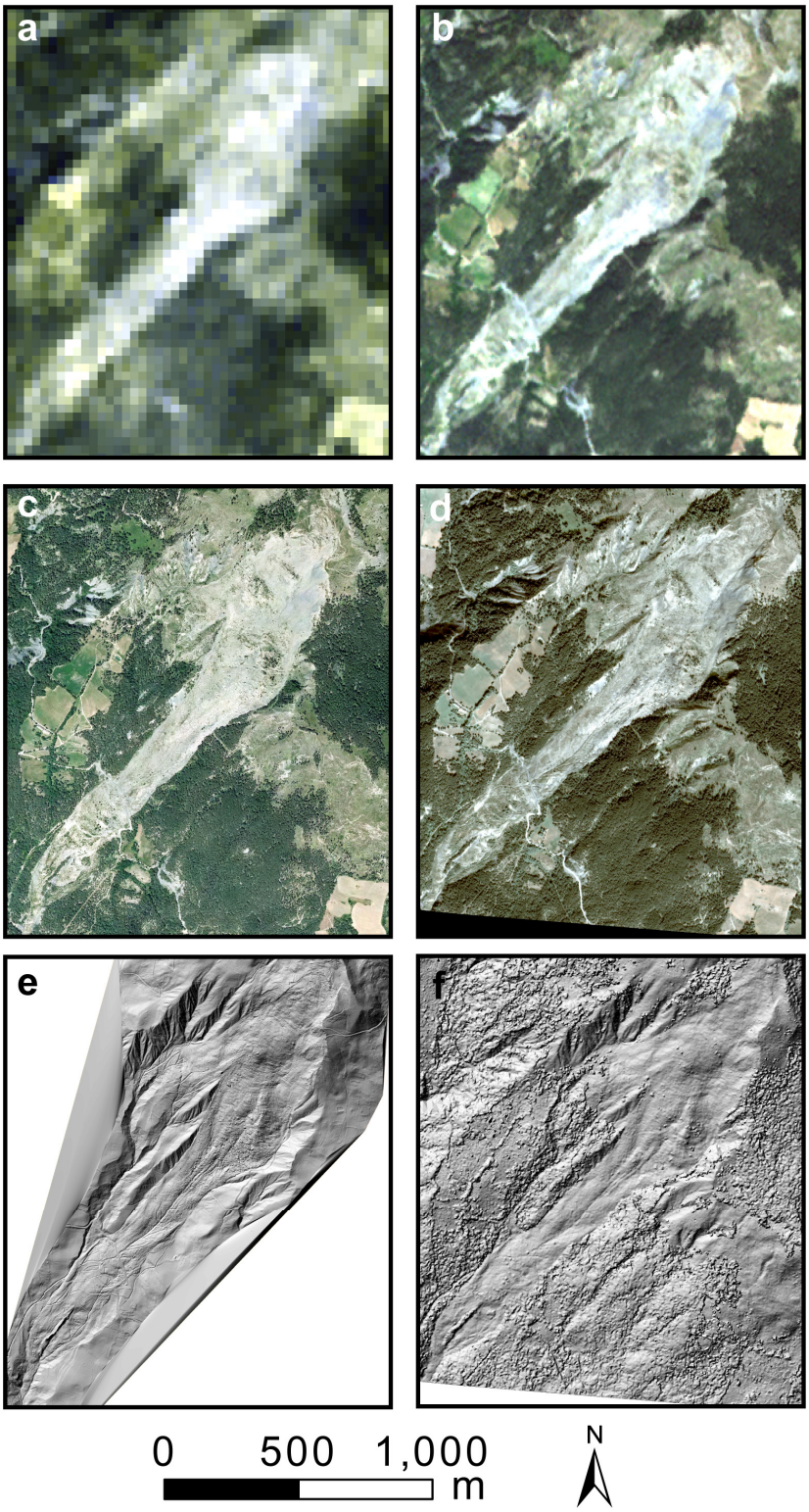


Figure 0-2: Examples of remote sensing datasets with different spatial resolutions for the La Valette landslide (South French Alps). (a) Natural-colour composite of a Landsat 7 ETM image (30 m) captured in 2009, (b) Natural-colour composite of a RapidEye image (5 m) captured in 2009, (c) Aerial photograph (IGN, 0.5 m) captured in 2009 image (d) Pan-sharpened Pleiades images (0.5 m) captured in 2012, (e) Shaded relief image derived from a LiDAR DTM (0.5 m) scanned in 2009, and (f) Hillshade image derived from a stereo-photogrammetric surface model (0.5) derived from Pleiades stereo-pairs captured in 2012.

Recently, emerging technologies that have some potential for the observation of landslides comprise hyperspectral remote sensing (Sterzai et al., 2010), airborne geophysics (Supper et al., 2013) and also bathymetry for the investigation of submarine landslides (Hough et al., 2011).

This historical overview shows the increasing number and **diversity of remote sensing techniques** that can be employed for landslide observation. It also demonstrates that changes in the practice of landslide investigations are not only driven by the increasingly better understanding of the underlying processes and resulting hazards, but often also by technological innovations.

Applications and limitations of modern remote sensing technologies for landslide investigations: from mapping and monitoring to services

Emerging remote sensing technologies are typically not designed specifically for landslide observation and, therefore, often require adaptation and validation to exploit their capabilities for research and operational applications. This process of adaptation has to take into account the application scenarios and general tasks that need to be realized for different stages of landslide investigations and management.

Based on the risk management cycle as a conceptual framework, three possible stages in the application of remote sensing techniques can be distinguished (Figure 0-3).

- **Rapid mapping** is the fast analysis of images and elaboration of maps depicting affected areas, the number of landslides and their size immediately after major events to support disaster response.
- **Characterization** comprises detailed investigations of the landslide types, sub-parts, activity state and volumes but also the collection of information on preparatory and triggering factors as a base for hazard assessment.
- **Monitoring** is the periodic and/or permanent acquisition and analysis of a series of observations over time to gather information about temporal changes of specific landslides or potentially unstable areas. Continuous monitoring is beneficial to progress in landslide process understanding, evaluation of the effectiveness of mitigation measures, and as a component of Early-Warning Systems (EWS).

Rapid mapping typically targets map creation at a regional scale while more detailed characterization and monitoring often requires site specific investigations.

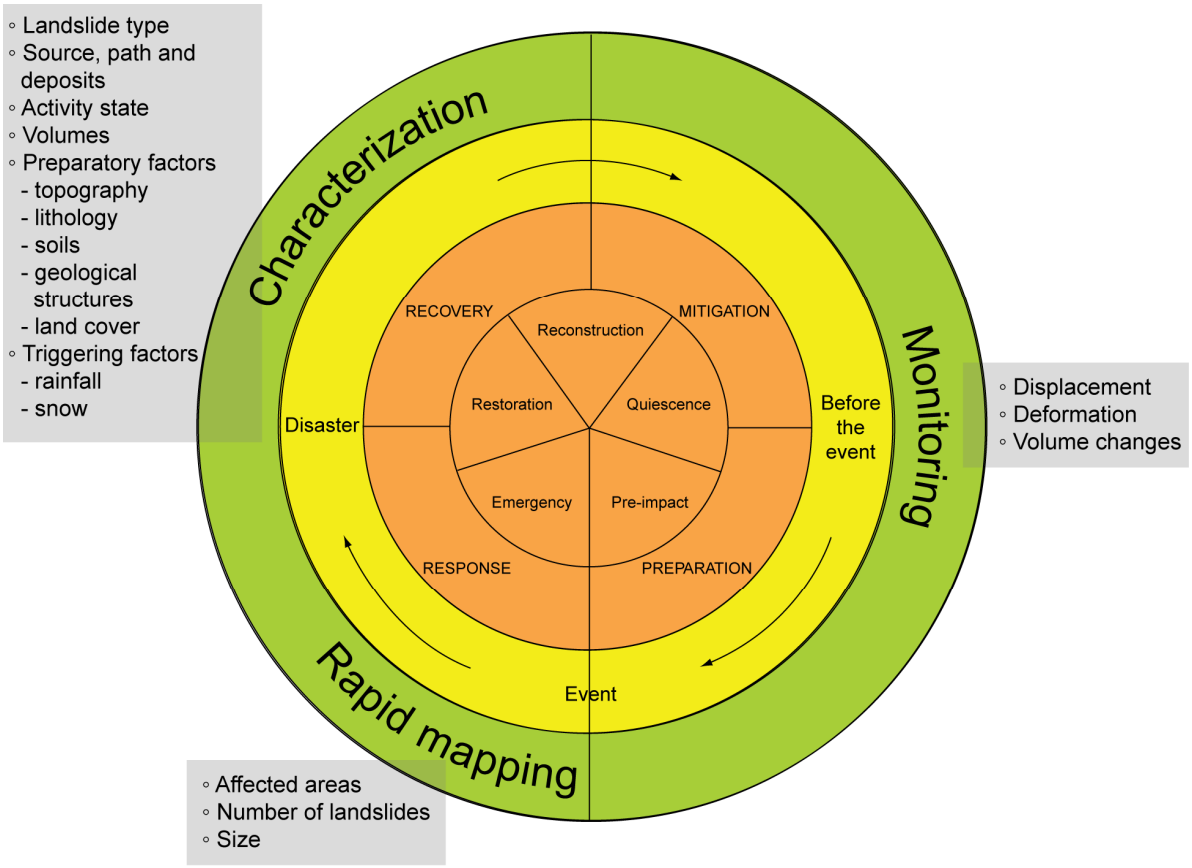


Figure 0-3: Rapid mapping, characterization and monitoring as stages of landslide remote sensing in the risk management cycle (modified after Alexander, 2002 and ; Glade et al., 2005). Priority tasks for landslide investigations (green circle) are shown in relationship to the temporal course of disasters (yellow circle) and the corresponding phases in the risk management cycle (orange circle).Relevant information that can be gathered at each stage with remote sensing techniques are listed within the grey boxes.

Numerous orbital satellites, airborne platforms and ground-based remote sensing systems are currently operational or planned for the near future, and in principle are able to provide valuable input for all three stages of a landslide investigation (Metternicht et al., 2005; Delacourt et al., 2007; van Westen et al., 2008; Guzzetti et al., 2012). Mantovani et al. (1996) argued that the use of remote sensing for landslide investigations is still not full exploited and Metternicht et al. (2005) renewed this point of view several years later. They concluded (1) that **manual stereoscopic image analysis was still the prevailing standard for landslide mapping**, (2) that kinematic surveys (monitoring) have been mostly undertaken with interferometric techniques, and (3) that data availability was still an issue. Due the multitude of available remote sensing systems, the issue of data availability has been largely resolved in recent years. However, the storage, distribution, harmonization, and **analysis of large image archives and terabytes of daily incoming data still pose major challenges** and are often the bottlenecks that hinder the dissemination of required information to end users. Considering the great diversity of landslide processes and site characteristics, it also remains difficult to evaluate the potential and limitations of different remote sensing technologies for specific sites and integrate them with *in-situ* and process-based models.

A series of recent European research projects have addressed the exploitation of new remote sensing datasets and processing techniques to furnish services for rapid mapping and emergency response (SAFER, 2007-2013; LAMPRE, 2013 - 2015) and the detection and monitoring of surface deformation (SLAM, 2003-2005; TERRAFIRMA, 2003-2012; PREVIEW, 2005 - 2008; DORIS, 2007-2013). On the long-term, those efforts target the **integration of rapid mapping, characterization, monitoring, and the direct dissemination of the results to end users**.

Most of the resulting services rely on the use of **SAR interferometry** (from ground-based and spaceborne systems) which has proven mature for operational use. SAR has the ability to operate weather independent and detect millimetre displacement with high accuracy. However, this technology provides only the Line-Of-Sight (LOS) component of the 3D surface displacement, has limited coverage (e.g. north facing slopes cannot be monitored) and is not adapted to displacement rates exceeding several centimetre per month or to rapid mapping tasks after large triggering events.

LiDAR systems have also become a mature technology for landslide investigation focusing mainly on structural analysis of rock slopes, the quantification of sediment budget and displacement between different time periods (Jaboyedoff et al., 2010). It must be distinguished here from (passive) optical remote sensing since LiDAR has the unique capability to penetrate dense vegetation and yields 3D point-cloud representations of the terrain with very high accuracy. However, the costs for hardware and operation of LiDAR are still relatively high and its application is currently limited to short-term measurements for local investigations or intervals of several years for regional surveys.

Visual image interpretation is still the most common use of **optical images** in landslide investigations. Though, methods for semi-automatic image analysis, displacement measurements and 3D reconstruction, have been developed they are not employed frequently and generally not ready for operational use in integrated services.

Potential and current limitations for the use of optical data in landslide investigations

For the investigation of landslides several characteristics of optical remote sensing systems suggest optical imaging as a valuable complement or alternative to established remote sensing techniques such as LiDAR and SAR. The capabilities of optical remote sensing systems are determined by features of the sensors, the supporting platforms and the image analysis techniques used to extract information.

In general optical sensors record the intensity of light in three bands of the visible spectra, whereas most satellites include also additional bands in the near-infrared and longer wavelength and are sometimes considered as optical sensors in a broader sense (Prasad et al., 2011). Recent progress in sensor manufacturing (CCD and complementary metal–oxide–semiconductor, CMOS) has **lowered the costs of high-resolution optical imaging** significantly during the last 20 years (Figure 0-4a).

Since the imaging process is passive little or no external power sources are required which allows the manufacturing of very compact and light-weight devices. Compared to SAR and LiDAR, optical imaging sensors are, therefore, more **easily supported by different platforms** (Figure 0-5) allowing to adapt the spatial scales, timing and mode of image acquisition with more flexibility.

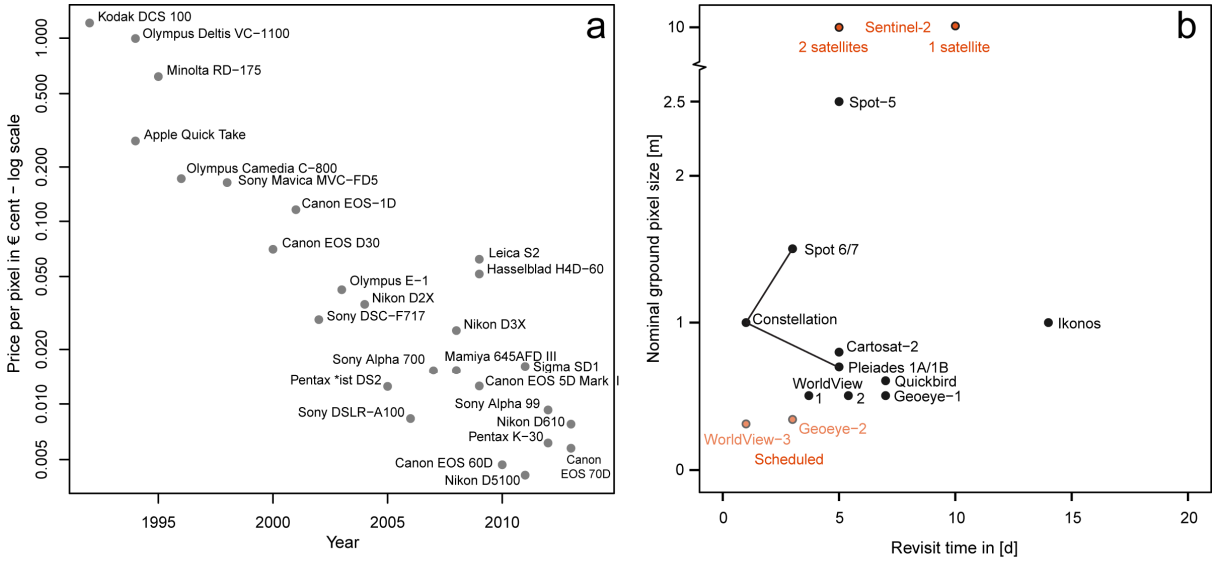


Figure 0-4: Some key dates for modern optical remote sensing systems. (a) Spatial resolution and revisit time of operational and scheduled medium resolution and VHR optical satellites (according to specification of space agency and private operators). The entry "Constellation" stands for the combination of Pleiades and Spot satellites which operate in the same orbital plane but phase shifted by 90°. (b) Evolution of costs (inflation adjusted) in €.pixel^{-1} for selected number of digital cameras since 1992 (compiled from different online resources).

Many different optical remote sensing systems and image products are nowadays available on the market. Systems with a significant potential for applications in landslide investigations comprise handheld cameras, airplanes (more recently also UAVs), and **most notably VHR satellites** (Figure 0-4b) with short revisit time and spatial resolutions that depict characteristic features of different landslide types and could also be exploited for displacement measurements.

An inherent limitation of optical imaging is its dependence on direct sight contact with the targeted object which does in general not allow investigating landslide under dense vegetation and limits the frequency of spaceborne observations in high latitudes and the tropics where cloud cover is more persistent throughout the year. In central Europe, the mean average cloud cover varies between 50% and 70% depending on the season and decreases towards the Mediterranean (Malberg, 1973).

Image processing techniques for the analyses of optical data comprise image classification, feature and change detection, image correlation and stereo-photogrammetry, which in principle allow rapid landslide mapping, displacement monitoring and volume estimation. Image correlation of VHR satellite images, for example, can be used to overcome limitations of SAR interferometry for the monitoring of displacement rates that exceed 1 m.year^{-1}. Comprehensive empirical studies also suggested aerial photogrammetry photogrammetric methods as a cost-efficient alternative to LiDAR providing denser point clouds and similar accuracies (Cramer and Haala, 2010; Leberl et al., 2010) .

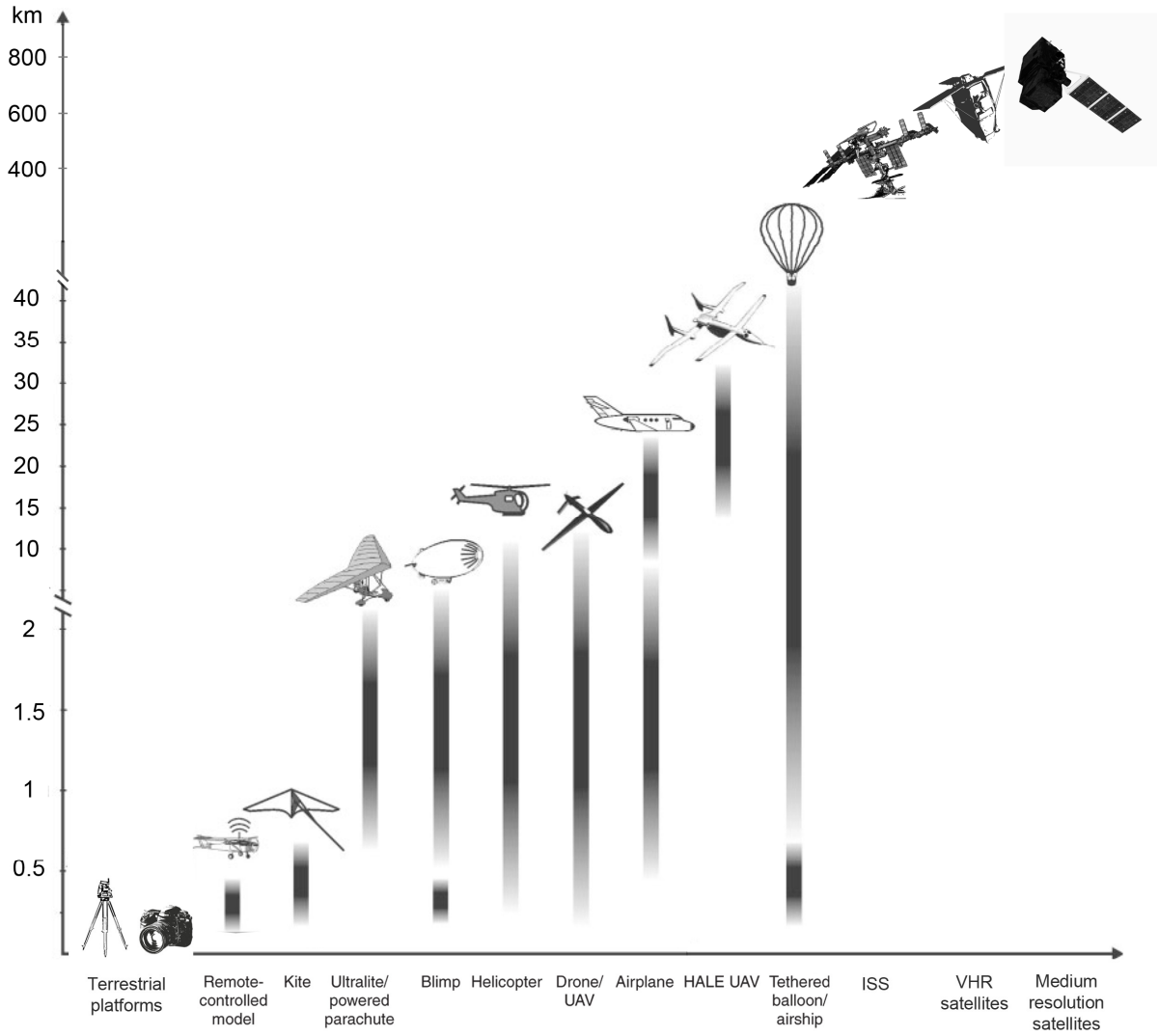


Figure 0-5: Overview of the approximate operational ranges relative to the earth surface of platforms for optical remote sensing (modified after Kerle et al., 2008). The bars below the platform indicate the operational ranges and darker grey shades imply more common application of the platform at a particular height.

Considering the outlined advantages it is somewhat surprising that, in contrast to SAR interferometry and LiDAR, **optical remote sensing is still not used as an operational tool for landslide mapping and monitoring**. The reasons for this shortfall can be traced back to a **lack of adapted image processing techniques** that address the specific information needs of landslide investigations while being sufficiently generic to remain applicable for a broad range of different landslide types and environmental settings.

Classical pixel-based image analysis methods are ill-suited to extract complex patterns from VHR images and **Object-Based Image Analysis (OBIA)** has evolved as a key concept to address shortcomings trough sequential image processing comprising segmentation and classification (Blaschke, 2010). Several studies have addressed the use of OBIA for landslide mapping and characterization with heuristic methods to translate expert knowledge into machine executable rule sets (Barlow et al., 2006; Moine et al., 2009; Martha et al., 2010a; Martha et al., 2011; Martha et al., 2012; Lacroix et al., 2013). Despite successful applications of rule-based classifiers in a number of case studies their elaboration remains time-consuming and subjective. Rule-based approaches provide very limited

transferability for the analysis noisy and variable input (Russell and Norvig, 2009) which is especially true when dealing with different study sites, variable landslide processes, and diverse image datasets. Generic machine learning algorithms have proven to be versatile tool to learn complex concepts automatically from labelled examples but their use in the context of landslide mapping has not been studied so far.

Precise sub-pixel image matching techniques have been developed, and applied to measure landslide surface displacement with airborne and spaceborne images (Kääb, 2002; Delacourt et al., 2004; Casson et al., 2005; Leprince, 2008; Debella-Gilo and Kääb, 2011) However, currently available techniques are specialized in the analysis of medium resolution images to measure large displacement (several meters) with limited accuracy or require aerial image datasets and ground control which are in many cases not available. A better exploitation VHR satellites images could yield more regular and accurate measurements and consequently better integration of the derived displacement fields with *in situ* measured time-series. The scientific literature currently provides no example for an end-to-end processing chain that allows displacement monitoring with VHR satellite images fully independent of external data sources. Potential pitfalls and accuracies are consequently unknown and impede further steps towards robust and precise operational tools.

Multi-temporal **high-resolution surface models** that enable the quantification of surface changes and 3D displacement can be obtained with terrestrial **stereo-photogrammetry**. Several case studies have demonstrated the value of terrestrial photogrammetry for rock mass characterization (e.g. Sturzenegger and Stead, 2009b, a) but, complex processing chains have hindered a frequent use of this technique for monitoring over longer time periods. Progresses in computer vision and photogrammetry have recently converged in a new class of algorithms (e.g. Hirschmuller, 2008; Snavely et al., 2008; Deseilligny and Clery, 2011) that greatly simplify the surface reconstruction from optical images but their potential and limitations for landslide monitoring are currently not well understood.

Research objectives

As a consequence of the potential benefits and current limitations for the operational use of remote sensing techniques outlined in the previous sections, this doctoral thesis is dedicated to the development and extensive validation of image processing methods for landslide mapping, characterization and monitoring with a focus on image analysis methods for optical data.

Considering the outlined tasks for remote sensing at different scales and the great diversity of landslide processes, site characteristics and available remote sensing technologies, the following research objectives are pursued:

- The development of a generic image processing chain for the rapid mapping of landslides from VHR optical images based on OBIA and machine learning algorithms;
- The formulation and implementation of an active machine learning algorithm to reduce the need for training data under special consideration of spatial constraints and ground truth uncertainty;
- The development of an image analyses workflows to detect landslide surface features which can be considered as indicators of landslide activity;

- The elaboration of processing chains for the measurement of landslide surface displacement with spaceborne VHR monoscopic and stereoscopic images;
- The adaptation and comparison of open-source photogrammetry libraries for landslide surface reconstruction and displacement monitoring with terrestrial photographs.

To tackle those objectives, this work takes an interdisciplinary approach combining methods of image classification and segmentation, machine learning, photogrammetry, computer vision, pattern recognition and statistical computing. The design, implementation and validation of image processing techniques should consider principles, paradigms, and ideas from remote sensing science, computer science, geomorphology and geophysics as well as the specific needs of end users such as risk managers.

Outline of the thesis

This thesis manuscript is subdivided in three chapters addressing successively (1) the establishment of criteria for the selection of remote sensing techniques for landslide monitoring, (2) the development of automated processing techniques for landslide mapping at regional scales, and (3) the development of automated processing techniques for landslide investigation at local scales. The chapters are organized along a series of published peer-reviewed papers and of manuscripts currently under review.

Chapter 1 provides a comprehensive review of recent innovations in the monitoring of landslides with remote sensing techniques. A catalogue of criteria is proposed to compare their capabilities and guidelines for the selection of the most appropriate technique for different landslide types, displacement rates, study scales, and environmental settings are provided. The criteria are furnished to aid operational decision making, and include information on the spatial resolution, accuracy and coverage, data and processing costs, and maturity of the method.

Chapter 2 presents the development, implementation and validation of image processing chains for rapid landslide mapping at the regional scale. The developed methods are based on a combination of image segmentation, feature extraction, and feature selection with the Random Forest (RF) machine learning algorithm. The RF algorithm is adopted as a base classifier for an active learning approach which combines criteria for sample uncertainty and diversity to reduce the amount of required training data. It is demonstrated that considering spatial constraints and actual costs at the side of the user provides significant performance improvements above approaches that optimize only the number of samples. Dedicated sampling schemes that resolve the potential bias in class-imbalanced problems are proposed and tested. Applications on VHR images captured mainly in the aftermath of large triggering events (Haiti earthquake, Wenchuan earthquake in China, Messina storm in Italy, Nova Friburgo storm in Brazil) are provided and validation is performed against expert mappings. Finally the uncertainties of expert mappings and relationships with the probabilistic outputs of the machine learning algorithm are investigated.

Chapter 3 presents three different approaches for the monitoring and detailed local characterization of landslides with optical imagery.

Section 3.1 presents a semi-automatic processing chain to map landslide surface fissures with multi-temporal UAV images. The developed processing chain combines Gaussian matched filters and an OBIA heuristic for the reduction of false positive detections and is assessed against results from expert image interpretation in terms of area, density, and orientation. The method can be used to derive indicators for landslide activity and an analysis of the pattern and evolutions of the detected fissures provides insights into underlying mechanical processes.

Section 3.2 introduces the combined use of satellite stereo-photogrammetry and sub-pixel image correlation for surface displacement measurements. The influence of the number of ground-control points is evaluated and the quality of the resulting surface models is compared to aerial LiDAR point clouds. The obtained displacement rates are compared with *in situ* measurements.

Section 3.3 addresses the use of Multi-View Stereo-photogrammetry (MVS) and Structure-from-Motion (SfM) for the 3D surface reconstruction and monitoring of an active landslide with photographs acquired at the ground. Different open-source libraries that combine approaches from computer vision and classical close-range photogrammetry are evaluated and the results of three different pipelines are compared with terrestrial and aerial LiDAR scans. Seasonal image acquisitions over a period of three years are used for multiple reconstruction and the resulting point clouds are compared with a cloud-to-cloud method to detect significant surface changes. The proposed pipeline provides insights into the seasonal dynamics of the geomorphological processes and reliable volume and 3D displacement measurements.

Research environment

The research carried out in the thesis has been embedded in two international research projects dedicated to the development of quantitative risk assessment tools for landslides at the European level (SafeLand, 2009-2011) and to the development of spatio-temporal data mining techniques for the monitoring of soil erosion (FOSTER, 2011-2014). It was supported by two host institutions at the University of Strasbourg, namely the *Laboratoire Image, Ville, Environnement* (LIVE, UMR 7362 CNRS) and the *Institut de Physique du Globe de Strasbourg* (IPGS, UMR 7516 CNRS), and by the *Faculty of Geo-Information Science and Earth Observation (ITC)* at the University of Twente (Enschede, Netherlands). Support was also provided by the Observatoire Multidisciplinaire des Instabilités de Versants (SNO-INSU OMIV).

During the course of this thesis several collaborations with researchers outside the host institutions have been established:

- *Laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie (ICube*, UMR 7357 CNRS), University of Strasbourg: T. Lampert, C. Kurtz, N. Lachiche, P. Gançarski;
- *Department of Earth Science,* University of Florence: P. Lu, N. Casagli, V. Tofani, S. Segoni;
- *Institute of Geophysics,* University of Stuttgart: U. Niethammer, S. Rothmund;
- *Laboratoire Méthodes d'Analyses pour le Traitement d'Images et la Stéréorestitution* (MATIS), Institut Géographique National en Ecole Nationale des Sciences Géographiques: M. Pierrot-Deseilligny

- *Laboratoire de Géologie de Lyon – Terre, Planètes, Environnement,* Université Claude Bernard Lyon : P. Allemand
- *Centre de Recherches en Environnement Terrestre* (CRET), Université de Lausanne (C. Michoud, A. Abellan, M. Jaboyedoff)

Chapter 1 - Setting the scene: Remote sensing techniques for landslide monitoring. Review and selection criteria for users

Objectives: This chapter provides a review of innovative remote sensing techniques that have emerged during the last decade and which are potentially useful or already operational for landslide monitoring. First, criteria are proposed to compare their capabilities and, second, guidelines for the selection of the most appropriate technique for monitoring different landslide types, displacement rates and environmental settings are provided. The guidelines are furnished to aid operational decision making, and include information on the spatial resolution, accuracy and coverage, data and processing costs, and maturity of the method.

This chapter is based on:

- Stumpf, A., Malet, J.-P., Kerle, N., Michoud, C., Tofani, V., Segoni, S., Michoud, C., Jaboyedoff, M., Casagli, N. In Review. Selecting appropriate remote sensing techniques for landslide monitoring: Review and selection criteria for users. *Earth Science Reviews.*
- Stumpf, A., Malet, J.-P., Kerle, N. (Eds.) 2011. Creation and updating of landslide inventory maps, landslide deformation maps and hazard maps as input for QRA using remote-sensing technology. SafeLand Deliverable D4.3. *Available at http://www.safeland-fp7.eu.*
- Stumpf, A., Kerle, N., Malet, J.-P. (Eds.) 2011. Guidelines for the selection of appropriate remote sensing technologies for monitoring different types of landslides. SafeLand Deliverable 4.4. *Available at http://www.safeland-fp7.eu*

1.1 *Introduction*

In recent years, several technical and scientific improvements have increased the potential use of remotely-sensed data (e.g. ground-based, airborne and spaceborne techniques) for mapping and monitoring landslides (Metternicht et al., 2005; Delacourt et al., 2007; Joyce et al., 2009; Jaboyedoff et al., 2010; Guzzetti et al., 2012). The reliability and effectiveness of the available instruments have been enhanced, providing a wide range of suitable solutions to measure different quantities such as displacement rates (Abellán et al., 2009; Heid and Kääb, 2012; Travelletti et al., 2012) or volumes (Chen et al., 2006; Martha et al., 2010b; Oppikofer et al., 2012). Relevant advances have also been made in the modelling and in the interpretation of ground movements (Clague and Stead, 2012) and the derivation of mechanical parameters from remote sensing observations (Booth et al., 2013; Travelletti et al., In Review). As a consequence, users can select the most appropriate observation strategy from a variety of technologies and analysis methods taking into account technical issues, economic budgets, environmental factors and specific features of the targeted landslide.

Among numerous improvements in sensor technology and analysis methods, several advances stand out. Recent launches of a new generation of optical and Synthetic Aperture Radar (SAR) satellites with shorter repeat-pass cycles and higher spatial resolutions (e.g. Pleiades 1A&1B, Spot 6&7, WorldView-2, Geoeye-1, TerraSAR-X, Cosmo-Skymed, Sentinel 1&2, ALOS-2) result in better capabilities to acquire data over wide areas shortly after major triggering events and disasters and at regular intervals. Related methods for the processing of optical (Guzzetti et al., 2012; Heid and Kääb, 2012; Poli and Toutin, 2012) and SAR (Hooper et al., 2012) satellite images enable to exploit the abundant data more efficiently and with greater accuracy. For active ground-based remote-sensing systems such a Light-detection and Ranging (LiDAR) and ground-based SAR (GB-SAR), several commercial systems with improved performance have become available over the past few years (Jaboyedoff et al., 2010; Intrieri et al., 2012). Unmanned Aerial Vehicles (UAVs) and consumer-grade digital cameras have become more affordable and related photogrammetric processing methods (Snavely et al., 2008; Furukawa and Ponce, 2010; Rothermel et al., 2012; Deseilligny et al., 2013) have reached a higher degree of automation accessible also to non-experts in the field of photogrammetry.

Many of these new developments have already been consolidated in operative monitoring scenarios, but enhancements are proposed at a fast rate as scientific research continuously pushes remote sensing techniques further, mainly because of the need to establish warning systems (Allasia et al., 2013; Intrieri et al., 2013) and services.

Despite technological advances, it remains complex to evaluate the applicability of different techniques for a monitoring campaign at a specific site and difficult to obtain a complete overview of their potential and limitations for the process at hand. One of the main limitations is the absence of established criteria for the selection of the most appropriate remote sensing data and processing methods. Another limitation is the commonly noticeable gap between the most advanced scientific improvements proposed at the academic level and the techniques currently employed in operational warning systems by the risk managers.

The motivation of this work is linked to the fast evolution of technologies observed in the last ten years which had a strong impact on landslide mapping and monitoring, and drove the current understanding of slope movements.

Guzzetti et al. (2012) have recently provided a detailed review of classical and innovative methods for regional landslide inventory mapping including remote sensing as well as field methods. Among many new valuable tools, they emphasized the use of VHR

satellite images and LiDAR but also the great need for standards and thorough uncertainty analysis of the resulting maps. In earlier reviews, Metternicht et al. (2005) treated the use of remote sensing for landslide hazard assessment while van Westen et al. (2008) provided an overview of geospatial datasets for landslide risk assessment. A brief summary of remote sensing techniques for the analysis of landslide kinematics has also been provided by Delacourt et al. (2007) and Jaboyedoff et al. (2010) gave a detailed review of the application of LiDAR.

In contrast to previous works, this article targets in particular technologies for landslide monitoring (as defined in section 1.2.1) and details criteria that should be considered for the selection of the most appropriate techniques. The proposed criteria (section 1.2.2) consider aspects of the landslide process (type, size, displacement rates, etc.), technological issues such as the measurement accuracy and the temporal resolution, as well as external factors such as financial constraints and risk management strategies. The article does not target to provide a comprehensive review of the immense body of literature on landslide monitoring but provides an overview of available techniques and the most recent innovations (section 1.3).

The outcome of this work is a set of inter-related graphs and guidelines that can be used by scientists and risk managers to obtain an overview of methods and technologies suitable for their particular needs (section 1.4). Some current applications and challenges are discussed in section 1.5.

1.2 *Definitions: landslide monitoring and decision-aiding criteria*

1.2.1 Landslide observation

Monitoring is essential to understand and predict the behaviour of landslides, and forecast the nature and timing of the slope deformation and potential catastrophic failures. Monitoring is defined here as the periodic acquisition and analysis of a series of observations over time to gather information on spatial and temporal changes of relevant parameters. Ideally, the spatial extent, the observation intervals (irregular or regular; campaigns or continuous operating systems) and the employed techniques are determined at the beginning of a monitoring campaign to achieve a spatially and temporally consistent coverage on a repetitive basis. The observation intervals may range from several months to near real-time observations (Figure 1-1a), and are in the latter case frequently integrated into warning systems. This definition excludes unique observations and most change detection techniques, which are generally used for landslide detection, mapping and fast characterization in response to large triggering events. Detection is understood as the recognition of new landslides from space- or airborne remote sensing data; event-based rapid mapping is defined as fast semi-automatic image processing for change detection and target detection; and fast characterization is defined as the retrieval of information on the failure mechanism, the volume involved, and the run-out length for a quick assessment as a base for hazard for emergency planning.

1.2.2 Criteria to select the most appropriate remote sensing techniques

The choice of the most appropriate monitoring techniques is conditioned by a number of different factors. Landslide-related criteria comprise the landslide type, size and expected displacement rates. External criteria include the configuration of the site, the surface

conditions, financial and logistic constraints, the current risk management phase and the scientific objectives of the study. Technological criteria correspond to the capability of particular remote sensing techniques. Many of them are interrelated (Figure 1-1b-c).

Technological criteria inherent to different monitoring techniques are the main focus of this review. The considered criteria are the measurement accuracy, measurement distances, ground sampling distances, spatial coverage, temporal resolution, quality of the obtained measurements, costs for data acquisition and processing, expected elaboration time, portability of the acquisition system and maturity of the technique. For each quantitative criteria, parameter values are provided where possible; otherwise, nominal scales are adopted for qualitative criteria such as the information content of the measurements and the portability of the system (Table 1-1). Typical ranges are provided to represent the variability that results from different sensor systems, processing algorithms, noise, operator experience and other factors.

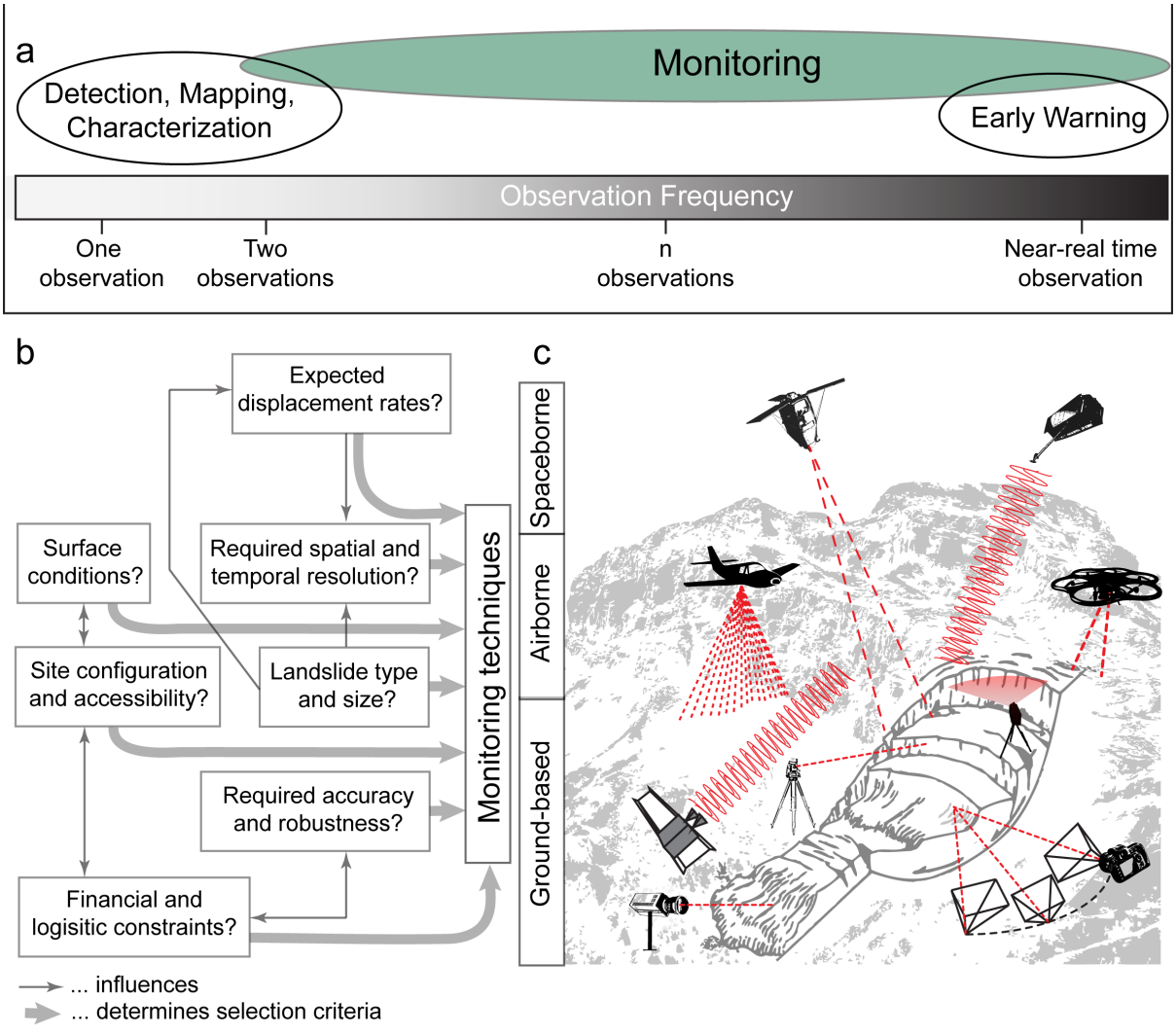


Figure 1-1: Definition, criteria and platforms for the remote sensing of landslides (a) Relationship between observation frequency and different tasks in landslide investigations. (b) External and landslide-related criteria that determine the most appropriate choice of (c) platforms and sensor systems for landslide monitoring techniques.

Table 1-1: Catalogue of evaluation criteria.

Technological criteria		
Criteria	**Scale range**	**Explanation**
Spatial coverage	Point, local (e.g. slope), regional	Typical scales at which the measurements are carried out.
Information type	1D	One component of the displacement or change along a spatial axis or along the Line-Of-Sight (LOS) of the sensor.
	2D	2D displacement (mostly horizontal components) at a point or spatially distributed 3D displacement field
	3D	3D displacement at a point or spatially distributed 3D displacement field.
	Volume	Volumes and volume changes of the moving mass.
	Surface features	Time and space evolution of surface features.
Spatial resolution	mm to hm	Typical spacing of individual measurements
Temporal resolution	seconds to months	Typical time lag between individual measurements.
Distance to target	m to km	Distances to target at which the measurement device can be employed. This category is not relevant for satellite systems that operate in fixed orbital heights.
Measurement accuracy	mm to dm, mm^3 to 10 m^3 for volumes	Accuracy of the measured quantities such as displacement rates, volumes and the location of surface features
Operation mode	Continuous	Automatic measurements can be carried without human intervention for long time periods.
	Campaigns	Measurements require regular human intervention and are thus typically carried at intervals of several days, weeks or months.
Approximate elaboration time	minutes to month	Approximate time lag between the measurement of the system and the final results.
Approximate costs	10 to 100.000 €	Typical costs including installation and operation of the system as well as the data processing
Technological maturity	Concept	Technical design and potential applications have been proposed.
	Prototype	Working prototypes have been tested in a limited number of experiments.
	Case-studies	Operating systems have been tested for landslide monitoring in the field for short time periods.
	Commercial	Working systems and processing software are commercially available and have been tested on several landslides for long time periods.
	Mature	Potential and limitations of the technique are well understood and it has been applied in established monitoring programs.

Table 1-1: Continued.

Landslide-related factors		
Landslide type Surface displacement rates Monitoring scale	Not applicable	The technique is not useful for this particular category.
	Probably not applicable	It is very unlikely that the technique is useful for this particular category, however exceptions may exist.
	Applicable in few cases	The technique could be applicable but restrictions must be expected. Possible alternatives should be considered.
	Suitable in many cases	The technique has been used in several case studies for the same landslide type/ displacement rate/ scale. Further criteria should be carefully checked before decision is made.
	Ideal in many cases	In many applications, this technique has provided excellent results for the same landslide type/displacement rate/scale.
External factors		
Surface conditions	-	Included in the discussion in section 1.3
Site configuration	-	Included in the discussion in section 1.3

To assess landslide-related factors, all available knowledge about the movement patterns should be compiled and a number of criteria for the choice of the remote sensing technique should be considered. Three landslide-related criteria can be distinguished. The scheme of Cruden and Varnes (1996) provides a process-based classification into fall, topple, rotational slide, translational slide, spreads and flows (Figure 1-2a). Complex landslides types combine several of those movement patterns. Understanding the underlying mechanical processes is important to evaluate in which parts of the slope displacement as well as volume- and surface changes are likely to occur. The applied remote sensing technique should capture at least the main component of the change which typically concentrates in a certain direction. Particular robust techniques should be adopted if a loss of coherence of the moving mass is likely. Since the size and spatial distribution of the targeted landslides vary considerably, the scale of the study typically constitutes a trade-off between spatial coverage and detail. Most available ground-based techniques target local measurements, whereas spaceborne techniques cover large areas (PST-A, 2008; Cuenca et al., 2011; Cigna, 2012).

The expectable displacement rates are closely linked to the landslide type and constitute a further criterion. Possible displacement rates (Figure 1-2b) vary by orders of magnitude ranging from a few millimetres per year to several meters per second (Cruden and Varnes, 1996) and, therefore, have a strong influence on the applicable technique and the choice of the observation frequency. Displacement rates often change over time and in response to rain, snowmelt or ground-shaking. The evolution of the slope movement (Figure 1-2c) should be reviewed in order to obtain a reliable estimate of the range of possible velocities. For previously in-active or non-investigated areas where little historic information is available, inventory maps and susceptibility maps (e.g. spatial occurrence) need to be prepared in advance to guide the monitoring efforts. The applicability of different remote sensing techniques according to the landslide type, expected displacement rates and the monitoring scale is evaluated with a rating scale including “not applicable”, “probably not applicable”, “applicable in few cases”, “suitable in many cases” and “ideal in many cases”.

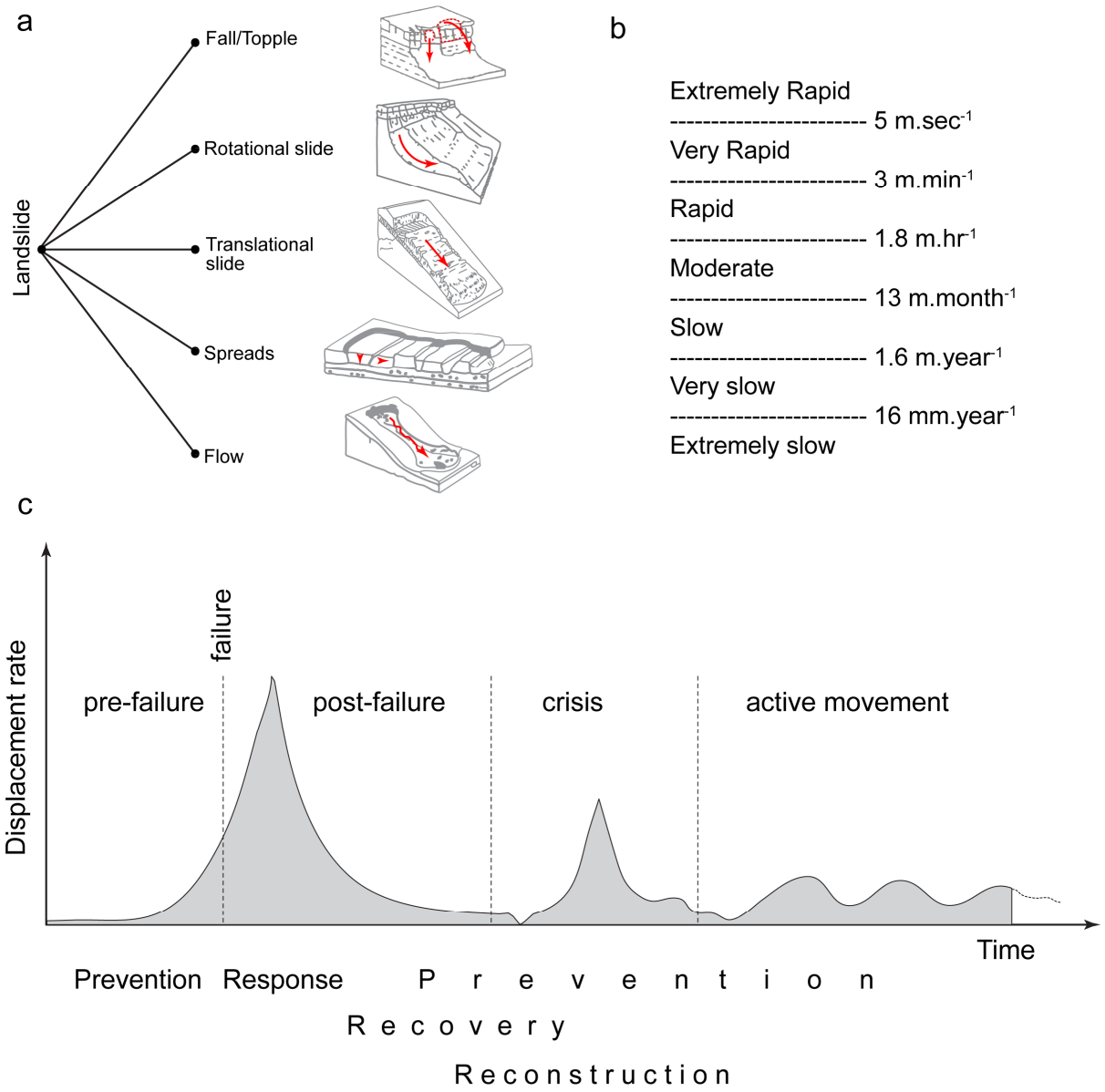


Figure 1-2: First order landslide classification according to (a) the type of movement and (b) the displacement rates (modified after Cruden and Varnes, 1996). Note that falls and topples are different failure mechanisms but are grouped together here since the site geometries and main component of the displacement vector are very similar. (c) Generalized phases of landslide activity (Vaunat and Leroueil, 2002) and typically corresponding risk management phases.

External factors are closely interlinked and include the site configurations, the surface conditions (vegetation, snow, and soil humidity), as well as financial and logistic constraints. Such issues are discussed together with the technological aspects in section 1.3.

In practice, the current risk management strategy and the scientific objectives of the study play an important role. The classical risk management phases are prevention, response, recovery and reconstruction (Alexander, 2002) and in the context of landslide investigations they are closely related to different phases of activity (Figure 1-2c). The usefulness of a technique for prevention (e.g. early warning) is governed by its capability to acquire measurements with high robustness, accuracy and temporal resolution, whereas during other management phases, criteria such as spatial coverage and instrument portability can gain

greater importance. In practice financial and logistic constraints are probably the most crucial aspects when setting up and operational monitoring systems over longer time periods. The development of generally applicable scenarios for the use of remote sensing techniques in different risk scenarios would have to involve complex economic, political, organizational and ethical aspects and is beyond the scope of this article. However, costs, elaboration time and portability of different remote sensing techniques are assessed and provide criteria that can be confronted with the available resources.

1.3 *Remote sensing techniques for landslide monitoring*

According to the utilized platforms, remote sensing can be grouped in spaceborne, airborne and ground-based techniques. Among hundreds of currently operational EO satellite systems, polar-orbiting SAR and optical satellites with high- and very high spatial resolution are the most relevant for landslide monitoring. Airborne platforms are very diverse and range from airships operating at up to 40 km altitude to remote controlled platforms flying only a few meters above the ground (Kerle et al., 2008). Classical airborne surveys continue to be a valuable tool for landslide mapping at regional scales, while UAVs are more adapted for local acquisition at narrow time-intervals that are often desirable for monitoring campaigns. Ground-based acquisitions can be carried out from fixed view points, with handheld devices or from mobile platforms (swimming vessels, cars). Remote sensing techniques such as sonar and seismic profiling are frequently used for the mapping and characterization of submarine landslides (Prior et al., 1982; Canals et al., 2004; Lastras et al., 2006; Talling et al., 2007; Hough et al., 2011). Since monitoring is typically not an issue for submarine landslides, these latter techniques are not included in this review.

1.3.1 Digital Image Correlation (DIC) from space- and airborne sensors

High-resolution optical and SAR images from satellite and airborne platforms are commonly used for landslide inventory mapping (Guzzetti et al., 2012). Related image analysis techniques exploit mainly multi-spectral data (Martha et al., 2010a; Stumpf and Kerle, 2011; Mondini et al., 2013; Stumpf et al., 2013a), time-series of panchromatic images (Martha et al., 2012; Lacroix et al., 2013) and in some cases also high-resolution SAR amplitude images (Singhroy and Molch, 2004). Though hyperspectral imaging has some potential to provide relevant indicators on the evolution of soil properties (Sterzai et al., 2010) and vegetation disturbances, the use of time-series for landslide analysis has not yet been explored.

For monitoring purposes, image matching can be used to derive 2D displacements from satellite and airborne images. If images from multiple view angles are available, it is also possible to reconstruct the topographic surface, quantify surface changes and measure 3D displacements. 2D displacements are commonly derived using Digital Image Correlation (DIC) techniques from series of panchromatic satellite images or high resolution SAR amplitude images *[spaceborne DIC]* and from series of aerial images *[airborne DIC]*.

Space- and airborne DIC with optical images has been employed in numerous case studies using both VHR satellite images and aerial photographs (Powers et al., 1996; Kääb, 2002; Casson et al., 2003; Delacourt et al., 2004; Leprince et al., 2008; Delacourt et al., 2009a; Debella-Gilo and Kääb, 2011; Mackey and Roering, 2011; Booth et al., 2013) to measure 2D displacement with decimetre accuracy. A comparison of several image matching algorithms for optical images is provided in Heid and Kääb (2012) The quality of the derived displacement fields varies significantly among different matching algorithms but the

geometric accuracy of input data, and post-processing of the derived displacement fields is in many cases more important for the final accuracy. For landslide monitoring, an image resolution of 2.5 m (e.g. SPOT 5) and higher should be considered. Automatic matching algorithms can theoretically provide the sub-pixel precision at 1/50 of a pixel but due to noise and co-registration errors a maximum accuracy of 1/5 of pixel should be expected in practical applications. DIC is more robust in tracking large displacements (several $m.yr^{-1}$) than other techniques (such as DInSAR for instance) but temporal decorrelation still becomes an issue if surface changes are too large (i.e. no homologous features are preserved). Optical image acquisition is weather-dependent and suitable images are often only available at intervals of several months or years. In principle, it is possible to combine images from different sensors (Delacourt et al., 2004) but, as a consequence, decorrelation is more likely and image series acquired with the same sensor are, therefore, preferable. Decorrelation also arises over vegetated areas even if the time-lag between the acquisitions is short (e.g. weeks).

Most DIC techniques use a window-based search to match areas with a similar appearance among two images. The window size typically ranges between 3x3 and 64x64 pixels. A larger window size is more robust (e.g. better correlation over areas with large vegetation patches) but also reduces the resolution of the obtained vector field. An important aspect of the image processing preceding DIC is precise co-registration and ortho-rectification which typically requires a number of well distributed tie points (image to image), ground-control points (GCPs, image to ground) with known 3D coordinates and an adequate surface geometry (Digital Elevation Models; DEM). Ideally, each image should be ortho-rectified with a DEM representing the surface at the time of the image acquisition. For practical reasons, DEMs are typically not updated at narrow time intervals (< 1 yr.) and therefore slope parallel displacement (no changes in z) is assumed when the same DEM is used for the ortho-rectification of multiple images. Especially at off-nadir view angles, a deviation from this hypothesis and inherent DEM errors reduce the accuracy of the displacement estimates (Van Puymbroeck et al., 2000). However, it has been demonstrated that generally those errors are not severe and accurate displacement measurements can be obtained even without GCPs and a coarse external DEM (Berthier et al., 2005; Stumpf et al., Submitted) if the relative orientation of the images is accurately modelled with tie points. Using slope parallel displacement as hypothesis allows converting 2D vectors into 3D displacement estimates through projection on the DEM but for many landslides important changes in the z-axis will introduce severe errors (Reeh et al., 2003)

For more than a decade, *spaceborne DIC* has also been applied for measuring co-seismic surface displacements (Michel et al., 1999; Erten et al., 2009) and glacier motion (Strozzi et al., 2002) from SAR amplitude images. The terminology SAR Offset tracking is then also used as an alternative to DIC. The basic principles of DIC applied on optical and SAR amplitude images are similar but two main differences are noteworthy. First, due to the distinct physics of SAR imaging, the correlation between image patches is sometime evaluated with a different similarity measure than for optical images (Michel et al., 1999; Erten et al., 2009). Second, the 2D displacements do not correspond to horizontal components but to the range and the azimuth. With most SAR satellite systems, the ground resolution in the azimuth is generally lower than in the range direction (~ factor 2) resulting in inferior results in the azimuth direction. For applications over wide area in high-mountain areas, a correction of topographic effects is recommendable (Kobayashi et al., 2009). For applications in flat terrain or areas with rather small topographic variations, such corrections are dispensable if the perpendicular baseline between the images is small (1-200 m, Raucoules et al., 2013). To avoid resampling artefacts, it is also preferable to estimate the

displacement first in the image geometry and project the derived measurements afterwards to the ground geometry (De Michele et al., 2010). SAR Offset tracking can be applied on SAR images in L-, C-, and X-band but denser (spatially and temporally) and more accurate measurement can be obtained with VHR X-band satellites. TerraSAR-X and CosmoSkymed are also more flexible in acquisition programming and thus have better capability to acquire images for the same site from different orbits, and thus view angles (Raucoules et al., 2013).

Ascending and descending orbits of SAR satellites can be combined to retrieve full 3D displacement (Rocca, 2003). While this idea has been realized for large scale co-seismic deformation (e.g. De Michele et al., 2010), such application to landslides is still limited. A first application to landslide monitoring is reported in Raucoules et al. (2013) measuring displacement of up to 20 $m.yr^{-1}$ with a precision of 0.1 m. Combining acquisitions from ascending and descending orbits allows measuring full 3D displacements, which is particularly valuable for landslides with complex displacement patterns.

Though several airborne SAR systems exist (Reigber et al., 2013), SAR Offset tracking has been used almost exclusively with satellite images. A promising example for measuring glacier motion from airborne SAR is given in Prats et al. (2009). High-resolution airborne SAR has a great potential for displacement monitoring (DIC and interferometry), estimation of soil parameters, and the investigation of the 3D structure of vegetation and man-made objects. However, currently the processing chain is still rather complex and the costs for repeated local and regional surveys are very high (Reigber et al., 2013).

1.3.2 Stereo-photogrammetry from space- and airborne sensors

Series of stereo-images can be exploited to reconstruct the surface topography and monitor its evolution over time. Airborne stereo-photogrammetry [*airborne SP*] has been applied in many studies to analyse archived aerial photographs typically for one particular landslide (Henry et al., 2002; Walstra et al., 2004; Casson et al., 2005; Brückl et al., 2006; Walstra et al., 2007; Baldi et al., 2008; Dewitte et al., 2008; Schwab et al., 2008; Kasperski et al., 2010; Prokešová et al., 2010; Fernández et al., 2012). The parallax shift between multiple stereo-pairs allows reconstructing the surface and vertical changes; associated orthophotographs can be used to track horizontal displacements. Such techniques typically achieve decimetre accuracy (sometimes metric with older aerial photographs) and provide very valuable information about horizontal displacement and vertical changes in the past. However, they are somewhat limited to monitor the current evolution since aerial surveys are only carried out at large time intervals (typically several years).

UAV photogrammetry can be considered as a special case of *airborne SP* and allows more flexible and cost-effective image acquisition than classical aerial surveys and several examples of UAV-based landslide monitoring have been provided (Eisenbeiss, 2009; Niethammer et al., 2010; Niethammer et al., 2011a; Rothmund et al., 2013). Niethammer et al. (2011a) reported a vertical RMSE of 0.3 m when compared to terrestrial LiDAR and uncertainty of ~0.5 m in the horizontal component. Employing low-cost UAV for landslide monitoring on a regular base and in response to disasters is still complicated by the fact that many GCP targets have to be distributed over the study site. The direct georeferencing with integrated GPS and IMU is an active field of research and recent studies show that decimetre accuracy of the georeferencing can be achieved (Turner et al., In Press). In this context, the additional costs for a larger UAV and GPS components may be worthwhile to circumvent the need for a large number of GCPs. An interesting aspect regarding the image processing is the increasing availability of open-source and low-cost software (e.g. Snavely et al., 2008; Furukawa and Ponce, 2010; Deseilligny et al., 2013) that enable also inexperienced users to

obtain high quality results. UAV systems that integrate other sensor types such as LiDAR and SAR are under active development (Nagai et al., 2009; Wallace et al., 2012) but to the best of our knowledge no such system has yet been employed for the investigation of landslides. The very high spatial resolution of UAV images (typically ~5 cm) does not only allow quantitative analyses of surface changes but also permits to analyse the distribution and evolution of small surface features (fissures, lobes, boulders) which are useful indicators for the underlying mechanical processes (Walter et al., 2009; Stumpf et al., 2013b). There is currently still some uncertainty about legal aspects for the operation of UAVs in public airspace and regulations (if in place) vary from country to country. A useful overview of national and international policies can be found in Maddalon et al. (2013) and national authorities might be consulted for clarification.

Spaceborne stereo-photogrammetry [*spaceborne SP*] follows similar principles than classical *airborne SP*, whereas the spatial resolution (currently 0.5 x 0.5 m), geolocation accuracy of the input images and results of the processing are generally lower than with *airborne SP*. Until recently, *spaceborne SP* has mainly been used for volume estimation after catastrophic failures and vertical errors (in the range 2 – 6 m) have been reported (Tsutsui et al., 2007; Martha et al., 2010b). Consequently, volume estimates with an uncertainty of ~ 20% can be achieved after additional co-registration of the extracted surface models (Martha et al., 2010b).

Despite promising case studies on *spaceborne SP* in the context of landslide monitoring, its application for long-term monitoring is, so far, limited. This can be attributed to the relatively long intervals between suitable images and the common assumption that a large number of GCPs is needed. However, it has been demonstrated that a single GCP is often sufficient to correct most of the bias in the sensor model (Fraser and Hanley, 2005) and in some cases even dispensable with only minor losses in the absolute positional error (Reinartz et al., 2006). It should also be noted that tie points which can be extracted automatically suffice to accurately recover the surface models and relative displacement for different epochs (Stumpf et al., Submitted). However, the absolute position of the derived surface will comprise errors of the sensor model which may pose problems for the validation and integration with ground measurements. The latest generation of VHR satellites have significantly shorter repeat-pass cycles (~ 5 days) and enhanced stereo capabilities at spatial resolutions of 0.5 m. Resulting datasets can be exploited to reduce the measurement interval to a few months with horizontal errors of ~0.1 m even without GCPs (Stumpf et al., Submitted).

The possibility to obtain 2D displacement fields and volume estimates makes DIC and stereo-photogrammetry complementary technologies especially for tracking displacement rates of > 1 $m.yr^{-1}$. Ideally, time-series should be acquired at constant sun- and view-angles. Before purchasing archived images or for planning acquisitions, high-resolution DEMs should be exploited to simulate and avoid cast shadows. However, despite considerable advances in stereo-photogrammetric algorithms, the processing still depends on highly specialized software and the processing of time-series cannot be easily automated. Open-source projects for image-matching and satellite photogrammetry are under active development and will provide good alternatives in the near future (Tinel et al., 2012; Deseilligny et al., 2013).

Radargrammetry can be considered as a special case of *spaceborne SP* since it exploits the parallax shift between stereo-pairs of SAR amplitude images. While the principles of this approach have been laid out several decades ago, only recent high resolution SAR satellites (Cosmo-Skymed, TerraSar-X, RADARSAT-2, and ALOS-2)

provide sufficient resolution to obtain high quality surface models. Several techniques to increase accuracy and alleviate the computational intensive process have been proposed recently but errors of at least 2 m in flat terrain (and worth in mountainous terrain) are observed (Balz et al., 2013; Capaldo, 2013; Nascetti, 2013) and limit the applicability for landslide monitoring.

1.3.3 DIC and stereo-photogrammetry from ground-based sensors

Similar to space- and airborne techniques, terrestrial stereo-photogrammetry relies on basic principles of image matching and stereo-vision. It targets the reconstruction of surfaces and displacement from multiple images acquired on the ground. Early works on the application of this technique to landslide monitoring date back to before the 1980s (see Chandler and Moore, 1989 ; Lane et al., 1993 and refrences therein) but still required metric cameras and a significant amount of expertise. The emergence of affordable consumer-grade SLR cameras greatly simplified image acquisition, storage and processing, but also raised new issues for the photogrammetric processing of images captured with non-metric cameras. Algorithms and software implementations were enhanced mainly by the photogrammetric and computer vision community focusing, respectively, on higher precision or greater automation (Fraser and Edmundson, 2000). In particular, self-calibration now enables an accurate calibration of SLR camera lenses (Remondino and Fraser, 2006) and the convergence of photogrammetric and computer vision techniques provides both high automation and accuracy (Barazzetti et al., 2009).

Image correlation of monoscopic time-series [*terrestrial DIC*] is commonly applied in soil and fluid mechanics for laboratory scale measurements (Adrian, 2005; Pan, 2011), whereas still relatively few field applications to landslides are reported in the scientific literature. Recently, Travelletti et al. (2012) demonstrated the possibility to monitor 2D surface displacement with a fixed installed camera including a detailed analysis of the sources of potential errors. The relative error of the measurement was generally below 20% of reference ground-based GNSS observations, and resulted mainly from the use of a mono date DEM used to convert monoscopic measurements from image into ground geometry. They showed that the camera programming should be adapted to the seasonal variations of the sun angles and quantified errors related to the scene depth, the view angle and the camera orientation. A similar monitoring approach was recently applied in Motta et al. (2013).

The problem of a mono date DEM can be resolved using a stereo- or multi-view camera setups (*terrestrial SP*) as already applied occasionally for glacier monitoring (Kaufmann, 2012; Rivera et al., 2012). Kaufmann (2012) reported an accuracy of ± 5 $cm.yr^{-1}$ indicating a great potential of time-lapse stereo for landslide monitoring. To the best of our knowledge no study has been conducted yet in this direction.

An alternative to fixed camera setups is the acquisition of photographs with a handheld camera from different view angles. This allows the construction of surface models at multiple time steps and resolves the dependency on external DEMs. The related approaches for 3D reconstruction are commonly referred as Structure-from-Motion and Multi-View stereo-photogrammetry (SfM-MVS). Images can be recorded at low cost and numerous software are available commercially and in the public domain (Snavely et al., 2008; Furukawa and Ponce, 2010; Rothermel et al., 2012; Deseilligny et al., 2013; Wu, 2013). The technique has been applied in several geomorphological studies (James and Robson, 2012; Westoby et al., 2012; Bretar et al., 2013) but applications to landslide monitoring have rather focused on UAV surveys (Section 1.3.2). An example for the application of terrestrial SfM-MVS for landslide monitoring has recently been provided by

(Stumpf et al., In Review-a) suggesting an RMSE between 0.05 and 0.2 m (with a ground-based LiDAR survey as a reference) of the resulting models. The study showed the possibility to quantify surface changes, displaced volumes and 3D displacement vectors. Achieving a relatively homogeneous accuracy over the entire slope still depends on some optimization of the acquisition geometry by the operators. However, the quality of the results closely resembles terrestrial and airborne LiDAR acquisitions and hence SfM-MVS should be considered as a versatile and cost-efficient alternative to costly laser scans.

For larger displacement rates (> $m.day^{-1}$), the temporal resolution of time-lapse photography is often to low and video cameras are employed instead. This technique has been used many times to study the velocity and behaviour of debris flows. Early works include the application of spatial filtering methods to measure surface velocity in image coordinates (Itakura and Suwa, 1989; Inaba et al., 2000). In Arattano and Marchi (2000), results of a simple method for projecting the image displacements into 3D ground geometry are presented. Compared to ultrasonic measurements, the error was never larger than 9% but also depends on the availability of an external surface model after the debris flow event. A bibliography of debris-flow studied with videos can be found in Itakura et al. (2005) and more recent examples in large scale experiments are given in Iverson et al. (2010), McCoy et al. (2010) and Kean et al. (2011). Videos also provide important qualitative evidence for the behaviour (surges, presence of boulders) of the events (Arattano and Marchi, 2008; Prochaska et al., 2008).

1.3.4 Electronic distance meters and total stations ground-based surveys

Electronic Distance Meters [*EDM*] and theodolites are classical tools for landslide monitoring and are continuously used for displacement measurements at high temporal resolution and centimetre accuracy. *EDM*s are based on the same principles than LiDAR systems but are significantly cheaper and can be operated automatically to measure displacement along the direction of the laser beam at distances up to 1000 m with specifically designed prism targets (Oswald, 2003; Manetti and Steinmann, 2007). Since EDMs only allow to measure distances in the direction of the laser beam, they are typically combined with robotized theodolites (also called tacheometer) in integrated systems [*Total stations*].
For automated measurements with high precision, *Total Stations* are often operated in combination with defined prism targets allowing to measure distances of more than 2000 m (Castagnetti et al., 2013; Giordan et al., 2013) and to obtain an accurate 3D estimate of the position..

Several modern image-assisted total stations can also be operated in a reflectorless mode that allows measuring 3D coordinates of up to 20 points per second (Abed and Abdullah, 2013). Such systems can be considered as a sort of low-density LiDAR that can be used to calculate sediment budgets (Tsai et al., 2012) or to asses global changes in the slope inclination (Yang et al., 2007). A closer integration of *Total Stations* with photogrammetric measurements has a great potential to automate near real-time 3D displacement measurements without the need of prism targets (Scherer and Lerma, 2009) but the implementations of such systems for landslide monitoring has so far been limited (Thuro et al., 2009; Paar et al., 2012). *EDM*s are also integrated in laser range finders which have proven useful for landslide mapping (Santangelo et al., 2010) but the accuracy of such systems is in the range of meters and, therefore, currently too low for monitoring applications.

1.3.5 LiDAR point-clouds from airborne and ground-based platforms

Light Detection and Ranging (LiDAR) technology is available for the geoscience community since the 1990s (Shan and Toth, 2009) and became a commonly used tool during the last ten years. A comprehensive review on the use of LiDAR in landslide investigations is given in Jaboyedoff et al. (2010). Therefore, only some key points and more recent developments are summarized here.

For *airborne LiDAR [ALS]* surveys at regional scale, the laser scanner is usually mounted on an airplane and the spatial resolution varies between meters and decimetres. For higher point density with local coverage, helicopters are often used with a steerable sensor that allows a better coverage of very steep slopes and vertical cliffs. For the scanning landslides under dense vegetation, a high point density is very important since fewer pulses will reach the topographic surface (Razak et al., 2011). The absolute accuracy of the point clouds depends on several factors such as the scanning distance, the incidence angle of the slope and the quality of the georeferencing which is typically based on the position and orientation of the sensor recorded with an on-board GPS and IMU during the flight (direct georeferencing). The absolute accuracy of scanning systems is often specified with 5-35 cm (Optech, 2013) but strongly depends on the flight height and the incidence angle of the laser pulse at ground (Ussyshkin et al., 2008). A detailed analysis of error sources and possible corrections in the context of glacier monitoring can be found in Joerg et al. (2012). A convenient way to correct systematic errors (translation, rotation) between multi-date scans is to use the Iterative Closest Point (ICP) algorithm over stable areas (Kumari et al., 2011) but generally decimetre uncertainties cannot be avoided.

Airborne LiDAR is one of the few remote sensing techniques that allow the investigation of landslide under dense vegetation (Razak et al., 2011; Van Den Eeckhaut et al., 2012) but its value for monitoring is limited by the fact that repeated surveys are cost-intensive. Examples for the monitoring of landslide dynamics with hourly frequency exist (Favalli et al., 2010) but for landslide monitoring the temporal resolution is typically rather yearly and constrained to specific sites (Roering et al., 2009; Mackey and Roering, 2011; Ventura et al., 2011). The temporal resolution of 3D measurements can be significantly increased by combining airborne and terrestrial LiDAR (Ghuffar et al., 2013).

Compared to airborne acquisitions, *terrestrial LiDAR [TLS]* can be employed with a much higher temporal and spatial resolution and has become an important tool for the monitoring of rockfalls, earthflows and debris flows (Jaboyedoff et al., 2010). The accuracy of the scanning device depends on the distance to the object (1-3000 m) and typically ranges between 1 and 2 cm. The detection of millimetre displacements is possible in some cases through averaging over a sufficiently large spatial neighbourhood (Abellán et al., 2009). TLS not only provides the possibility to quantify displacement rates and volumes but also to infer the geometry of past and potential future failures through the analysis of structural features (Jaboyedoff et al., 2009).

Abellán et al. (2013) provide a comprehensive overview of acquisition, pre-processing and analyses steps of TLS for the characterization, volume estimation and displacement monitoring on rock slopes. The most common application of TLS is indeed the quantification of rockfall volumes on steep cliffs (Rosser et al., 2005; Abellán et al., 2011), which, if repeated at high temporal resolution, can be used to establish magnitude-frequency curves (Rosser et al., 2007; Barlow et al., 2012).

For slow-moving landslides, TLS can be employed to derive 1D-3D displacement vectors and different analysis techniques have been developed for this purpose. A simple approach is the interpolation of continuous surfaces and differencing of the components to

measure changes in the vertical direction. This method poses problems if the considered surface is not globally flat and more advanced algorithms have been proposed to compute the displacements along the surface normals of interpolated meshes (EDF-R&D, TelecomParisTech, 2013) or along the surface normals directly estimated on the point clouds (Lague et al., 2013). However, resulting displacements are still only 1D and do not fully exploit the 3D nature of the scanned data.

2D displacement can be measured using DIC techniques on a shaded relief of the interpolated surface (Travelletti et al., 2008) of from synthetic images derived from a projection of the LiDAR point cloud into a common plane (Aryal et al., 2012). A drawback of those methods is the underlying assumption that all displacements are parallel to the slope (or the defined plane) and still not all components can be recovered.

The currently most commonly used technique to recover 3D displacement is a piecewise rigid alignment of point clouds using ICP (Teza et al., 2007) or a least square approach (Monserrat and Crosetto, 2008). Both methods are based on the assumption that some rigid features are preserved in the point-clouds over time. Such approaches can provide displacement measurements (translation and rotation) with centimetre accuracy but usually require significant manual intervention to identify areas that can be matched. Displacement can also be measured manually through the identification of homologous surface features among the point clouds time-series (Oppikofer et al., 2008) but the associated errors might be too high if centimetre displacement should be detected (Oppikofer et al., 2009).

An alternative approach to derive 3D displacement from TLS time-series has been proposed recently in Travelletti et al. (In Review). The method is based on the application of DIC on a 2D image of the interpolated surface and the projection of the derived displacement on the real ground geometry. It yields accuracies equivalent to those obtained with ICP but requires significantly less manual intervention. A further interesting method for 3D motion estimation has recently proposed in Ghuffar et al. (2013) adopting a hierarchical framework for range flow estimation on relative coarse DEMs (> 1m. pixel^{-1}) derived from airborne and terrestrial LiDAR..

1.3.6 SAR interferometry from spaceborne platforms

Spaceborne SAR interferometry is commonly applied to monitor landslides at both regional and slope scales. The SAR imaging process is significantly less weather sensitive than optical systems and yields complementary information (phase, amplitude and polarization) of the radar beam backscattered at the ground (Richards, 2009). Relevant information for landslide monitoring can be extracted with Differential SAR interferometry (DInSAR) and image matching techniques exploiting the phase and amplitude information, respectively.

Classical SAR interferometry *[Classical DInSAR]* exploits the phase difference between a pair of radar images with a temporal baseline that depends on the repeat pass-cycle of the satellite (46 days for ALOS PALSAR to 4 days with Cosmo-Skymed). Numerous case studies have shown that this technique can be used to measure landslide deformation along the Line-Of-Sight (LOS) with accuracies in the range of centimetres to millimetres (Strozzi et al., 2005; Delacourt et al., 2009b; Roering et al., 2009; García-Davalillo et al., 2013; Nikolaeva et al., 2013; Schlögel et al., in elaboration)Some studies have demonstrated the use of DInSAR with C-band (Squarzoni et al., 2003; Bulmer et al., 2006; Riedel and Walther, 2008) and X-band satellite sensors (Motagh et al., 2013) for landslide monitoring. L-band is less sensitive to decorrelation and, therefore, more suitable for classical DInSAR processing (García-Davalillo et al., 2013; Nikolaeva et al., 2013). C-band and X-band are generally more prone to temporal decorrelation resulting from surface changes and large displacements. The

maximum displacement that can be recovered theoretically without ambiguity depends on the wavelength and is 46.8 cm.year^{-1} for L-band, 14.6 cm.year^{-1} for C-band and ~25cm.year^{-1} (11 days revisit time) for X-band. Additionally L-band can penetrate sparse vegetation. At the moment, however, no L-band SAR satellite is operational and until the launch of ALOS-2 (scheduled for 2014) and only historical archives (ALOS 2006-2012) are available. It is possible to overcome those theoretical limits with simple hypothesis on the kinematics of the landslide (only downslope movement) but careful manual intervention in the phase unwrapping is required (Crosetto et al., 2013a).

To overcome some of those issues, *advanced DInSAR* methods have been developed. Since its original introduction by (Ferretti et al., 1999, 2001), many variants of the Permanent Scatterers approach (PS-InSAR) have been proposed (see Colesanti and Wasowski, 2006 and references therein). They all have in common that many interferograms are constructed using one master image out of a time-series of at least 15-20 SAR images. Pixels with a coherent scattering mechanism over time (PS points) typically correspond to infrastructures or rock faces oriented toward the sensor that are used to measure ground displacement with millimetre accuracy. Thereby some of the shortcomings of classical DInSAR such as temporal decorrelation and atmospheric noise can be circumvented. The number of PS points is typically very high in urban environments and reduces over rural and vegetated areas. SBAS (Berardino et al., 2002) is a further approach for interferometric processing that does not rely on the presence of PS points. Instead of selecting one master image the algorithm selects image pairs with small temporal and satellite baseline to construct a series of interferograms that can be combined via singular value decomposition. Compared to classical DInSAR the technique is more robust to temporal and decorrelation and provides accuracies similar to those of PS-InSAR. Indeed PS-InSAR and SBAS are complementary and methods to combine them have been proposed (Hooper et al., 2012). Recent algorithmic enhancements of PS-InSAR (Ferretti et al., 2011) can significantly increase the points density by exploiting also pixels that qualify as distributed scatterers and first applications for landslide investigations have been reported (Tamburini et al., 2011; Iasio et al., 2012). Other techniques are dedicated to the correction of atmospheric effects in small baseline approaches (NSBAS, Jolivet et al., 2011) and denoising of interferograms time-series (MulSAR, Pinel-Puyssegur et al., 2012). It has been demonstrated that their combination can significantly enhance the result of interferometric processing chains for the monitoring of tectonic uplift (Grandin et al., 2012) but yet those techniques have not been applied for landslide monitoring.

InSAR techniques are generally dependent on the presence of targets with a temporal coherent backscattering mechanism. If the natural environment does not comprise such objects artificial targets, corner reflectors can be installed. The most commonly used targets are metallic trihedral corner reflectors. Their in-situ installation involves additional costs and several considerations of the orientation of the slope, the orientation of the reflectors and the distance between the reflectors in relation to the displacement rates (Crosetto et al., 2013a). Several studies have employed corner reflectors using C-band and X-band interferometry for landslide monitoring (Xia et al., 2004; Fu et al., 2010; Bovenga et al., 2012a; Crosetto et al., 2013a).

Both *Classical and Advanced InSAR* techniques are applicable in the different microwave bands but the latest generation of SAR satellites provides several advantages that are noteworthy. TerraSAR-X and CosmoSkymed are satellite constellations operating in X-band and can provide much higher spatial and temporal resolution than previous missions. At a minimum (somewhat exceptional) revisit time of 4 days (CosmoSkymed), the maximal displacement that could be measured without ambiguity may exceed 60 cm.yr^{-1}. The main

advantage of X-band SAR is the enhanced resolution in azimuth and range directions which can reach up to 1 m and significantly increases the density of measurement points. So far, landslide monitoring studies have exploited X-band SAR at average revisit times between 11 and 29 days and displacement rates between 10 $mm.yr^{-1}$ and 12 $cm.yr^{-1}$ have been identified (Notti et al., 2010; Ling et al., 2011; Bovenga et al., 2012b; Herrera et al., 2013; Motagh et al., 2013).

In summary the choice of input data and processing methods should consider the expected displacement rates, the site configuration and the coherence of the movement. However, several common limitations of InSAR techniques must be taken into account for the planning of a monitoring campaign. Though L-band SAR provides capabilities to penetrate sparse vegetation, the results of DInSAR generally degrade significantly in vegetated areas. Similarly no meaningful measurements can be obtained for snow-covered surfaces and areas with strong surface changes (e.g. erosion, accumulation). As mentioned above there is some ambiguity if the displacement rate of the landslide exceeds a quarter of the wavelength of the microwaves. For relatively large coherent movements with well-known displacement patterns, this condition can sometimes be relaxed with manual intervention in the phase unwrapping but requires additional processing time and expertise.

One of the most crucial aspects for the applicability of DInSAR in specific study is the geometry of the site. The polar-orbiting trajectory of the satellites makes the measurement insensitive to any displacement that occurs in the North-South direction. Relatively steep slopes (e.g. > 23°, depending on the LOS geometry, Colesanti and Wasowski, 2006) will suffer from shadowing and foreshortening. For optimal coverage, data from ascending and descending orbits should be combined. Topographic datasets can be exploited in advance to analyse if a suitable imaging geometry is possible (Plank et al., 2012) and additional databases for landcover and infrastructure can be exploited to pre-compute the expected density of PS points (Plank et al., 2013).

Many considerations regarding the site characterisitcs and displacement rates are rather fundamental and will also apply to future satellite missions. However, the upcoming Sentinel-1 mission will provide publically available C-band images at a repeat-pass cycle of only 6 days (at least over Europe) and significantly extend the capabilities to measure faster displacement rates with higher precision and at lower costs (Rucci et al., 2012).

1.3.7 SAR interferometry from ground-based platforms

Most currently available ground-based SAR (GB-SAR) systems are operated in the Ku-band, while also systems with longer wavelength (X-band, C-band, S-band) are being tested (see Iwe, 2012 for a recent review). Longer wavelength are significantly more robust to temporal decorrelation, atmospheric affects and vegetation but provide lower spatial resolution and decrease the accuracy of the distance measurements from millimetre to centimetre (Luzi, 2010). Operational commercial GB-SAR systems are available but the development is on-going and proposed upgrades will significantly lower the systems costs and enhance the imaging frequency (Tarchi et al., 2013). In contrast to real-aperture radar, current SAR systems are faster and lighter (Pieraccini, 2013), whereas developing real-aperture prototypes (Strozzi et al., 2012) may narrow this performance gap in the near future.

Current limitations in the use GB-SAR are decorrelation at high displacement rates and atmospheric disturbances as well as the 1D (LOS) nature of the measurement. A recently proposed analysis method for GB-SAR adopts a classical window-based matching technique on the amplitude images and thereby enables to measure 2D displacement in azimuth and

range direction. The technique has already been applied to monitor landslide displacement and yields sub-pixel accuracy corresponding to centimetre errors (Crosetto et al., 2013b).

The first field campaigns using GB-SAR interferometry for landslide monitoring have been carried out more than a decade ago (Pieraccini et al., 2003; Antonello et al., 2004) and many subsequent studies enhanced the performance of the processing methods (Noferini et al., 2005; Pipia et al., 2008) and extended its use for long-term monitoring and integration into early warning systems (e.g. Casagli et al., 2010; Intrieri et al., 2012). A fixed installation of one entire system is preferable and standardized setups have been proposed (Tofani et al., 2013b) to avoid temporal decorrelation and co-registration errors. Continuous monitoring on surfaces with sparse vegetation and even permanent snow cover is possible (Casagli et al., 2010) but millimetre accuracy is only achieved for bare soil and rock surfaces. Averaging of interferograms and multi-look processing can be used to reduce atmospheric effects and noise from surface changes but compromises the temporal and spatial resolution.

1.3.8 Geophysical measurements from airborne platforms

Airborne geophysics consists in a variety of remote sensing techniques such as gamma-ray spectrometry, airborne magnetics, gravimetry and airborne electromagnetics (Thomson et al., 2007). Most of those techniques have been initially developed for mineral and ground-water exploration and only recently are being exploited for landslide research (Supper et al., 2008; Pfaffhuber et al., 2010; Baroň et al., 2013; Supper et al., 2013).

The helicopter-based system recently used in Supper et al. (2013) and Baroň et al. (2013) includes an electromagnetic measurement system (to provide electrical resistivity within the subsurface down to a maximum 100 m in depth), a gamma ray spectrometer (to measure the concentration of radioactive elements in the first 0.30 m of the subsurface), and a passive microwave antenna. Since the soil water content is closely related to microwave emission the measurements can be used as a proxy for the soil moisture in the first 0.10 m of the subsurface (Supper et al., 2013).

Airborne geophysics is still a rather new technique for the investigation of landslides. In contrast to other remote sensing techniques it does not target to measure displacement rates and volumes but has the advantage to provide valuable information about the depth of the moving mass and the structure of the sub-surface. Supper et al. (2013) were for instance able to localize present and future unstable slopes, explore the general structure of a landslide and the regional geological context. The added value of repeated surveys for monitoring might be limited and constrained by the limited number of world-wide operational systems.

1.4 *Guidelines for the selection of the appropriate remote sensing techniques*

This section provides a detailed evaluation of the proposed criteria (Section 1.2.2) that can be used as rule sets for the choice of the most appropriate technique for different monitoring scenarios.

Most available remote sensing techniques for landslide monitoring can provide local and regional coverage and information about the 3D displacement and volume changes. A notable exception is spaceborne SAR interferometry [*classical and advanced DInSAR*] which only provides displacement measurements in the LOS direction (Figure 1-3a, b). The information that can be derived from airborne geophysics is unique in the sense that a detailed image of the subsurface structures, volumes and physical parameters can be

obtained. Besides quantitative information (e.g. point clouds), LiDAR and optical remote sensing techniques also yield images of the surface that can be interpreted manually or semi-automatically to observe diagnostic features at the surface (fissures, lobes, boulders, vegetation disturbance).

Ground-based techniques yield higher spatial and temporal resolution for local monitoring. For a better interpretation of the spatial resolution (Figure 1-3c), it is important to consider that it refers to horizontal resolution for *spaceborne DIC*, *spaceborne SP* and *airborne DIC*, to the resolution in azimuth and range for SAR techniques, to the 3D points spacing for *ALS*, *TLS* and *terrestrial SP* and to 3D cell resolution for airborne geophysics. While the spatial and temporal resolutions (Figure 1-3c, d) for spaceborne systems are determined by the specifications of the sensor, the resolution of ground-based techniques is variable and only limited by the distance to the target (Figure 1-3e) and, eventually, the available resources for data transmission, storage and processing.

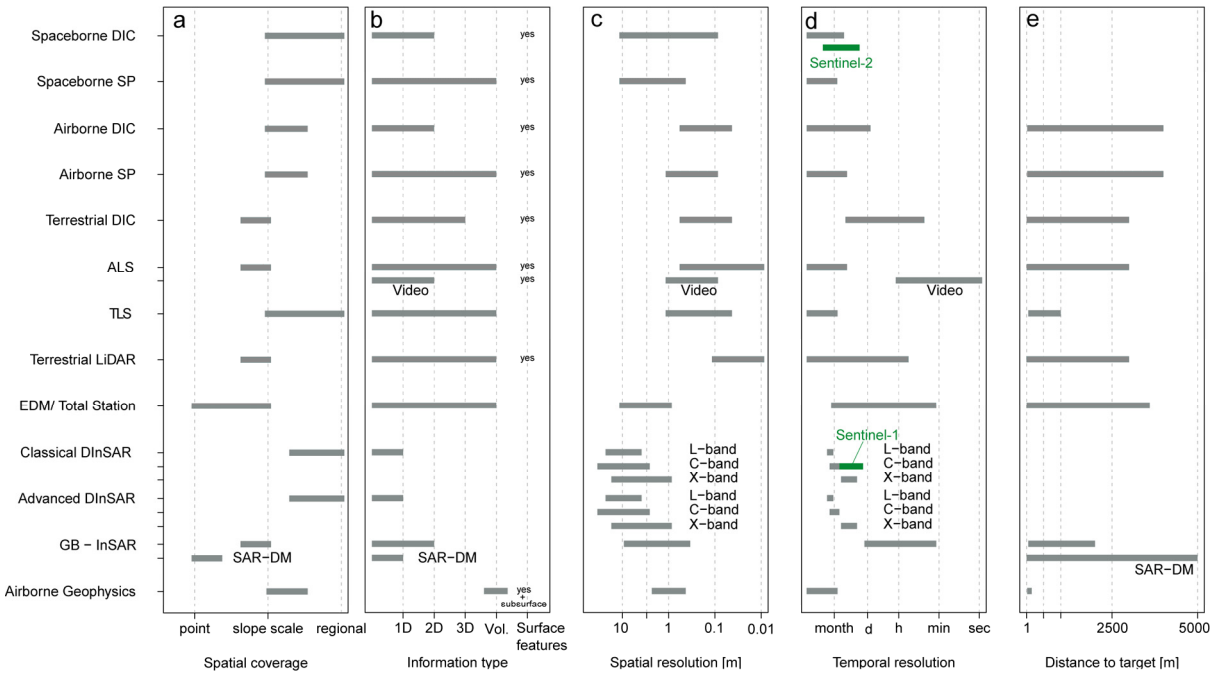


Figure 1-3: Capability of remote sensing techniques for landslide monitoring regarding (a) spatial coverage, (b) resulting information type, (c) spatial resolution, (d) temporal resolution and (e) distance to the target. Sub-categories such as SAR distance meter (*SAR-DM*) and different microwave bands are distinguished for criteria where their value range differs.

Among many relevant satellite missions which are planned or scheduled for the near future (ALOS-2, Cartosat-3, WorldView-3), the ESA Sentinel mission stands out since all captured images will be publicly available without charges (ESA-Copernicus, 2013) and because of its high temporal frequency. Sentinel-2 (scheduled for launch in late 2014) provides optical images with a maximum spatial resolution of 10 m and a repeat pass cycle of 5 days. A comparatively high co-registration accuracy of better than 0.3 pixel is foreseen (Drusch et al., 2012) yielding continuous time-series whose value for DIC monitoring of slow and moderate moving landslide should be explored (Figure 1-4a).

Regarding the measurement accuracy (Figure 1-4a), three main groups of remote sensing techniques can be distinguished. First, ground-based and spaceborne SAR as well as Total Station surveys can achieve measurements with only millimetre and sub-millimetre

errors and are therefore especially valuable for the monitoring of small displacements. X-band SAR is currently one of the most accurate techniques, whereas the costs of data and satellite scheduling are relatively high (Figure 1-4d). C-band SAR with the forthcoming Sentinel-1 (scheduled to be launched in early 2014) has the potential to yield sub-millimetre accuracy (Rucci et al., 2012) at significantly reduced costs (Figure 1-4d). Second, *ALS*, *TLS*, and *terrestrial and airborne* (UAV) *SP* typically achieve centimetre and decimetre accuracy. Third, accuracy of *spaceborne DIC* and *spaceborne SP* is usually not higher than several decimetres. Note that the accuracies refer to the measured quantities such as LOS displacements, heights, 3D point positions and 3D displacements and must be extrapolated to obtain derivates such as volumes. *Airborne geophysics* provides very precise measurements of resistivity, gamma and microwave radiation but the delineation of structural units can comprise uncertainties in the range of a few metres.

The possibility to operate a remote sensing technique continuously without significant intervention (Figure 1-4b) is conditioned by the capability of the sensor to provide continuous time-series and the availability of automatic processing chains. Due to short repeat-pass cycles and automated processing chains, spaceborne SAR systems qualify for continuous measurements. Also, ground-based systems including optical cameras, *GB-InSAR* and *Total Stations* can record data continuously over longer time periods (several months and even longer). Other systems for *TLS* and *spaceborne SP* can only be used for permanent or campaign surveys since issues such as the data volume and the required manual intervention in the processing chain do currently not allow automatic operation.

Figure 1-4c provides an overview of the approximate elaboration time which marks the time lag between the acquisition of data and the completion of the final results. Since the elaboration time depends on the experience of the operator, the size of the dataset (e.g. number of SAR images, size of LiDAR point clouds), requirements of external data sources (e.g. GCPs) and the computational resources, only indicative ranges are provided. It shows that the analysis phase requires several days or weeks with most remote sensing techniques which limits their use for early warning. Automated *Total Stations* and *GB-InSAR*, at the opposite, provide results already a few seconds or minutes after the measurement process and are frequently integrated into Earl-Warning Systems (Casagli et al., 2010; Cardellini and Osimani, 2013; Michoud et al., 2013).

The approximate costs of the technique (Figure 1-4d) include costs for the instrumental equipment, the processing software, the field installation and the required manpower. The costs were evaluated based on publicly available (in 2013) price lists of instruments and sensors, data providers and space agencies, inquiries at private companies and the experience of the partners within the SafeLand European project. In general, the costs represent the current state-of-the-art and are likely to reduce as new spaceborne, airborne and ground-based systems enter the market. Especially free data from the forthcoming Sentinel missions 1 & 2 will nullify data costs for the acquisition of SAR images (Figure 1-4d). For historical analysis, techniques, (such as spaceborne SAR, *spaceborne DIC* and *spaceborne SP* can greatly benefit from large image archives that sometimes span over several decades (e.g. ERS-1/2, airborne photographs after 1945) and are available at low costs or even free of charge. For continuous monitoring with state-of-the-art satellite systems (e.g. TerraSAR-X, Pleiades), costs are still relatively high especially if scheduled acquisitions are requested. This generally also applies for *ALS* surveys and airborne geophysical surveys.

It is however important to note that those techniques are rather designed for regional applications (Figure 1-7) and are therefore relatively cost-efficient when price per area is concerned.

The expenses for the use of *GB-InSAR*, *TLS* and *Total Stations* are dominated by the costs of the hardware and leasing. Shared use among institutions and/or study areas is often appropriate for such systems.

UAVs and especially optical cameras can be considered as low-cost systems and entry-level models are already available for a few hundred EUR. Additional costs for *terrestrial DIC* and *SP* typically arise from fixed installations which require power supplies, data transmission systems and sometimes additional data loggers. The prices for high-end civil UAVs easily exceed 10.000€ but can carry higher payloads that allow the integration of navigational systems and high quality SLR cameras reducing the need for GCPs and generally providing significantly more accurate results.

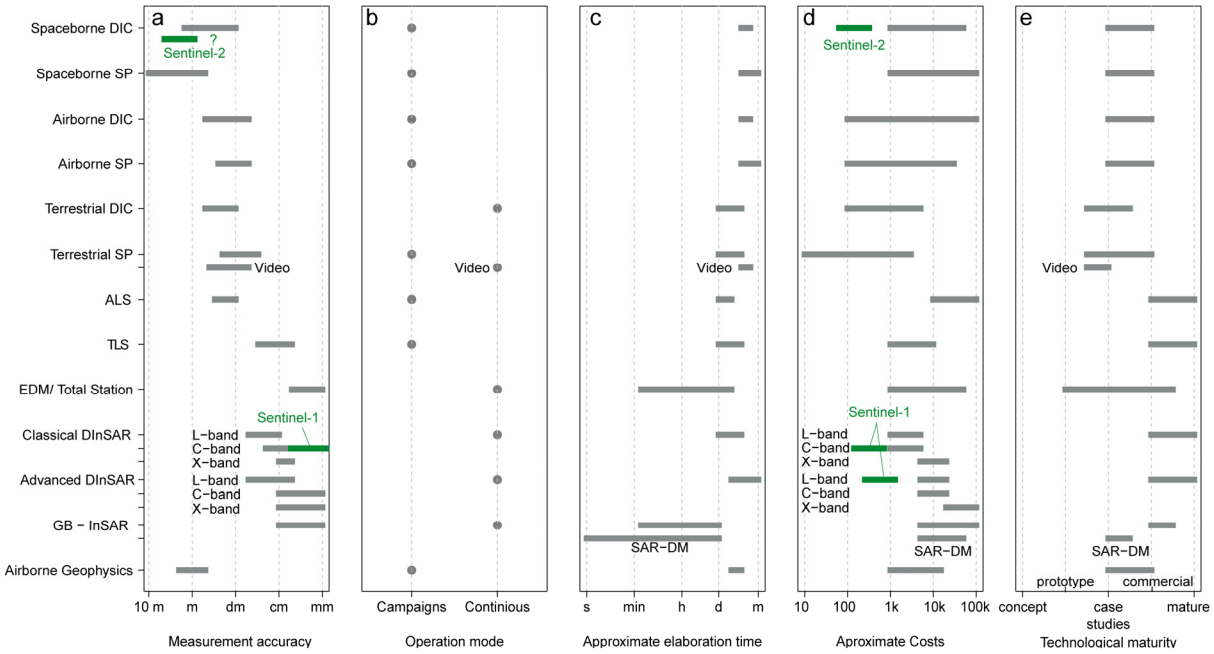


Figure 1-4: Capability of remote sensing techniques for landslide monitoring regarding (a) measurement accuracy, (b) operation mode, (c) approximate elaboration time, (d) approximate costs and (e) technological maturity. Sub-categories such as SAR distance meter (*SAR-DM*) and different microwave bands are distinguished for criteria where there value range differs.

The applicability of remote sensing techniques to a particular landslide type (Figure 1-5) has to involve several aspects. Especially, the geometry of the measurement should be able to capture the principal component of the movement and the largest extension of the moving mass. Consequently many spaceborne and airborne techniques are for example less suitable for the monitoring of falls and topples (Figure 1-5a) since only a small fraction of the moving surface is visible from a quasi-vertical view point and the resolution and accuracy of the techniques are generally too low. On the contrary, spaceborne SAR and *GB-InSAR* are very well adapted to measure very small displacement preceding falls and topples. In many cases, they are also suitable for the monitoring of rotational slides where the movement is distributed over various components (Figure 1-5b). It has been demonstrated that *Classical DInSAR* techniques can be used to detect and characterize also translational landslides (Nikolaeva et al., 2013) but the most important component of the movement along the slope is not very well depicted.

Different ground-based techniques are applicable to a variety of landslide types but their use is problematic for the monitoring of spreads (Figure 1-5d) which mainly occur in flat or low gradients terrains where it is typically difficult to obtain a near-vertical view on the surface. In this assessment, *ALS* is depicted as an ideal tool for most landslides types since the high costs for frequent surveys are not considered at this point.

Translational and rotational slides display rather coherent displacement with little deformation, whereas flows and spreads typically comprise strong deformation of the moving mass along with significant changes of the surface. The latter can yield decorrelation when using SAR techniques, and to a lesser extent *terrestrial* and *airborne DIC* (Figure 1-5e).

The suitability of remote sensing techniques for different displacement rates (Figure 1-6), is mainly determined by their temporal resolution (Figure 1-3d) and the accuracy of the measurement (Figure 1-4a). In general, most available techniques can only be employed to measure extremely slow to slow movements (Figure 1-6a-c) but possible choices for moderate to extremely rapid velocity are limited (Figure 1-6d, e). Notable exceptions are *GB-InSAR*, robotized *Total Stations*, and *terrestrial SP* and *DIC* based on videos. It should be noted in this context that video-based *SP* and *DIC* are mature measurement techniques in experimental mechanics (Sutton et al., 2009) but are rarely used for landslide monitoring. This must be attributed to the fact that rapid slope failures also exceed the reaction time for the installation of such specialized devices and hence rapid movements have in practice been rarely observed with dedicated cameras (Arattano and Marchi, 2000). The high measurement accuracies of SAR techniques make them particularly well adapted for the monitoring for extremely slow and very slow movements (Colesanti and Wasowski, 2006). Due to phase ambiguity and repeat pass cycles, spaceborne SAR systems cannot provided reliable measurement, at displacement rates exceeding ~1.5 $m.year^{-1}$ (Figure 1-6a, b), whereas *GB-InSAR* systems have a much higher temporal resolution (several minutes) and are therefore applicable even for displacement rates up to several decimetres per hour (Figure 1-6d).

Figure 1-7a provides an overview of the suitability of the evaluated techniques for monitoring applications at different scales considering aspects such as spatial coverage (Figure 1-3a), measurement accuracy and costs (Figure 1-4a, e). Most techniques are dedicated for surveys at the local (or slope) scale (Figure 1-7b). Exceptions are EDMs, Total Stations and GB-InSAR which can be employed efficiently for point-wise observation (Figure 1-7a), and Spaceborne SAR techniques which have also been used for regional surveys (Figure 1-7c). Airborne DIC and SP, Airborne LiDAR as well as Spaceborne DIC and SP are applicable in some cases for monitoring at the regional scale (Figure 1-7c) but comprise disadvantages regarding costs and/or accuracy.

Figure 1-5: Suitability of remote sensing techniques for the monitoring of different landslide types including (a) falls and topples, (b) rotational slides, (c) translational slides, (d) spreads and (e) flows.

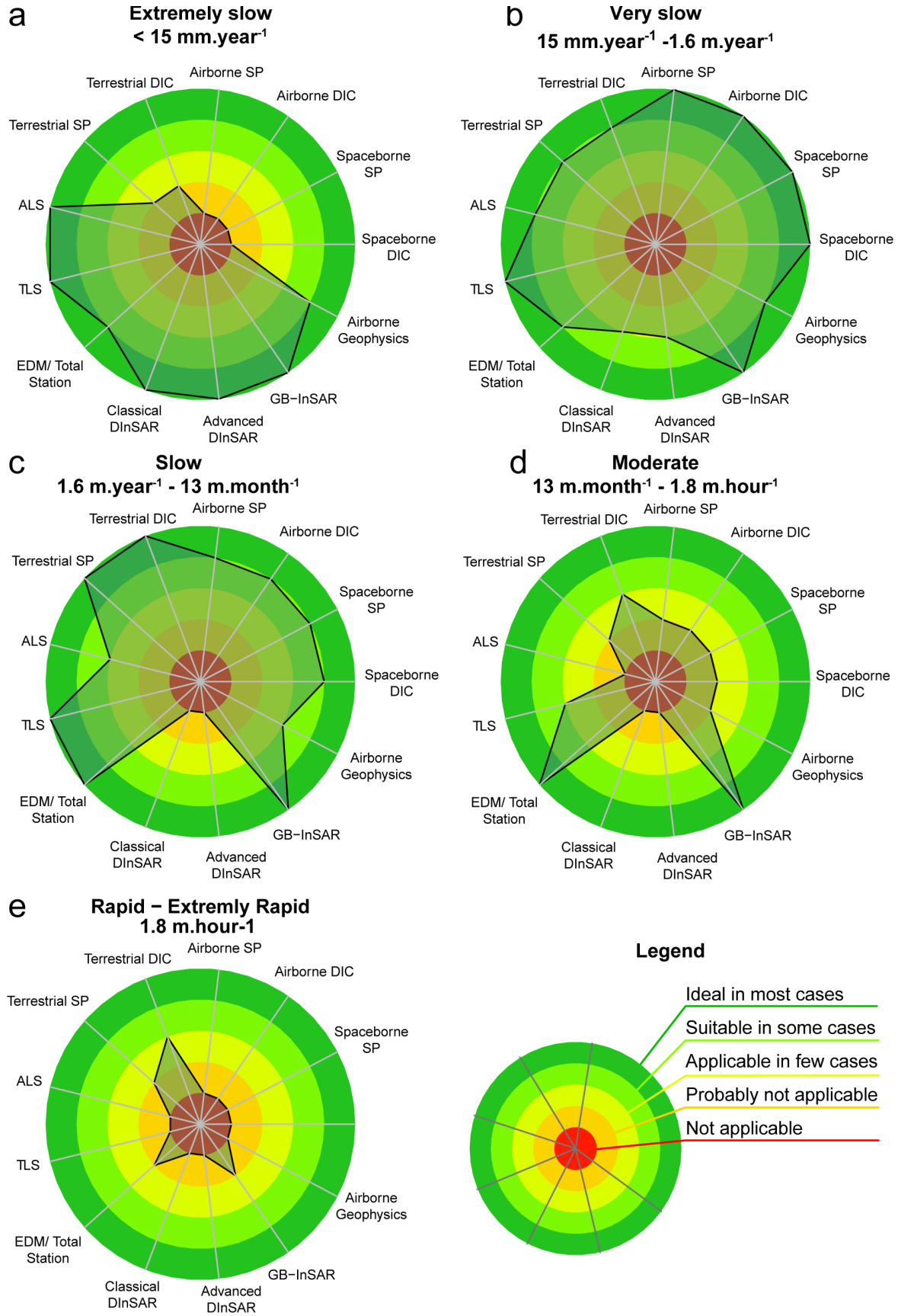


Figure 1-6: Suitability of remote sensing techniques for the monitoring of different displacement rates including (a) extremely slow, (b) very slow, (c) slow (d) moderate (e) rapid – extremely rapid.

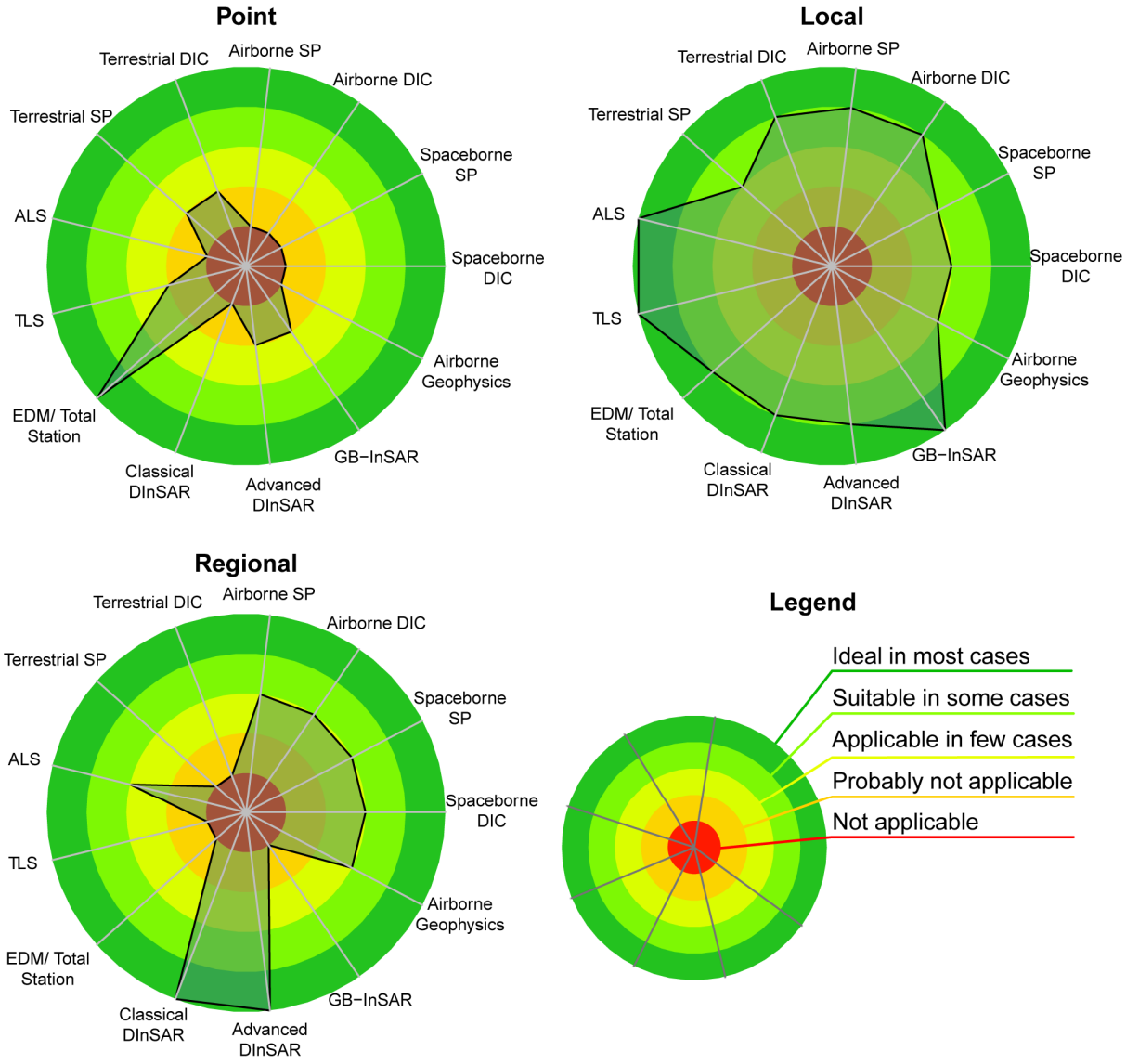


Figure 1-7: Suitability of remote sensing techniques for landslide monitoring for (a) point-wise surveys, at (b) local scale and (c) regional scale.

1.5 *Current applications and challenges*

In this section, the types of information and current limitations that can be expected from the use of different techniques are discussed and exemplified with selected case studies.

1.5.1 Long term measurement of 1D displacements

Figure 1-8 illustrates two typical examples of the use of SAR interferometry for landslide monitoring. (Casagli et al., 2006) detail one of the first operational applications of *GB-InSAR*. The *GB-InSAR* system was placed at an average distance of 1800 m from the monitored slope and operated for a short period of 5 days. This allowed to follow the evolution of an entire rockslide at a great level of detail (Figure 1-8a) and to prepare warning scenarios for the local administration. More recently, it has been demonstrated that continuous measurements can be extended over periods of several weeks to years (Herrera et al., 2009; Casagli et al., 2010).

Since LOS displacement maps can be obtained in near-real time, the integration into EWS is becoming more and more common (e.g. Intrieri et al., 2012). In general, GB-InSAR is a very versatile technique and limitations such as the LOS geometry and high costs of the system are often outweighed by the high accuracy and resolution of the measurements. Compared to spaceborne SAR techniques *GB-InSAR* is much more robust to temporal decorrelation and measurements even with permanent snow cover have been reported (Casagli et al., 2010).

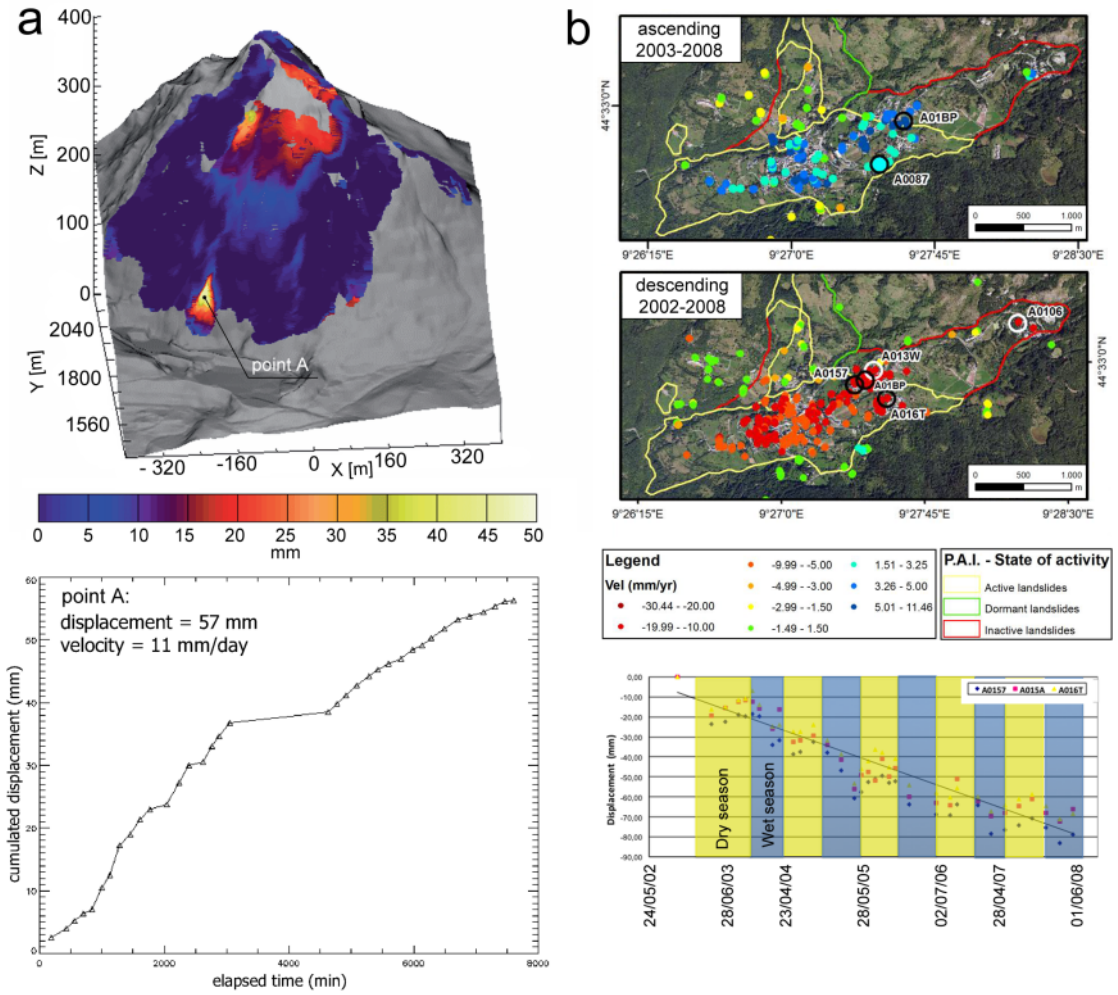


Figure 1-8: LOS displacement rates derived from (a) *GB-InSAR* at the Monte Beni (modified after Casagli et al., 2006) and, (b) *Advanced InSAR* (PS-InSAR with ENVISAT times series) at Santo Stefano d'Aveto landslide (modified after Tofani et al., 2013a)

Several further techniques such as airborne and *space-borne SP* can provide 1D measurements (through DEM differencing), whereas spaceborne interferometry is today probably the most commonly used technique in this context. *Classical* and *Advanced DInSAR* start to bridge the gap between classical mono-date inventory maps and continuous monitoring allowing not only inferring different activity states but also providing spatially distributed quantitative information on the displacement (Figure 1-8b). The LOS nature of the measurement and reduced resolution on slopes facing away from the satellite can be overcome to some extent by combining ascending and descending orbits. The great advantage of spaceborne DInSAR is the combination of high accuracy and large spatial coverage. However, several limitations should be kept in mind. First, due to the geometry of

the radar images, not all parts of the slopes are captured and in mountainous terrain the coverage does often not exceed about third of all known landslides (Cascini et al., 2009, 2010; Righini et al., 2011). Second, some advanced techniques inherently assume a linear displacement model and therefore systematically underestimate displacement exceeding one quarter of the microwave wavelength (Tofani et al., 2013a). This issue has also been illustrated in a remarkable benchmark study on the monitoring of non-linear deformation associated with mining activities by testing different PS-InSAR approaches (Raucoules et al., 2009).

The launch of the first satellite in the Sentinel series is now foreseen for spring 2014 (ESA, 2013) and will provide free access to C-band SAR data at repeat pass cycles of 12 days (6 days with two satellites). Such short time intervals will yield ideal conditions for interferometric LOS measurements with millimetre accuracy (Rohmer and Raucoules, 2012; Rucci et al., 2012) but also pose new challenges in terms of data volume that will require advanced infrastructure for efficient processing and possibly new tools for the analysis of the derived displacement fields (e.g. Julea et al., 2011).

1.5.2 Long-term measurement of 2D displacements

While 2D displacement fields have been derived with *terrestrial DIC* techniques (Travelletti et al., 2012) and through analyses of derivates of *ALS/TLS* point clouds (Daehne et al., 2011), the most commonly used sources for input data are currently airborne (Figure 1-9a) spaceborne optical (Figure 1-9b) and SAR amplitude images. For long-term observations, archives of airborne photographs can be exploited and allowed for example the reconstruction of the historical evolution of La Clapière landslides over several decades (e.g. Casson et al., 2005). This provided quantitative insights in the evolution of the slope and the derive displacement fields can also be exploited to derive further physical parameters on strain, geometry and rheology of a moving mass (Booth et al., 2013). Also (Mackey et al., 2009) exploited time series of aerial photographs and reconstructed the movement of an active landslide over a period of more than 40 years. To measure of present day landslide displacement at shorter time intervals several case studies have explored the use of VHR satellite images (Delacourt et al., 2004; Debella-Gilo and Kääb, 2012) but only from bi-temporal datasets and with metre accuracies. The need for GCPs is commonly considered as a bottleneck for the construction of longer time series from VHR satellites and only recently it has been demonstrated in (Stumpf et al., Submitted) that decimetre accuracy can be achieved without ground control (Figure 1-9b).

Currently operational VHR satellites have similar resolution, radiometric characteristics and sensor models. Combining optical images from airborne platforms and different satellite platforms could significantly enhance the currently rather limited temporal resolution. Available software solutions for aero-triangulation, stereo-photogrammetry and ortho-rectification are highly specialized and not designed to easily combine images from different platforms. Further development is needed in this context to establish end-to-end processing chains that enable to exploit archived datasets and increasing the increasing amount of daily incoming data flows (e.g. CNES 4 satellite constellation comprising Pleiades 1A/1B and SPOT 6/7). The upcoming Sentinel-2 will provide images at a maximum resolution of 10 m which could be exploited for measurements of displacement rates above 1 m. Since the instrument yields publicly available data with favourable radiometric and geometric accuracies at no costs it's the applicability of DIC techniques should be exploited.
Also terrestrial DIC with optical images is generally sensitive to weather conditions but provides significantly higher temporal and spatial resolution than space- and airborne systems,

and therefore a better possible integration with other physical parameters measured *in-situ*. For qualitative monitoring of displacement patterns, a single monoscopic camera can be used (Travelletti et al., 2012). To alleviate the requirement of an external DEM, stereo-camera setups frequently used in experimental mechanics are currently being installed operationally on several landslides (Gance et al., In Press).

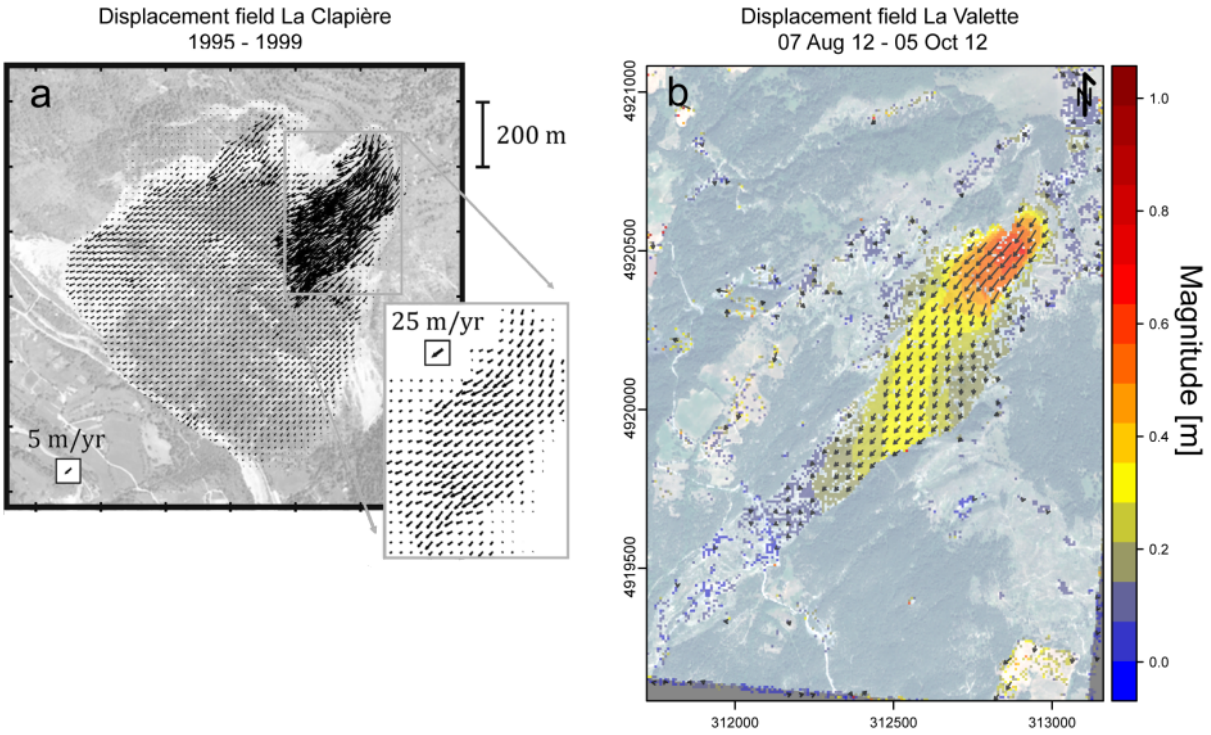


Figure 1-9: 2-D displacement fields for (a) the Clapière landslide derived from ortho-rectified airborne photographs (modified after Booth et al., 2013) and (b) the La Valette landslide derived from ortho-rectified Pleiades images.

1.5.3 Long-term measurement of 3D displacements and volume changes

The currently most commonly applied remote sensing techniques capable of capturing all three components of the movement are based on LiDAR acquisitions (in particular *TLS*). The measurements are usually based on point-cloud matching and several case studies have demonstrated centimetre (in some cases even millimetre) accuracy for different types of landslides (Teza et al., 2007; Monserrat and Crosetto, 2008; Oppikofer et al., 2008; Travelletti et al., 2008; Abellán et al., 2009; Oppikofer et al., 2009). However, point-matching algorithms depend on the preservation of rigid features among point-clouds from different dates. Several algorithms to address this issue have been recently proposed (Wu et al., 2012; Ghuffar et al., 2013; Travelletti et al., In Review) and should be further evaluated in case studies to fully exploit the 3D information contained in very dense point clouds. Figure 1-10b provides an example of innovative 3D displacement measurements obtained from the analysis of derivates (depth map) of very dense point clouds acquired by *TLS*.

First results regarding the measurement of 3D landslide displacement from space (through the correlation of amplitude images) have been presented in Raucoules et al. (2013) and are encouraging in terms of accuracy and spatial coverage (Figure 1-10a). The analysed dataset, originated from a specifically scheduled acquisition campaign of TerraSAR-X

scenes, allowed a very short time lag between ascending and descending orbits. Sentinel-1will have a time-lag between ascending and descending orbits of approximately 12 days. While it has been argued in Rucci et al. (2012) that the azimuth resolution of Sentinel-1 will be to coarse for 3D measurements, (Jung et al., 2013) provided evidence that decimetre accuracy in the azimuth direction is possible.

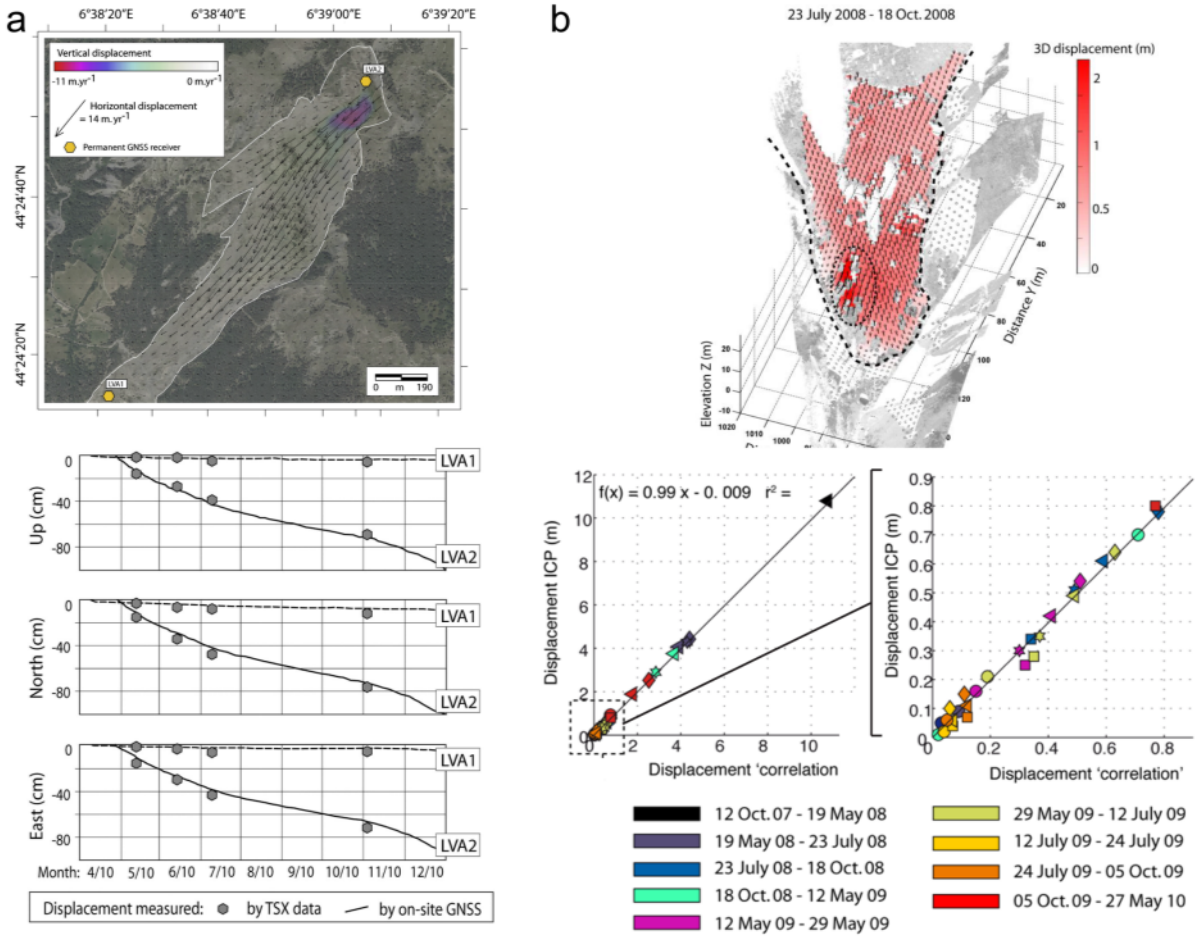


Figure 1-10: 3D-displacement fields derived from (a) correlation of TerraSAR-X amplitude image at the La Valette landslide (modified after Raucoules et al., 2013) and (b) correlation of depth maps derived from multi-temporal terrestrial LiDAR derived at the toe of the Super-Sauze landslide (modified after Travelletti et al., In Review).

Total stations are standardized and accurate instruments for measurements of 3D displacements and volumes. Recent enhancements lead to a new generation of *robotized total stations* that can provide quick point-wise 3D measurements without the need of reflectors at the surface. A largely automated tracking of natural 3D points (no prism reflectors) is currently not possible but dedicated prototypes are being developed (Reiterer et al., 2009).

Measurements of volumes can be accomplished with a wide range of techniques including spaceborne SP (e.g. Tsutsui et al., 2007; Martha et al., 2010b), airborne SP (e.g. Kerle, 2002; Dewitte et al., 2008), and terrestrial SP and TLS (e.g. Chen et al., 2006; Baldo et al., 2009; James and Robson, 2012; Stumpf et al., In Review-a). Currently, TLS (Figure 1-10b,Figure 1-11a) and terrestrial SP (Figure 1-11a) are the most interesting techniques for the

quantification of sediment budget since they allow a temporal resolution of up to a few days and are generally more accurate than space- and airborne techniques. Sediment budgets with high temporal frequency are especially important to establish magnitude-frequency laws for rockfall events (Barlow et al., 2012; Rosser et al., 2013) which can be used for probabilistic forecasts.

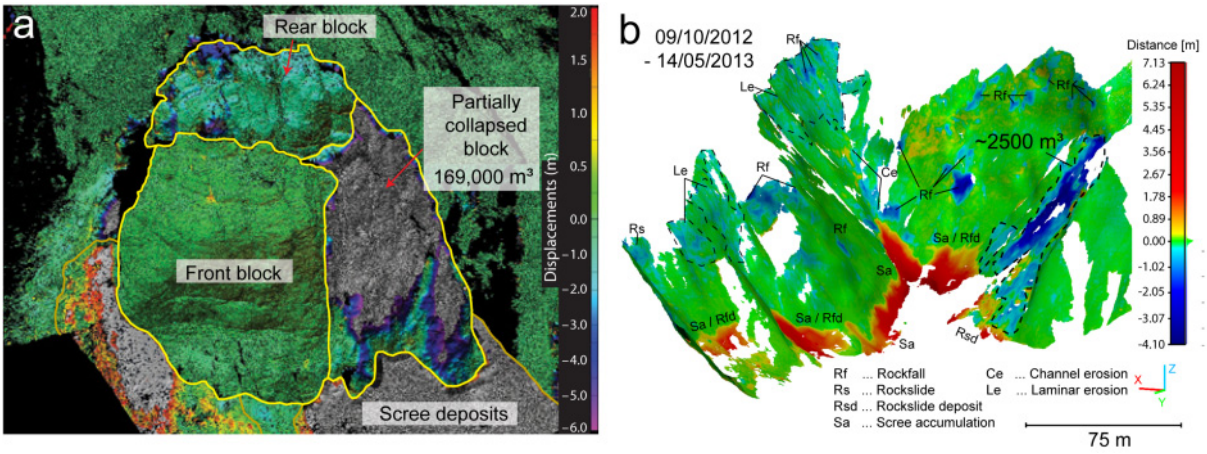


Figure 1-11: Exemplary results of techniques with 3D measurement capabilities. (a) 3-D displacement maps and volume estimates for a collapse at the Eiger massive derived from multi-temporal terrestrial LiDAR scans (modified after Oppikofer et al., 2008), and (b) surface changes and volume estimation for the main scarp of the Super-Sauze landslide based on multi-temporal terrestrial photogrammetry.

1.5.4 Surface and subsurface features

Beside quantitative measurements of surface displacements and volumes, remote sensing images also convey a large amount of qualitative information such as the presence and evolution of surface fissures, and compression lobes on clayey slopes or fractures and dipping planes on rock slopes. While in practice such qualitative parameters are difficult to integrate into threshold-based EWS and real-time monitoring, they are important for a better understanding of the underlying mechanical processes. It is for example still difficult to derive quantitative information from coincidental video recordings of debris flows but they provide unique documents that help to understand natural processes which otherwise can only be studied with dedicated experiments. Other examples include the mapping an interpretation of surface fissures in VHR optical images (Figure 1-12a) for which also semi-automatic approaches have been suggested (Stumpf et al., 2013b), and the structural analysis of high-resolution LiDAR DTMs (Figure 1-12b) for which dedicated software tools have been developed (Jaboyedoff et al., 2009). Airborne geophysics (Figure 1-12c) is still a rather new technique for the investigation of landslides and unique regarding the possibility to obtain information about the subsurface and technique can provide valuable information about the depth and structure of the moving mass (Baroň et al., 2013; Supper et al., 2013). Further case studies are required to explore its potential for repeated surveys and the landslide research community can certainly learn a lot from previous experiences in fields such as agriculture (George and Woodgate, 2002) and hydrogeological studies (Paine and Minty, 2005).

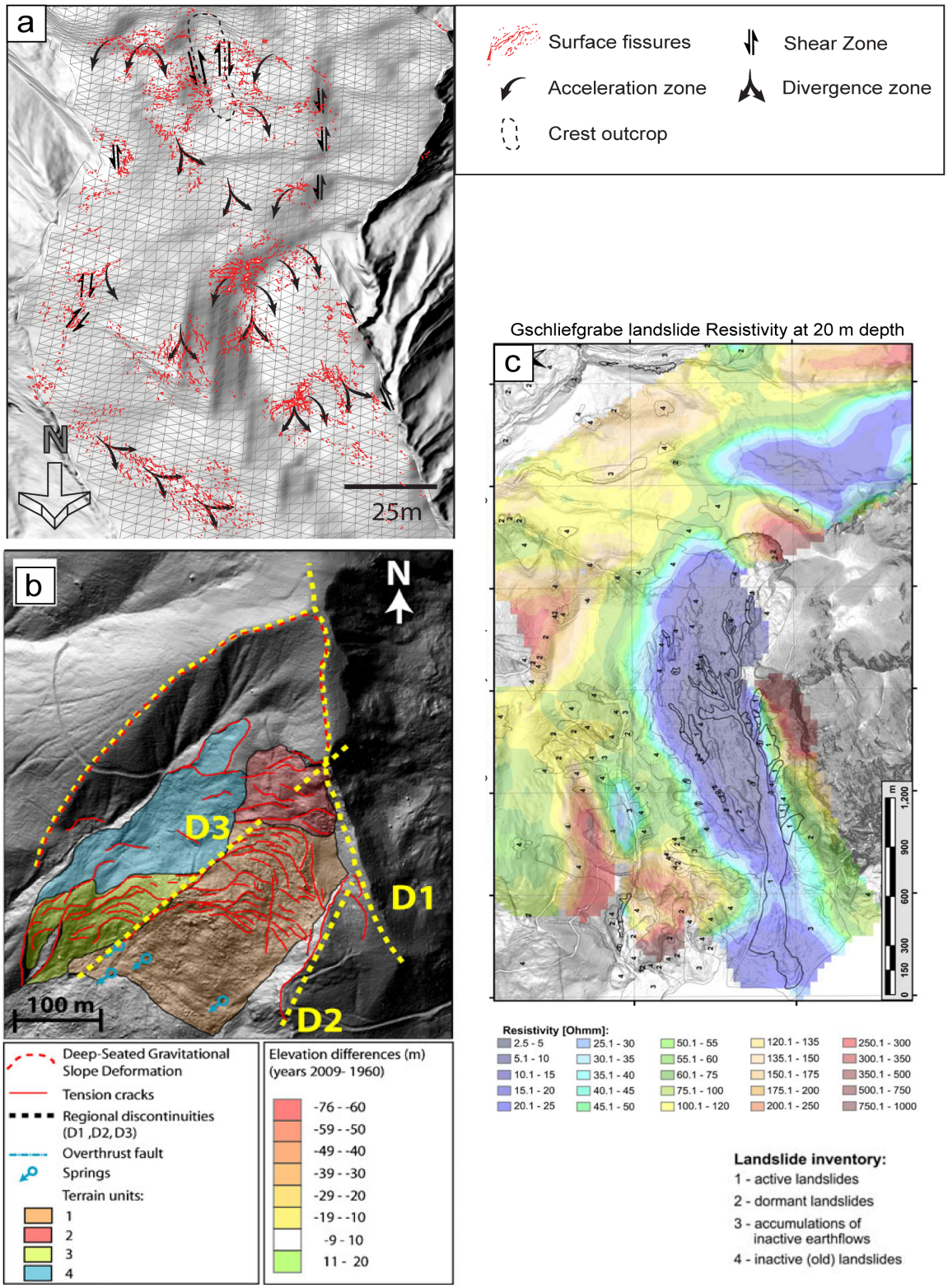


Figure 1-12: (a) Example for the semi-automatic detection of surface fissures in UAV images (modified after Stumpf et al., 2013b), (b) the visual interpretation of a high-resolution Airborne LiDAR DTM to locate superficial and deep-seated discontinuities (modified after Travelletti et al., 2013) and (c) subsurface resistivity derived from and airborne geophysical surveys (modified after Supper et al., 2013).

1.6 *Conclusion*

In this work, we reviewed the current state-of-the-art of remote sensing techniques for landslide monitoring and developed a catalogue of criteria that can be adopted as guidelines to select the most appropriate techniques in different monitoring scenarios. The catalogue includes several technological aspects such as spatial coverage, temporal resolution, measurement accuracy and costs, but also their suitability for different landslide types, displacement rates and targeted scale of a study. The different types of information that can be expected from different techniques have been illustrated with a number of selected examples. The fact that different techniques can yield very different types of information often suggests a complementary use of techniques, especially when the same type of input data can be exploited (e.g. interferometry applied on SAR phase values, and DIC applied on amplitude values).

Despite the diversity of the evaluated techniques, a few general conclusions can be drawn. For the use of remote sensing in EWS, there are currently only few techniques (Total Station, GB-InSAR, terrestrial stereo-photogrammetry only at daylight) which are sufficiently fast, accurate and robust to be useful. For other options, such as LiDAR or spaceborne techniques it is still necessary to increase the temporal resolution by at least one order of magnitude, which cannot be expected at reasonable costs in the near future. In any case, remote sensing techniques should rather be regarded as a supplementary tool in EWS to obtain spatially distributed information complementing point-wise *in-situ* measurements.

All reviewed techniques have significantly advanced in the last decade and are very valuable for long-term observations that improve both decision making and scientific understanding. Due to the greater variety of available remote sensing systems, the access and acquisition is now less problematic but at the same time the development of automated analysis techniques and standardized products and services is becoming more challenging. While some advanced InSAR techniques already provide capabilities to process large datasets, time-series analysis of multi-sensor optical data, for example, is still associated with significant manual intervention. Also the size of datasets resulting from VHR optical satellite or high-resolution LiDAR surveys reaches, in many cases, already the limit of commercial and public domain software and often requires to subset or downsample the original dataset. To resolve such issues, a closer collaboration of the landslide research community with computer scientists has to be envisaged. Open-source based initiatives like the Orfeo Toolbox (OTB Development Team, 2013) for the development of automated image analysis tools seem a promising approach in this direction.

Several remote sensing techniques can provide full 3D information about volumes and in principle many further parameters such as strain, geometry and rheology can be derived from such type of information. Recent studies (Booth et al., 2013; Travelletti et al., 2013) are taking first steps towards the inversion of mechanical parameters from remotely sensed displacement and further integration of remote sensing and landslide modelling is desirable to fully exploit the acquired data.

Many of the criteria evaluated in this review are linked to technological aspects of currently available sensors, platforms and analysis algorithms which change quickly as technology evolves. An overview of the main evaluated criteria and additional information regarding available systems, data providers, research institutions, private companies and new case studies should therefore be established and maintained to provide up-to-date criteria for operational and scientific landslide monitoring.

Acknowledgements
This work was supported by the project SafeLand 'Living with landslide risk in Europe: assessment, effects of global change, and risk management strategies' (Grant Agreement No. 226479) funded by the 7th Framework Programme of the European Commission., and by the EUR-OPA European and Mediterranean Major Hazards Agreement of Council of Europe (Strasbourg) as part of the research activities of the European Centre on Geomorphological Hazards (CERG). The authors thanks all the landslide monitoring experts of Analisi e Monitoraggio del Rischio Ambientale (AMRA Scarl), Bureau de Recherches Géologiques et Minières (BRGM), Centre National de la Recherche Scientifique (CNRS), Faculty for Geo-information Sciences and Earth Observation (ITC), Geological Survey of Austria (GSA), Geological Survey of Norway (NGU), Geological Survey of Slovenia (GeoZS), Joint Research Centre (JRC, Ispra), Norwegian Geotechnical Institute (NGI), Università degli studi di Firenze (Unifi), Università degli studi di Salerno (UNISA), Universitat Politècnica de Catalunya (UPC), Université de Lausanne (UNIL).

Chapter 2: Very High Resolution optical satellite images for landslide mapping at the regional scale

Objectives: Rapid landslide inventory mapping after major triggering events is essential for immediate disaster response and the assessment of future hazards. It is a tedious and time-consuming task if approached with filed work and manual image interpretation and still poses great challenges for automatic image analysis. The objective of this chapter is the development and validation of accurate and robust image processing chains to facilitate rapid landslide mapping with VHR optical satellite images at the regional scale.
Considering the great diversity of landslide types, environmental settings and image sources, the domain-specific objectives are:

- The development of a generic image processing method for rapid and accurate mapping of landslides after major triggering events;
- The quantification of uncertainties in the mapping;
- The assessment of time requirements for the use of the image processing chain.

For the implementation of the image processing chain, our research targets as much as possible the use of universal landslide features, and the provision of automated tools for the reduction of manual labor. More specifically, the technical objectives are:

- The design of relevant objet-based image features combining spectral, shape, textural and topographic information;
- The proposition of a framework to rank and select relevant features based on labeled examples;
- The development of stratified and Active Learning (AL) sampling schemes that accounts for class-imbalance and reduces the need for training data.

Methods: A combination of image segmentation, feature extraction, feature selection and machine learning techniques is employed to implement the processing chain. The Random Forest (RF) algorithm is adopted as a base algorithm and dedicated stratified- and AL sampling schemes are developed to address class-imbalance and the reduction of required training data, respectively. The AL algorithm combines criteria for sample diversity to select the most beneficial sampling areas and a statistical analysis of spatial auto-correlation is considered to optimize the size of the sampling area.

This chapter is based on:

• Stumpf, A., Kerle, N., 2011. Object-oriented mapping of landslides using Random Forests. Remote Sensing of Environment 115, 2564–2577.

• Stumpf, A., Lachiche, N., Malet, J.-P., Kerle, N., Puissant, A., In Press. Active learning in the spatial domain for remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 16 pages, 10.1109/TGRS.2013.2262052.

2.1 *Landslide mapping at regional scales: user requirements*

Landslide inventory maps should depict the location, the type and ideally also the timing of slope failures. They are typically prepared from different sources (images, field work) to characterize the affected areas, and to initiate hazard assessment. The spatial and temporal patterns of landslides are also studied to quantify long-term evolution of slope movements in mountainous landscapes, and underline relationships to tectonic and hydro-climatic processes (Guzzetti et al., 2012). Visual image interpretation and field surveys are the prevailing methods for inventory mapping but require several months or even years of manual labour. Semi-automatic image analysis methods have been proposed but are not frequently used in practice since their quality generally lacks behind expert mappings (Guzzetti et al., 2012).

No common standards for the preparation of landslide inventory have been established yet (Guzzetti et al., 2012) which is probably related to the fact that different applications require different types of information.

Firstly, for analysing the spatial distribution of landslides and its relation to triggering factors (abundant rainfall, snowmelt or fault rupture processes), knowledge on the location and size of landslides is typically sufficient (e.g. Guzzetti et al., 2002; Harp et al., 2010; Gorum et al., 2013). Reliable volume measurements over large areas are generally difficult but approximations can be derived from empirical relationships between area and volume (Parker et al., 2011).

Secondly, for characterizing landslide damages after severe triggering events and in the case of disasters, it is important to provide immediate information of the location and extent of the affected areas rather than a detailed characterization of the phenomena.

Thirdly, for landslide hazard analysis, detailed information on landslide types and on landslide sub-units (source area, run out area) is necessary for proper assessments (Fell et al., 2008; Van Den Eeckhaut and Hervás, 2012).

Lastly, large disagreements in the mappings are generally observed among experts both in terms of total landslide affected areas (Carrara, 1993; Ardizzone et al., 2002; Galli et al., 2008; Fiorucci et al., 2011) and in terms of landslide classification. Such disagreements have to be critically quantified to ensure reproducible results, and provide measures of the uncertainties.

Considering the absence of established map standards, the immediate need for maps of affected areas, and the unknown degree of uncertainty in the differentiation of landslide types, the mapping of landslides is treated here as a binary problem (i.e. an area is affected by a landslide or not).

The domain-specific objectives of this chapter are:

- The development of a generic image processing method for rapid and accurate mapping of landslides after major triggering events;
- The quantification of uncertainties in the mapping;
- The assessment of time requirements for the use of the image processing chain.

The processing chain is developed and tested on different mountain environments recently affected by landslides of different type and size and triggered by different causes (large earthquakes, intense rainstorms). The selected study cases are depicted in Figure 2-1.

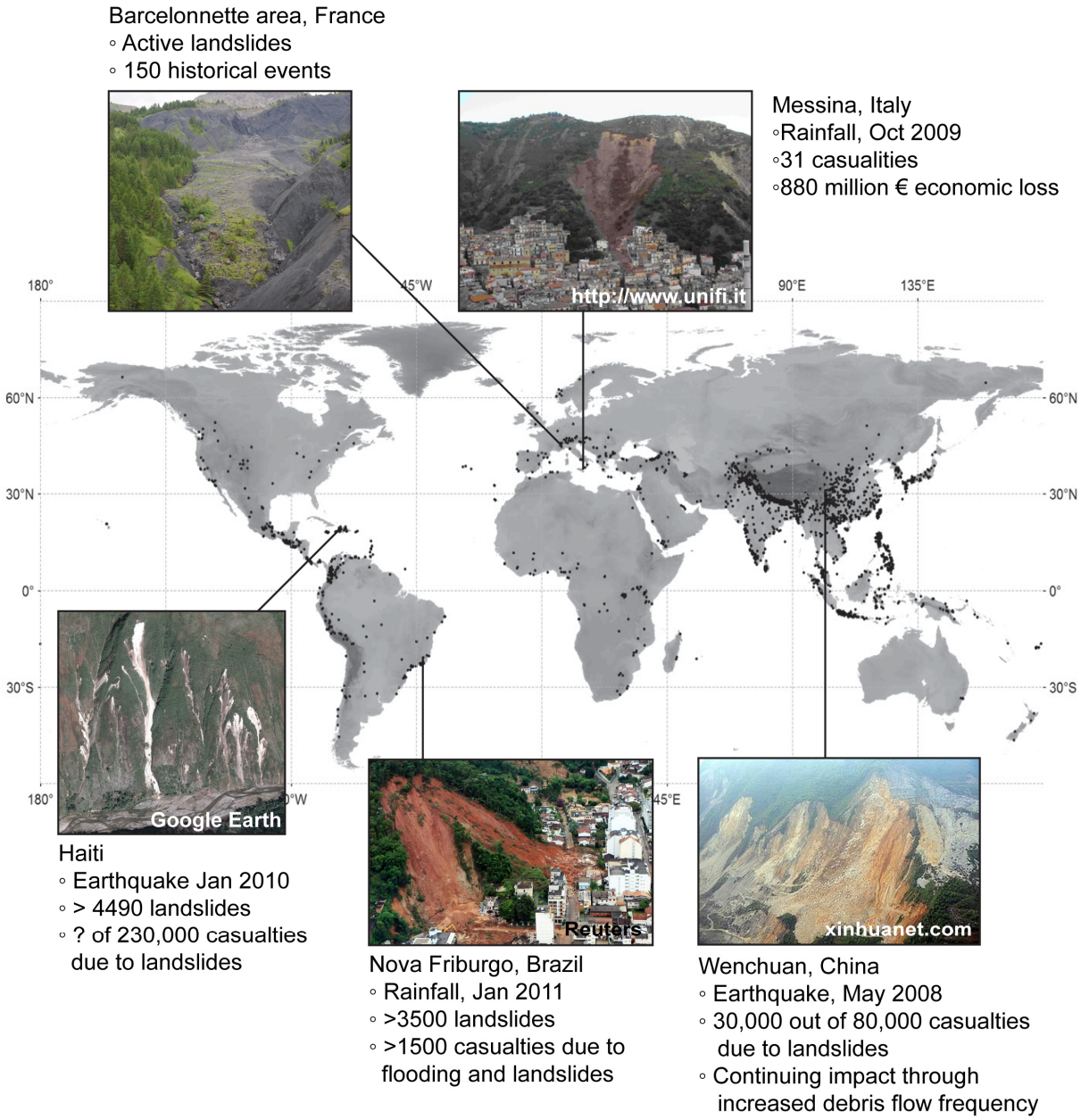


Figure 2-1: Location and key dates for the events/sites investigated in chapter 2. The base map displays the distribution of fatal landslides (black dots) between 2003 and 2010 (Petley, 2012).

2.2 *Classification of optical imagery for landslide mapping: current limitations*

VHR satellite and aerial images convey rich spectral, textural and contextual information about objects on the ground. An important indicator of landslide occurrence in a landscape is the disturbance or removal of vegetation and the exposure of bare soils; however, other features regarding the morphology of the surface and the spatial context are typically required for an accurate identification.

Classical per-pixel classifiers based on spectral values are not suitable to exploit the rich geometric information conveyed in VHR images. Furthermore, the intra-class variability

of spectral values augments and interclass variability reduces with increasing spatial resolution leading to deteriorating class separability if only pixel values are considered (Bruzzone and Carlin, 2006)

To better exploit spatial and contextual information, sequential image processing by segmentation, feature extraction and classification was proposed by Kettig and Landgrebe (1976) and has strongly evolved in recent years under the name Object-Based Image Analysis (OBIA), (Blaschke, 2010). Several studies have addressed the use of OBIA for landslide investigations and employed heuristic methods to translate expert knowledge into machine executable rule sets (Barlow et al., 2006; Martha et al., 2010a; Martha et al., 2011; Martha et al., 2012; Lacroix et al., 2013). Advantages of such rule-based approaches are the possibility to benefit from domain knowledge of geoscientists, efficient execution of rule sets, and a high degree of transparency since most rules are formulated explicitly. Considering the great diversity of landslide types, environmental conditions and image types, it is, however, hard to generate rule sets that generalize to all possible situations, and the elaborated rule sets still strongly depend on the subjective judgment of the operator. OBIA allows the extraction of hundreds of image features (spectral values, gradients, shapes, spatial relationships, hierarchical representations, topographic variables, etc.) and even for trained experts it is general difficult to select and prioritize features and arrange rules in such a high dimensional space.

Early research in Artificial Intelligence (AI) in the 1960s put great expectation in the development of rule-based systems that imitate human problem-solving protocols. Such systems can find solutions to constrained and well-structured problems but it has been realized that *"When information extraction must be attempted from noisy or varied input .... It is too hard to get all the rules and their priorities right; it is better to use a probabilistic model rather than a rule-based model"*(Russell and Norvig, 2009).

Machine learning offers a wide range of generic probabilistic and discriminative models to learn complex concepts from labelled data automatically and thereby overcomes many limitations of rule-based systems. In our work, the Random Forest (RF) algorithm (Breiman, 2001) is adopted as a generic classifier which requires very little tuning of free parameters and provides inherent measures to rank the importance of the input features. The classification accuracy achieved with this algorithm depends on several factors such as the representation of data in the feature space, class-overlap and class-imbalance (He and Garcia, 2009) as well as representativeness of available training data.

Considering these issues, the specific technical objectives of this chapter are:

- The design of relevant objet-based image features combining spectral, shape, textural and topographic information;
- The proposition of a framework to rank and select relevant features based on labelled examples;
- The development of stratified and Active Learning (AL) sampling schemes that accounts for class-imbalance and reduces the need for training data.

2.3 *Object-oriented mapping of landslides from VHR optical imagery using Random Forests*

This section is based on: Stumpf, A., Kerle, N., 2011. Object-oriented mapping of landslides using Random Forests. Remote Sensing of Environment 115, 2564–2577.

This section describes the development of a generic supervised image processing chain for landslide mapping from VHR satellite and aerial images in combination with topographic data. Algorithms for image segmentation, feature extraction, feature selection and supervised learning are adapted and combined in a processing chain. The developed technique is applied and validated at different study sites. Relationships between the segmentation scale, the importance of different object features and the accuracy of the resulting maps are studied in detail.

Remote Sensing of Environment 115 (2011) 2564–2577

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Object-oriented mapping of landslides using Random Forests

André Stumpf [a,b,*], Norman Kerle [a]

[a] ITC-Faculty of Geo-Information Science and Earth Observation of the University of Twente, Department of Earth Systems Analysis, Hengelosestraat 99, P.O. Box 6, Enschede, 7500 AA, The Netherlands

[b] École et Observatoire des Sciences de la Terre, Institut de Physique du Globe de Strasbourg, UMR 7516 CNRS, Université de Strasbourg, 5, rue René Descartes, 67084 Strasbourg Cedex, France

ARTICLE INFO

Article history:
Received 19 November 2010
Received in revised form 18 May 2011
Accepted 19 May 2011
Available online 24 June 2011

Keywords:
Landslide mapping
VHR satellite images
Image segmentation
Object-oriented
Random Forest

ABSTRACT

Landslide inventory mapping is an indispensable prerequisite for reliable hazard and risk analysis, and with the increasing availability of very high resolution (VHR) remote sensing imagery the creation and updating of such inventories on regular bases and directly after major events is becoming possible. The diversity of landslide processes and spectral similarities of affected areas with other landscape elements pose major challenges for automated image processing, and time-consuming manual image interpretation and field surveys are still the most commonly applied mapping techniques. Taking advantage of recent advances in object-oriented image analysis (OOA) and machine learning algorithms, a supervised workflow is proposed in this study to reduce manual labor and objectify the choice of significant object features and classification thresholds. A sequence of image segmentation, feature selection, object classification and error balancing was developed and tested on a variety of sample datasets (Quickbird, IKONOS, Geoeye-1, aerial photographs) of four sites in the northern hemisphere recently affected by landslides (Haiti, Italy, China, France). Besides object metrics, such as band ratios and slope, newly introduced topographically-guided texture measures were found to enhance significantly the classification, and also feature selection revealed positive influence on the overall performance. With an iterative procedure to examine the class-imbalance within the training sample it was furthermore possible to compensate spurious effects of class-imbalance and class-overlap on the balance of the error rates. Employing approximately 20% of the data for training, the proposed workflow resulted in accuracies between 73% and 87% for the affected areas, and approximately balanced commission and omission errors.

1. Introduction

During the last century (1903–2004) approximately 16,000 people were killed by landslides in Europe (Nadim et al., 2006), while in other parts of the world even single events can have comparable dimensions (20,000 in Peru, 1970, 29,000 in China, 2008)(Kjekstad & Highland, 2009; Petley, 2009). The mean annual costs of landslides in Italy, Austria, Switzerland and France are estimated between USD 1–5 billion for each of the countries (Kjekstad & Highland, 2009). The assessment of associated risks, a prerequisite for disaster mitigation, is still a difficult task, with comprehensive landslide inventories being the most commonly used source for quantitative landslide hazard and risk assessment at regional scales (van Westen et al., 2006).

Landslide inventories have traditionally been prepared combining the visual interpretation of aerial photographs and field work, which to date remains the most frequently followed approach for the elaboration of inventory maps in scientific studies and by administrative bodies (Hervás & Bobrowsky, 2009). Despite its time-consuming and labor intensive nature, however, results still include a large degree of subjectivity (Galli et al., 2008), and incur the risk of omissions due to limited site access or aerial survey campaigns only being mounted with some delay, when landslide traces are starting to disappear.

Notable advances are being made in the detection of surface-displacements from active (e.g. Cascini et al., 2010) and passive (e.g. Debella-Gilo & Kääb, 2011) spaceborne sensors, allowing for detailed monitoring of ground-deformations. Those techniques depend on a coherent signal over time and are applicable for the mapping of slow to extremely slow moving landslides (<13 m/month after Cruden & Varnes, 1996) with a sparse vegetation cover. For the automated mapping of dormant landslides under forest high-resolution surface models from airborne laser scans provide new opportunities (e.g. Booth et al., 2009). However, most hazardous landslides reach considerable velocities and can typically only be mapped in a post-failure stage, for which optical airborne and satellite images are the commonly chosen data sources. Large events with thousands of individual landslides such as recently in China (earthquake, 2008),

* Corresponding author at: ITC-Faculty of Geo-Information Science and Earth Observation of the University of Twente, Department of Earth Systems Analysis, Hengelosestraat 99, P.O. Box 6, Enschede, 7500 AA, The Netherlands.
E-mail addresses: stumpf24883@itc.nl (A. Stumpf), kerle@itc.nl (N. Kerle).

0034-4257/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2011.05.013

Haiti (earthquake, 2010) and Brazil (rainfall, 2011) illustrate the immense challenges posed for any non-automated mapping approach.

The large fleet of existing and planned very high resolution (VHR) satellites allows to record inexpensive imagery within days or even hours after a given landslide event, and a number of studies have already addressed the development of more automatic techniques for landslide mapping with VHR images (Barlow et al., 2006; Borghuis et al., 2007; Hervás & Rosin, 1996; Joyce et al., 2008; Lu et al., 2011; Martha et al., 2010; Nichol & Wong, 2005; Rau et al., 2007; Whitworth et al., 2005). Most of them targeted the mapping of fresh features after rapid slope failures, but a few works also demonstrated the potential of optical data for the identification of slow-moving and dormant landslides (Hervás & Rosin, 1996; Whitworth et al., 2005).

Proposed approaches may be generally classed into pixel-based and object-based techniques, both including methods for the analysis of monotemporal and multitemporal imagery, and often making use of ancillary datasets such as digital elevation models (DEMs). Pixel-based approaches include unsupervised (Borghuis et al., 2007) and supervised classification (Joyce et al., 2008), as well as change detection techniques (Hervás et al., 2003; Nichol & Wong, 2005; Rau et al., 2007). Although those techniques consider to some extent additional geometric constrains, such as minimum size, minimum slope or non-rectangular shapes, they rely mainly on the spectral signal of individual pixels. To exploit better the information content of local pixel neighborhoods, Hervás and Rosin (1996) conducted a systematic statistical evaluation of texture measures for landslide mapping and found texture features after Haralick et al. (1973) especially useful to highlight hummocky surfaces often associated with landslides. Similarly, more recent studies concluded that the integration of texture improves the image classification and may yield more accurate maps (Carr & Rathje, 2008; Whitworth et al., 2005).

In general there is an emerging agreement in the remote sensing community that unsatisfactory results of per-pixel analysis can often be attributed to the fact that geometric and contextual information contained in the image is largely neglected (e.g. Blaschke, 2010). This is especially true at higher resolutions, with a higher spectral variance leading to increased intra-class variability and typically lower classification accuracies (Woodcock & Strahler, 1987). Further challenges arise due to the typically lower number of spectral bands of modern VHR sensors and a higher sensitivity to co-registration errors at higher resolutions. To address such issues object-oriented analysis (OOA), also often referred to as object-based image analysis (OBIA), became a widely spread concept for many geoscientific studies to exploit geometric and contextual image information of multi-source data (Blaschke, 2010).

Image segmentation and classification resemble human cognition to some degree and have inspired a number of researchers to transfer existing knowledge in machine executable rule sets. Such rule sets have already been used for landslide mapping as a self-contained classification scheme (Barlow et al., 2003), prior to supervised classification (Barlow et al., 2006), for the post-processing of pixel-based classification (Danneels et al., 2007), and for change detection with multi-temporal images (Lu et al., 2011). Martha et al. (2010) emphasized the importance of exploiting a range of features as widely as possible, and developed a complex set of decision rules, including 36 particular thresholds, to detect and classify landslides of 5 different types in the High Himalayas.

Expert rule sets are a very transparent solution for the exploitation of domain knowledge but comprise two main limitations: (i) the difficulty to decide which descriptive features are actually significant, and (ii) their restricted generic applicability for different input data types and under variable environmental conditions. Professional OOA software solutions readily provide hundreds of potentially useful object metrics, and further customized features enrich this great variety. They allow the user high flexibility in setting up efficient automated processes, but the selection of significant features remains a challenging and time-consuming task.

Feature selection in high-dimensional datasets is an important task in many fields such as bioinformatics (Saeys et al., 2007) or hyperspectral remote sensing (e.g. Guo et al., 2008), and typically targets a better performance of the algorithm classifying the data and/or the investigation of causal relationships. A few object-oriented studies already addressed statistical feature selection for land cover mapping from VHR imagery (e.g. Laliberte & Rango, 2009; Van Coillie et al., 2007), but no such efforts have been in the context of landslide mapping. Little is known about the robustness, efficiency, scale-dependency and generic applicability of the object-features and thresholds proposed in individual studies. Considering the great variety of landslide types, environmental conditions and available imagery this largely prevents the transferability of proposed methods and the development of operational workflows.

Machine learning algorithms, such as Random Forests (RF, Breiman, 2001), have demonstrated excellent performance for the analyses of many complex remote sensing datasets (Gislason et al., 2006; Lawrence et al., 2006; Watts et al., 2009). RF is based on ensembles of classification trees and exhibits many desirable properties, such as high accuracy, robustness against over-fitting the training data, and integrated measures of variable importance (Diaz-Uriarte & Alvarez de Andres, 2006). However, like many other statistical learning techniques RF is bias-prone in situations where the number of instances is distributed unequally among the classes of interest. Under class-imbalance in fact most classifiers tend to be biased in favor of the majority class, and vice versa may underestimate the number of cases belonging to the minority class (He & Garcia, 2009). Experiments on synthetic datasets suggest that such biases are combined effects of class imbalance and an overlap of the classes in feature space (e.g. Denil & Trappenberg, 2010). As landslides typically cover only minor fractions of a given area, class-imbalance is an inherent issue that affects the probabilistic assessments of slope susceptibility (Van Den Eeckhaut et al., 2006), and may complicate the application of machine learning algorithms for image-based inventory mapping.

The objective of this study was to investigate the applicability and performance of the RF learning algorithm in combination with OOA to reduce the manual labor in landslide inventory mapping with VHR images. Assuming that a sample-based framework combining both techniques could be a flexible and efficient solution for many real-world scenarios, VHR imagery recorded by state-of-the-art systems (Geoeye-1, IKONOS, Quickbird, and airborne) at four different sites was analyzed. To achieve an accurate and robust image classification it was of particular interest to determine which image object metrics efficiently distinguish landslide and non-landslide areas. Training and testing samples were derived from existing landslide inventories, and a RF-based feature selection method (Diaz-Uriarte & Alvarez de Andres, 2006) was adopted to evaluate the capability of a broad set of object metrics (color, texture, shape, topography) and their sensitivity to changing scales of the image segmentation. Class-imbalance and -overlap were expected to be critical points for the application of the RF, and we further investigated if an iterative resampling scheme could be used to design training sets that lead to a balance between commission and omission errors. The efficiency of this approach was evaluated at each test site with different segmentation scales and in scenarios where 20% of the image objects would be available for training.

2. Study sites and data

VHR images collected in the immediate aftermath of two recent major earthquakes, as well as from two sites affected by non-seismic landslides, were used in this study (Table 1). The areas are characterized by a great diversity of environmental settings, landslide processes and image acquisition conditions, and in this manner

Table 1
Overview of analyzed images and topographic data.

Test site	Haiti	Wenchuan	Messina	Barcelonnette
Sensor	Geoeye-1	IKONOS	Quickbird	Aerial photograph
Spectral bands	4-band multispectral	4-band multispectral	4-band multispectral	3-band natural Color
Pixel size (multispectral/panchromatic) [m]	2/0.5	4/1	2.4/0.61	0.5/–
Sensor Tilt [°]	2.7	15.7	3.1	n.a.
Nominal collection azimuth [°]	343.8	62.7	343.3	n.a
Solar zenith angle [°]	45.6	19.2	45.6	–
Sun angle azimuth [°]	150.2	119.3	161.7	–
Date (days after the event)	13/01/2010 (1)	23/05/2008 (11)	10/8/2009 (8)	07/2004 (n.a.)
Test area [km²]	1	4	1	1
landslide affected areas [%]	9.6	15.1	19.6	8.7
DEM resolution (Source resolution)	10 m (1 m LiDAR DSM)	10 m (20 m contour lines)	10 m (1 m LiDAR DSM)	10 m (1 m IFSAR DSM)

simulate realistic test cases with imagery that is typically available shortly after major events.

2.1. Test site 1: Momance River — Haiti

On January 12, 2010 an earthquake with a moment magnitude of 7.0 struck Haiti. It was caused by the rupture of a previously unknown system of a blind thrust fault (Hayes et al., 2010) and claimed approximately 230,000 victims. Landslides caused extensive yet unquantified damage at several locations (Eberhard et al., 2010), and an increased frequency of slope failures and debris flows can be expected during future rainfall events. The study site is located at the Enriquillo fault line, which forms a tectogenetic valley followed by the Momance River. The slopes are between 20 to 50° steep and show a large number of shallow debris and rock slides. Most of the gentler terrain sections are under agricultural use by hundreds of scattered family farms. Due to erosion bare soils are exposed at several locations, and the valley bottom is covered by fluvial gravel bars and fresh landslide deposits. Geoeye-1 imagery was recorded one day after the event.

2.2. Test site 2: Wenchuan town — China

The rupture of the Longmenshan fault system on May 12, 2008 ($M_L = 8.0$) triggered more than 60,000 individual slope failures (Gorum et al., in press), and approximately 30,000 of the 80,000 casualties can be attributed to the impact of landslides (Tang et al., 2010). The county capital, Wenchuan town, is located on both sides of the Min River at 1330 m.a.s.l. and is surrounded by steep terrain with average slopes of approximately 30°. The town and its surroundings were seriously affected by a large number of mainly shallow translational landslides, which are concentrated on the steepest slopes in proximity to the drainage lines. Already before the event those terrain units were rather sparsely vegetated and showed bedrock outcrops at several locations. The main land cover types are degraded mountain forest and terraced field crops, which extend to slopes of up to 35°. Because the harvest was underway at the time of the initial rupture, many fields were barren and showed similar spectral characteristics as newly triggered landslides. IKONOS imagery was acquired 11 days after the main shock.

2.3. Test site 3: Messina — Italy

On the 1st of October 2009 a series of debris flows struck several catchments a few kilometers south of the city of Messina/Sicily. The debris flows were triggered by extraordinarily intense rainfall in the afternoon of that day, which had been preceded by prolonged intense rainfall at the end of September. Thirty-one people were killed during the event and the direct economic loss was estimated as almost US$ 825 million (Civil-Protection-Sicily, 2010). The affected area comprises ten small and medium size catchments that rise from sea level to about 700 m in the Peloritani Mountains. The present land cover types comprise bare ground, crop-, shrub- and grassland, deciduous forest and rural built-up areas. Most of the landslides were initiated as shallow debris flows or slides at the upper slopes, and evolved into rapid hyper-concentrated flows along their way through the drainage network. The Quickbird imagery was recorded 7 days after the event.

2.4. Test site 4: Barcelonnette Basin — France

The Barcelonnette Basin is located in the South French Alps and characterized by a mountain climate with Mediterranean influence. The area is known for the large number of slow-moving active landslides, and in the present study a small subset comprising the Super Sauze active slow-moving mudslide (Malet, 2003) was examined. The task here was mainly to distinguish the landslide body from the surrounding badlands, and since the affected area is one compact object this rather corresponds to an image segmentation task. The available imagery is a natural color aerial photograph recorded in summer 2004.

2.5. Landslide inventories

The reference inventories for Wenchuan, Messina and Barcelonnette are based on field work and visual interpretation of aerial photographs as well as VHR satellite imagery. As detailed field investigations of the earthquake-induced landslides in Haiti have not yet been completed, the corresponding inventory is based on the interpretation of remote sensing products only. To minimize the risk of miss-mapping we considered pre-and post-event VHR satellite imagery from multiple sensors (IKONOS, Geoeye-1, WorldView-2) and a post-event LiDAR DEM for the manual delineation of affected areas.

3. Methods

At each test site we selected subsets (Fig. 1) that include landslides and spectrally similar objects, such as river plains, urban areas, roads, badlands and barren fields. A scalable segmentation algorithm (Section 3.1) was applied on the images from each area, and a comprehensive set of object metrics was calculated (Section 3.2). These processing steps (Fig. 2 a) were performed with eCognition® software, which implements nearest neighbor interpolation to resample coarser image layers to the resolution of the finer panchromatic layers. Subsequent to segmentation and metric calculation the landslide inventories compiled from field work and visual image interpretation (Section 2.5) were used to create a sample database with all objects assigned either as landslide objects (O_{LS}) or non-landslide objects (O_{NLS}, Fig. 2 b). Each image object containing at least 50% of landslide-affected area was labeled as O_{LS}, and all others

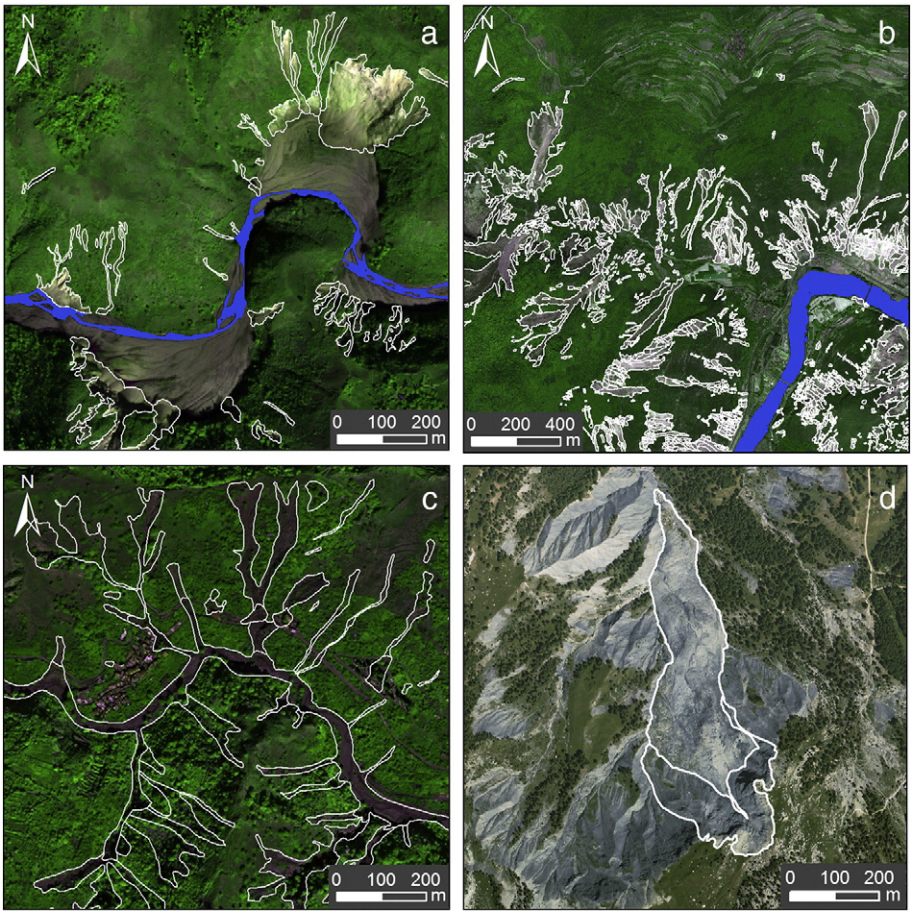


Fig. 1. Analyzed areas at the different test sites. a) Momance River, Haiti (Momance river in blue), b) Wenchuan, China (Min River in blue), c) Messina, Italy, d) Barcelonnette basin, France. White outlines indicate the landslide areas.

as O_{NLS}. Such a majority criterion was considered as the most logical choice because it minimizes the overall amount of miss-labeled areas, while retaining also marginal cases that may provide useful information for the classifier training. To evaluate a comprehensive set of object metrics (Table 2) for the discrimination of landslides and unaffected areas, all O_{LS} and an equally sized random sample of O_{NLS}, were used at all test sites and scales, respectively. They were introduced in the RF-based approach for feature evaluation and reduction (Fig. 2 c) proposed by Diaz-Uriarte and Alvarez de Andres (2006), and described in greater detail in Section 3.3.1.

Non-relevant features were subsequently removed and the data were split into training and testing sets (Fig. 2 d1). To account for spurious effects of class-imbalance and class-overlap, an iterative scheme for the adjustment of the training set was developed and tested (Fig. 2 d2, Section 3.3.2.). The classification accuracy of the approach was finally assessed on a test set comprising 80% of all image objects (Fig. 2 d3).

3.1. Image segmentation

Image segmentation generates the building blocks of OOA, and the delineation quality of the target objects has a direct influence on the accuracy of the subsequent image classification. Numerous image segmentation algorithms have been developed in the last decades and applied in remote sensing image analysis (Dey et al., 2010), all of them aiming at the delineation of relatively homogeneous and meaningful segments.

The multi-resolution image segmentation (MRIS) implemented in eCognition® software is a frequently used algorithm in Earth science studies (Blaschke, 2010). MRIS is a region-growing segmentation algorithm which, starting from individual pixels, merges the most similar adjacent regions, as long as the internal heterogeneity of the resulting object does not exceed the user defined threshold scale factor (Benz et al., 2004). Proposed statistical optimization methods (e.g. Drăguţ et al., 2010) may allow an objectification of the choice of the scale parameter if the targeted objects or processes exhibit a single operational scale. However, slope failures and surrounding land cover elements feature several orders of magnitudes in volume and area, prompting other researchers to look for automatic optimisation at multiple scales (Martha et al., in press).

To evaluate the impact of changing segmentation scales on the feature space and class separability, image segmentation was performed at 15 different scales (10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 70, 80, 90, and 100). The segmentation results depend on data characteristics such as spatial resolution, the number of bands, image quantization and the scene characteristics. The same scale factor does not necessarily yield comparable objects in different scenes, but increasing the scale factor for the segmentation of the same dataset will generally lead to larger object sizes. Thus, it is possible to emulate increasingly coarser representations of the same scene and compare resulting trends among the tested sites.

The MRIS framework offers the possibility to assign different weights to spectral bands and shape of segments. All multi-spectral bands (blue, green, red, and near-infrared [NIR]) were equally weighted with a value of one, while the panchromatic channel of the satellite images was assigned a weight of four, allowing a balance of multispectral and finer panchromatic data in the segmentation. The shape criteria were weighted with zero and, consequently, not considered in the segmentation.

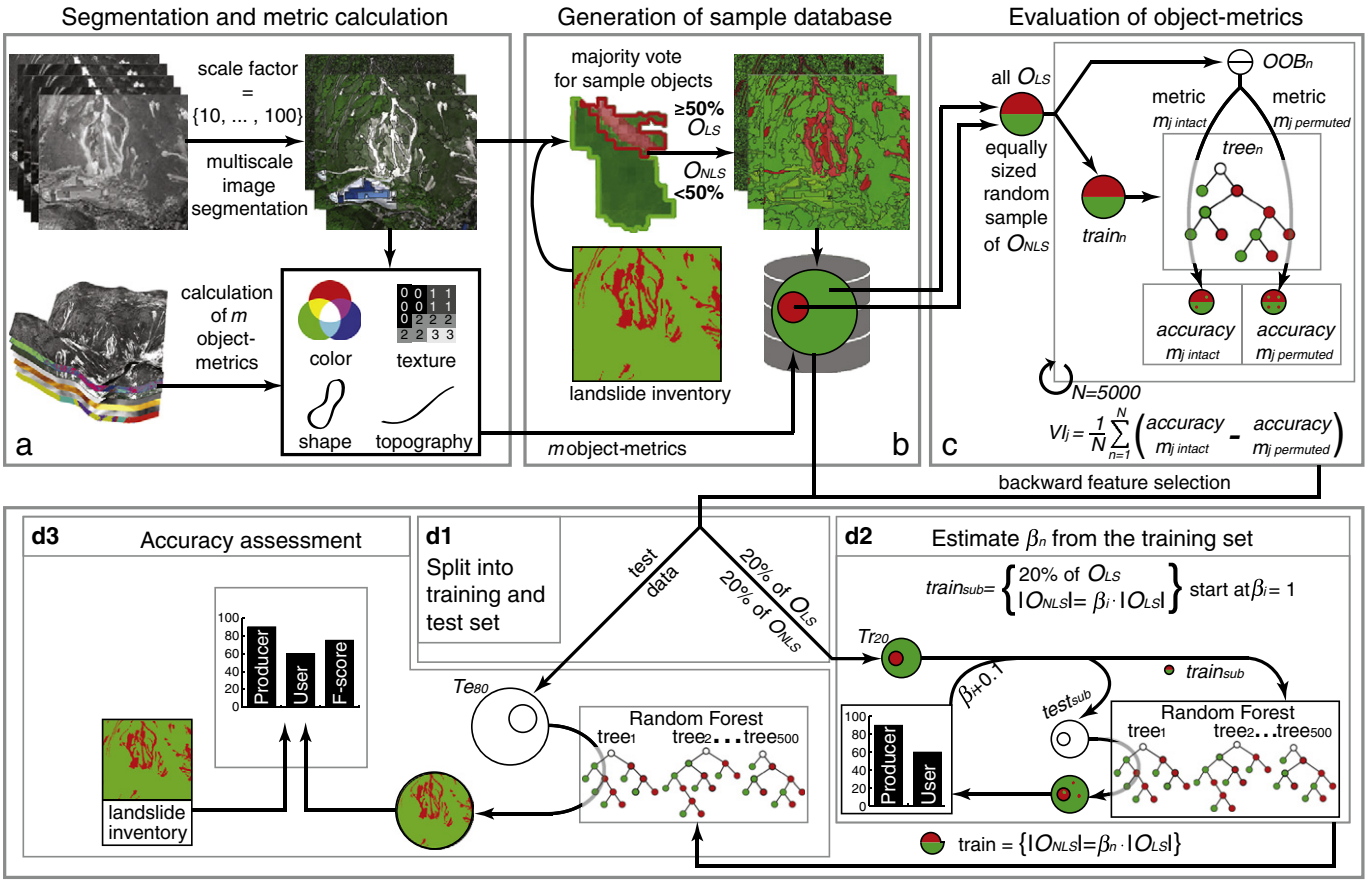


Fig. 2. Overview of the processing steps followed in this study. Explanations are given throughout the text in Section 3.

3.2. Calculation of image object metrics

This section provides an overview of features adopted from previously studies (Table 2), and introduces a number of further object metrics that were calculated subsequent to the image segmentation (Fig. 2 a). Spectral features previously recommended in the literature (Table 2) comprise band intensities, band ratios, principal component (PC) transform and brightness, and respective mean values were calculated per image object. The mean brightness (B) was defined as the sum of the object means in the visible and panchromatic band ($\bar{c}_{i(vis)}$) divided by the number of corresponding bands (n_{vis}).

$$B = \frac{1}{n_{vis}} \sum_{i=1}^{n_{vis}} \bar{c}_{i(vis)}$$

The same bands were considered to calculate *MaxDiff* for each object, defined as the absolute value of the difference of the minimum

Table 2
Overview of features used to identify landslides in previous works and adopted for this study. Most of the studies combined several attributes and are listed only exemplarily. Number in brackets indicates the number of features used with the aerial photographs.

	Tested features	No.	Case study
Spectral information	Spectral bands	5 (3)	(e.g. Nichol & Wong, 2005)
	PC	4 (3)	(Forsythe & Wheate, 2003)
	Band ratios (blue/green, green/red, red/NIR)	3 (2)	(e.g. Rau et al., 2007)
	Brightness	1	(Martha et al., 2010)
	MaxDiff	1	This study
Texture	$GLCM_{all\ dir.}$ (Ent., Mean, Cor., Con., Stdv.)	25 (15)	(Carr & Rathje, 2008; Hervás & Rosin, 1996; Martha et al., 2010; Whitworth et al., 2005)
Geometric	Shape index, compactness, roundness	3	(Moine et al., 2009)
	Length–width ratio	1	(Martha et al., 2010; Martha et al., in press)
Auxiliary data	Hillshade	1	(Martha et al., 2010)
	Slope	1	(Borghuis et al., 2007; Danneels et al., 2007)
Combined metrics	Object direction/flow direction	1	(Martha et al., 2010; Martha et al., in press)
	$GLCM_{flow.dir.}$ (Ent., Mean, Cor., Con., Stdv.)	25 (15)	This study
	GLCM (Ent., Mean, Cor., Con., Stdv.)	25 (15)	This study

object mean ($min(\bar{c}_{i(vis)})$) and the maximum object mean ($max(\bar{c}_{i(vis)})$), divided by the object brightness B.

$$MaxDiff = \frac{\left|min\left(\bar{c}_{i(vis)}\right) - max\left(\bar{c}_{i(vis)}\right)\right|}{B}$$

To quantify surface textures a variety of derivatives of the Grey Level Co-occurrence Matrix (*GLCM*) has been adopted in previous landslide studies (Table 2). Considering their large computational burden and frequent reports on strong correlations among several *GLCM* derivatives (Hall-Beyer, 2007; Laliberte & Rango, 2009), a subset of five texture measures was selected for this study. Those are *contrast* (*Con.*), *correlation* (*Cor.*), *entropy* (*Ent.*), *standard deviation* (*Stdv.*) and *Mean*. For a detailed formulation of the *GLCM* and derivatives we refer to Haralick et al. (1973) and here only recall that the co-occurrence frequencies of grey-levels are typically counted in symmetric matrices for pixels neighboring directly at 0° (N–S), 45° (NE–SW), 90° (E–W) or 135° (SW–NE), respectively. Rotation-invariance of a *GLCM* derivative can be achieved by calculating its mean or minimum value among all four directions (e.g. Pesaresi et al., 2008), or by summing up the four directional *GLCMs* ($GLCM_{all\ dir.}$) before the calculation of the derivative. The latter technique is implemented in eCognition (Trimble, 2011) and was used in this study to calculate five rotation-invariant texture measures per band directly for each image object.

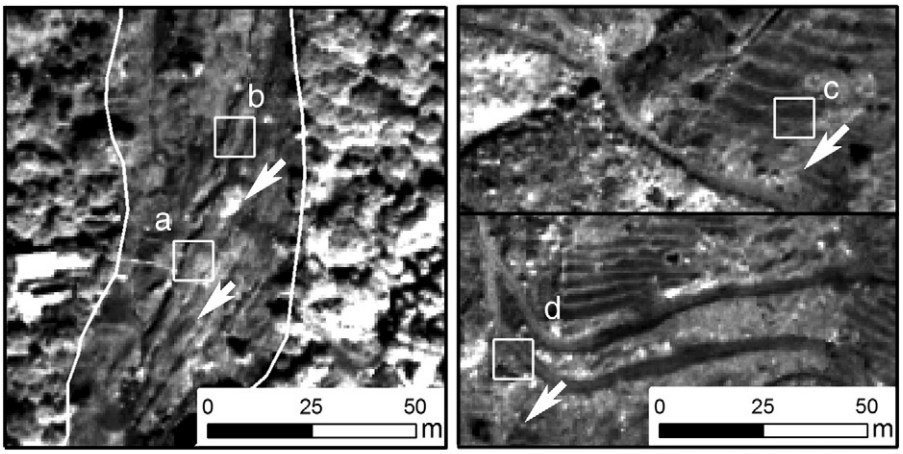
Rotation-invariance is desirable for many applications but fails to capture directional patterns in the grey-value distribution. Landslide-affected surfaces often show downslope-directed texture patterns that are potential diagnostic features to distinguish them from surfaces with texture patterns oriented at the strike of the slope (Fig. 3). In order to quantify such patterns better, additional directional texture measures were derived from two directional *GLCMs*; one computed along the hydrological flow direction ($GLCM_{flow\ dir.}$) and one perpendicular to it ($GLCM_{\perp\ flow\ dir.}$). For this purpose flow direction rasters (Jenson & Domingue, 1988) were derived from the respective DEMs (10 m resolution, Table 1) and their latticeswere superimposed on the images. For each resulting 10×10 m grid cell two directional *GLCMs* were calculated according to the direction (and the normal) indicated in the flow direction raster. Fig. 3 shows this exemplarily for *GLCM Correlation*, where the flow direction in each squared cell is aligned at 45°, and the two directional *GLCMs* consequently consider the grey-levels of pixels neighboring at 45° ($GLCM_{flow\ dir.}Cor.$) or 135° ($GLCM_{\perp\ flow\ dir.}Cor.$), respectively.

Ratio features ($GLCM_{\perp/\parallel}$) were subsequently calculated for each squared cell simply as the quotient of the texture measures computed in flow direction and their counterparts computed in the perpendicular direction. *Contrast*, *correlation*, *entropy*, *standard deviation* and *Mean* from $GLCM_{flow\ dir.}$ and their respective *GLCM* ratios are also referred to as topographically-guided texture measures. They were computed on all image bands in a 10×10 m grid and finally converted into raster layers with a pixel size of 10 m. This corresponds to 10 additional layers per band, where each image object (Section 3.1) obtains the mean layer value within its extent. Together with the texture measures from $GLCM_{all\ dir.}$ and a number of object metrics characterizing mean spectral values, shape and topographic metrics, a total of 96 and 62 features per image object were calculated for the satellite imagery and aerial photograph, respectively (Table 2).

3.3. Random Forests

Since the fundamental works on ensemble decision trees (e.g. Breiman, 2001), Random Forests (RF) have already provided promising results in fields such as genomics (Diaz-Uriarte & Alvarez de Andres, 2006), ecology (Cutler et al., 2007) and remote sensing (Lawrence et al., 2006; Watts et al., 2009).

Small changes in the training data induce a high variance in single classification trees and often lead to rather low classification accuracies (Breiman, 1996). The underlying idea of RFs is to grow multiple decision trees on random subsets of the training data and related variables. For the classification of previously unseen data, RFs take advantage of the high variance among individual trees, letting each tree vote for the class membership, and assigning the respective class according to the majority of the votes. Such ensembles demonstrate robust and accurate performance on complex datasets with little need for fine-tuning and in the presence of many noisy variables. Furthermore, integrated procedures for variable assessment and selection, and freely available high-



	Object	$GLCM_{all\ dir.}$ Cor.	$GLCM_{flow\ dir.}$ Cor.	$GLCM_{\perp flow\ dir.}$ Cor.	$GLCM_{\perp/\parallel}$ Cor.
a	debris flow	0.877	0.879	0.812	1.083
b	debris flow	0.870	0.892	0.780	1.144
c	field	0.881	0.864	0.959	0.901
d	street	0.879	0.751	0.783	0.959

Fig. 3. Exemplary comparison between the rotation-invariant and topographically-guided *GLCM Cor.* at the Messina test site. The texture measures are calculated on the panchromatic channel. White arrows indicate the hydrological flow direction within the measured cells. For linear structures along the flow direction (debris flows) values of $GLCM_{flow\ dir.}$ *Cor.* tend to be lower and values of $GLCM_{\perp flow\ dir.}$ *Cor.* tend to be higher. Hence, their ratio ($GLCM_{\perp/\parallel}Cor.$) is typically lower for linear structures aligned perpendicular to the flow direction (e.g. fields, streets).

quality software implementations, make RFs an interesting tool to be combined with OOA. In this work we extensively used the `randomForest` package (Liaw, 2010) and its extension for variable selection `varSelRF` (Diaz-Uriarte, 2010) implemented in the `R` statistical programming environment (R-Development-Core-Team, 2009).

3.3.1. Evaluation and selection of object metrics

As a starting point we were interested in object metrics that are actually helpful to distinguish landslides from other image objects, and in understanding how their performances depends on the scale of the image segmentation. For this purpose a RF–based variable importance measure was used to evaluate the object metrics at each test site with 15 different segmentation scales (10–100). RF offers a number of internal measures to estimate the importance of employed variables for the accuracy of a given classification. The properties of those measures have been intensively studied in recent years, and the so-called permutation importance is considered a computationally tractable choice for the screening of large datasets (Nicodemus et al., 2010). The permutation importance, subsequently termed variable importance (*VI*), is calculated as follows.

The original training data are resampled randomly (with replacement) to create a training set ($train_n$, Fig. 2 c) and build a classification tree. Considering a total number of m extracted object-features (Fig. 2 a) at each tree node a subset $\sqrt{m}$ features is randomly selected and tested for the best split. Approximately one third of the instances are left out of the training set and remain as *out-of-bag* sample (OOB_n, Fig. 2 c) that can be used to assess the classification accuracy of the tree. The importance of a feature m_j for the correct classification is estimated by permuting the feature values within the OOB_n sample and calculating the difference of prediction accuracies before and after the perturbation. The *VI* of the variable m_j (VI_j, Fig. 2 c) results from averaging the permutation importance of m_j over a large number of trees (N = 5000, Fig. 2 c). In the present study it provided a measure for the utility of the different object metrics to distinguish between O_{LS} and O_{NLS}. In order to give equal weight to both classes, at each test site all O_{LS} and an equal number of randomly sampled O_{NLS} were taken into account. *VIs* were calculated for all variables at the 15 different segmentation scales (Section 3.1), where the overall number of sample objects varied between a few hundred at the coarsest scale and more than 60,000 at the finest scale.

Diaz-Uriarte and Alvarez de Andres (2006) proposed to compute the *VI* from a large RF (N = 5000) to obtain an initial variable ranking and then proceed with an iterative backward elimination of the least important variables. In each iteration the least important 20% of the features are dropped, a new RF (N = 2000) is trained with the remaining feature set, and the *OOB* sample is used to assess its miss-classification rate (*OOB error*). The final features set is selected according to the RF that produces the lowest *OOB error* (Fig. 4). In the present study this procedure was used to determine the set of object metrics that were used for the construction of the final RF classifiers (Fig. 2 d1–3).

3.3.2. Balancing of error rates and accuracy assessment

At all four test sites landslides covered only minor fractions of the scene (Table 1). This is a typical situation leading to an imbalance between O_{LS} and O_{NLS}, and potentially introduces a bias of the classification towards the over-represented non-affected area. Preliminary test runs adopting naturally imbalanced training sets indeed demonstrated serious underestimation of the landslide class, suggesting the presence of the class-imbalance problem. Such biases are undesirable in any manual or automated landslide mapping, because an over- or underestimation of the affected areas would generally lead to a respective over- or underestimation of the associated hazards and risks.

Numerous methods to account for such effects have been proposed in the context of different statistical learning techniques. For logistic regression they may involve prior corrections and weighting methods (King & Zeng, 2001) or asymptotical coefficient estimates for infinite

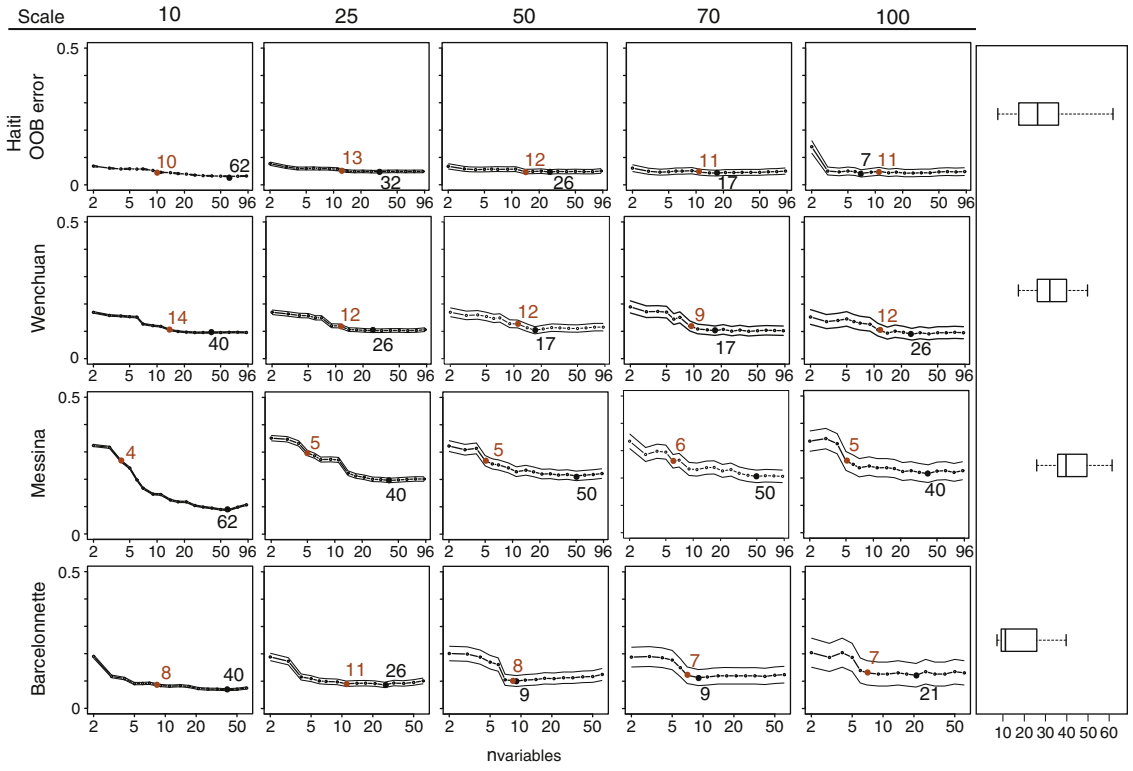


Fig. 4. Feature selection histories at the four test sites shown exemplarily for 5 of the 15 segmentation scales. The *OOB error* is drawn with a margin of one standard error against the number of selected variables. Black dot: Model with the smallest *OOB error* and number of selected metrics. Brown dot: Highest ranked texture measure with the respective rank. Boxplots indicate the variability of the number of selected features among all 15 segmentation scales.

class imbalance (Owen, 2007). Many approaches have also been developed for nonparametric learning algorithms such as RF, and may be grouped into resampling, cost-sensitive learning and kernel methods (He & Garcia, 2009). None of the methods proposed so far provides a universal advantage in all situations, but it has been demonstrated that an undersampling of the majority class is a beneficiary strategy in many different setups (Blagus & Lusa, 2010; Burez & Van den Poel, 2009). The sampling of a balanced training set might in some situations be sufficient to compensate class-imbalance when using RF (e.g. Fusaro et al., 2009), but the optimal class distribution will generally depend on the specific method and studied case (Burez & Van den Poel, 2009).

For the analysis of class imbalance and the final accuracy assessment only the previously selected object metrics were used. 20% of each dataset were sampled randomly for training (Tr_{20}, Fig. 2 d1–2) and the remaining 80% were used as test set (Te_{80}, Fig. 2 d3). To estimate the class ratio in the training sample that leads to a balance of commission and omission errors an iterative procedure was implemented and tested, where Tr_{20} was split repeatedly into subsets for training ($train_{sub}$) and testing ($test_{sub}$, Fig. 2 d2). The parameter β_i was defined as the ratio of O_{LS} and O_{NLS} in the current $train_{sub}$, and changed systematically to approximate a target value β_n yielding a balance between user's and producer's accuracy on the $test_{sub}$. In each iteration 20% of the O_{LS} and β_i-fold number of O_{NLS} were sampled randomly from Tr_{20} to train a RF ($N = 500$) and assess the classification accuracies on the remainder $test_{sub}$ (Fig. 2 d2). The procedure started from a balanced class distribution ($\beta_i = 1$) and in each step β_i increased by 0.1 (Fig. 2 d2). The underlying assumption was that the estimated β_n could be applied to adjust the class-balance for the entire training set Tr_{20}, and would also yield balanced user's and producer's accuracies in the classification of the actual test set Te_{80}.

For each β_i the procedure was repeated ten times using replicates of $train_{sub}$ and $test_{sub}$ randomly sampled from Tr_{20}. Mean error rates and their standard deviations were calculated from ten runs, and in cases where the observed variance in the resulting learning curves were too high for the determination of a unique β_n the number of random replicates was increased (Section 4.3.1).

To assess the accuracy of the described RF framework, RFs ($N = 500$) were trained with β_n-adjusted subsets of Tr_{20}, and applied on the remainder 80% test sample (Te_{80}, Fig. 2 d3). At each test site the sample balancing and accuracy assessments were performed exemplarily at a fine, medium and small segmentation scale (10, 30, and 70), and compared to reveal the effects of the segmentation on the user's, producer's and overall accuracies.

4. Results and discussion

4.1. Effects of scale on variable importance and selection

In none of the examined cases the *OOB error* reduced if more than 77 object metrics were introduced, and on average only about one third of the pre-selected metrics were detected as useful. In most cases the *OOB error* remained rather stable or increased if all variables were used. Especially at the test sites Haiti and Barcelonnette many of the object metrics provided only minor further enhancements. This is reflected by flat parts of the respective curves in Fig. 4, where slight changes of the object characteristics can have a stronger impact on the position of the *OOB error* minima, which was the criterion for the model selection. Consequently, among all segmentation scales there is a high variability in the observed overall number of selected features (boxplots Fig. 4), which coincides with those flat parts of the curves. Larger segmentation scales generally yield fewer sample objects, and consequently the standard error of the *OOB error* estimate increased (Fig. 4). It should be considered that in situations where the number of samples becomes much smaller than the number of features, the feature selection method can deteriorate strongly (Diaz-Uriarte & Alvarez de Andres, 2006; Yu et al., 2006). However, this was not an issue in the present study because even at the largest segmentation scales the number of sample objects was at least twice the number of features.

Selecting the model with the lowest *OOB error* is a rather conservative strategy that may retain some redundant and partially correlated variables. However, it was suitable for the present study in order to retain all useful features and targeting a maximal predictive accuracy. For applications where the smallest set of features with causal relationships is important (e.g. Diaz-Uriarte & Alvarez de Andres, 2006) a further reduction might be desirable, but no further enhancements of the predictive accuracy can be expected.

Although the absolute number of selected object metrics strongly depended on the particular test site and segmentation scale, some features emerged as significant in most cases and should be further highlighted. Unsurprisingly, metrics related to spectral information resulted as the most important ones for all test cases and scales (Table 3). The band ratios and PC that depict the contrast between vegetated and non-vegetated areas ranked with a particularly high variable importance (*VI*). Object means of the slope and hillshade layers significantly reduced the error rates, but in most cases their relative importance decreased with larger segmentation scales (Fig. 5). Shape metrics displayed a rather contrary behavior (Fig. 6), and generally contributed little to the reduction of the error rates. Only for larger segmentation scales at Wenchuan and Messina, where the segments more closely approached the elongated shape of the landslides, shape metrics were selected by the selection procedure. They have been reported as useful after initial spectral classification steps (Martha et al., 2010; van der Werff & van der Meer, 2008), but provide little additional information within the tested sample-based framework.

However, the *VI* ranks of the most important spectral and textural metrics exhibited low variability among the different segmentation scales (σ in Table 3) and were not subject to a persistent trend. The topographically-guided *GLCM Con.*, *Cor.* and *Ent.* helped to reduce the *OOB error* at all tested sites and largely outperformed the rotation-invariant *GLCMs*. Furthermore, the topographically-guided *GLCM Con.* was apparently more efficient when derived from the higher resolution panchromatic channels. Both rotation-invariant and topographically-guided versions of *GLCM Mean* and *Stdv.* were frequently included in the selected models, but the rotation-invariant versions were in most cases ranked higher, indicating that the topographic control did not enhance the significance of *GLCM Mean* and *Stdv*. Although *GLCMs* have been previously adopted for landslide mapping (e.g. Martha et al., 2010) the proposed topographic control on their calculation provides significant enhancement (Fig. 4), and makes such object metrics potentially useful for the automated mapping of various geomorphological processes.

Although the optimal choice of the texture measures depends to a certain degree on the application, it is interesting to note that Clausi (2002) highlighted *Con.*, *Cor.* and *Ent.* as particularly useful *GLCM* derivatives for the recognition of sea ice, and Laliberte and Rango (2009) concluded that *Con.*, *Ent.* and *Stdv.* are the most suitable texture measures for rangeland mapping.

4.2. Effects of the feature reduction on the predictive accuracy

The *OOB errors* reported during the feature selection process (Fig. 4) are not suitable to assess the predictive accuracies of the models because (i) in a real case only a fraction of the O_{LS} would be available for training, (ii) the set of optimal features may differ among subpopulations (Diaz-Uriarte & Alvarez de Andres, 2006), and (iii) the overall *OOB error* does not inform about commission and omission errors.

Those facts motivated a further experiment in which the training sets included only 20% of all O_{LS} (number of O_{LS} in Table 4) and an

Table 3
The 20 object metrics with the highest average variable importance rank among all 15 tested scales and at each respective test site. The number of scales at which the variable has been selected (n_{sel}), and the standard deviation of the rank among all 15 scales (σ_{rank}), are provided as indicators for the stability of the variable importance.

Messina		Haiti		Wenchuan		Barcelonnette	
Feature	n_{sel}/σ_{rank}	Feature	n_{sel}/σ_{rank}	Feature	n_{sel}/σ_{rank}	Feature	n_{sel}/σ_{rank}
Red/NIR	15/0.0	Red/NIR	15/0.0	Red/NIR	15/0.0	Blue/Green	15/0.0
NIR	15/0.2	Slope	15/0.2	PC 2	15/0.0	PC 2	15/0.0
PC 1	15/0.4	Green/red	15/0.4	Red	15/0.5	Max. Diff.	15/1.0
Max. Diff.	15/1.2	Red	15/1.2	Green/red	15/1.0	Blue	15/0.8
$GLCM_{flow.dir.}$ Con. PAN	15/1.6	PC 1	15/1.6	Blue	15/0.5	PC 3	15/1.8
$GLCM_{flow.dir.}$ Cor. PAN	15/2.2	Blue	14/2.2	Blue/green	15/1.0	Slope	15/2.6
Blue/Green	15/3.9	Blue/green	15/3.9	Green	15/0.6	Hillshade	15/1.4
GLCM Cor. PAN	15/2.4	Green	14/2.4	PC 1	15/1.1	PC 1	9/2.9
GLCM Con. PAN	15/3.0	PC 2	14/3.0	Brightness	15/1.0	$GLCM_{flow.dir.}$ Con. Blue	9/2.4
GLCM Cor. Red	15/4.4	PAN	14/4.4	Slope	15/2.8	Brightness	9/2.1
GLCM Cor. Green	15/4.3	$GLCM_{flow.dir.}$ Con. PAN	15/4.3	PAN	15/1.9	Green/red	8/2.5
Blue	15/6.8	NIR	13/6.8	*GLCM* Con. PAN	15/1.3	$GLCM_{flow.dir.}$ Con. Red	9/3.4
Slope	15/4.4	Hillshade	12/4.4	*GLCM* Cor. Green	13/2.7	$GLCM_{flow.dir.}$ Cor. Blue	7/2.4
GLCM Cor. blue	15/4.9	$GLCM_{flow.dir.}$ Cor. PAN	13/4.9	*GLCM* Cor. blue	14/3.0	$GLCM_{flow.dir.}$ Con. Green	8/3.3
GLCM Con. blue	15/3.8	Max. Diff.	14/3.8	*GLCM* Con. blue	13/2.9	$GLCM_{flow.dir.}$ Con. red	9/3.4
PC 2	14/7.8	Brightness	13/7.8	*GLCM* Cor. red	12/4.0	Green	6/2.8
GLCM Con. red	15/3.9	*GLCM* Con. PAN	12/3.9	Max. Diff.	14/4.0	$GLCM_{flow.dir.}$ Con. green	7/2.8
$GLCM_{flow.dir.}$ Cor. blue	15/3.9	PC 3	11/3.9	$GLCM_{flow.dir.}$ Cor. PAN	13/2.4	Red	7/2.2
$GLCM_{flow.dir.}$ Con. blue	15/5.1	*GLCM* Cor. PAN	10/5.1	NIR	10/8.5	*GLCM* Con. blue	5/5.4
GLCM Con. green	15/4.7	$GLCM_{flow.dir.}$ Ent. blue	9/4.7	*GLCM* Con. green	11/3.7	$GLCM_{flow.dir.}$ Ent. blue	4/2.5

equal number O_{NLS}, which were randomly sampled from the entire population. The training sets consequently comprised between 3% (Barcelonnette) and 11% (Wenchuan) of the datasets, while the classification accuracies were assessed on the remaining test sets in terms of correctly classified objects.

RFs (N = 500) were trained and tested using once all object features and once only the previously selected feature subsets (Section 4.2). As expected the F-measures, which are the harmonic means of user's and producer's accuracies, indicated a generally lower predictive power than the *OOB errors*, but also enhanced accuracies if only the previously selected object metrics were used (Fig. 7). Especially for the cases Messina and Barcelonnette, with rather low overall accuracies, the feature reduction enhanced the F-measures by up to 5%.

More importantly, the test revealed that a balanced training sample did not provide balanced user's and producer's accuracies, and the RF overestimated the landslide area in all cases (Fig. 7). It could be argued that for hazardous processes such as landslide an over-detection might be easier to accept than omission. However, uncertainties in landslide inventories propagate forward into susceptibility assessment (e.g. Galli et al., 2008), and a high error of commission would lead to unrealistic overestimates of the associated hazard and risks. Under the assumption that in a real case it might be an acceptable additional labor to provide further O_{NLS} samples, the

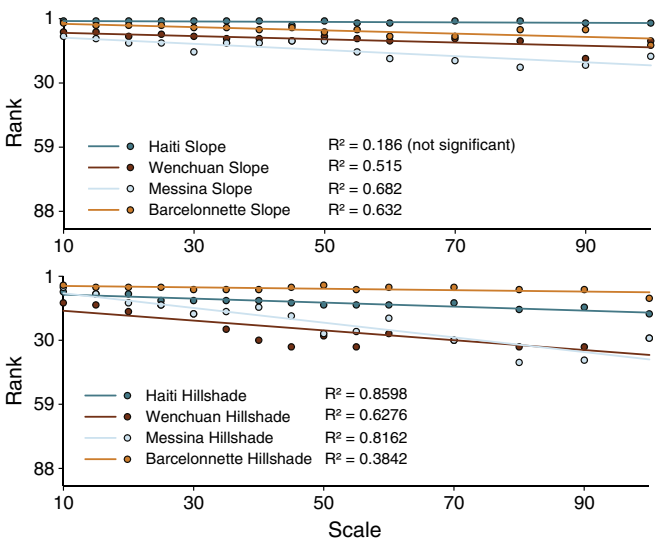


Fig. 5. Relationships between the VI-ranks of slope and hillshade and the segmentation scales. Linear regression lines fitting the data series show the overall trends, and their significance was tested at $p<0.05$ level.

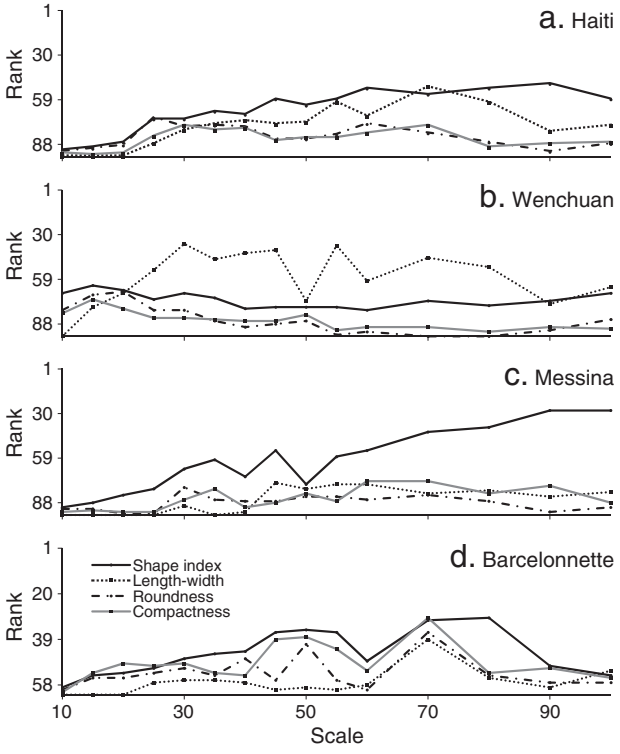


Fig. 6. Dependency of the variable importance of shape-metrics on the segmentation scale for the different test sites. Especially with small scale factors (<25) the segmentation did not generate landslide objects and non-landslide objects with distinguishable shapes. Only the length–width ratio (b, between scale 30 and 80) and the shape index (c, scale>55) had some impact on the accuracy.

Table 4
Final accuracy assessment for all test sites at three exemplary segmentation scales. Accuracies show the average performance of RFs (N = 500), trained with 20% of the O_{LS} and β_n–fold amount of O_{NLS}, applied to the test set Te_{80}.β_o is the original class-ratio of the entire population. The mean accuracies and their standard deviations were calculated over 50 randomly resampled replicates of Tr_{20}. The best results for each test site are indicated with bold numbers.

	Scale	$\beta_n(\beta_o)$	User's accuracy [%]	Producer's accuracy [%]	F_{area} [%]	F_{obj} [%]	β_n-adjusted Tr_{20}		
							O_{LS}	O_{NLS}	% of all objects
Haiti	10	3.0 (5.8)	**88.8 ± 0.1**	**85.7 ± 0.2**	**87.1 ± 0.1**	89.7 ± 0.1	4512	13536	11.7
	30	2.3 (4.2)	82.8 ± 1.2	87.1 ± 0.9	84.9 ± 0.7	88.3 ± 0.3	564	1297	12.8
	70	2.6 (4.0)	88.5 ± 1.1	72.4 ± 1.3	79.6 ± 0.7	88.5 ± 0.5	149	387	14.3
Wenchuan	10	2.7 (3.4)	**81.3 ± 0.1**	**81.1 ± 0.1**	**81.2 ± 0.1**	80.5 ± 0.1	6535	17645	17.0
	30	2.5 (3.0)	81.2 ± 0.4	77.1 ± 0.5	0.791 ± 0.2	80.3 ± 0.2	570	1425	17.4
	70	2.0 (2.6)	77.7 ± 0.9	75.3 ± 1.1	76.5 ± 0.6	79.9 ± 0.6	125	250	16.5
Messina	10	1.8 (4.2)	**72.9 ± 0.3**	**74.6 ± 0.2**	**73.7 ± 0.1**	73.0 ± 0.1	6135	11043	10.8
	30	1.9 (4.1)	69.0 ± 1.2	60.9 ± 0.9	64.7 ± 0.4	59.2 ± 0.4	663	1260	11.3
	70	1.9 (3.7)	64.3 ± 2.0	59.8 ± 1.3	62.0 ± 0.8	60.5 ± 1.1	125	238	11.9
Barcelonnette	10	4.7 (9.5)	**77.8 ± 1.0**	**78.0 ± 0.5**	**77.9 ± 0.4**	76.5 ± 0.2	1810	8507	10.8
	30	5.5 (11.5)	74.7 ± 2.1	75.9 ± 1.8	75.2 ± 1.0	67.4 ± 0.8	237	1304	10.1
	70	4.9 (12.1)	63.3 ± 5.6	88.6 ± 2.3	73.3 ± 3.5	65.3 ± 2.7	46	226	8.9

next section of this paper examines a procedure to balance user's and producer's accuracy.

4.3. Accuracy assessment

The balancing of under- and over-detection and the final accuracy assessment (Fig. 2 d1–3) were performed at three exemplarily selected scales (10, 30, and 70) with the previously selected features and in a scenario where 20% of the data would be available for training. The datasets were split (Fig. 2 d1) into a training subset (Tr_{20}), used for the estimation of the class balance and the classifier construction, and a testing subset for the final accuracy estimate (Te_{80}).

4.3.1. Estimates of β_n from the training samples (Tr_{20})

For all cases we observed a strong over-prediction of landslide areas if a class-balanced training sample was employed. The over-prediction problem was more pronounced for Messina and Barcelonnette, where already visual examination of the images suggested a higher class-overlap than in the two other areas. In controlled experiments such a behavior of classifiers has been explained by a higher density of positive examples in the class-overlap region (e.g. García et al., 2007).

Nevertheless, the iterative increase of β_i described in Section 3.3.2 (Fig. 2 d1), which corresponds to a relative increase of O_{NLS} in the training sample ($train_{sub}$), was an efficient strategy to adjust the balance of user's and producer's accuracies in the test sets ($test_{sub}$). At all test sites the estimated ratios of β_n (Fig. 8) resembled solutions that were a trade-off between the natural class-distribution (Table 4) and a completely balanced sample. The highest β_n estimates were obtained for the Barcelonnette dataset, where also the over prediction problem was most prominent.

Larger segmentation scales generally lead to a smaller number of image objects, and the 20% benchmark for the proportion of training data consequently translated into a reduced number of training samples over constant areas. Fig. 8 shows that the reduced number of sample objects resulted in an increasingly large variability in the test set accuracies, and yielded larger uncertainties in the estimation of β_n. It can be demonstrated that in such cases an increased number of random replications for each β_i still led to smoother converging curves with one unique crossing (Fig. 8). However, it should also be stressed that the estimation of β_n still only provides an intelligent guess on the design of the training sample for the classification of "unknown" image objects. The efficiency of the estimated β_n, to generate an RF with balanced user's and producer's accuracies was examined for the test set Te_{80} as described in the final section of this paper.

4.3.2. Estimation of the accuracy on the test set (Te_{80})

The majority class (O_{NLS}) in the training sample (Tr_{20}) was under-sampled according to the estimated ratio β_n. A RF (N = 500) was constructed from the β_n-adjusted Tr_{20} and applied to the remainder test set Te_{80} to assess the efficiency of the β_n estimate and the overall

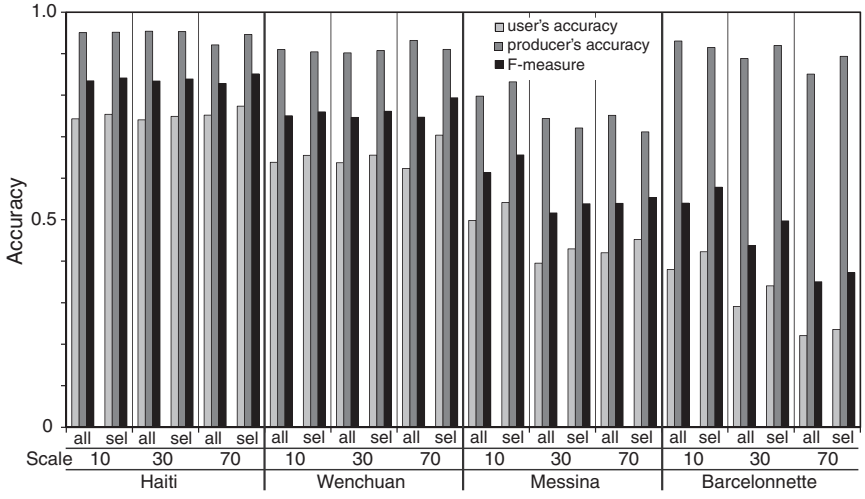


Fig. 7. Comparison of the accuracies of correctly classified objects before (all) and after (sel) variable selection at the different test sites and three exemplarily selected segmentation scales. 20% of all O_{LS} and an equal number O_{NLS} are used for the training of a RF (N = 500).

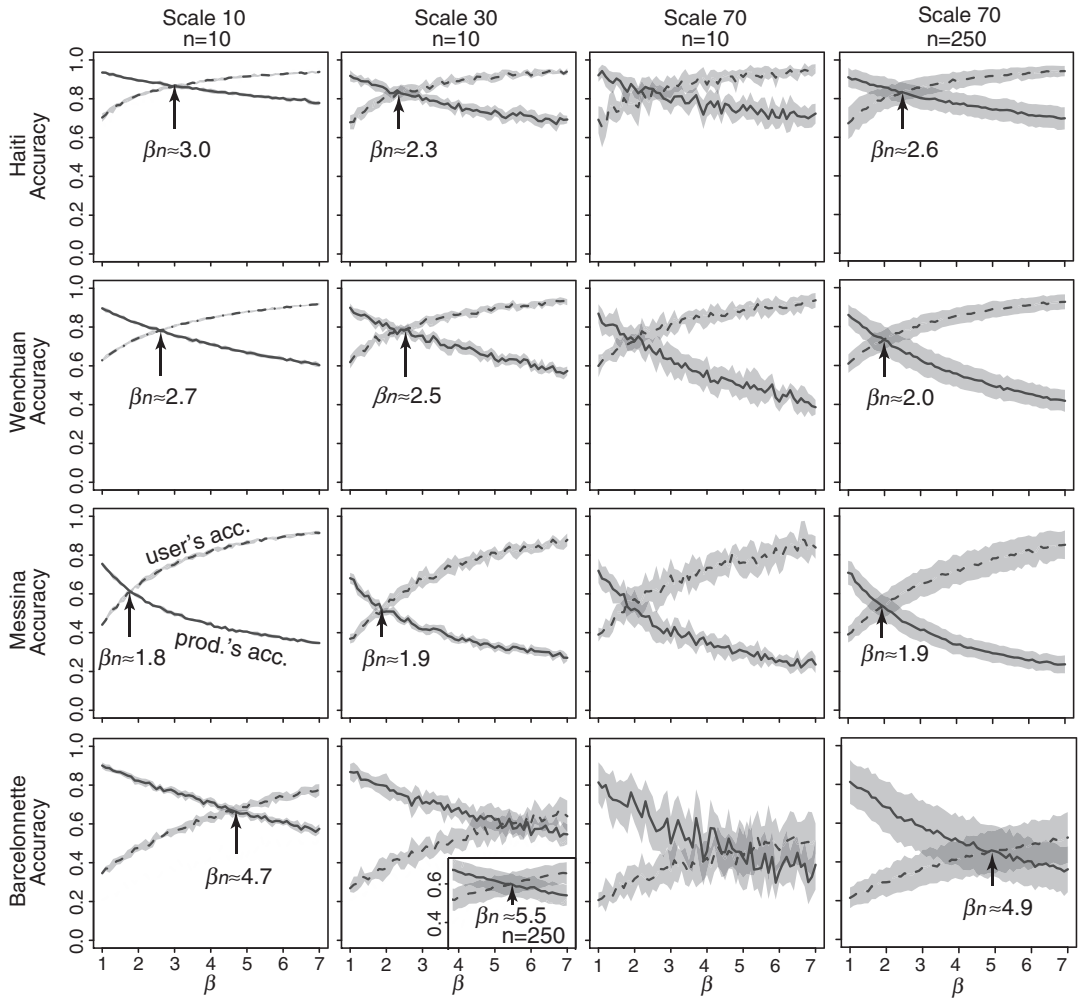


Fig. 8. Estimates of the class balance (β_n) in the training sample that lead to a balance of the mean user's (dashed black line) and mean producer's accuracies (solid black line). Accuracies are expressed in terms of image objects. The means of the accuracies for each β were calculated from 10-fold random replicate runs (n = 10). The grey margins show the corresponding standard deviations. For learning curves with high variance additionally figures from 250 random replicate runs (n = 250) are presented.

accuracy (Fig. 2 d3). The accuracy was assessed in terms of objects (Table 4, F_{obj}) and, to provide a final accuracy estimate for the entire approach, furthermore by comparing the classified areas with the landslides and non-landslide areas in the manually elaborated inventories (Table 4, user's accuracy, producer's accuracy, F_{area}). Each test was repeated with 50 β_n-adjusted randomly sampled replicates of Tr_{20}. The means and standard deviations of the achieved accuracies were calculated from the 50 runs and are displayed in Table 4. Although it did not solve the problem entirely, the strategy provided a significantly better balance between user's and producer's accuracies than could be achieved with the natural class distribution or an *ad hoc* balanced training sample (Fig. 7).

The accuracies in terms of correctly predicted area generally decreased for larger segmentation scales. At the test sites Haiti and Wenchuan this must be attributed to an increasing misfit between segmented object boundaries and the reference inventory leading to greater impurities within mixed objects. This means that the misclassified area increased due to a stronger generalization of the segments with a larger scale factor, the predictive accuracy of the RF (expressed by F_{obj}, Table 4) remained nearly constant among the different scales.

Conversely, for Messina and Barcelonnette F_{obj} was consistently lower than F_{area} (Table 4), and the classifier performance decreased significantly with larger scale factors. The comparatively higher areal accuracy can be explained by the fact that the average size of correctly classified objects was greater than those of misclassified objects. Spectral confusion and hence the importance of additional textural and topographic features was higher for the classification of the datasets from Barcelonnette and Messina (Table 3, Fig. 4). Leaving such features unconsidered during the segmentation may contribute to a higher class-overlap and a consequently lower F_{obj} at larger scales.

The general observations for the Messina test site confirm once more that omission is an especially likely error for the mapping of debris flows (Barlow et al., 2006; Lu et al., 2011), due to a high probability of occlusions in the local topography and beneath the remaining vegetation. At the Barcelonnette site most of the spectrally very similar badlands (Fig. 1 d) were successfully distinguished (Fig. 9 d) through the combination of spectral, textural and morphological features. Spatial clustering of missed areas at the crown and the toe of the landslide (Fig. 1 d, Fig. 9 d) indicates that such a landslide complex might be still better treated as a multi-class problem.

In summary, the RF classifier provided relatively high accuracies of up to 87% for the test sites Haiti and Wenchuan, while in the case of Messina the best model reached an accuracy of 73%. Those figures are in a similar range as the results of other recent studies on landslide mapping from optical imagery (Barlow et al., 2006; Lu et al., 2011; Martha et al., 2010). Though the quantities of employed samples are not always explicitly mentioned (Barlow et al., 2006; Nichol & Wong, 2005), all proposed solutions depend on the availability of some sort of training data. Once the samples are provided, the framework

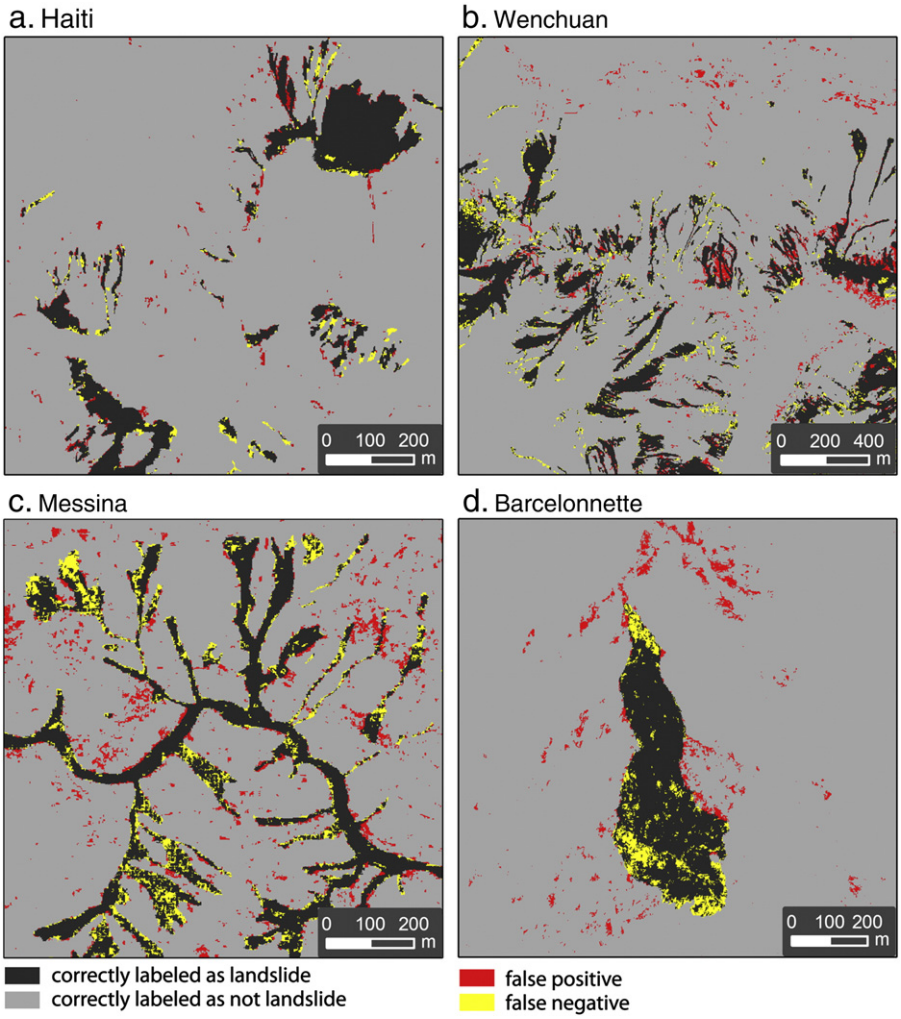


Fig. 9. Results with a segmentation scale of 10, after feature selection and balancing of the error rates as indicated in Table 4 at a, Haiti b, Wenchuan c, Messina and d, Barcelonnette. Correctly classified areas include the samples used for training.

presented in this paper has the potential to run fully automated with different image types, and liberates the user from the selection of appropriate features and thresholds. At each of the four test sites a medium resolution DEM, one VHR resolution image, and the reported numbers of training objects were sufficient for an efficient performance of the RF classifiers. In most practical situations such kind of data will be available, and the described algorithms may provide a generic approach to map the overall affected area more efficiently before site and data specific tasks, such as the classification of landslide types (Barlow et al., 2006; Martha et al., 2010), are targeted. In situations where more data (e.g. pre-event imagery) are available the proposed framework is suitable to accommodate a large variety of additional datasets and object metrics, which may be used to further increase the mapping accuracies.

In order to differentiate individual landslides and provide map products with less dispersed class distributions (Fig. 9) the current architecture still needs enhancements. This is closely related to the observed fact that high-level features such as shape are better exploited on larger scales (Section 4.1). The design of a hierarchical algorithm that robustly and efficiently incorporates sample data and relevant features in the classification, and delineation of image objects among a number of different scales, remains a major challenge, and with potential benefits for many remote sensing applications.

It also has to be noted that at this point we only explored the technical aspects of the supervised framework in relatively small test areas. A detailed analysis of the impact of sample quality and quantity provided by different users and over larger areas was beyond the scope of the present study, while research in this direction is certainly desirable before an operational use of the technique.

5. Conclusions

Previously proposed methods for object-oriented mapping of landslides from VHR images are highly reliant on manual thresholding and a subjective selection of suitable features, making it difficult to adapt them to new locations and datasets. To overcome such issues this study investigated the use of image segmentation and the Random Forest framework for feature selection and image classification. A variety of VHR remote sensing images and different landslide processes was analyzed with the RF data-mining technique to evaluate useful image object metrics, the influence of the segmentation scale, and the consequences of class-imbalance.

Although the optimal set of object metrics varies considerably from case to case, a number of spectral, topographic, and textural features are generally useful. Rotation-invariant and topographically-guided GLCMs provide complementary information to distinguish affected from non-affected areas, while topographically-guided GLCM derivatives introduced in this paper provide more significant enhancements. They also appear potentially useful for the automated mapping of other geomorphological processes such as gully erosion

(Shruthi et al., in press) or fluvial sediments. The range of potentially useful object metrics for landslide mapping and other applications seems still not fully exploited, and data mining techniques such as RF are valuable tools to ease feature selection for machine learning, or to guide experts during the elaboration of knowledge-driven rule sets. Our results indicate that feature reduction leads to an improved image classification, but also that not all significant features can be fully exploited with one particular segmentation scale.

Class-imbalance and class-overlap caused severely imbalanced error rates at all test sites. An iterative scheme to estimate a compensating class balance for the training data was found to enhance substantially the balance of user's and producer's accuracies. In the presented setup, accuracies between 73% and 87% were achieved when 20% of the total area was provided for training.

In the short term further enhancements are certainly possible through the integration of ancillary datasets such as pre-event imagery or the exploration of additional object metrics. More research is needed to optimize the segmentation process, which at present is based on spectral information solely. An initial sample-based estimate of the variable importance might thereby be an interesting tool to decide which further layers should be included in the segmentation. The processing time for a small test area can be streamlined to a few hours on a standard desktop PC, while for larger areas the RFs can easily be implemented for parallel processing, and the scale factor may provide an interesting parameter to trade between accuracy and processing time.

Acknowledgments

The work described in this paper was supported by the project SafeLand "Living with landslide risk in Europe: Assessment, effects of global change, and risk management strategies" under Grant Agreement No. 226479 in the 7th Framework Programme of the European Commission. This support is gratefully acknowledged.

References

Barlow, J., Franklin, S., & Martin, Y. (2006). High spatial resolution satellite imagery, DEM derivatives, and image segmentation for the detection of mass wasting processes. *Photogrammetric Engineering and Remote Sensing*, *72*, 687–692.

Barlow, J., Martin, Y., & Franklin, S. E. (2003). Detecting translational landslide scars using segmentation of Landsat ETM + and DEM data in the northern Cascade Mountains, British Columbia. *Canadian Journal of Remote Sensing*, *29*, 510–517.

Benz, U. C., Hofmann, P., Willhauck, G., Lingenfelder, I., & Heynen, M. (2004). Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *ISPRS Journal of Photogrammetry and Remote Sensing*, *58*, 239–258.

Blagus, R., & Lusa, L. (2010). Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *11*, 523.

Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, *65*, 2–16.

Booth, A. M., Roering, J. J., & Perron, J. T. (2009). Automated landslide mapping using spectral analysis and high-resolution topographic data: Puget Sound lowlands, Washington, and Portland Hills, Oregon. *Geomorphology*, *109*, 132–147.

Borghuis, A. M., Chang, K., & Lee, H. Y. (2007). Comparison between automated and manual mapping of typhoon-triggered landslides from SPOT-5 imagery. *International Journal of Remote Sensing*, *28*, 1843–1856.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*, 4626–4636.

Carr, L., & Rathje, E. (2008). The use of remote sensing to identify landslides caused by the 2004 Niigata-ken Chuetsu earthquake in Japan. *6th International Workshop on Remote Sensing for Disaster Management Applications. Pavia, Italy*.

Cascini, L., Fornaro, G., & Peduto, D. (2010). Advanced low- and full-resolution DInSAR map generation for slow-moving landslide analysis at different scales. *Engineering Geology*, *112*, 29–42.

Civil-Protection-Sicily (2010). Landslide and mud food emergency Messina province, Itlay, October 1st 2009. In A.f.a.f.t.E.U.S. Fund (Ed.), *Regione Siciliana — Presidenza, Dipartimento dellla Protezione Civile* (pp. 83).

Clausi, D. A. (2002). An analysis of co-occurrence texture statistics as a function of grey level quantization. *Canadian Journal of Remote Sensing*, *28*, 45–62.

Cruden, D. M., & Varnes, D. J. (1996). Landslides types and processes. In A. K. Turner & R.L. Schuster (Eds.), *Landslides: Investigation and Mitigation* (pp. 36–75). Transportation Research Board, National Academy of Sciences: Washington D.C.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random Forest for classification in ecology. *Ecology*, *88*, 2783–2792.

Danneels, G., Pirard, E., & Havenith, H. -B. (2007). Automatic landslide detection from remote sensing images using supervised classification methods. *Geoscience and Remote Sensing Symposium. Barcelona*. Spain: IGARSS.

Debella-Gilo, M., & Kääb, A. (2011). Sub-pixel precision image matching for measuring surface displacements on mass movements using normalized cross-correlation. *Remote Sensing of Environment*, *115*, 130–142.

Denil, M., & Trappenberg, T. (2010). Overlap versus imbalance. In A. Farzindar & V. Kešelj (Eds.), *Advances in Artificial Intelligence* (pp. 220–231). Heidelberg: Springer Berlin.

Dey, V., Zhang, Y., & Zhong, M. (2010). A review on image segmentation techniques with remote sensing perspective. In W. Wagner & B. Székely (Eds.), *ISPRS TC VII Symposium — 100 Years ISPRS* (pp. 31–42). Vienna, Austria: IAPRS.

Diaz-Uriarte, R. (2010). *varSelRF: Variable selection using random forests. R package version 0.7-2*.

Diaz-Uriarte, R., & Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*, 3.

Drăguţ, L., Tiede, D., & Levick, S. R. (2010). ESP: A tool to estimate scale parameter for multiresolution image segmentation of remotely sensed data. *International Journal of Geographical Information Science*, *24*, 859–871.

Eberhard, M. O., Baldridge, S., Marshall, J., Mooney, W., & Rix, G. J. (2010). The MW 7.0 Haiti earthquake of January 12, 2010. *USGS/EERI Advance Reconnaissance Team report* (pp. 58). : U.S. Geological Survey Open-File Report.

Forsythe, K. W., & Wheate, R. D. (2003). Utilization of Landsat TM and digital elevation data for the delineation of avalanche slopes in Yoho National Park (Canada). *Geoscience and Remote Sensing, IEEE Transactions on*, *41*, 2678–2682.

Fusaro, V. A., Mani, D. R., Mesirov, J. P., & Carr, S. A. (2009). Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat Biotech*, *27*, 190–198.

Galli, M., Ardizzone, F., Cardinali, M., Guzzetti, F., & Reichenbach, P. (2008). Comparing landslide inventory maps. *Geomorphology*, *94*, 268–289.

García, V., Mollineda, R., Sánchez, J., Alejo, R., & Sotoca, J. (2007). When overlapping unexpectedly alters the classimbalance effects. In J. Martí, J. Benedí, A. Mendonça, & J. Serrat (Eds.), *Pattern Recognition and Image Analysis* (pp. 499–506). Heidelberg: Springer Berlin.

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random Forests for land cover classification. *Pattern Recognition Letters*, *27*, 294–300.

Gorum, T., Fan, X., van Westen, C.J., Huang, R.Q., Xu, Q., Tang, C., & Wang, G. (in press). Distribution pattern of earthquake-induced landslides triggered by the 12 May 2008 Wenchuan Earthquake. *Geomorphology*. doi:10.1016/j.geomorph.2010.12.030.

Guo, B., Damper, R. I., Gunn, S. R., & Nelson, J. D. B. (2008). A fast separability-based feature-selection method for high-dimensional remotely sensed image classification. *Pattern Recognition*, *41*, 1653–1662.

Hall-Beyer, M. (2007). GLCM Tutorial. http://www.fp.ucalgary.ca/mhallbey/tutorial.htm [15 May 2010].

Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural features forimage classification. *IEEE Transactions on Systems, Man, and Cybernetics*, *3*, 610–621.

Hayes, G. P., Briggs, R. W., Sladen, A., Fielding, E. J., Prentice, C., Hudnut, K., Mann, P., Taylor, F. W., Crone, A. J., Gold, R., Ito, T., & Simons, M. (2010). Complex rupture during the 12 January 2010 Haiti earthquake. *Nature Geoscience*, *3*, 800–805.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *Ieee Transactions On Knowledge And Data Engineering*, *21*, 1263–1284.

Hervás, J., Barredo, J. I., Rosin, P. L., Pasuto, A., Mantovani, F., & Silvano, S. (2003). Monitoring landslides from optical remotely sensed imagery: The case history of Tessina landslide, Italy. *Geomorphology*, *54*, 63–75.

Hervás, J., & Bobrowsky, P. (2009). Mapping: Inventories, susceptibility, hazard and risk. In K. Sassa & P. Canuti (Eds.), *Landslides — Disaster Risk Reduction* (pp. 321–349). Heidelberg: Springer.

Hervás, J., & Rosin, P. L. (1996). Landslide mapping by textural analysis of ATM data. *11th Thematic Conference on Applied Geologic Remote Sensing* (pp. 394–402). Las Vegas, USA.

Jenson, S. K., & Domingue, J. O. (1988). Extracting topographic structure from digital elevation data for geographic information system analysis. *Photgrammetric Engineering & Remote Sensing*, *54*, 1593–1600.

Joyce, K. E., Dellow, G. D., & Glassey, P. J. (2008). Assessing image processing techniques for mapping landslides. *IEEE International Geoscience and Remote Sensing Symposium* (pp. 1231–1234). Boston, USA.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*, 137–163.

Kjekstad, O., & Highland, L. (2009). *Economic and social impacts of landslides* (pp. 573–587).

Laliberte, A., & Rango, A. (2009). Texture and scale in object-based analysis of subdecimeter resolution unmanned aerial vehicle (UAV) imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *47*, 761–770.

Lawrence, R. L., Wood, S. D., & Sheley, R. L. (2006). Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). *Remote Sensing of Environment*, *100*, 356–362.

Liaw, A. (2010). *randomForest: Breiman and Cutler's random forests for classification and regression, Version 4.5-36*.

Lu, P., Stumpf, A., Kerle, N., & Casagli, N. (2011). Object-oriented change detection for landslide rapid mapping. *Geoscience and Remote Sensing Letters, IEEE* (pp. 701–705).

Malet, J. P. (2003). Les 'glissements de type écoulement' dans les marnes noires des Alpes du Sud. Morphologie, fonctionnement et modélisation hydro-mécanique. *Institut de Physique de Globe* (pp. 364). Strasbourg: Université Louis Pasteur.

Martha, T., Kerle, N., van Westen, C., Jetten, V., & Kumar, V. (in press). Segment optimisation and data-driven thresholding for knowledge-based landslide detection by object-oriented image analysis. *IEEE Transaction in Geoscience and Remote Sensing*. doi:10.1109/TGRS.2011.2151866.

Martha, T., Kerle, N., van Westen, C. J., & Kumar, K. (2010). Characterising spectral, spatial and morphometric properties of landslides for semi-automatic detection using object-oriented methods. *Geomorphology*, *116*, 24–36.

Moine, M., Puissant, A., & Malet, J. -P. (2009). Detection of landslides from aerial and satellite images with a semi-automatic method. Application to the Barcelonnette basin (Alpes-de-Haute-Provence, France). In J. -P. Malet, A. Remaître, & T. Bogaard (Eds.), *International Conference 'Landslide Processes'* (pp. 63–68). Strasbourg, France.

Nadim, F., Kjekstad, O., Peduzzi, P., Herold, C., & Jaedicke, C. (2006). Global landslide and avalanche hotspots. *Landslides*, *3*, 159–173.

Nichol, J., & Wong, M. S. (2005). Satellite remote sensing for detailed landslide inventories using change detection and image fusion. *International Journal of Remote Sensing*, *26*, 1913–1926.

Nicodemus, K., Malley, J., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, *11*, 110.

Owen, A. B. (2007). Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, *8*, 761–773.

Pesaresi, M., Gerhardinger, A., & Kayitakire, F. (2008). A robust built-up area presence index by anisotropic rotation-invariant textural measure. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, *1*, 180–192.

Petley, D. (2009). On the occurrence of fatal landslides in 2008. *Geophysical Research Abstracts*, *11* (EGU2009-9030).

R-Development-Core-Team (2009). *R: A language andenvironment for statistical computing — Version 2.10.0*. Vienna, Austria: R Foundation for Statistical Computing.

Rau, J. -Y., Chen, L. -C., Liu, J. -K., & Wu, T. -H. (2007). Dynamics monitoring and disaster assessment for watershed management using time-series satellite images. *Geoscience and Remote Sensing, IEEE Transactions on*, *45*, 1641–1649.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*, 2507–2517.

Shruthi, B.V.M., Kerle, N., & Jetten, V. (in press). Object-based gully feature extraction using high resolution imagery. Geomorphology.

Tang, H., Jia, H., Hu, X., Li, D., & Xiong, C. (2010). Characteristics of landslides induced by the great Wenchuan earthquake. *Journal of Earth Science*, *21*, 104–113.

Trimble (2011). *eCognition developer 8.64.0 reference book*. München: Germany Trimble Documentation.

Van Coillie, F. M. B., Verbeke, L. P. C., & De Wulf, R. R. (2007). Feature selection by genetic algorithms in object-based classification of IKONOS imagery for forest mapping in Flanders, Belgium. *Remote Sensing of Environment*, *110*, 476–487.

Van Den Eeckhaut, M., Vanwalleghem, T., Poesen, J., Govers, G., Verstraeten, G., & Vandekerckhove, L. (2006). Prediction of landslide susceptibility using rare events logistic regression: A case-study in the Flemish Ardennes (Belgium). *Geomorphology*, *76*, 392–410.

van der Werff, H. M. A., & van der Meer, F. D. (2008). Shape-based classification of spectrally identical objects. *ISPRS Journal of Photogrammetry and Remote Sensing*, *63*, 251–258.

van Westen, C. J., van Asch, T. W. J., & Soeters, R. (2006). Landslide hazard and risk zonation — Why is it still so difficult? *Bulletin of Engineering Geology and the Environment*, *65*, 167–184.

Watts, J. D., Lawrence, R. L., Miller, P. R., & Montagne, C. (2009). Monitoring of cropland practices for carbon sequestration purposes in north central Montana by Landsat remote sensing. *Remote Sensing of Environment*, *113*, 1843–1852.

Whitworth, M., Giles, D., & Murphy, W. (2005). Airborne remote sensing for landslide hazard assessment: A case study on the Jurassic escarpment slopes of Worcestershire, UK. *Quarterly Journal of Engineering Geology & Hydrogeology*, *38*, 285–300.

Woodcock, C. E., & Strahler, A. H. (1987). The factor of scale in remote sensing. *Remote Sensing of Environment*, *21*, 311–332.

Yu, W., Wu, B., Huang, T., Li, X., Williams, K., & Zhao, H. (2006). Statistical methods in proteomics. In H. Pham (Ed.), *Handbook of Engineering Statistics* (pp. 623–636). London: Springer.

2.4 *Active learning for the reduction of labeling costs in image classification experiments*

This section is based on: Stumpf, A., Lachiche, N., Malet, J.-P., Kerle, N., Puissant, A., In Press. Active learning in the spatial domain for remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 16 pages, 10.1109/TGRS.2013.2262052.

As demonstrated in Section 2.3, supervised machine learning is a versatile tool that can be adapted to different image datasets and domain objectives. The drawback of the approach developed in Stumpf and Kerle (2011) is the extensive need for training samples which must be a good representation of the underlying classes and their distribution in the feature space. Preparing the training data still depends on visual image interpretation and/or field surveys whose implementation is costly especially to cover large areas.

Therefore, this section addresses the formulation, implementation and performance of a new active learning (AL) heuristic targeting the reduction of labeling costs and its application to landslide mapping.

The first occurrence of AL in the machine learning literature can be traced back to the 1990's (Zhang and Veenker, 1991) and refers to the hypothesis that a machine learning algorithm will perform better with less training data if some control on the selection of the training data is given. AL methods have already been introduced for remote sensing image classification (Tuia et al., 2009) but it will be shown that important aspects of the sample distribution and user interaction have been largely neglected so far. In particular a region-based strategy is proposed to identify the most valuable and interesting sample region on which the users should focus their efforts (Figure 2-2).

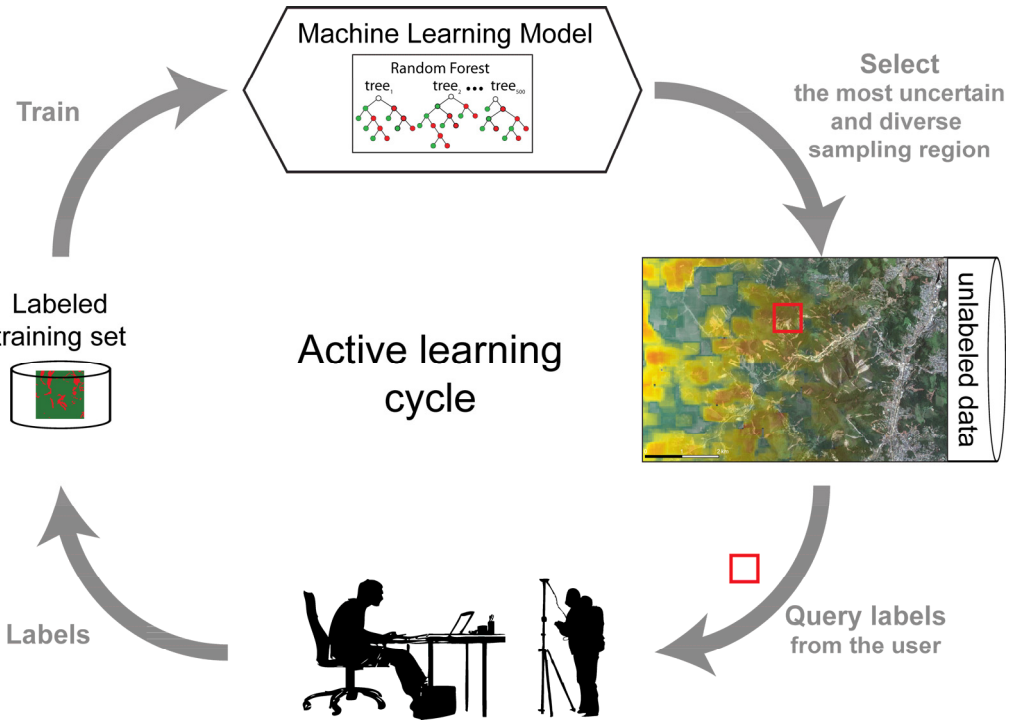


Figure 2-2: Simplified representation of the AL heuristic developed in our work (modified after Settles, 2010).

Active Learning in the Spatial Domain for Remote Sensing Image Classification

André Stumpf, Nicolas Lachiche, Jean-Philippe Malet, Norman Kerle, and Anne Puissant

***Abstract*—Active learning (AL) algorithms have been proven useful in reducing the number of required training samples for remote sensing applications; however, most methods query samples pointwise without considering spatial constraints on their distribution. This may often lead to a spatially dispersed distribution of training points unfavorable for visual image interpretation or field surveys. The aim of this study is to develop region-based AL heuristics to guide user attention toward a limited number of compact spatial batches rather than distributed points. The proposed query functions are based on a tree ensemble classifier and combine criteria of sample uncertainty and diversity to select regions of interest. Class imbalance, which is inherent to many remote sensing applications, is addressed through stratified bootstrap sampling. Empirical tests of the proposed methods are performed with multitemporal and multisensor satellite images capturing, in particular, sites recently affected by large-scale landslide events. The assessment includes an experimental evaluation of the labeling time required by the user and the computational runtime, and a sensitivity analysis of the main algorithm parameters. Region-based heuristics that consider sample uncertainty and diversity are found to outperform pointwise sampling and region-based methods that consider only uncertainty. Reference landslide inventories from five different experts enable a detailed assessment of the spatial distribution of remaining errors and the uncertainty of the reference data.**

***Index Terms*—Active learning (AL), batch-mode, class imbalance, ground truth uncertainty, image classification, landslide inventory mapping, spatial information.**

TABLE OF SYMBOLS

H	Vote entropy.
p_i	Fraction of votes for the ith class.
$W^{\times}$	Queried region.
w	Edge length of the squared query region.
M_H	Entropy map.
μ_H	Mean local vote entropy.
x, y	x, y coordinates on the regular search grid.
G	Search grid.
g	Cell size of the search grid.
X	Training set.
n	Number of iterations.
U	Unlabeled pool.
S	Set of queried samples.
s	Individual samples in different sets.
fun	Selected diversity function.
c	Individual samples in a candidate set.
σ_d	Standard deviation of the feature space distances between the candidate batch and the training set.
$\lvert W^m \rvert$	Cardinality of the candidate set.
$\rho_k(X, c)$	Euclidean distance between a candidate sample and its nearest training point in feature space.
m	Number of candidate regions.
W^m	Candidate set.
$H^{\times}$	Cross-entropy.
v_D	Volume of the unit ball.
$\mathbb{R}^D$	D-dimensional feature space.
D	Number of features.
ψ	Digamma function.
$\lvert X \rvert$	Cardinality of the training set.
k	Order of the nearest neighbor search.
t_i	Minimum variable importance threshold.
F	F-measure.

I. INTRODUCTION

MACHINE learning algorithms have become important tools for the extraction of environmental information from remote sensing images. State-of-the-art supervised algorithms, such as support vector machines (SVMs), artificial neural networks, and ensemble-based learning methods [1], have already been developed, among others, for land cover analysis [2], [3], biophysical parameter estimation [4], [5], change and anomaly detection [6], [7], and geomorphological mapping [8], [9].

Manuscript received September 16, 2012; revised January 18, 2013 and March 19, 2013; accepted April 18, 2013. This work was supported in part by the project FOSTER "Spatio-temporal data mining: application to the understanding and monitoring of soil erosion" funded by the French Research Agencya under Contract ANR Cosinus, 2011–2013; by the project SafeLand "Living with landslide risk in Europe: assessment, effects of global change, and risk management strategies" under Grant Agreement 226479 funded by the 7th Framework Programme of the European Commission; and by the project "Landslide mapping at various spatial scales" funded by the EUR-OPA Major Hazards Open Partial Agreement of Council of Europe.

A. Stumpf is with the Laboratoire Image, Ville, Environnement, Centre National de la Recherche Scientifique UMR 7362, University of Strasbourg, 67000 Strasbourg, France, and also with the École et Observatoire des Sciences de la Terre—Institut de Physique du Globe de Strasbourg, Centre National de la Recherche Scientifique UMR 7516, University of Strasbourg, 67084 Strasbourg Cedex, France (e-mail: andre.stumpf@live-cnrs.unistra.fr).

N. Lachiche is with the Image Sciences, Computer Sciences and Remote Sensing Laboratory, Centre National de la Recherche Scientifique UMR 7005, University of Strasbourg, 67412 Strasbourg Cedex, France (e-mail: nicolas.lachiche@unistra.fr).

J.-P. Malet is with the École et Observatoire des Sciences de la Terre—Institut de Physique du Globe de Strasbourg, Centre National de la Recherche Scientifique UMR 7516, University of Strasbourg, 67084 Strasbourg Cedex, France (e-mail: jeanphilippe.malet@unistra.fr).

N. Kerle is with the Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7500 Enschede, The Netherlands (e-mail: kerle@itc.nl).

A. Puissant is with the Laboratoire Image, Ville, Environnement, Centre National de la Recherche Scientifique UMR 7362, University of Strasbourg, 67000 Strasbourg, France (e-mail: anne.puissant@live-cnrs.unistra.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2013.2262052

Supervised algorithms are adaptable to a broad range of problems but depend on the availability of training data, which, in remote sensing, are typically obtained through cost- and labor-intensive field work or time-consuming visual image interpretation. Active learning (AL) has evolved as a key concept to reduce annotation costs and generally refers to systems where the learning algorithm receives some control over the selection of additional training data during several iterations [10], [11]. Common strategies are to query the labels for samples with high uncertainty, which can be measured by the ambiguity of the posterior probabilities [12], the distance to the decision boundary [13], or the disagreement of a classifier committee [14].

Recently, several AL heuristics have been introduced for remote sensing applications and demonstrated promising results in the classification of multi- and hyperspectral images [15]–[17]. Iterative retraining of the learning algorithm is typically a computational bottleneck and can be addressed with batch-mode query functions [18], [19] that select more than one sample per iteration. In addition to uncertainty criteria, these models also incorporate sample diversity in order to reject candidates that are highly uncertain but are largely redundant among each other [15].

Most AL approaches assume that annotation costs linearly depend on the size of the required training set (e.g., pixels in most cases) and, therefore, aim to reduce the overall number of queried samples. However, recent studies on practical applications of AL have demonstrated that real annotation costs can vary considerably depending on how the labeling is performed, among instances and annotators, and may change with the number of queried instances [20], [21]. Leaving real annotation costs unconsidered in many cases still hinders a successful transfer of AL algorithms to real-world applications [20].

The acquisition of training data for remote sensing applications depends to a large degree on the spatial distribution of the queried instances; however, very few studies have integrated spatial constraints into the AL algorithm [22], [23]. Most proposed AL algorithms select samples only according to their position in feature space, which typically yields a pointwise dispersed distribution of the training data in geographic space and incurs the risk of revisiting (during image interpretation or field work) the approximately same spatial location several times. Human scene interpretation generally involves the assessment of high-level contextual features [24], whereas pointwise queries do not exploit full interpreter knowledge of the spatial context around a particular point. This suggests that queries, which allow focusing on certain spatial subsets of an area of interest, are a strategy that is more aligned with human perception.

To address such issues, we propose region-based query strategies that select compact spatial batches with high sample uncertainty and diversity. Some key ideas of this approach have already been discussed in [25], whereas here, a more complete formulation, including new query criteria, considerations of class imbalance and expert uncertainties, and a thorough experimental evaluation, is provided. The proposed region-based AL is employed for a two-class problem and tested on very-high-resolution (VHR) optical remote sensing images depicting geographic sites affected by large-scale landslide events. The visual interpretation of VHR images is a common approach for the collection of training data for applications such as land cover classification, and it is still the prevailing standard for landslide inventory mapping typically performed by trained experts who annotate affected areas with marker tools on the image. Landslide inventory mapping is, therefore, a challenging problem for the development of semiautomatic image analysis techniques and a suitable example to investigate the effects of different labeling strategies.

This paper is organized in seven sections. Section II introduces closely related works on AL, and Section III details the developed methodology and baseline methods used for comparison. Section IV describes the data sets used for a series of experiments. Section V details the experimental design to evaluate the performance of the AL heuristics and the labeling time required by users. Section VI discusses corresponding results, and Section VII concludes on the outcomes of this work.

II. Related Works on AL

A review of advances in AL methods has been recently given in [11], and a comprehensive overview of AL algorithms developed for the analysis of remote sensing images was provided in [15]. As in the present study, most approaches assume that all target classes are known *a priori* to initialize the AL algorithm, whereas probabilistic approaches that also allow for the discovery of new classes are more suitable if this assumption is no longer fulfilled [26], [27]. The general underlying idea of most AL approaches is to initialize a machine learning model using a small training set and to exploit the model state and/or the data structure to iteratively select the most valuable samples that should be labeled by the user and added to the training set. With relatively few queries and labeled samples, an AL strategy should ideally yield at least the same accuracy as an equivalent classifier trained with many randomly selected samples.

Large-margin heuristics based on SVMs are frequently adopted for the design of AL algorithms in remote sensing applications [15] and have already been used as a base learner for a conceptual open-source software implementation to combine the benefits of object-oriented image analysis and AL [28]. Committee-based heuristics that select samples according to the maximum disagreement of an ensemble of classifiers have achieved promising results in a recent AL benchmark challenge [29] but are less frequently adopted for remote sensing [17]. For both approaches, iterative retraining of the classifier is typically a computational bottleneck for large data sets, and it has been demonstrated that batch-mode query functions, which consider the uncertainty and diversity of the samples [18], [19], are able to reduce the number of iterations significantly.

A very different approach to AL has been taken in recent studies for semantic image segmentation by iteratively exploring hierarchical data representations [30], [31]. They highlighted the benefits of integrating additional spatio-contextual features into the feature vector but did not explicitly consider the distribution of the queried samples in geographic space. Pasolli *et al.* [23] considered spatial distances among points to enhance margin-based sampling for pointwise queries.

Increasing the distances among the queried points beyond the range of spatial autocorrelation generally encourages querying less correlated and more representative samples. Liu *et al.* [22] formulated an AL heuristic as a traveling salesman problem to minimize the travel distances to the most uncertain points and, thereby, attempt to reduce the overall traveling costs. They showed that incorporating the uncertainty at each location as a form of reward for the traveled distances performed better than baseline methods considering only the distance. More recently, Demir *et al.* [32] have addressed the issue that traveling costs in field surveys are typically not directly related to the Euclidean distance but depend on more complex variables such as the terrain and route network. They demonstrated that considering auxiliary information on the terrain and the route network helps reduce the traveling time compared with batch-mode AL [18] without considering spatial costs.

In contrast to the aforementioned works, this study is focused on labeling time to identify samples during visual image interpretation rather than on traveling time needed in field surveys.

III. Proposed Methods

The developed AL algorithm follows the query-by-committee (QBC) strategy, where the next sample is chosen according to the maximum disagreement of an ensemble [14]. The random forest (RF) algorithm [33] is used to generate tree ensembles with 500 fully grown trees. The disagreement of the ensemble is quantified with vote entropy H, which is computed as

$$H = -\sum_{i=(0,1)} p_i * \log(p_i) \tag{1}$$

where p_i denotes the fractions of the trees voting for the respective classes $i \in \{0, 1\}$. Here, only a binary classification problem (0—nonlandslide; 1—landslide) is considered, but the vote entropy can be easily extended to multiple probabilities in a multiclass setting, and it has been suggested to normalize the measure with respect to the number of classes [34]. It should also be noted that other uncertainty measures can be used when the learner is a probabilistic or large-margin-based classifier [11]. Based on the notion of uncertainty, a simple strategy would be to iteratively select the sample with the highest vote entropy to be labeled by the user. However, since classifier retraining is computationally expensive, it is generally desirable to query samples in batches, and for the aforementioned reasons, it appears advantageous to focus labeling efforts on compact spatial subsets of the area of interest.

A. Region-Based QBC (QBC_R)

The problem of finding the most uncertain region $W^{\times}$ on the map is addressed by sliding window W with a desired window size w ($w_x = w_y$) on entropy map M_H to compute the mean local vote entropy μ_H at each position (x, y), as expressed in

$$\mu_{H(x,y)} = \frac{1}{w^2} \sum_{i=-wx}^{w_x} \sum_{j=-w_y}^{w_y} M_H(x+i, y+j). \tag{2}$$

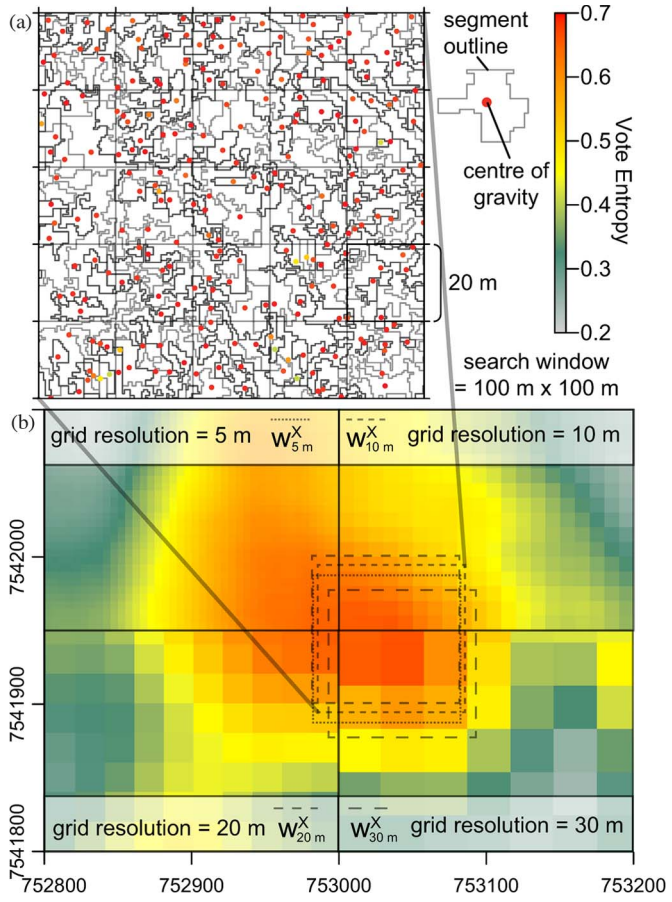


Fig. 1. Projection of the vote entropy values from (a) the centers of gravity of the segments on (b) a regular search grid at four different resolutions. For the example, the size of the search window was held constant at 100 m × 100 m to illustrate the uncertainty of the location of $W^{\times}$ at different search grid resolutions.

In (2), i and j denote indexes on a regular raster grid. The sliding window could be directly employed on the regular grid of an input image if the classification process is performed at the pixel level; however, with large images ($> 1000 \times 1000$ pixels), computation will be rather slow and will hinder a near real-time user interaction. If learning and classification are performed on the level of segments resulting from a presegmentation of the image (see Section IV-B), image segments can be represented by their centers of gravity, marking the spatial point coordinates to which the entropy values of the segments are assigned. For an efficient search of the local maximum, the centers of gravity are first projected onto a regular grid G (search grid) with a certain grid cell size g ($g_x = g_y$) to compute the mean entropy of all points per grid cell (see Fig. 1). The sliding window is subsequently applied to the search grid to locate the global maximum. Fig. 1(b) shows the effect of changing the resolution of the search grid at a constant search window size ($w = 100$ m) and illustrates the uncertainty in the position of the maximum, which typically is $<= g/2$. As a tradeoff between good location accuracy and fast computation, we set $g = 20$ m except for search window sizes $w < 60$ m, where $g = w/3$ was used.

Tests with finer search grids did not reveal any enhanced learning performance. Based on this sliding-window method, a region-based query function can be formulated to query, in

each iteration, the labels of all samples contained in region $W^\times$ maximizing μ_H. The corresponding query function is formulated as

$$W^\times = \arg \max_{W \in M_H} \mu_H(W). \tag{3}$$

Subsequently, all samples with a center of gravity that is spatially contained within $W^\times$ are queried. For a binary classification, it is then convenient to ask the user to identify only the positive examples and automatically assign all nonlabeled samples as negative examples. This method is further referred to as QBC_R and described in the pseudocode under Algorithm 1. Note that step 4 creates a buffer around all W queried in previous iterations to avoid querying previously labeled parts of the image. The algorithm yields queries with high average uncertainty of the samples within the batch but does not guarantee that the queried samples provide complementary information. In an extreme case, all samples in the batch can be highly uncertain but carry largely redundant information. One of the key requirements for batch-mode AL is, therefore, to enforce diversity among the samples queried in each iteration [15], [18], [29].

Algorithm 1: $\mathbf{QBC_R}$

Inputs:
X^1: Initial training set
n: Number of iterations
w: Size of the region/spatial-batch
U: Pool of unlabeled samples
g: Grid cell size of search grid G
Output:
S^i: Set of unlabeled samples to be included in the training set in each iteration

FOR $i \leq n$ DO
1. Train an RF with the current training set X^i
2. Use the RF to cast votes and compute the vote entropy for each unlabeled sample $s \in U$
3. Project the location of all $s \in U$ on regular search grid G with a desired resolution g to compute μ_H with a sliding window of size w
4. IF $i > 1$ THEN Mask out all parts of G within a distance $< w/2$ from any queried W in earlier iterations
5. Find region $W^\times$ according to (3)
6. Query the labels for all samples S^i contained within $W^\times$
7. Add all samples $s \in S^i$ to training set X^{i+1} and remove them from the unlabeled pool U

ENDFOR

B. Region-Based QBC Considering Sample Diversity (QBC_{RD})

The combination of criteria for sample uncertainty and diversity has already been addressed in SVM-based studies, using clustering techniques to partition candidate sets of uncertain samples and avoid sampling of redundant instances that fall into the same clusters [18], [19], [35]. The spread of queried samples in the input space and along the separating hyperplane is thereby increased, which leads to a better representation of the data space and the decision boundary. For the region-based query strategy developed in this work, the composition of the batches is preconstrained by the coverage of the spatial window, and the proposed clustering techniques are, therefore, not directly applicable. However, distance metrics used for clustering can also be employed to directly measure the dispersion of samples in feature space and thereby quantify the diversity of the batch. To this end, two distinct measures are proposed. The first one, which is formulated in (4), is σ_d, which represents the standard deviation of the Euclidean distances $\rho_k(X, c)$ between each of the samples in the candidate batch $(c \in W^m)$ and their respective nearest training point $(s \in X)$. Thus

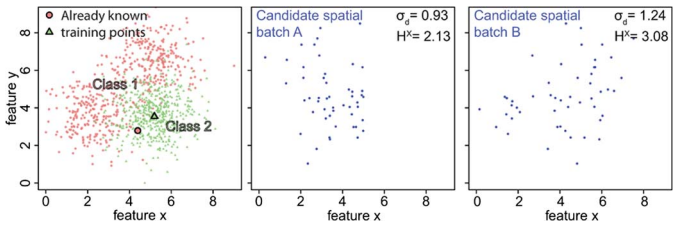


Fig. 2. Example calculations of diversity measures σ_d and $H^\times$ for two candidate batches in a simplified 2-D feature space. Candidate batch B is preferable since contained samples have a higher spread away from the already known training points and provides a better representation of the boundary between the two classes.

$$\sigma_d = \sqrt{\frac{1}{|W^m| - 1} \sum_{c \in W^m} \left(\rho_k(X, c) - \overline{\rho_k(X, c)}\right)^2}. \tag{4}$$

Here, $|W^m|$ denotes the cardinality of the candidate set. In general, a larger σ_d indicates a higher feature space spread of the contained samples in relation to the already acquired training data (see Fig. 2). The corresponding query function formulated in (5) can be used to select the region with the highest standard deviation out of m candidate batches. Thus

$$W^\times = \arg \max_{W^m \in M_H} \sigma_d(W^m). \tag{5}$$

A second measure is the cross-entropy $H^\times$ between the training and the candidate set. The calculation of $H^\times$ depends mainly on the mean logarithmic distances of the samples in the candidate batch $(c \in W^m)$ to their k-nearest neighbors (kNNs) in the training set $(s \in X)$ and is defined as

$$H^\times(W^m, X) = \log\left(v_D |X|\right) - \psi(k) + \frac{D}{|W^m|} \sum_{c \in W^m} \log \rho_k(X, c) \tag{6}$$

where v_D is the volume of the unit ball in $\mathbb{R}^D$, D is the number of features, ψ is the digamma function, and $|X|$ denotes the cardinality of the training set [36]. The measure $H^\times$ is proposed here as a utility criterion to quantify the distinctiveness of the samples in the candidate batch with respect to the already available training data. The corresponding query function is formulated as

$$W^\times = \arg \max_{W^m \in M_H} H^\times(W^m). \tag{7}$$

The implementation in [37] was used to compute $H^{\times}$ for each candidate batch setting $k = 1$. Fig. 2 illustrates that candidate regions with a higher σ_d or $H^{\times}$ are generally preferable since the contained samples yield a better exploration of previously undersampled parts of feature space.

To avoid the exploration of feature dimensions, which are less relevant or even useless for the separation of the considered classes, it is possible to constrain the distance computation to only those features that are beneficial according to the current model state. The RF algorithm offers an internal measure for variable importance, which is based on the instances that are left unconsidered [or out-of-bag (OOB)] in the bootstrap samples used for the construction of the individual classification trees [33]. The properties of different variants of this measure have been analyzed in recent studies with artificial and real-world data sets [38]–[40]. Here, the *mean decrease in accuracy* when a variable is randomly permuted is adopted, since it is more robust in the presence of many potentially correlated variables [41]. Threshold t_i is introduced to define the minimum variable importance for which a feature will still be included in the calculation of the kNN distances.

Algorithm 2: $\mathbf{QBC_{RD}}$

Inputs:
- X^1: Initial training set
- n: Number of iterations
- w: Size of the region/spatial-batch
- m: Number of preselected candidate regions
- t_i: Minimum variable importance threshold for the inclusion of a feature in the distance calculation
- fun: Diversity function (5) or (7)
- U: Pool of unlabeled samples
- g: Grid cell size of search grid G

Output:
- S^i: Set of unlabeled samples to be included in the training set in each iteration

FOR $i \leq n$ DO
1. Train an RF with the current training set X^i
2. Use the RF to cast votes and compute the vote entropy for each unlabeled sample $s \in U$
3. Project the location of all $s \in U$ on regular search grid G with a desired resolution g compute μ_H with a sliding window of size w
4. IF $i > 1$ THEN Mask out all parts of G within in a distance $< w/2$ from any queried W in earlier iterations
6. Find the m nonoverlapping candidate regions W^m according to (3)
7. Select out of m candidate regions the final region $W^{\times}$ that maximizes the selected diversity function fun computed on all features with a variable importance $> t_i$
8. Query the labels for all samples S^i contained within $W^{\times}$
9. Add all samples $s \in S^i$ to training set X^{i+1} and remove them from the unlabeled pool U

ENDFOR

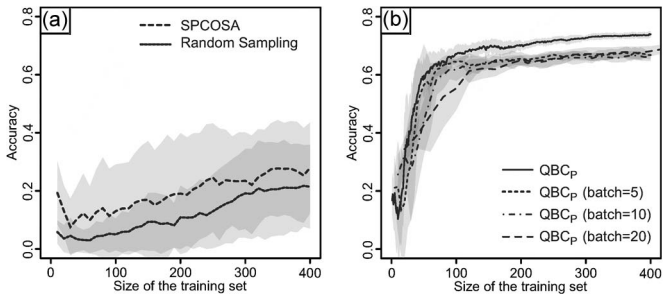


Fig. 3. Learning curves of an RF classifier (average and standard deviations over ten runs) in dependency of the size of the training set sampled (a) with SPCOSA and simple random sampling and (b) with the pointwise AL scheme (QBC_P) at different batch sizes.

In [18], [19], and [35], the criteria for uncertainty and diversity are combined by preselecting a number of m candidates with high uncertainty and by choosing among the candidates the instances with the highest diversity. A similar method is adopted in this work, whereas out of m nonoverlapping candidate regions (W^m), the one with the highest diversity (σ_d or $H^{\times}$) is selected for the final query. This algorithm integrates criteria for uncertainty and diversity and from now on will be referred to as QBC_{RD} (Algorithm 2), where $QBC_{RD\sigma}$ and $QBC^{\times}_{RDH}$ are the two different versions depending on the adopted diversity criteria.

C. Baseline Methods for Comparison

The ability of AL heuristics to reduce the annotation costs is commonly assessed through comparison with simple random sampling. Nevertheless, systematic or stratified sampling designs are more frequently applied in remote sensing studies since they usually yield more accurate estimates of environmental variables than simple random sampling [42]. A stratification of the study area according to previously existing maps or preliminary image analysis is often adopted to encourage a representative sampling of all target classes. In cases where it is difficult to obtain *a priori* information about the suitable strata, sampling schemes that target a homogenous spatial distribution of the sampling points can be applied without prior information. An implementation of such a spatial coverage sampling (*SPCOSA*) scheme has been recently proposed [43]. It makes use of the k-means algorithm to divide a given area into spatially homogenous clusters and provides additional points at the most undersampled localitions. Fig. 3(a) indicates the amount of performance increase of an RF classifier trained with samples randomly derived and with *SPCOSA* and demonstrates that *SPCOSA* is preferable to assess the effectiveness of AL heuristics in the spatial domain.

The second baseline method used in this study is a simple pointwise AL heuristic (QBC_P), which, initialized with just one sample per class, queries in each iteration the unlabeled instance with the highest vote entropy of the RF at each iteration. This method is very similar to the query-by-bagging algorithm proposed in [44], whereas the training of an RF involves not only bagging but also random feature selection during the construction of each tree. Fig. 3(b) illustrates that QBC_P run in batch-mode yields only suboptimal results and that QBC_P run with single queries per iteration is the more interesting benchmark.

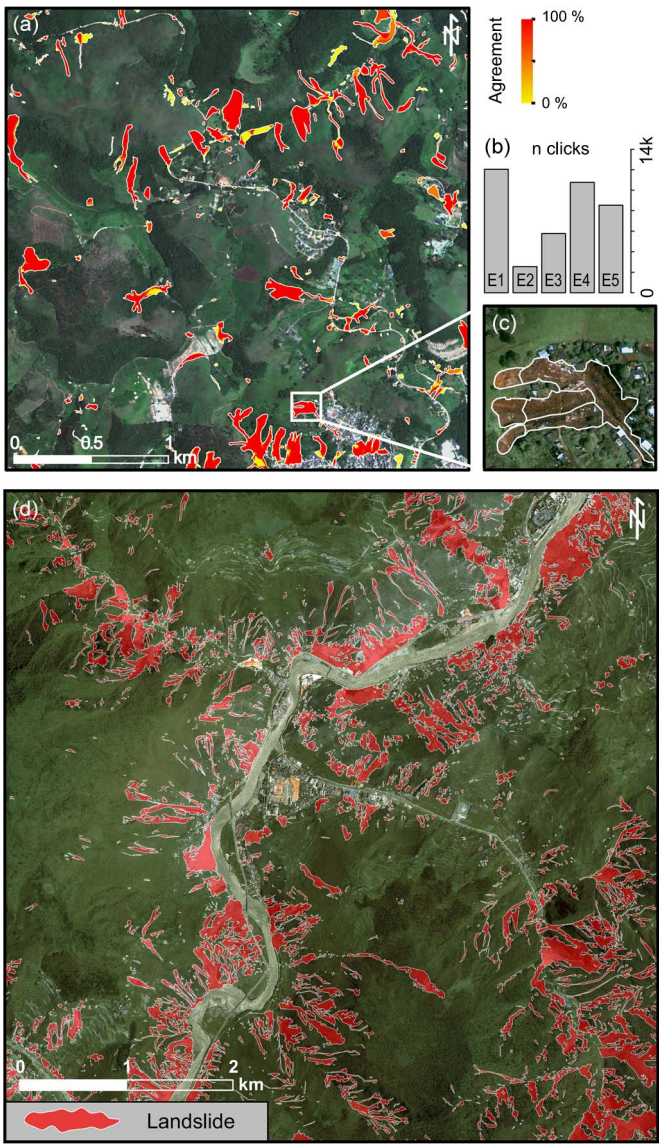


Fig. 4. (a) Geoeye-1 image of the BRAZIL study site (01/20/2011) and agreement rate of the landslide inventories of five experts. (b) Overall number of clicks per expert (E1–E5). (c) Example landslide as outlined (white) by E1. (d) Ikonos-2 image of the CHINA study site (06/01/2008) and the corresponding single expert landslide inventory.

IV. Study Sites

The first study site (Brazil) used for the experimental evaluation of the methods is located in the Serrana Mountains around Nova Friburgo (Brazil) and extends over 10.5 km^2. On January 11–12, 2011, the region was affected by heavy thunderstorms, triggering thousands of shallow landslides [45]. The second study site (China) is located at the eastern margin of the Tibetan Plateau around Wenchuan County and extends over 36 km^2. On May 12, 2008, a Mw 7.9 earthquake occurred in the region, triggering approximately 60 000 landslides [46].

A. Geospatial Data Sets

Geoeye-1 satellite imagery of the Brazil site was acquired after the event on January 20, 2011 [see Fig. 4(a)] and before the event on May 26, 2010. The ground spatial resolution of the data set is 2 m for the multispectral and 0.5 m for the panchromatic bands. A further pre-event image with a coarser resolution (30 m) was acquired by the Landsat 5 satellite on September 4, 2010. A pre-event image that resembles optical sensors with intermediate resolution (e.g., such as the current SPOT-5 or the forthcoming Sentinel-2 [47]) was simulated by resampling the Geoeye-1 image of May 26, 2010 to a resolution of 10 m using bicubic interpolation. All images were orthorectified using the ASTER Global Digital Elevation Model [48, ASTER GDEM] with a spatial resolution of 30 m. Between the pre- and post-event Geoeye-1 images, 22 tie points were manually selected to update the rational polynomial function (RPC) and thereby enhance the coregistration. The residual root mean square error (RMSE) of the updated RPC model was 14.7 m.

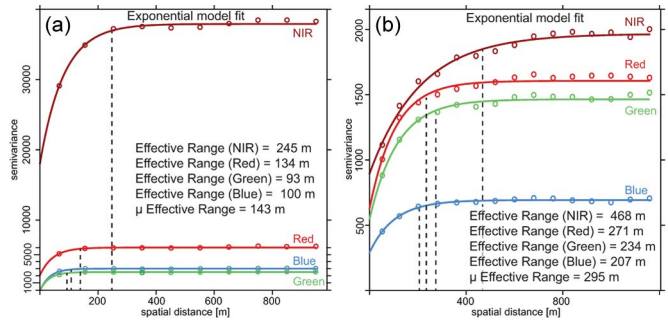


Fig. 5. Variograms derived from 5% subsamples of the four bands of the postlandslide image for (a) the BRAZIL data set and (b) the CHINA data set.

For the China site, Ikonos-2 imagery was acquired after the event on July 1, 2008 [see Fig. 4(d)]. The ground spatial resolution of Ikonos-2 is 4 m for the multispectral bands and 1 m for the panchromatic band. A pansharpened (1-m pixel resolution) and orthorectified product with an expected geolocation RMSE of 5 m is used. At both sites, a number of terrain variables (slope, hillshade at the times of image acquisition, hydrological flow direction, and the average distances to the mountain ridge [47]) were calculated.

Using a semivariogram analysis with an exponential model fit [49], it is possible to measure the spatial autocorrelation of the gray values within the images and gain quantitative insights into the spatial structure of the observed areas. Fig. 5 shows that, depending on the spectral bands, the effective range of the spatial autocorrelation varies between 93 and 245 m for the Brazil data set and between 207 and 468 m for the China data set. The significantly increased range of the autocorrelation in the near-infrared (NIR) band is related to large vegetated areas with relatively homogenous NIR reflectance, whereas the reflectance in the visible bands captures changes in the surface color with higher spatial frequencies. It provides an indication for the minimum size of the sampling windows, which should be sufficiently large to capture spatial variability beyond locally autocorrelated characteristics. Note that only random subsamples of the images (5%) were used for the semivariograms (see Fig. 5); however, no significant changes were observed when using different random subsamples.

In order to obtain reliable estimates of the performance and variance of AL routines, it is typically necessary to repeat

the routine several times with variable starting conditions. Therefore, it is mostly not feasible to integrate user interaction directly in the evaluation of AL algorithms, and most studies rely on readily prepared ground truth data sets. It has, however, been documented that landslide mapping by multiple experts is subjected to considerable uncertainties [50], [51], and it is generally desirable to assess uncertainties of ground truth data used for algorithm evaluation [52], [53]. To address this issue for the Brazil data set, five alternative landslide inventories (E1–E5) are independently prepared by five expert geomorphologists through joint interpretation of the two available VHR resolution images and the DEM.

Even in the absence of an ultimate ground truth, the availability of multiple-expert maps allows for the estimation of a lower bound for the mapping error. Optimistically assuming that the majority vote is correct, expert judgements that contradict the majority vote can be regarded erroneous, and the lower error bound can be estimated as a function of the expert disagreement [52]. For the five expert inventories, a lower error bound of 1.1% can be estimated for any of the five maps. This low error rate applies to the full map and results, to some degree, from the sparsely distributed landslide class. A higher error rate of 16.1% results if only pixels marked by at least one of the experts would be considered. To analyze further the interexpert variability, the F-measure F, which is not prone to class imbalance [54], was computed for each pair of experts. The highest agreement between two experts was indicated by $F = 0.86$ (and the lowest by $F = 0.71$). Considering that a learning algorithm that perfectly predicts the mapping of one expert would inevitably commit errors with respect to the mapping of a second expert, the expert agreement suggests an upper bound for the accuracy achievable with any automatic method on this particular problem. In this context, the uncertainty of each expert can be regarded as a type of inherent stochastic noise [55], which cannot be fitted without sacrificing the generalization to the maps of the other experts. A possible strategy to reduce this type of noise would be to combine the different ground truths (e.g., by majority vote). However, this would also lead to positively biased classification results in comparison to a real-world scenario where multiple ground truths are typically not available.

In order to select one of the Brazil inventories for the experiments, the number of clicks used by the experts is assessed and compared as a proxy for the degree of detail of the mapping. While it is generally difficult to assess the level of expertise, the amount of clicks can serve as an indicator for the level of detail and time investment by the expert. Fig. 4(b) illustrates that the most clicks were used for the elaboration of inventory E1, which was therefore chosen as a reference in the benchmarking of the different AL heuristics. For the significantly larger China data set, it was only feasible to obtain one reference landslide inventory, which is displayed in Fig. 4(d).

B. Segmentation and Feature Extraction

The multiresolution image segmentation algorithm [56, MRIS] was adopted, considering only spectral information of the post-event images and equal weights for all spectral bands. The region-growing algorithm comprises a scale parameter, which is a threshold for the maximum allowed increase in the segment's variance during a merging operation. It has been demonstrated that statistical methods can help constrain the choice of the scale parameter for knowledge-based image classification [57], whereas oversegmentation with a small scale factor was found to yield higher classification accuracy in a supervised framework [9]. Consequently, the scale parameter was set to 20, corresponding to strong oversegmentation of landslides and most other landscape features. For all segments, a broad set of features describing the spectral characteristics, texture, shape, topographic variables and contrasts to neighbors are calculated. The object features are selected according to previous studies on the object-oriented mapping of landslides [9], [58]. Additional features, such as flow accumulation [59], average distance to the ridge [60], mean contrast to neighbors [61], and certain band ratios [62], are selected considering the known prevalence of landslides in certain topographic positions and typical spectral characteristics such as high contrast to surrounding vegetated areas.

Only certain combinations of all available features are used in the experiments to simulate typical scenarios of accessible data sets (see Section V-A). The class membership of the segments is assigned according to the maximum overlap with the reference landslide inventory maps (E1 for the BRAZIL data set). For the spatial queries described in the next section, segments are represented by their center of gravity. The segmentation results, the extracted features, and the `R` implementation of the methods are available at http://eost.unistra.fr/recherche/ipgs/dgda/dgda-perso/andre-stumpf.

V. Experimental Design

All experiments are carried out using the RF algorithm with 500 trees. To address class imbalance, stratified bootstrap sampling [63] is used for the implementation of the region-based methods so that each tree in the ensemble is built on a balanced training sample. The advantages of this strategy are discussed in Section VI-D. The number of maximum AL iterations is defined before the start of each run, and each run is repeated at least ten times with random seeds in order to estimate the mean prediction accuracy values and standard deviations on the unlabeled set. The initiation of each run is performed through stratified random sampling in order to ensure the presence of at least one example per class for the first iteration.

All segments of the data sets are labeled, and therefore, learning can be performed on the full map. This is different from the experimental designs of recent AL studies in remote sensing, where the data sets comprised reference labels only for subsets of the image [15], [17]–[19]. To avoid preconstraining the pool and to provide the query function with access to the full unlabeled pool, random prepartitioning of the data set into an unlabeled pool and test set was also avoided. Instead, for each iteration, the query functions move instances from the unlabeled pool into the training set, and the test set error is assessed on the full unlabeled pool of samples. For all experiments, the test set error, the OOB error, the system runtime, and the identifier of queried samples are recorded. In addition,

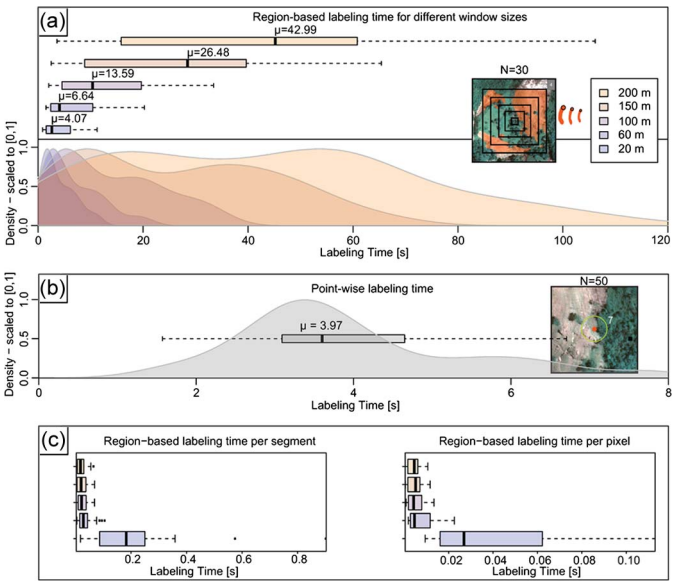


Fig. 6. Expert labeling time for the Brazil data set with (a) marker-based labeling of regions ($N = 30$) according to different window sizes and (b) with pointwise labeling ($N = 50$). (c) Marker-based labeling times per segments and pixels contained within each region.

the labeling time required by the user was experimentally evaluated, as described in the next section. Altogether, this permits the analysis of three important aspects of an AL system: 1) How often will the user be asked (number of iterations)? 2) How long will it take the user to answer the query (labeling time)? 3) How long will the user have to wait between the queries (computational time)?

A. Labeling Time

The costs of field surveys are highly dependent on the particular application and environmental conditions. It is consequently difficult to estimate related costs *a priori*, and repeated field experiments that would be required to support the development of an algorithm are not practical. However, for many applications, sampling is based on image interpretation, and the time requirement for image interpretation can be easily assessed. Two alternative labeling strategies for image interpretation are considered, one being the pointwise annotation of individual samples queried by a pointwise AL method, and the second being the annotation with a brush-like tool within compact regions queried by a region-based AL method. To evaluate the required labeling time for both strategies, sequences of post-event image subsets around points ($N = 50$) queried by a pointwise AL strategy (QBC_P), and around regions ($N = 30$) queried by a region-based AL method (QBC_R) are presented to a domain expert. For the labeling of regions, the expert is provided with a brush-like marker tool with adjustable size and asked to mark all the landslide area, assuming that the unmarked area will be automatically assigned as background. For the Brazil data set, the expert was asked to mark five nested subsets (20, 60, 100, 150, and 200 m) of each displayed region successively from the smallest to the largest nested subset [see Fig. 6(a)]. The required annotation time for each region and each of the respective subsets is recorded in order to assess not only the overall time per region but also the relationship between the window size and required time. For the China data set, the experiment is performed at only one window size of 200 m. The choice of the window size is guided by the mean effective range of the spatial autocorrelation as determined through the semivariogram analysis in Section IV-B. For the pointwise queries, labeling is performed with a single click, signaling a positive or a negative example. For both labeling strategies, the expert is given the possibility to zoom in and out as needed, and the annotation time is recorded from the moment the subset was displayed until the expert completed the identification.

Fig. 6 provides a comparison of the labeling time for regions and individual points in the Brazil data set. In both cases, the probability density distributions of the labeling times show a bimodal distribution [see Fig. 6(a) and (b)] that can be linked to two categories of queries, namely, simple regions (unique homogenous land cover, no landslides) and more complex regions (abundance of landslides, complex landslide boundaries).

B. Sensitivity Analysis for Different Sampling Window Sizes

Experiments on the Brazil data set are dedicated to analyzing the effects of different window sizes on the performance of the region-based AL heuristics and to compare their performance with the pointwise query function and *SPCOSA*. Considering the range of spatial autocorrelation, it can be assumed that a sufficiently large window size would encourage the inclusion of uncorrelated samples in one batch and, at the same time, guarantee a distance between sampling regions beyond the range of autocorrelation. In order to test this assumption, the region-based AL heuristics (see Section III-A and B) are tested at five different window sizes with edge lengths $w = \{20, 60, 100, 150, 200\ \text{m}\}$. The diagonals of these squared windows range from 28.2 to 282.4 m and covered a distance range in which relevant semivariance changes are observed within the Brazil study site (see Fig. 5). The parameters for the QBC_{RD} heuristics that determine the number of candidate regions and the minimum variable importance are kept constant at $m = 3$ and $t_i = 0.01$, respectively. For these experiments, a stacked feature vector including object features from the pre- and post-event Geoeye-1 imagery (see Table I) is used.

C. Sensitivity Analysis of Parameter Settings in QBC_{RD}

Further experiments are carried out to evaluate the sensitivity of the parameters that control the number of candidate regions and the minimum variable importance for the calculation of σ_D and $H^{\times}$ within QBC_{RD}. Four respective experiments are dedicated to test the influence of the preselection of more or fewer candidate regions with $m = \{2, 3, 4, 5\}$, keeping the window size constant at an intermediate value of $w = 100$ m and $t_i = 0.01$. In a second setup, five different values of $t_i = \{0, 0.01, 0.02, 0.03\}$ are tested, keeping the number of candidate regions and the window size constant at $m = 3$ and $w = 100$ m, respectively. Learning is performed with all features from pre- and post-event Geoeye-1 images (see Table I) for the Brazil data set, and each AL routine is repeated 30 times

TABLE I
OVERVIEW OF THE DATA SETS RESULTING FROM IMAGE SEGMENTATION AND FEATURE EXTRACTION

		Brazil		China	
n segments		394,885		275,360	
landslide	non-landslide	22,703	372,182	42,321	233,039
Spectral features (post-event Geoye-1/Ikonos-2)		10		9	
Texture features (post-event Geoye-1/Ikonos-2)		65		56	
Shape features (post-event Geoye-1/Ikonos-2)		7		7	
Contrast to neighbours (post-event Geoye-1/Ikonos-2)		6		12	
Spectral features (pre-event Geoye-1)		7		-	
Spectral features (pre-event SPOT like)		7		-	
Spectral features (pre-event Landsat 5)		11		-	
Topographic features		6		5	

TABLE II
OVERVIEW OF THE DIFFERENT TESTED POSTDISASTER SCENARIOS AT THE BRAZIL STUDY SITE

	Description	n features
Scenario 1	Post-event VHR image + topographic data	93
Scenario 2	Post-event VHR image + pre-event Landsat type image + topographic data	105
Scenario 3	Post-event VHR image + pre-event Sentinel/SPOT type image + topographic data	101
Scenario 4	Post-event VHR image + pre-event VHR image + topographic data	101

starting from randomly seeded runs to assess mean prediction accuracy values and standard deviations.

D. Performance on Different Data Sets

A final set of experiments is carried out in order to assess the robustness of the proposed methods on different data sets. These included tests on the China data set at a window size of $w = 200$ m and a simulation of data scenarios typically encountered in disaster response mapping using the Brazil data set at a window size of $w = 100$ m. With the wide range of operational and forthcoming steerable VHR satellites, postdisaster images with submeter resolution can be acquired within hours after a given event. Pre-event images, contrarily, are often only available in historical archives and from sensors with coarser spatial resolution dedicated to long-term environmental monitoring. In this context, four different scenarios of data availability (see Table II) are evaluated. On the China data set, tests were performed to compare *SPCOSA*, QBC_P, $QBC_{R\mu}$, and $QBC^{\times}_{RDH}$. The abovementioned experiments on the Brazil data set already evaluated the performance of those different methods, and hence, only $QBC^{\times}_{RDH}$ and $QBC_{RD\sigma}$ are tested to evaluate their robustness in the different scenarios. The learning curves are averaged over 30 random seeded runs, and the parameters for the diversity criteria are kept constant at $t_i = 0.01$ and $m = 3$.

VI. RESULTS AND DISCUSSION

A. Sampling Window Size

Fig. 7 depicts the outcome of the experimental comparisons of the baseline methods (*SPCOSA*, QBC_P) and the proposed region-based AL heuristics on the Brazil data set. For all tested window sizes ($w = \{60, 150\}$ are not displayed in Fig. 7), the region-based queries generally yield steeper learning curves, outperforming *SPCOSA* and pointwise query method QBC_P in terms of runtime, labeling time, and number of iterations.

The region-based algorithms query a significantly greater overall number of samples and are computationally more complex than pointwise queries, leading to longer runtimes per iteration. However, since the overall number of iterations is lower when querying a limited number of spatial batches, significantly less overall computational runtime is required for the same accuracy level [see Fig. 7(a)].

Multiplying the mean time requirements for pointwise labeling and marker-based labeling of regions [see Fig. 6(a) and (b)] with the number of iterations, the time expenditure of a user of the proposed methods is estimated. Fig. 7(b) shows that all three region-based queries ($QBC_{R\mu}, QBC_{RD\sigma}$, and $QBC^{\times}_{RDH}$) yield steeper learning curves at all window sizes and result in a reduction in labeling time compared with QBC_P. Depending on the window size, the average labeling time was reduced from 37.9% ($w = 200$ m) to 72.2% ($w = 60$ m) when comparing the learning curves of $QBC^{\times}_{RDH}$ to the pointwise heuristic QBC_P over all accuracy levels. At intermediate window sizes ($20\text{ m} < w < 200$ m), these gains tended to be more important (from 57.9% to 72.2%).

It is shown in Fig. 7(a) and (b) that at $w = 200$ m, the region-based heuristics quickly reach a plateau, whereas QBC_P approximates and slightly exceeds the same accuracy level in later iterations. Consequently, the proposed methods still yield important overall time savings in early iterations but provide slightly lower final accuracy values when selecting a large window size. This is probably because a large window size ($w = 200$ m, diagonal $= 283$ m) significantly beyond the range of spatial autocorrelation (143 m; see Fig. 5) does not help increase the variance captured per query region. The fact that the batch size (window size) cannot be arbitrarily increased without sacrificing the performance of the AL model is consistent with other studies using batch-mode AL with uncertainty and diversity criteria on remote sensing data sets [18], [19] and points toward a lower sample variance per batch when the ratio between $|X|$ and the number of model updates (iterations) decreases.

Since larger sampling windows result in fewer iterations and, consequently, lower computational runtimes, changing the window size constitutes a tradeoff parameter that enables the

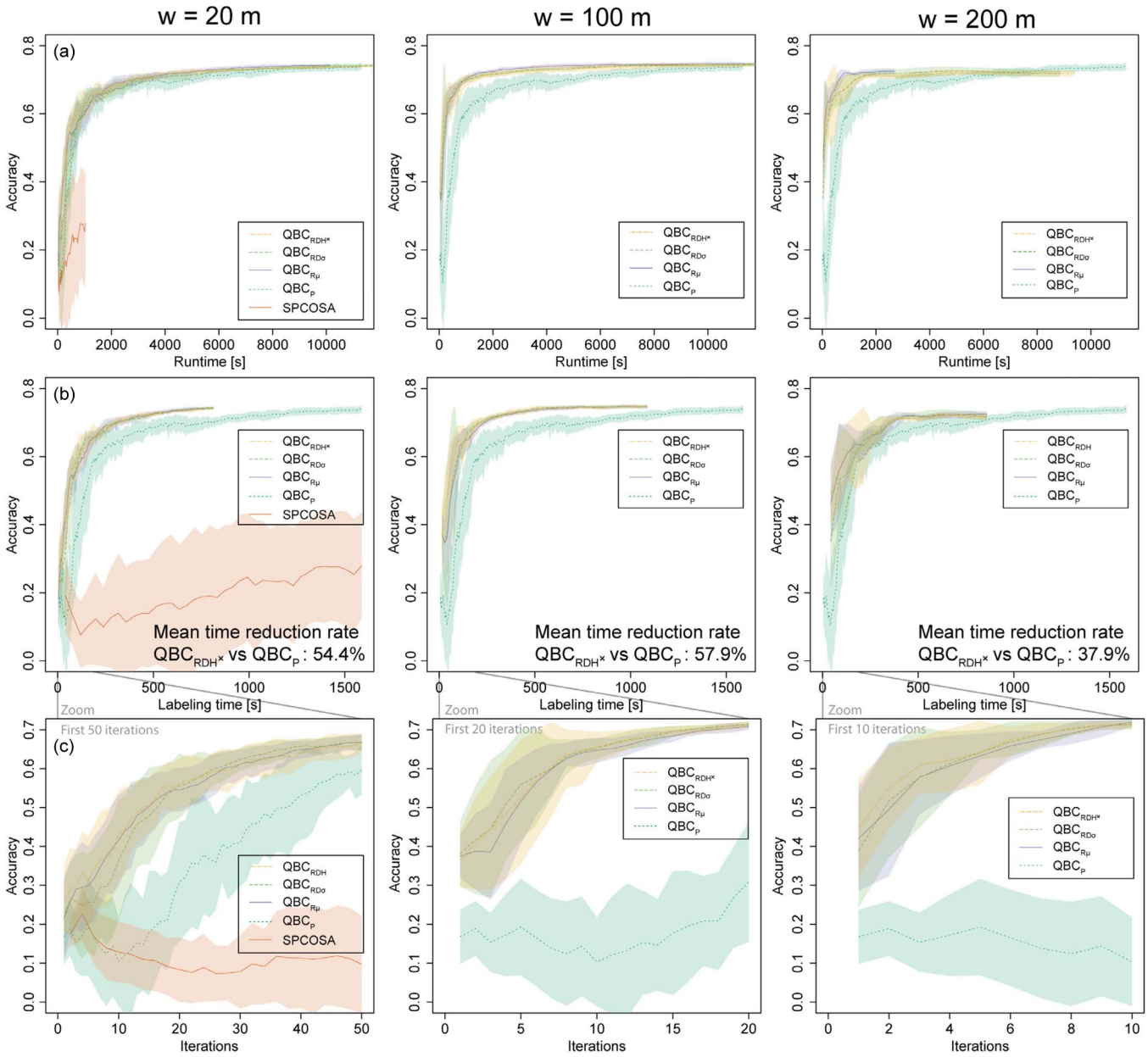


Fig. 7. Learning curves for comparison between the baseline sampling methods (SPCOSA and QBC_P) and the proposed region-based queries ($QBC_{R\mu}$, $QBC_{RD\sigma}$, and $QBC^{\times}_{RDH}$) on the Brazil data set for three of the five tested window sizes. Accuracy gains are plotted with respect to (a) the computational runtime, (b) the labeling time derived by multiplying mean labeling times as detailed in Section V-A with the number of iterations, and (c) the number of iterations (early iterations, including 20 additional runs for the estimation of mean and variance). The mean time reduction rate was computed over all accuracy levels as the mean difference between the $QBC^{\times}_{RDH}$-curve and the nearest corresponding points on the QBC_P-curve. For better visibility, the learning curves for SPCOSA are only presented in the first column.

user to choose between more frequent labeling on small batches (smaller window size) or accepting longer waiting times between the labeling of fewer larger and more time consuming but fewer spatial batches (larger window size). In the experiments, a good tradeoff was provided by $w = 100$ m (diagonal $= 141$ m), which approximately coincides with the mean effective range of the autocorrelation measured on the gray values of the image bands (143 m; see Fig. 5).

Regarding performance differences between $QBC_{R\mu}$, $QBC_{RD\sigma}$, and $QBC^{\times}_{RDH}$, Fig. 7(c) shows the learning curves for the early iterations from 30 random seeded runs. At the smallest window size ($w = 20$) $QBC^{\times}_{RDH}$ and $QBC_{R\mu}$ yield very similar learning curves, whereas $QBC_{RD\sigma}$ performs the worst over the first 20 iterations. This suggests that σ_D becomes a less robust diversity measure when computed with fewer samples in the training and candidate set. In addition, $QBC^{\times}_{RDH}$ appears to be more robust to changes in the window size and outperformed $QBC_{R\mu}$ and $QBC_{RD\sigma}$, particularly at $w = \{150, 200$ m$\}$.

On the one hand, the integration of an explicit diversity criteria enables further reduction in the labeling costs, whereas on the other hand, the higher computational complexity of $QBC_{RD\sigma}$ and $QBC^{\times}_{RDH}$ results in longer runtimes when compared with the simpler $QBC_{R\mu}$ heuristic [see Fig. 7(a)].

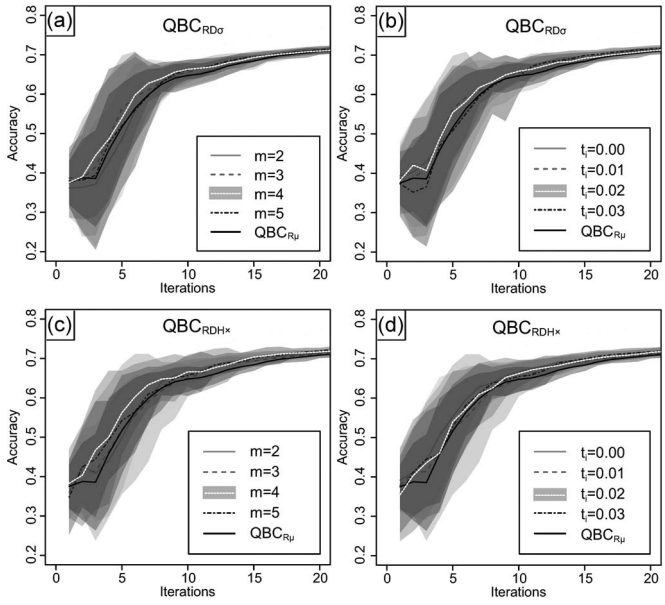


Fig. 8. Learning curves on the Brazil data set for (a) and (b) $QBC_{RD\sigma}$ and (c) and (d) $QBC^{\times}_{RDH}$ in dependency of changes in parameters t_i and m compared with QBC_{Ru} at $w = 100$ m.

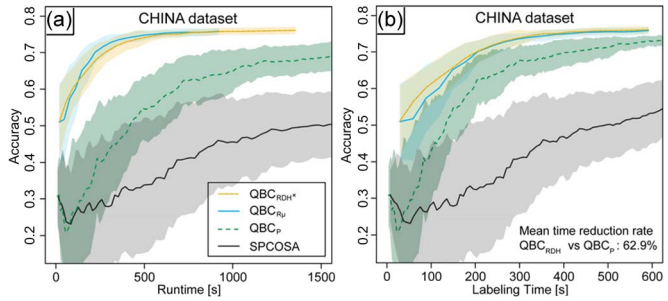


Fig. 9. Learning curves for the comparison of *SPCOSA*, QBC_P, $QBC_{R\mu}$, and $QBC^{\times}_{RDH}$ on the China data set in terms of (a) computational runtime and (b) labeling time.

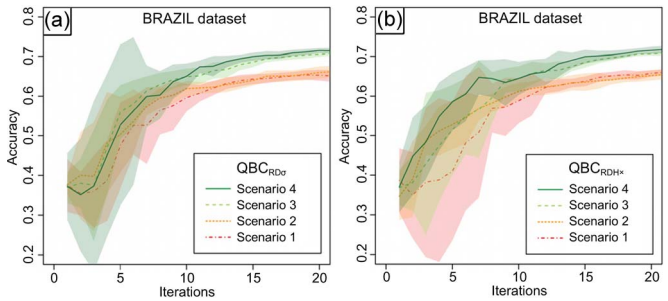


Fig. 10. Comparison of the performance of (a) $QBC_{RD\sigma}$ and (b) $QBC^{\times}_{RDH}$ in the four different data set scenarios for the Brazil data set.

Fig. 7(c) also includes a comparison of the proposed region-based methods ($QBC_{R\mu}, QBC_{RD\sigma}$, and $QBC^{\times}_{RDH}$) with QBC_P and *SPCOSA* in terms of iterations. While it should be kept in mind that comparison of region-based and pointwise queries in this way ignores the different labeling and computational costs, it illustrates the considerably lower number of queries sent to the user, which is an important issue for the design of an AL system.

B. Influence of Parameters m and t_i and m in QBC_{RD}

Fig. 8 illustrates the performance of $QBC_{RD\sigma}$ and $QBC^{\times}_{RDH}$ on the Brazil data set according to changes in the main parameters m and t_i. The AL heuristic, which does not consider the sample diversity ($QBC_{R\mu}$), marks the lower bound of the obtained learning curves. This confirms that the proposed diversity criteria are useful in further reducing the labeling time in the early iterations of the training process.

Furthermore, it demonstrates that, within the evaluated parameter range, the performances of $QBC_{RD\sigma}$ and $QBC^{\times}_{RDH}$ are at least competitive with $QBC_{R\mu}$, no matter which values are set for m and t_i.

For the tested parameter values, both heuristics tend to perform better with a higher number of m candidate regions [see Fig. 8(a) and (c)]. For the choice of t_i, the results give no clear indication that discarding features with low variable importance yields any advantage. For $QBC^{\times}_{RDH}$, the best performance is achieved using all available features in each iteration for the computation of the cross-entropy (see Fig. 8(d); $t_i = 0.00$). This suggests the use of a simpler version of $QBC^{\times}_{RDH}$ without the computation of the variable importance and parameter t_i. However, the remaining experiments are carried out with the default settings introduced above ($m = 3; t_i = 0.01$) to not positively bias the final results.

C. Performance on Different Data Sets

Outcomes of the evaluation of *SPCOSA*, QBC_P, $QBC_{R\mu}$, and $QBC^{\times}_{RDH}$ on the China data set are presented in Fig. 9 and show a significantly better learning performance in terms of runtime and labeling time of the proposed region-based methods than *SPCOSA* and QBC_P. In terms of labeling time, $QBC^{\times}_{RDH}$ reduces the required time (mean time reduction over all accuracy levels) by 56% when compared with the pointwise queries [see Fig. 9(b)] and outperform the purely uncertainty-based $QBC_{R\mu}$ method. $QBC^{\times}_{RDH}$ also yields the highest overall accuracy resulting in $F = 0.76$. Due to its lower complexity, $QBC_{R\mu}$ provides slight advantages in terms of computational runtime [see Fig. 9(a)]. The performance increase with the developed region-based methods is consistent with the experimental results on the Brazil data set.

Fig. 10 displays the results of the two region-based AL heuristics that integrate diversity criteria ($QBC_{RD\sigma}$ and $QBC^{\times}_{RDH}$) when tested in the four different scenarios for the Brazil data set. Unsurprisingly, both algorithms reach the best performance when pre- and post-event images are VHR (scenario 4). Degrading the pre-event images to a 10-m resolution (scenario 3) yielded a rather small loss in accuracy (F-measure), which was, on average, less than 1% beyond the tenth iteration. The use of medium resolution (scenario 2) or the complete absence of pre-event imagery (scenario 1), however, led to a significant loss in accuracy, which was, on average, 5.5% and 5.7% beyond the tenth iteration.

Within the first ten iterations, $QBC^{\times}_{RDH}$ yielded greater improvement (steeper learning curve) from scenario 4 to scenario 3 compared to $QBC_{RD\sigma}$. This demonstrates that $QBC^{\times}_{RDH}$ heuristics exploit the additionally available data more

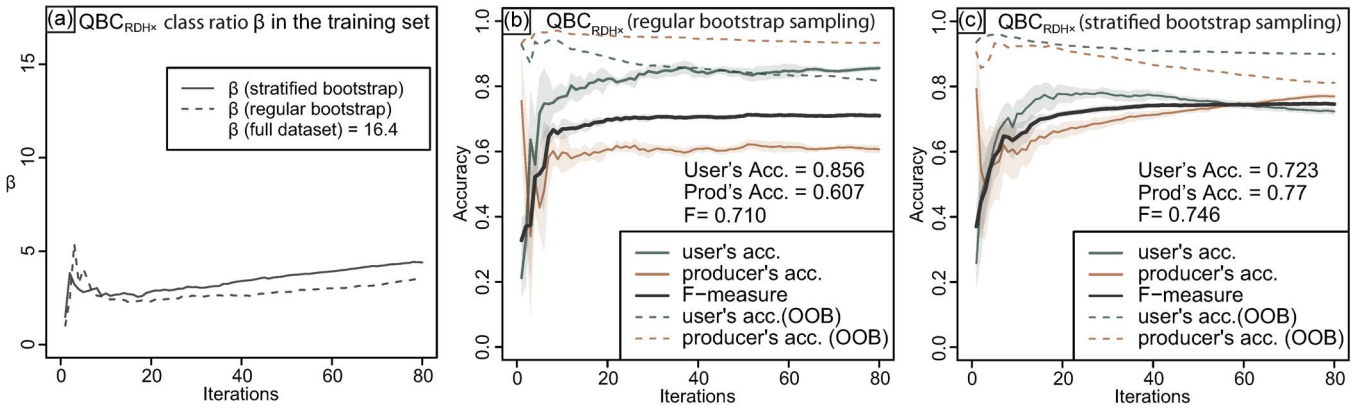


Fig. 11. (a) Ratio of landslide and nonlandslide samples (β) in the training set queried with regular and stratified bootstrap sampling. Effects of class imbalance with (b) regular bootstrap sampling and (c) stratified bootstrap sampling where the majority class is downsampled to the size of the minority class in each iteration. The displayed results were obtained applying $QBC^{\times}_{RDH}$ at $w = 100$ m on the Brazil data set but are representative for positive effects of stratified bootstrap sampling observed in all experiments.

efficiently in the first iterations. Beyond the tenth iteration, no significant performance difference between the two heuristics could be noted. In general, the results suggest that both AL heuristics ($QBC_{RD\sigma}$ and $QBC^{\times}_{RDH}$) remain efficient in reducing labeling costs when using lower resolution pre-event images or post-event images only. However, the achievable overall accuracy decreases with coarser or missing pre-event imagery due to general loss of information. Considering the rather small losses in overall accuracy with the 10-m resolution, pre-event images, and free access to such type of images in the near future [47], scenario 3 appears as the most feasible for the application of the proposed algorithm in practice.

D. Effects of Class Imbalance and Stratified Bootstrap Sampling

Class imbalance is a frequent challenge for the application and design of machine learning algorithms [54] and intrinsic to landslide mapping where affected areas usually cover only minor fractions of the landscape.

In certain cases, iterative resampling can be effective to approximate a class balance in the training set that will lead to a balance of user's and producer's accuracy on the test set [9]. The underlying assumption is that the training set was sampled at random and, therefore, represents well the underlying class distribution. However, this does not typically apply to a training set obtained with an AL algorithm, since the queries are intentionally biased toward the regions in feature space where class overlap and class imbalance become less pronounced. It has been previously reported that this bias typically yields a more balanced training sample and can alleviate the class imbalance problem [54]. Fig. 11(a) illustrates that this also applies in our experiments, where the class imbalance in the queried training sets is considerably lower than in the full data sets. It is observed that the queries are particularly directed toward regions with abundant occurrences of landslides, which must be regarded as a positive attribute of the AL heuristics. Fig. 11(a) also shows that class imbalance in the training set becomes slightly higher when using stratified bootstrap sampling [63] for the tree construction (all described experiments).

Fig. 11(b) displays the results of $QBC^{\times}_{RDH}$ when using regular bootstrap sampling, yielding a considerable bias toward the nonlandslide class, an underdetection of the affected area, and a relatively low producer's accuracy. Integrating stratified bootstrap sampling in the tree construction during the AL iterations leads to convergence during the learning process and provides a significantly higher F-measure than training with the regular bootstrap sampling [see Fig. 11(c)]. The inversion of the balance between user's and producer's accuracy when stratified bootstrap sampling is applied can be explained by the combined effect of queries and downsampling, which yields an overrepresentation of the landslide class in the overlap region of the two classes in feature space [64].

E. Spatial Distribution of Errors and Uncertainties

To investigate the causes of the remaining classification errors, 40 iterations of $QBC^{\times}_{RDH}$ are performed with the initially chosen standard parameter settings ($m = 3; t_i = 0.01; w = 100$ m) on the Brazil data set. The resulting landslide map and related errors are displayed in Fig. 12(a). The spatial batches queried during 40 iterations [see Fig. 12(c)] cover 0.4 km^2 and correspond to 3.8% of the entire study area. The achieved accuracy values [see Fig. 12(b)] are consistent with the mean accuracy values estimated through cross-validation in previous experiments, and $F = 0.73$ is above the lowest ($F = 0.71$) and below the highest ($F = 0.86$) interexpert agreement rate. Assuming that all queried samples were correctly labeled, the accuracy of the full map would be $F = 0.81$. In this context, it is important to note that a pointwise labeled training set contributes significantly less to the overall accuracy of the entire map.

The obtained landslide map is used to analyze the distribution of the classification errors in dependence of a few factors that are suspected to favor misclassification, namely, proximity to the landslide boundary, occurrence of shadows, and certain terrain positions. Fig. 12(a) illustrates that errors are clustered in certain regions and particularly along landslide boundaries. This is illustrated through quantitative comparison of the frequency density distributions of the distances to

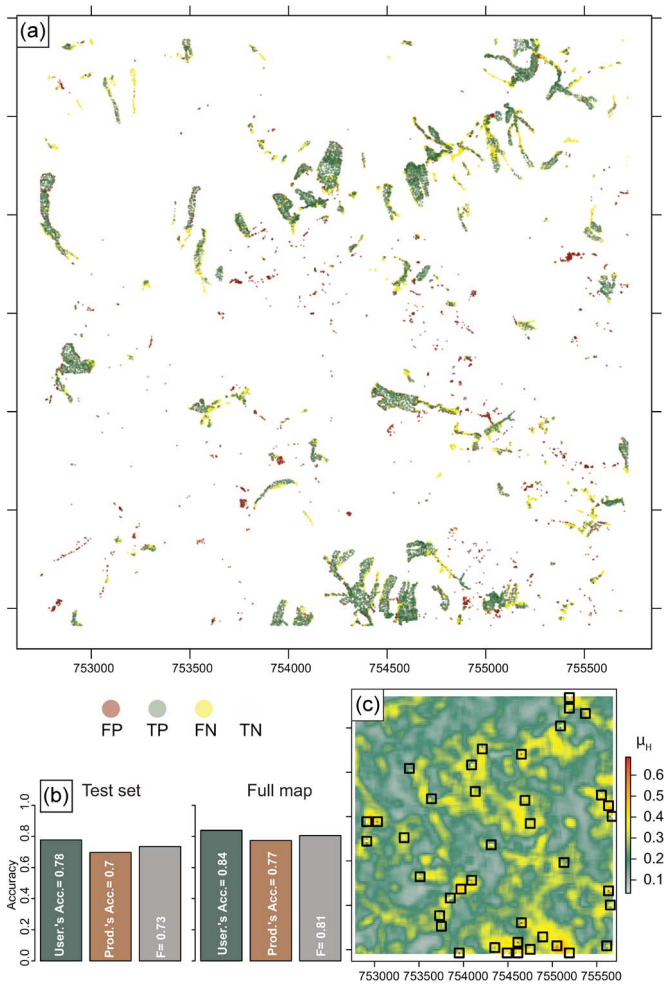


Fig. 12. (a) Final map results for the Brazil study site after 40 iterations with $QBC^{\times}_{RDH}$ ($m = 3$, $t_i = 0.01$, and $w = 100$ m). (b) Accuracy measure for the classification of the test set and the full map, assuming that all queried samples have been correctly labeled. (c) All 40 regions queried for training.

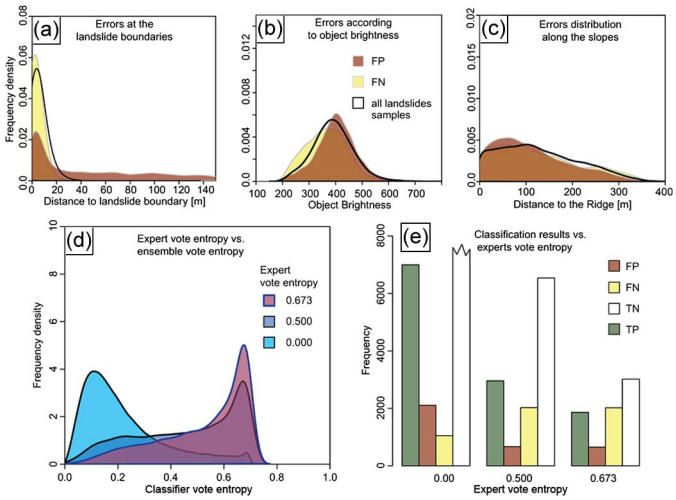


Fig. 13. Frequency distribution of classification errors for the map in Fig. 12(a) in relationship to (a) the distance to the landslide boundary, (b) the object brightness, (c) the distance to the most proximate ridge along the slope, and (e) the vote entropy of the expert geomorphologists. (d) Relationship between expert vote entropy and ensemble vote entropy.

landslide boundaries of false positives (FPs), false negatives (FNs), and all landslide samples [see Fig. 13(a)]. FNs concentrate along the boundaries of larger clearly visible landslides and, while introducing errors in the overall area, do not typically hinder the detection of the respective landslides. On the contrary, FPs are rather located some distant from actual landslides.

The presence of shadows [see Fig. 13(b)] could not be clearly related to an increased misclassification rate, whereas FPs had a slightly higher object brightness related to infrastructure and other anthropogenic features with high reflectance. Many of such FPs are located in the upper parts of the slopes, whereas FN errors display slight prevalence at the lower slopes [see Fig. 13(c)].

Analogous to the ensemble vote entropy (1), the five available expert mappings [see Fig. 4(a)] enable the computation of a map of the vote entropy of expert geomorphologists. The vote entropy of the experts can be considered as a measure for the uncertainty of the reference data, and it is interesting to compare this inherent uncertainty with the uncertainty of the RF classifier and the distribution of errors. Fig. 13(d) clearly shows that image objects with high expert vote entropy are also much more frequently evaluated as uncertain by the classifier. Fig. 13(e) complements this picture, showing that most of the correctly classified objects coincide with areas of low expert uncertainty, whereas FNs are particularly concentrated in areas with high expert vote entropy. It can be inferred that the true class of the latter is highly uncertain, and efforts to increase classification accuracy for those samples will be prone to failure because an improvement with respect to one expert map will result in a performance decrease with respect to the opposing opinion of another expert. On the contrary, more than half of the FPs overlap with areas where all experts agreed [see Fig. 13(e)]. The fact that all experts agreed on those samples suggests that they exploit relatively unambiguous image features in their judgement, making those samples interesting candidates for the design of enhanced object features. This group of samples might also be more reliable to evaluate algorithmic improvements since class membership does not depend on the respective ground truth.

VII. Conclusion

This work was dedicated to the integration of spatial constraints in the design of an AL algorithm to avoid a spatially dispersed distribution of queried samples and consequently reduce the sampling costs for remote sensing applications. Using the RF classifier and the QBC framework, three different AL methods are proposed to select interesting regions rather than individual sampling points. The developed heuristics make use of criteria for sample uncertainty and diversity and are evaluated for landslide mapping for two reference data sets using mono- and multitemporal images from different satellite sensors. To address class imbalance, which is an inherent issue for landslide mapping and many other remote sensing applications, all region-based AL heuristics are implemented using stratified bootstrap sampling for the construction of individual trees in the ensemble. This strategy helps achieve significantly

better balance between the user's and the producer's accuracy and, furthermore, reduce the computational time for training and classification.

The overall performance is quantified by accuracy gains (F-measure) in dependency of the number of iterations, runtime, and labeling time and compared with spatial coverage sampling and a pointwise query function. The comparison demonstrated that querying compact spatial batches significantly reduced the required number of iterations and, therefore, the labeling time and the computational runtime. The size of the spatial search window can be regarded as a tradeoff parameter that controls the size of the queried batches. A semivariogram analysis is adopted to assess the range of the spatial autocorrelation as an *a priori* estimate of potentially suitable window sizes. A systematic assessment indicates that region-based heuristics are efficient over a range of different window sizes. More favorable results are achieved with window sizes close to the mean effective range, whereas significantly larger window sizes rather lead to performance decay.

Two of the proposed region-based AL heuristics combined ($QBC_{RD\sigma}$ and $QBC^{\times}_{RDH}$) sample uncertainty with additional measures to select the most diverse batch among uncertain candidate regions. Particularly in early iterations, they frequently provide steeper learning curves than purely uncertainty-based heuristics ($QBC_{R\mu}$). $QBC_{RD\sigma}$ and $QBC^{\times}_{RDH}$ comprise two tuning parameters, namely, the number of candidate regions m and a variable importance threshold t_i. The impact of parameter changes on the algorithm performance is tested through experimentation showing that both algorithms benefited from a higher number of candidate regions ($m > 2$), whereas the omission of less important features ($t_i > 0.0$) does not lead to an enhanced performance. In terms of required labeling time, $QBC^{\times}_{RDH}$ provides the best performance (particularly in early AL iterations), whereas the purely uncertainty-based $QBC_{R\mu}$ requires less computational time and provides competitive results (particularly in later AL iterations). In comparison to pointwise queries, $QBC^{\times}_{RDH}$ helps reduce labeling time by 56.0% on the China data set and by up to 78.2% on the Brazil data set.

The results of 40 iterations of $QBC^{\times}_{RDH}$ on the Brazil data set are examined to gain a better understanding of the causes and distribution of remaining errors. Querying 3.8% of the study area for training provided a test set error of $F = 0.73$. Remaining errors are found to be spatially clustered and particularly FNs concentrated at the boundaries of manually mapped landslides. Although it is well understood that reference data sets from experts and other sources typically bear considerable uncertainties, their impact on the evaluation of image classification and feature detection still remains unconsidered in most remote sensing studies [53], [65]. The availability of five expert inventories at the Brazil study site enables comparison of the uncertainty of the expert judgment with the uncertainty of the classification and the distribution of remaining errors. A strong positive relationship between expert and classifier uncertainty is found as well as prevalence of FNs in areas where experts disagreed, showing that the corresponding samples are inherently difficult and should be considered as uncertain rather than as real errors. This issue and the fact that the classification accuracy already exceeded the agreement of the most disagreeing experts suggest that further studies must address uncertainties in the ground truth to achieve real enhancements for similar types of problems.

The region-based AL methods reduce labeling time when sampling through a visual interpretation of VHR images and are also a valuable tool to guide field sampling surveys that typically bear significantly higher costs for the analysis of low- and medium-resolution images. The employed criteria for sample uncertainty and diversity can be easily adapted for applications to multiclass problems, and the use of the proposed framework for applications such as land cover mapping could be an interesting direction for further research.

ACKNOWLEDGMENT

The authors would like to thank T. A. Lampert for proofreading and the fruitful discussions on ground truth uncertainty, and B. Koenig for his support in the elaboration of the landslide inventory maps.

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning—Data Mining, Inference, and Prediction*. New York, NY, USA: Springer-Verlag, 2009.

[2] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random Forests for land cover classification," *Pattern Recogn. Lett.*, vol. 27, no. 4, pp. 294–300, Mar. 2006.

[3] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.

[4] L. Bruzzone and F. Melgani, "Robust multiple estimator systems for the analysis of biophysical parameters from remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, pp. 159–174, Jan. 2005.

[5] G. Camps-Valls, L. Bruzzone, J. L. Rojo-Alvarez, and F. Melgani, "Robust support vector regression for biophysical variable estimation from remotely sensed images," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 3, pp. 339–343, Jul. 2006.

[6] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. L. Rojo-Alvarez, and M. Martinez-Ramon, "Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, pp. 1822–1835, Jun. 2008.

[7] A. Banerjee, P. Burlina, and C. Diehl, "A support vector method for anomaly detection in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 8, pp. 2282–2291, Aug. 2006.

[8] A. Brenning, "Benchmarking classifiers to optimally integrate terrain analysis and multispectral remote sensing in automatic rock glacier detection," *Remote Sens. Environ.*, vol. 113, no. 1, pp. 239–247, Jan. 2009.

[9] A. Stumpf and N. Kerle, "Object-oriented mapping of landslides using Random Forests," *Remote Sens. Environ.*, vol. 115, no. 10, pp. 2564–2577, Oct. 2011.

[10] D. Cohn, "Active Learning," in *Encyclopedia of Machine Learning*, C. Sammut and G. Webb, Eds. New York, NY, USA: Springer-Verlag, 2011, pp. 10–14.

[11] B. Settles, "Active Learning Literature Survey," Univ. Wisconsin–Madison, Wisconsin, MI, USA, Computer Sciences Tech. Rep. 1648, 2010.

[12] D. Lewis and W. Gale, "A sequential algorithm for training text classifiers," in *Proc. 17th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Dublin, Ireland, 1994, pp. 3–12.

[13] G. Schohn and D. Cohn, "Less is more: Active learning with support vectors machines," in *Proc. 17th Int. Conf. Mach. Learn.*, Stanford, CA, 2000, pp. 839–846.

[14] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proc. 5th Workshop Comput. Learn. Theory*, San Mateo, CA, 1992, pp. 287–294.

[15] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE J. Select. Topics Signal Process.*, vol. 5, no. 3, pp. 606–616, Jun. 2011.

[16] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 4, pp. 1231–1242, Apr. 2008.

[17] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, Jul. 2009.

[18] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 1014–1031, Mar. 2011.

[19] M. Volpi, D. Tuia, and M. Kanevski, "Memory-based cluster sampling for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 8, pp. 3096–3106, Aug. 2012.

[20] B. Settles, "From theories to queries," in *Proc. JMLR: Workshop Conf.*, Sardinia, Italy, 2011, pp. 1–8.

[21] B. Settles, M. Craven, and L. Friedland, "Active learning with real annotation costs," in *Proc. NIPS Workshop Cost-Sensitive Learn.*, 2008, pp. 1–10.

[22] A. Liu, G. Jun, and J. Gho, "Spatially cost-sensitive active learning," presented at the 9th SIAM Int. Conf. Data Mining Sparks, Sparks, NV, USA, 2009.

[23] E. Pasolli, F. Melgani, D. Tuia, F. Pacifici, and W. J. Emery, "Improving active learning methods using spatial information," in *IEEE IGARSS*, 2011, pp. 3923–3926.

[24] J. M. Henderson and A. Hollingworth, "High-level scene perception," *Annu. Rev. Psychol.*, vol. 50, pp. 243–271, Feb. 1999.

[25] A. Stumpf, N. Lachiche, J.-P. Malet, N. Kerle, and A. Puissant, "Active learning in the spatial-domain for landslide mapping in remote sensing images," presented at the ALRA: Active Learning Real-world Appl. Workshop ECML-PKDD, Bristol, U.K., 2012.

[26] A. J. Joshi, F. Porikli, and N. P. Papanikolopoulos, "Scalable active learning for multiclass image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2259–2273, Nov. 2012.

[27] T. S. F. Haines and T. Xiang, "Active learning using Dirichlet processes for rare class discovery and classification," in *British Machine Vision Conf. Univ. Dundee*, Dundee, U.K., 2011.

[28] J. Michel, J. Malik, and J. Inglada, "Lazy yet efficient land-cover map generation for HR optical images," in *Proc. IEEE IGARSS*, 2010, pp. 1863–1866.

[29] A. Borisov, E. Tuv, and G. Runger, "Active batch learning with stochastic query-by-forest (SQBF)," in *Proc. JMLR Workshop Active Learn. Exp. Des.*, 2011, pp. 50–69.

[30] J. Muñoz-Mariá, D. Tuia, and G. Camps-Valls, "Semisupervised classification of remote sensing images with active queries," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3751–3763, Oct. 2012.

[31] D. Tuia, J. Muñoz-Marí, and G. Camps-Valls, "Remote sensing image segmentation by active queries," *Pattern Recognit.*, vol. 45, no. 6, pp. 2180–2192, Jun. 2012.

[32] B. Demir, L. Minello, and L. Bruzzone, "A cost-sensitive active learning technique for the definition of effective training sets for supervised classifiers," in *Proc. IEEE IGARSS*, 2012, pp. 1781–1784.

[33] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[34] L. Copa, D. Tuia, M. Volpi, and M. Kanevski, "Unbiased query-by-bagging active learning for VHR image classification," in *Proc. SPIE Remote Sensing Conf.*, Toulouse, France, 2010, p. 78 300K.

[35] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang, "Representative sampling for text classification using support vector machines," in *Proc. 25th Eur. Conf. Inf. Retrieval Res.*, Pisa, Italy, 2003, pp. 393–407.

[36] S. Boltz, E. Debreuve, and M. Barlaud, "kNN-based high-dimensional Kullback–Leibler distance for tracking," in *Proc. 8th Int. WIAMIS*, 2007, p. 16.

[37] R package version 0.6-3 S. Li, FNN: Fast nearest neighbor search algorithms and applications 2012, R package version 0.6-3.

[38] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformat.*, vol. 8, p. 25, Jan. 2007.

[39] K. Nicodemus, J. Malley, C. Strobl, and A. Ziegler, "The behaviour of random forest permutation-based variable importance measures under predictor correlation," *BMC Bioinformat.*, vol. 11, p. 110, Feb. 2010.

[40] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," *Pattern Recognit.*, vol. 44, no. 2, pp. 330–349, Feb. 2011.

[41] K. K. Nicodemus, "Letter to the Editor: On the stability and ranking of predictors from random forest variable importance measures," *Briefings Bioinformat.*, vol. 12, no. 4, pp. 369–373, Jul. 1, 2011.

[42] W. G. Cochran, *Sampling Techniques*. NewYork, NY, USA: Wiley, 1977.

[43] D. J. J. Walvoort, D. J. Brus, and J. J. de Gruijter, "An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means," *Comput. Geosci.*, vol. 36, no. 10, pp. 1261–1267, Oct. 2010.

[44] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," in *Proc. 15th Int. Conf. Mach. Learn.*, Madison, WI, USA, 1998, pp. 1–9.

[45] A. S. Avelar, A. L. C. Netto, W. A. Lacerda, L. B. Becker, and M. B. Mendonça, "Mechanisms of the recent catastrophic landslides in the mountainous range of Rio de Janeiro, Brazil," presented at the 2nd World Landslide Forum, Rome, Italy, 2011.

[46] T. Gorum, X. Fan, C. J. van Westen, R. Q. Huang, Q. Xu, C. Tang, and G. Wang, "Distribution pattern of earthquake-induced landslides triggered by the 12 May 2008 Wenchuan earthquake," *Geomorphology*, vol. 133, no. 3/4, pp. 152–167, Oct. 2011.

[47] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini, "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sens. Environ.*, vol. 120, pp. 25–36, May 2012.

[48] ASTER-GDEM-VALIDATION-TEAM, ASTER global digital elevation model version 2—Summary of validation results, METI/ERSDAC, NASA/LPDAAC, USGS/EROS2011.

[49] E. J. Pebesma, "Multivariable geostatistics in S: The gstat package," *Comput. Geosci.*, vol. 30, no. 7, pp. 683–691, Aug. 2004.

[50] F. Ardizzone, M. Cardinali, A. Carrara, F. Guzzetti, and P. Reichenbach, "Impact of mapping errors on the reliability of landslide hazard maps," *Nat. Hazards Earth Syst. Sci.*, vol. 2, no. 1/2, pp. 3–14, 2002.

[51] M. Galli, F. Ardizzone, M. Cardinali, F. Guzzetti, and P. Reichenbach, "Comparing landslide inventory maps," *Geomorphology*, vol. 94, no. 3/4, pp. 268–289, Feb. 2008.

[52] P. Smyth, "Bounds on the mean classification error rate of multiple experts," *Pattern Recognit. Lett.*, vol. 17, no. 12, pp. 1253–1257, Oct. 1996.

[53] T. A. Lampert, A. Stumpf, and P. Gançarski, "An Empirical Study into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation," arXiv:1307.0426 [cs.CV], 2013.

[54] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Know. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[55] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, "Learning from data: A short course," *AMLBook*, 2012.

[56] M. Baatz and A. Schäpe, "Multiresolution Segmentation—An optimization approach for high quality multi-scale image segmentation," presented at the Angewandte Geographische Informationsverarbeitung XII, Salzburg, Austria, 2000.

[57] T. R. Martha, N. Kerle, C. J. van Westen, V. Jetten, and K. V. Kumar, "Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 4928–4943, Dec. 2011.

[58] T. Martha, N. Kerle, C. J. van Westen, and K. Kumar, "Characterising spectral, spatial and morphometric properties of landslides for semi-automatic detection using object-oriented methods," *Geomorphology*, vol. 116, no. 1/2, pp. 24–36, Mar. 2010.

[59] S. K. Jenson and J. O. Domingue, "Extracting topographic structure from digital elevation data for geographic information system analysis," *Photgramm. Eng. Remote Sens.*, vol. 54, no. 11, pp. 1593–1600, 1988.

[60] T. K. Tesfa, D. G. Tarboton, D. W. Watson, K. A. T. Schreuders, M. E. Baker, and R. M. Wallace, "Extraction of hydrological proximity measures from DEMs using parallel processing," *Environ. Modell. Softw.*, vol. 26, no. 12, pp. 1696–1709, Dec. 2011.

[61] *eCognition Developer 8.64.0 Reference Book*, Trimble, München, Germany, 2011, Trimble Documentation.

[62] J. Southworth, "An assessment of Landsat TM band 6 thermal data for analysing land cover in tropical dry forest regions," *Int. J. Remote Sens.*, vol. 25, no. 4, pp. 689–706, Feb. 2004.

[63] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," Dept. Statistics, Unive. California, Berkeley, CA, USA, 2004.

[64] V. García, R. Mollineda, J. Sánchez, R. Alejo, and J. Sotoca, *When Overlapping Unexpectedly Alters the Class Imbalance Effects*, J. Martí, J. Benedí, A. Mendonça, and J. Serrat, Eds. Berlin, Gemany: Springer-Verlag, 2007, pp. 499–506.

[65] G. M. Foody, "Assessing the accuracy of land cover change with imperfect ground reference data," *Remote Sens. Environ.*, vol. 114, no. 10, pp. 2271–2285, Oct. 2010.

2.5 *Discussion and perspectives*

It has been demonstrated that a combination of image segmentation, feature extraction, feature selection and supervised learning provides a generic framework for accurate landslide mapping in diverse environmental settings and with a great variety of VHR images from airborne and satellite systems. It was found that a small segmentation scale (over-segmentation) is preferable to obtain high classification accuracies corroborating previous findings for land cover classification (Smith, 2010). A RF-based feature selection method allowed selecting the most relevant features and permitted to reduce the complexity of the machine learning model while yielding comparable or even higher classification accuracies. It is also noteworthy that image segmentation reduces the number of samples by two orders of magnitude (compared to the number of pixels) and thereby greatly reduces the computational complexity of training and classification. Specifically designed textural and topographic-based features were found to significantly improve the class separability. The developed stratified resampling scheme allowed compensating the biases resulting from inherent class-imbalances.

A new AL heuristic was proposed which, contrariwise to previous approaches, considers spatial constraints on the sample selection, and costs in terms of labeling time. The developed region-based query function reduced significantly the labeling time required by the user when compared to point-wise queries, spatial-coverage sampling and random sampling. An evaluation of different scenarios in terms of data availability enabled to quantify the benefits of combining pre- and post-event VHR images (Lu et al., 2011, Appendix 1). A thorough assessment of the uncertainties of expert mappings reveals that the errors of the elaborated processing chain are not higher than the disagreement among multiple expert maps.

The experimental evaluation of different segmentation scales in this study suggests that over-segmentation provides higher accuracies when supervised learning is employed. Techniques for segmentation optimization with respect to the spectral variance of the resulting segments (Drăguţ et al., 2010; Martha et al., 2011) provide no guarantee to enhance the class separability for supervised learning. However, exploiting labelled samples and user interactions also during the segmentation (Tuia et al., 2012; Kurtz et al., In Press, Appendix 2) appears as a promising approach to extend the developed processing chain in the future.

The employed set of object-based features was specifically designed for landslide mapping and has proven generic in a variety of different scenarios. Further enhancements could be achieved through the use of hierarchical object features based on multi-level segmentations and integration of auxiliary maps depicting land cover and geological units. In this context, the employed feature selection framework constitutes a versatile tool to evaluate the relevance of new features. Since the design of object-specific features requires considerable expertise and time, emerging approaches for unsupervised feature learning (e.g. Coates and Ng, 2012; Coates et al., 2013) appear as a promising approach to exploit the constantly growing archives of optical earth observation data.

The proposed AL heuristic is, in principle, ready for operational use but would benefit from a further optimization to reduce computational costs and the time lag for the feedback loops. Other high-performance parallel implementations of the Random Forest algorithm are now publicly available (e.g. Schwarz et al., 2010) and should be tested in this context. In a recent study, we observed that support vector machines (SVM) provided slightly higher classification accuracies (Lampert et al., In Review, Appendix 3) but comprise drawbacks of

initial tuning of the hyper-parameters and scale less efficiently than the RF algorithm to large datasets.

The evaluation of the elaborated AL algorithm showed a significant reduction of the time required by the user for image interpretation and it can be expected that also sampling based on field surveys would benefit from the possibility to focus on a few relevant sampling areas rather than on hundreds of distributed points. Further efforts to take AL forward to practical remote sensing applications should consider the design and implementation of user-friendly interfaces and experimental evaluations in the framework of field surveys. The AL algorithm described in 3.3 could be extended with strategies to optimize the initialization of the sampling (Pasolli et al., 2013) and compared with cost-functions that takes the road network and different modes of transport into account (Demir et al., In Press).

Tests in real world applications should be addressed before further algorithmic developments since it may well turn out that the final accuracy and practical constraints (e.g. the point of departure in the field) are more important than gains of a few hours for the collection of the samples. Care should be taken to minimize the amount of free parameters that a potential user has to adjust, and to interact with experts in field surveys and image interpretation during the design of the system.

For reasons explained in 2.1, in this work, landslide mapping has been addressed as a binary problem. This is still an important limitation of the presented approach if inventory maps that comprise information about the landslide type or different landslide units are required. The RF algorithm is a multi-class classifier and hence directly applicable for the differentiation of multiple categories as long as training data for each class are provided. The features proposed in 2.3 and 2.4, however, would have to be extended and refined substantially to depict subtle differences among different landslide types. The problem is ill-posed since many features that are relevant for the identification of particular landslide type cannot be exploited before the affected areas have been identified. Two- or multi-stage strategies that implement mapping and characterization phases in separate steps could help to resolve this issue, and the processing chains developed in this thesis could be used as the first mapping step. We also addressed this issue in a further study with a hierarchical top-down framework that is based on successive analysis of multi-resolution remote sensing images and provided promising results (Figure 2-3) for the delineation of landslide parts (i.e. source, transport zone, toe) (Kurtz et al., In Press, Appendix 2).

The RF and AL framework proposed in this chapter have been tested with multiple image types and topographic datasets. Significant accuracy gains were observed when combining pre-and post-event images and further enhancements can be expected through the integration of time series that could capture seasonal trends and long-term land cover changes. Medium resolution satellites (Landsat 8, upcoming Sentinel-2) provide free access to such time-series and could be integrated as additional spectral features in the proposed framework. Such an approach should consider radiometric and atmospheric corrections to assure consistency of the absolute values over time and might also benefit from similarity measures that take into account the temporal structure of the time series (Petitjean et al., 2012).

Finally, it has been demonstrated in this chapter that a combination of specifically designed object-based features and machine learning algorithms allows the mapping of landslide affected areas with accuracies that are comparable to the variance among multiple experts. Indeed the uncertainty of the image classification strongly correlates with the uncertainty of the expert mappings. Endeavours towards further enhancements should not neglect this issue. A subsequent study was conducted to quantify the effects of ground truth

uncertainty and evaluate different methods for combining multiple-expert annotations (Lampert et al., In Review, Appendix 3). It showed that the performance ranking of different detection methods depends to a large degree on the considered ground truth dataset and hence the performance of the expert who created it. Consequently, whenever possible, multiple ground truth datasets should be collected to assess the intrinsic variance of the mapping and thereby the maximum accuracy that can be expected from an automatic mapping. This is probably not feasible in practical applications (e.g. landslide mapping after major events) and uncertainty maps (e.g. vote entropy) resulting from the classification process can be used as a proxy to spatially visualize the quality of the landslide mapping.

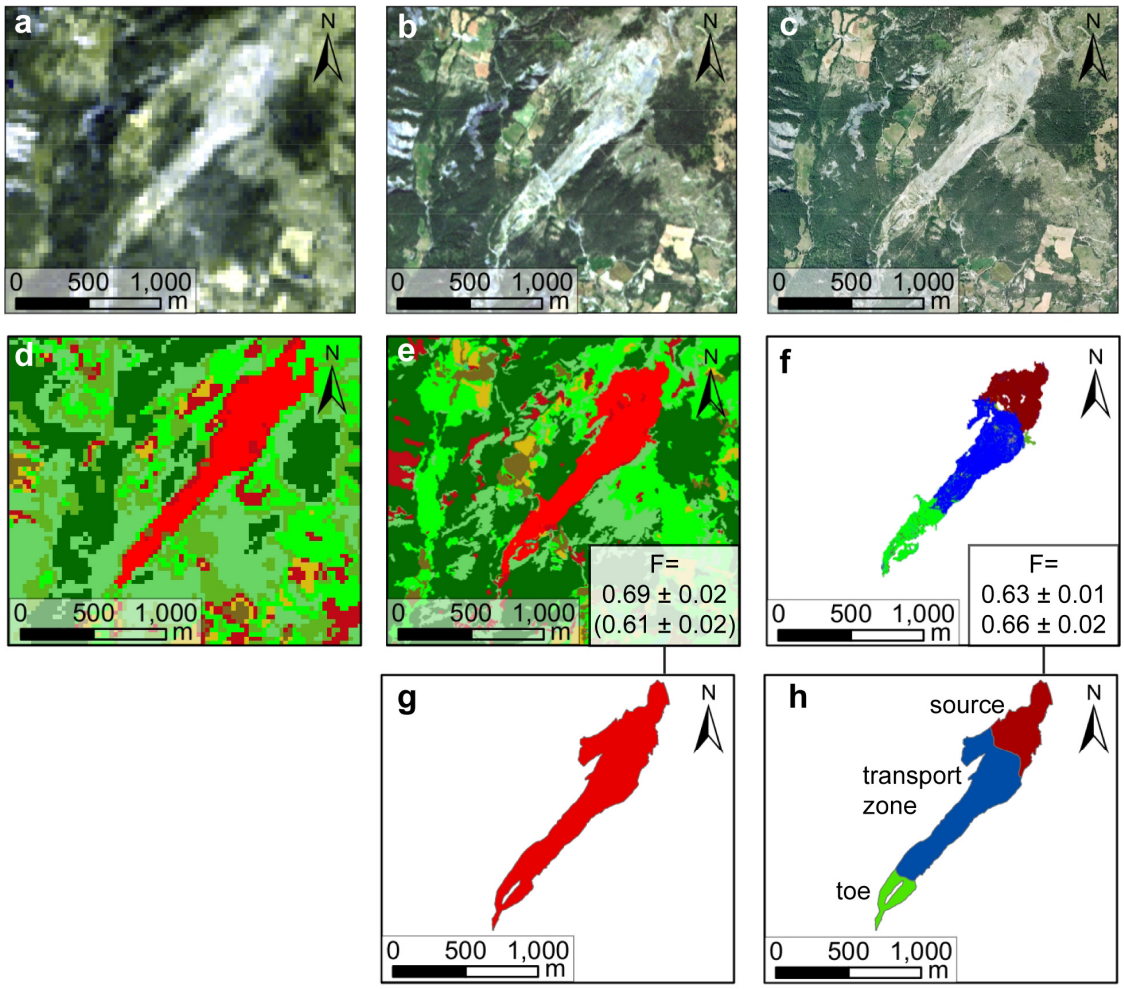


Figure 2-3: Exemplary results obtained with a top-down hierarchical image analysis approach for the delineation of different sub-parts of the La Valette landslide. Input images are from (a) LANDSAT TM (30 m), (b) RapidEye (5 m) and (c) an aerial survey (0.5 m). Classification results derived at (d) level 1, (e) level 2 (landslide delineation) and (f) level 3 (sub-part delineation) are shown. The accuracy (F-measure) of the classification was assessed against reference maps (g-h) elaborated by experts. The F-measures in brackets indicate the accuracies when the segmentation was learned at other landslides that occurred in the same geological setting and transferred to the La Valette landslide.

Synthesis of the research findings on landslide mapping at regional scales:

Main results: The research presented in this chapter was dedicated to the creation of a generic image processing chain for accurate and rapid landslide mapping at the regional scale with VHR optical images. Particular emphasizes was given to the reduction manual labor and time requirements at the side of the user and the assessment of mapping uncertainties and errors. The elaborated processing chains combine algorithms for image segmentation, feature extraction, feature selection and supervised learning and their general applicability was validated with a broad set of VHR images at several different study sites. A variety of general and tailored object features was evaluated with an innovative feature selection method and helped to identify a broad set of features that are generically applicable for operational use. Developed stratified and active learning sampling schemes were shown to effectively compensate class-imbalance and reduce the need for training data significantly. Mapping errors of the elaborated processing chain are not higher than the disagreement among multiple expert maps and encourage their use for operational landslide mapping especially in the aftermath of intense triggering events.

Perspectives: The developed processing chain currently addresses landslide mapping as a binary classification problem. The employed machine learning algorithms are suitable for multi-class classification and would enable the extension of the proposed framework for the differentiation of different landslide types. To allow a better exploitation of higher level object features multi-stage strategies differentiating detection and classification are a promising approaches. The feature selection method presented in this work is suitable for screening high-dimensional representations of the image data, whereas the possible range of relevant object features is still not fully exploited. Further attributes encoding high-level image information and trends in satellite image time-series are likely to promote a further reduction of the mapping error. Unsupervised feature learning methods should be considered as a possible tool to facilitate the engineering of new features automatically from archived data.
Our research demonstrated the significant impact of ground truth uncertainty on the evaluation of automatic image analyses methods and further investigations should quantify such uncertainties whenever possible. The definition of common standards for landslide inventory mapping and uncertainty measures should be targeted to ensure reproducible and comparable results among different study sites and investigations.
The developed processing chain has a great potential to be used as an operational tool for landslide mapping but still requires some optimization of the computational efficiency (e.g. parallel execution), the design of a user friendly interface and tests in real-time applications. Care should be taken to minimize the amount of free parameters that a potential user has to adjust, and to interact with experts in field surveys and image interpretation during the design of the system.

Chapter 3: Multi-platform optical data for landslide monitoring and characterization at the local scale

Objectives: Dedicated image analysis techniques are required for the characterization and quantification of landslide kinematics and underlying mechanical processes at the local scale. The objective of this chapter is the development of processing chains for acquiring spatially distributed information on surface displacements and related surface features testifying the landslide activity. In this context, the domain-specific objectives are:

- The generation of fissure maps from low-altitude aerial surveys to estimate landslide activity and enhance the understanding of the relationship between fissure formation, bedrock geometry and deformation in soft rock slopes;
- The quantification of horizontal displacement fields from VHR satellite images to extend ground-based point-wise measurements;
- The quantification of 3D surface displacement and transported volumes with terrestrial multi-view photogrammetry

These objectives are tackled through the elaboration of three dedicated image processing chains applied to VHR satellites images, UAV images, and terrestrial photographs. More specifically, the technical objectives are to:

- Develop a computer vision approach for the mapping of surface fissures in UAV image time-series;
- Develop a processing chain combining spaceborne stereo-photogrammetry and Digital Image Correlation (DIC) to generate high-resolution surface models and measure the horizontal surface displacement of several active landslides;
- Develop a processing chain to obtain multi-date dense point clouds and surface models from multi-view terrestrial photographs, as an alternative to the acquisition of LiDAR point clouds.

Methods: The elaboration of the image processing chains in this chapter departs from existing algorithms for line detection, image segmentation, as well as hierarchical- and subpixel image correlation. A medical image analysis algorithm for the detection of blood vessels in the human retina is adapted for the detection of surface fissures and complemented with an OBIA heuristic to exploit high-level image information. Algorithms for bundle-block adjustment, hierarchical DSM extraction, orthorectification and sub-pixel image correlation are combined to measure surface displacement from VHR satellite images, and extended with a dedicated post-processing routine for image denoising. Based on open-source libraries for tie point extraction, bundle-adjustment and dense matching, three pipelines for the extraction of dense point clouds from terrestrial photographs are implemented. An innovative cloud-to-cloud change detection method is adopted to assess the quality of the resulting models against LiDAR scans and detect significant surface changes.

This chapter is based on:

- Stumpf, A., Malet, J.-P., Kerle, N., Niethammer, U., Rothmund, S., 2013. Image-based mapping of surface fissures for the investigation of landslide dynamics. Geomorphology 186, 12-27, 10.1016/j.geomorph.2012.12.010.

- Stumpf, A., Malet, J.-P., Allemand, P., Deseilligny, M.-P., Skupinski, G., In Review. Terrestrial multi-view photogrammetry for landslide monitoring, Journal of Geophysical Research, 48 p.

- Stumpf, A., Malet, J.-P., Allemand, P., Submitted. Stereo-photogrammetry and displacement monitoring with Pleiades VHR satellite images, ISPRS Journal of Photogrammetry and Remote Sensing, 33p.

3.1 *Landslide monitoring and characterization at the site scale*

Surface displacement rates and the volumes involved are key parameters (1) to estimate the hazard associated with a specific landslide, (2) to complement long-term monitoring experiments for the determination of thresholds for early warning, and (3) to estimate the geometry and mechanical properties of unstable slopes.
Most commonly used techniques for displacement measurements are *in-situ* instruments such as GNSS (Figure 3-1b), inclinometers, extensometers and crack-meters (Malet et al., 2002). The resulting time series can be correlated with other physical parameters such as pore water pressures and rainfall amounts (Figure 3-1c) and lead to a better understanding of the relationship between displacement and environmental triggering factors.

Frequently used remote sensing techniques comprise Total Station measurements (Figure 3-1a), GB-SAR and TLS surveys. Several studies have also demonstrated the possibility to derive horizontal displacement measurements from optical satellite images and aerial photographs (Delacourt et al., 2004; Casson et al., 2005; Leprince et al., 2008; Debella-Gilo and Kääb, 2012) and more recently also 3D displacement from SAR amplitude images (Raucoules et al., 2013). Currently those measurements are of limited use for the integration in long-term monitoring experiments (e.g. landslide observatories) because of coarse image resolutions resulting in low accuracies or long-time intervals between subsequent observations. However, with the latest generation of VHR optical satellites, observation intervals of only a few days and decimetre accuracy get into reach, and could complement point-wise *in-situ* measurements with spatially distributed information on the displacement field.

Landslides typically feature complex 3D displacement and deformation patterns, and the displaced volumes can vary by several orders of magnitudes. The quantification of all displacement components and volumes of slope failures is consequently necessary to arrive at a comprehensive understanding of the underlying geomechanical processes. Monoscopic satellite images permit to recover the horizontal component of the displacement and photogrammetry with VHR stereo-pairs is currently limited to meter accuracy. TLS and ALS surveys permit to obtain dense surface representations (point clouds) with centimetre to millimetre accuracy and have become a frequently used technique (Jaboyedoff et al., 2010). Limiting factors for the application of LiDAR are still the relatively high cost of the equipment, the large data volumes that hinder fixed permanent installations and additional expenses if aerial surveys are desired.

Terrestrial photogrammetry permits detailed 3D surface reconstruction and displacement measurements and several case studies have demonstrated the value of this technique for rock mass characterization (Sturzenegger and Stead, 2009b) and historical reconstructions of landslide evolution (Kaufmann, 2012). Recently, Travelletti et al. (2012) demonstrated that a fixed installation of a consumer-grade SLR camera permits to obtain time series of the 3D displacement field (Figure 3-1d) with intervals of only a few days. Resulting displacement fields permit a more detailed description of the spatial and temporal patterns of acceleration and mass accumulation than point-wise measurements, and thereby help to refine conceptual and numerical landslide models. With only one camera, the approach is not fully 3D and still depends on the availability of frequently updated surface models but could be extended with stereo and multi-view camera setups.

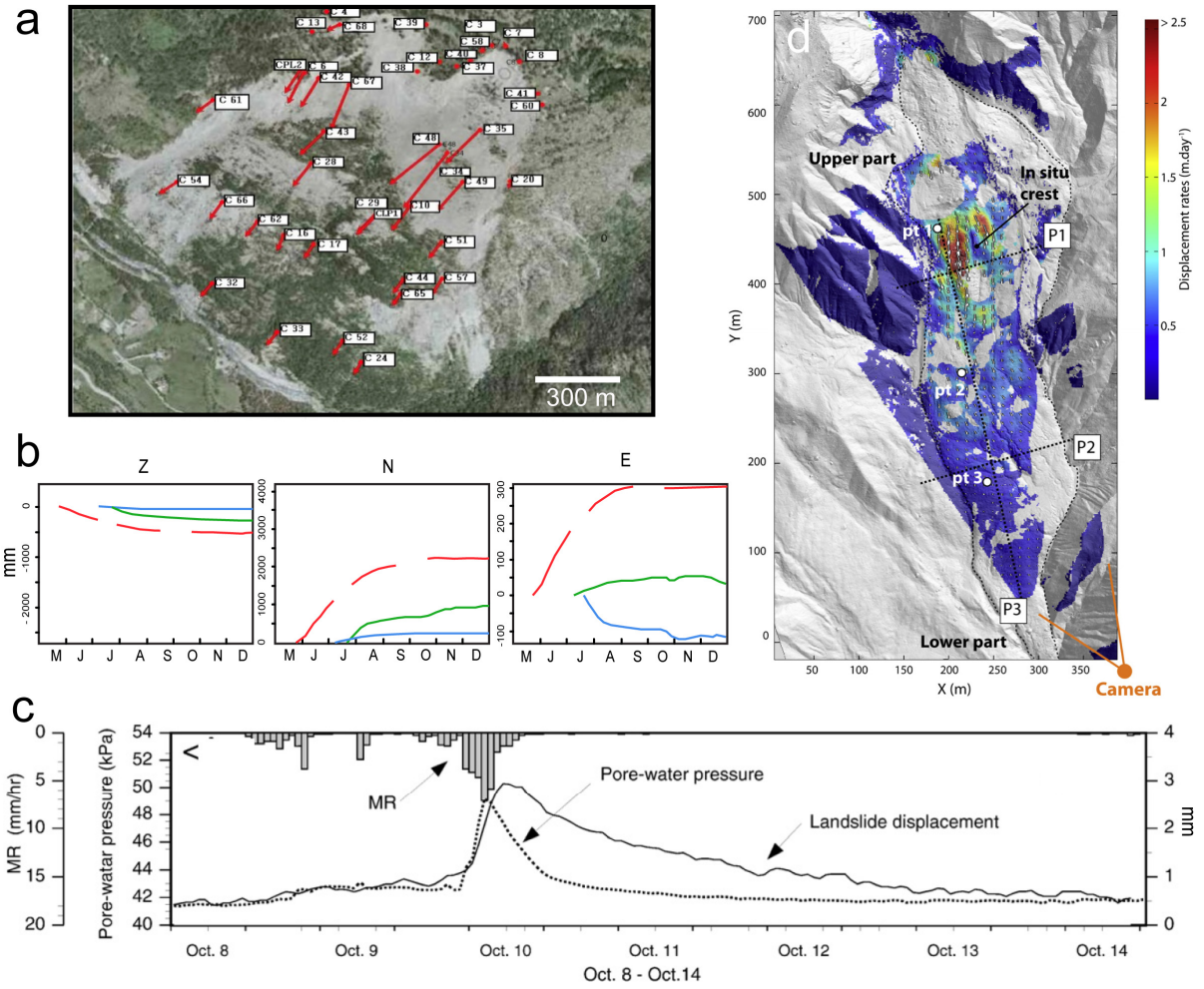


Figure 3-1: Examples illustrating the diversity of displacement time series and fields derived from local monitoring systems. (a) Mean annual velocity of the la Clapière landslide between May 2003 and January 2006 derived from Total Stations surveys (omiv.osug.fr), (b) XYZ displacement components of the Super-Sauze landslide from May to December 2011 with three permanent GNSS stations (http://eost.u-strasbg.fr/omiv/), (c) 7 day time series of displacement measurements derived with an inclinometer and comparison with time series of the pore-water pressure and melt- and rain water (Matsuura et al., 2008), and (d) displacement field of the Super-Sauze landslide for the period 1–4 June 2008 derived from correlation of terrestrial photographs (Travelletti et al., 2012).

Other indicators of landslide activity can be used in long-term observation experiments. For instance, in soft-rock slopes and sediments, the analysis of surface fissures may indicate the development of future failures (Krauskopf et al., 1939; Shreve, 1966; Chowdhury and Zhang, 1991; Abramson et al., 2001; Khattak et al., 2010) and their observation and interpretation can contribute to a better understanding of the controlling physical processes (McCalpin, 1984; Fleming and Johnson, 1989; Parise, 2003). Surface fissures also influence the sub-surface slope hydrology which in turn affects the ground-water system and the kinematic response of the slopes (Malet et al., 2005; Krzeminska et al., 2009; van Asch et al., 2009). Maps of surface fissures and other deformation features (Figure 3-2) can be obtained by extensive field surveys either through the direct visual observation of the topography (Fleming et al., 1999; Meisina, 2006) or through the indirect measure of seismic wave

propagation in tomography setups (Grandjean et al., 2011). While Sowers and Roster (1978) stated that aerial photographs do not reach sufficient spatial resolution, the increasing availability of UAV surveys has changed this situation (Niethammer et al., 2011a).

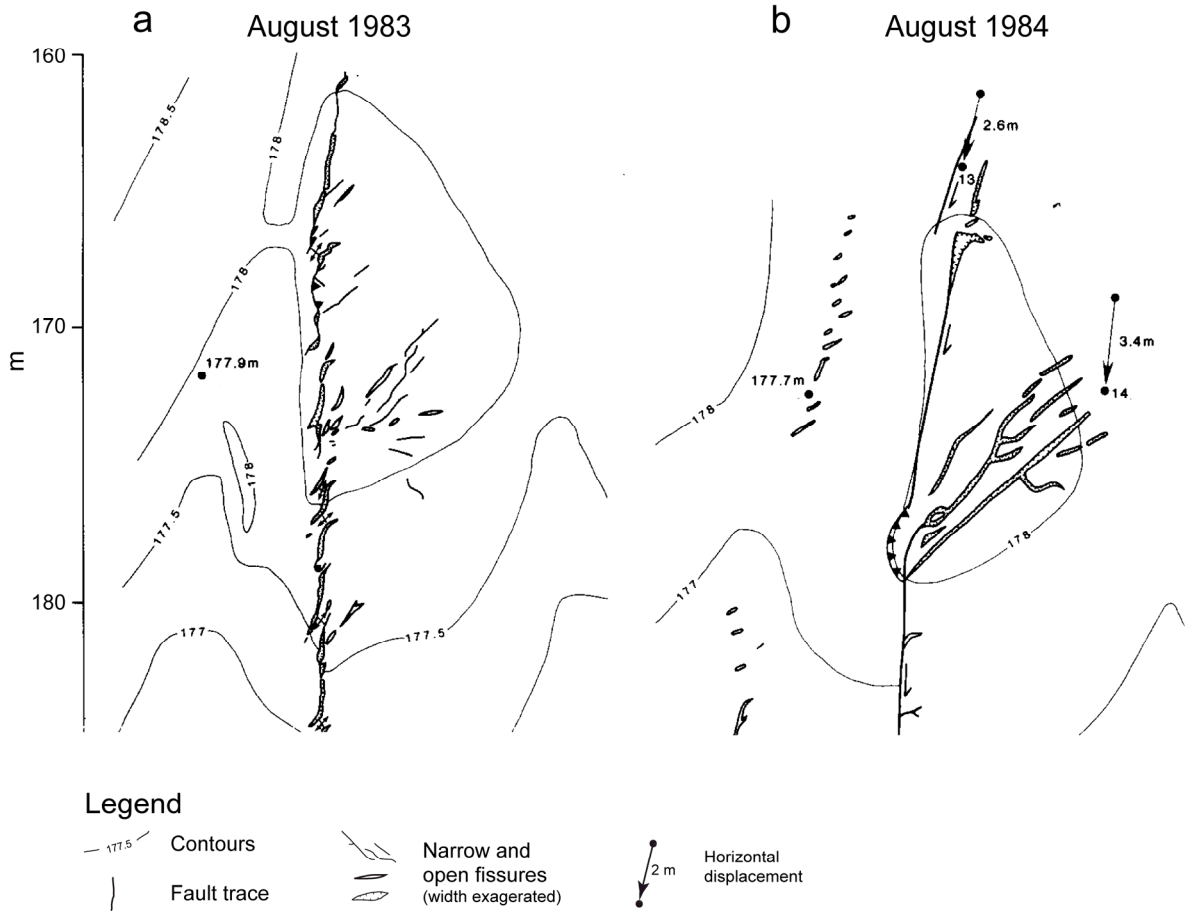


Figure 3-2: Maps depicting fissures at the right flank of the Aspen Grove landslide in (a) August 1983 and (b) August 1984 (modified after Fleming and Johnson, 1989). The fissure pattern indicates shear stress and the evolution of two sub-parallel strike-slip faults at the boundary of the landslide.

The domain-specific objectives of this chapter are:

- The generation of fissure maps from low-altitude aerial surveys to estimate landslide activity and enhance the understanding of the relationship between fissure formation, bedrock geometry and deformation in soft rock slopes;
- The quantification of horizontal displacement fields from VHR satellite images to extend ground-based point-wise measurements;
- The quantification of 3D surface displacement and transported volumes with terrestrial multi-view photogrammetry

The study sites investigated in this chapter are located in the South French Alps (Figure 3-3). Image processing chains are developed and tested in particular at three large landslides developed in clay-shales (Super-Sauze, La Valette, Poche) and monitored by the INSU Landslide Observatory OMIV (Observatoire Multi-Discplinaire des Instabilités de Versants).

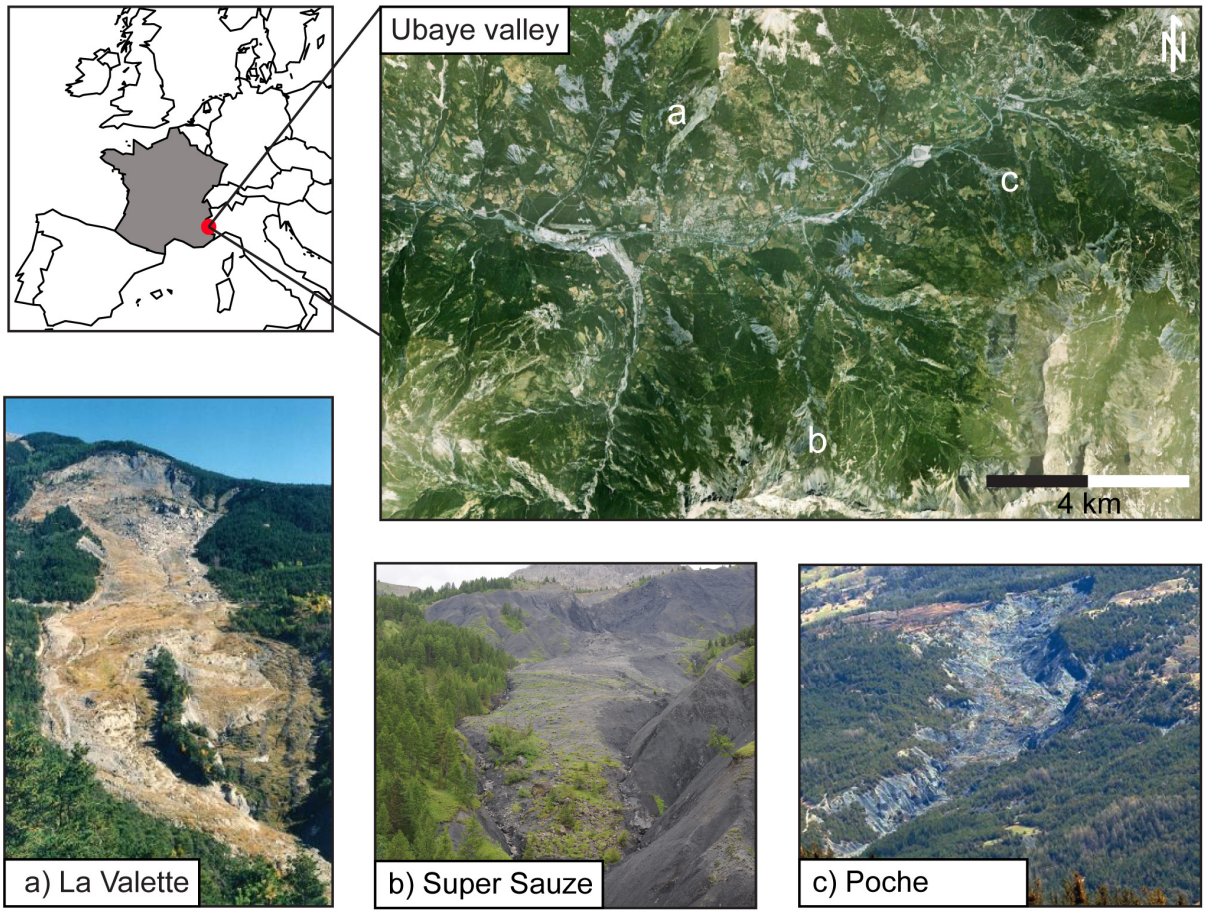


Figure 3-3: Locations of the study site in the South French Alps and overview of the three landslides investigated in this chapter.

3.2 *Image matching and feature detection for landslide monitoring: current limitations*

The detection and extraction of linear features from digital images is probably one of the most fundamental problems addressed in the field of pattern recognition. First general approaches for edge detection have been proposed in the 1980s (Marr and Hildreth, 1980; Canny, 1986) but the detection of linear features remains an active field of research in many disciplines such as medical research (Fraz et al., 2012), earth science (Shao et al., 2011) and signal processing (Lampert and O'Keefe, 2010). In particular the detection of blood vessels in images of the human retina has been studied intensively for many years now since the structure of the vessels is an important indicator for cardiovascular and ophthalmologic diseases and manual delineation of the fine vessel structures is a tedious and time-consuming task (Fraz et al., 2012). Also the detection and mapping of fine fissure structures depicted in UAV images would benefit from automated detection techniques that support visual image interpretation. However, the characteristics of UAV images and the surface fissures are significantly different from medical images depicting blood vessels and hinder a direct application of available detection algorithms.

Sub-pixel image matching algorithms have been developed and applied in many studies (Kääb, 2002; Delacourt et al., 2004; Casson et al., 2005; Leprince, 2008; Debella-Gilo and Kääb, 2011) to measure landslide surface displacement with airborne and spaceborne images (Figure 3-4). Theoretically DIC can yield sub-pixel accuracy but, image orientation, co-registration, decorrelation and especially the accurate modelling of topographic distortions still pose several limitations and high demands on the quality of the input datasets (e.g. up-to-date high-resolution topographic data, airborne stereo-pairs).

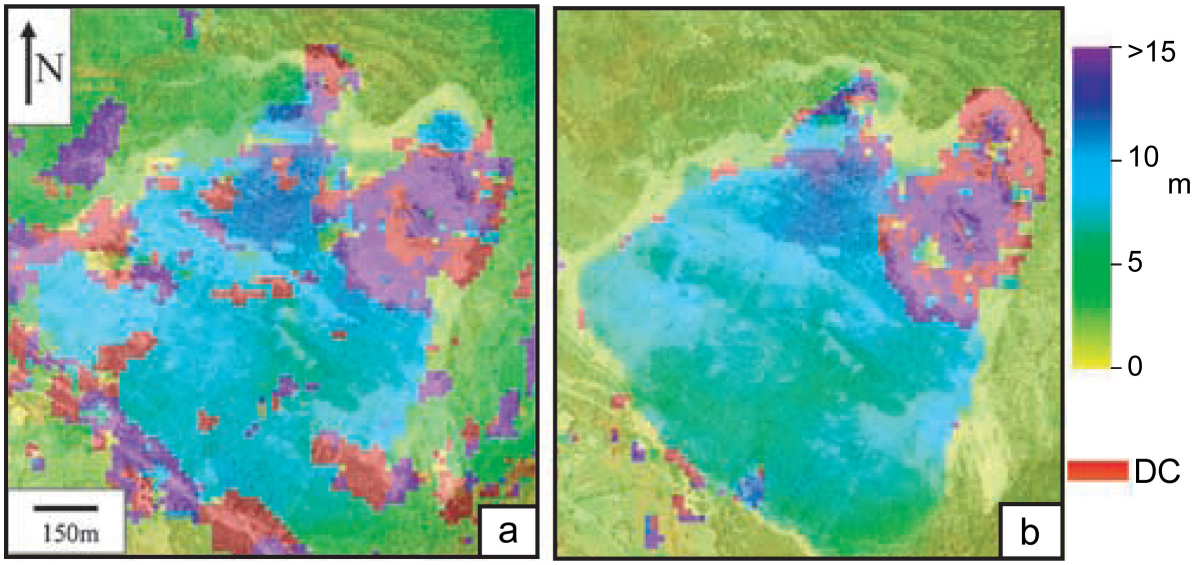


Figure 3-4: North-south deformation map of the La Clapière landslide derived from image correlation of (a) an aerial and a Quickbird image (1999 and 2003) and (b) two aerial images between 1995 and 1999. DC stands for decorrelation and marks areas where no significant measurement could be obtained (Delacourt et al., 2004).

The latest generation of agile VHR satellites provide enhanced geolocation accuracy and capabilities for the acquisition of monoscopic and stereo images at shorter time intervals. Stereo-photogrammetry and orthorectification can be used to extract high-resolution surface models and distinguish the topographic component (parallax shift) from the surface displacement preceding sub-pixel image correlation. Such processing strategies have been explored with medium resolution satellite images and aerial photographs for measurements of displacement exceeding several meters. (Kääb, 2002; Delacourt et al., 2004; Casson et al., 2005). One of the most agile and accurate satellite systems currently operational is the Pleiades constellation of the French CNES (Greslou et al., 2013) but its use for displacement measurements has not been explored yet. Limited knowledge about potential pitfalls, accuracies and the demands on ground control for the exploitation of resulting VHR images impedes further steps towards robust and precise operational tools.

Several case studies have demonstrated the value of terrestrial photogrammetry to measure surface displacements of landslides (Cardenal et al., 2008; Fernández et al., 2012; González-Díez et al., 2013) or rock glaciers (Kaufmann, 2012). Applications for the characterization of rock masses were also developed (Sturzenegger and Stead, 2009b, a). However, the proposed methods are often very case specific, and require significant expertise and manual intervention. As a consequence multi-date applications are difficult and the time lag between two observations is often more than one year.

Progress in computer vision and photogrammetry have recently converged in a new class of algorithms (e.g. Hirschmuller, 2008; Snavely et al., 2008; Deseilligny and Clery, 2011) that greatly simplify the surface reconstruction from multiple optical images. Those new algorithms and techniques are commonly referred as Structure-from-Motion (SfM) and Multi-View Stereo-photogrammetry (MVS) and have already been implemented in a number of commercial, open-source and web-based processing tools. Several geoscience researcher have already employed commercial and open-source implementations (Westoby et al., 2012) focusing mainly on the processing of UAV images. However, possible benefits of those new tools for landslide monitoring with terrestrial photographs have not yet been explored and the accuracy and limitations of different processing pipelines are poorly understood.

To address these current limitations for displacement measurements and mapping of surface features, the technical objectives of this chapter are to:

- Develop a computer vision approach for the mapping of surface fissures in UAV image time series;
- Develop a processing chain combining spaceborne stereo-photogrammetry and Digital Image Correlation (DIC) to generate high-resolution surface models and measure the horizontal surface displacement of several active landslides;
- Develop a processing chain to obtain multi-date dense point clouds and surface models from multi-view terrestrial photographs, as an alternative to the acquisition of LiDAR point clouds.

3.3 *Characterization of landslide dynamics through image-based mapping of surface fissures from low-altitude aerial surveys for the investigation*

This section is based on: Stumpf, A., Malet, J.-P., Kerle, N., Niethammer, U., Rothmund, S., 2013. Image-based mapping of surface fissures for the investigation of landslide dynamics. Geomorphology 186, 12-27, 10.1016/j.geomorph.2012.12.010.

This section describes the development and application of an image processing chain to detect and map landslide surface fissures at the surface of slow-moving landslide from multi-date low-altitude aerial surveys. The implemented processing chain exploits low-level image information with Gaussian-matched filters and morphological filtering with an OBIA post-processing routine that take into account higher-level image information, topographic data and the formative mechanical processes. Five images recorded with UAVs and other aerial platforms over a period of four years are analysed and the resulting fissure maps are compared to expert mappings in terms of area, density, and orientation for accuracy assessment. It is found that the proposed processing chain produced relatively few false positives and true positive rates up to 65%. The analysis is further extended for mappings on the slope scale and permitted to establish relationships between an increased fissure density and increasing displacement rates. The multi-temporal analysis shows that the similar fissure patterns recurrently emerge at the same positions indicating stress patterns that closely correlate with the geometry the underlying stable sliding surface.

Geomorphology 186 (2013) 12–27

Contents lists available at SciVerse ScienceDirect

Geomorphology

journal homepage: www.elsevier.com/locate/geomorph

Image-based mapping of surface fissures for the investigation of landslide dynamics

André Stumpf [a,b,c], Jean-Philippe Malet [b,*], Norman Kerle [c], Uwe Niethammer [d], Sabrina Rothmund [d]

[a] Laboratoire Image, Ville, Environnement, CNRS ERL 7230, University of Strasbourg, France
[b] Institut de Physique du Globe de Strasbourg, CNRS UMR 7516, University of Strasbourg, France
[c] Faculty of Geo-Information Science and Earth Observation, ITC, University of Twente, Enschede, The Netherlands
[d] Institut für Geophysik, University of Stuttgart, Stuttgart, Germany

ARTICLE INFO

Article history:
Received 22 February 2012
Received in revised form 17 September 2012
Accepted 11 December 2012
Available online 22 December 2012

Keywords:
Landslide
Fissure
Geomorphological mapping
Remote sensing
Image filtering
UAV

ABSTRACT

The development of surface fissures is an important indicator for understanding and forecasting slope movements. Landslide investigations therefore frequently include the elaboration and interpretation of maps representing their spatial distribution, typically comprising intensive field work and instrumentation. It is only recently that aerial photography with sub-decimetre spatial resolution is becoming more commonly available and opens a window to analyse such features from a remote sensing perspective. While these data are in principle helpful to elaborate maps from image interpretation techniques, there is still no image processing technique available to extract efficiently these geomorphological features. This work proposes a largely automated technique for the mapping of landslide surface fissures from very-high resolution aerial images. The processing chain includes the use of filtering algorithms and post-processing of the filtered images using object-oriented analysis. The accuracy of the resulting maps is assessed by comparisons with several expert maps in terms of affected area, fissure density and fissure orientation. Under homogenous illumination conditions, true positive rates up to 65% and false positive rates generally below 10% are achieved. The resulting fissure maps provide sufficient detail to infer mechanical processes at the slope scale and to prioritize areas for more detailed ground investigations or monitoring.

1. Introduction

Observations of features and structural patterns of earth surface landforms can reveal information on the origin and mechanisms controlling the geomorphological processes. Structural geology and geomorphology have developed comprehensive concepts to delineate geomorphological units and structure types from remote sensing images, and infer about mechanical processes without necessarily measuring displacement, deformation or the applied forces directly (Melton, 1959; Davis and Reynolds, 1996; Passchier and Trouw, 2005; Pollard and Fletcher, 2005). Surface discontinuities observed in rocks and sediments have proven to be valuable indicators of the deformation history and stress pattern of the slope. For landslide analysis, their observation and interpretation can contribute to a better understanding of the controlling physical processes and help in the assessment of the related hazards (McCalpin, 1984; Fleming and Johnson, 1989; Parise, 2003). In hard-rock slopes, the analysis of structural discontinuities (faults, bedding planes, joints, and fractures) allows us to characterize potentially unstable areas (Hoek and Bray, 1981; Matheson, 1983; Priest, 1993; Selby, 1993; Günther et al., 2004; Jaboyedoff et al., 2004; Glenn et al., 2006). In soft-rock slopes and sediments, the analysis of surface fissures may indicate the development of future failures (Krauskopf et al., 1939; Shreve, 1966; Chowdhury and Zhang, 1991; Abramson et al., 2001; Khattak et al., 2010) and is often considered as a geo-indicator of the activity stage of a landslide. In sediments, the surface fissure characteristics also influence water infiltration and drainage, which in turn affect the ground-water system and the kinematic response of slopes to hydrological events (Malet et al., 2003, 2005a; van Asch et al., 2009).

Maps of surface deformation features can be obtained by extensive field surveys either through the direct visual observation of the topography (Fleming et al., 1999; Meisina, 2006) or through the indirect measure of seismic wave propagation in tomography setups (Grandjean et al., 2011; Bièvre et al., 2012). Relatively large fissures on landslides may also be discernible in Very-High-Resolution (VHR) spaceborne images (Glenn et al., 2006; Youssef et al., 2009), but typically, those structures reach widths in the decimetre-range and at present only airborne photographs provide sufficient detail for their detection in the centimetric range. Recent studies (Eisenbeiss, 2009; Niethammer et al., 2011a) have shown that VHR images acquired from unmanned aerial vehicles (UAVs) are cost-efficient data sources for the monitoring of landslide surfaces with sub-decimetric image resolution. Especially small UAVs with payloads below 5 kg and operating altitudes below 2000 m are expected to be employed much more frequently in coming years (Frost and Sullivan Co., 2007) though more specific regulations for their operational use are being discussed at national and international levels (Prats et al., 2012; Watts et al., 2012).

Visual interpretation of VHR imagery is a classical method in geomorphology, but it remains subjective, and rather impractical for

* Corresponding author.
E-mail address: jeanphilippe.malet@unistra.fr (J.-P. Malet).

0169-555X/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.geomorph.2012.12.010

repetitive observations or the inspection of large areas. An increasing number of studies therefore targeted the development of automated techniques to extract relevant features from imagery (Graham et al., 2010; Martha et al., 2010; Stumpf and Kerle, 2011). Although the detection and extraction of linear features is a fundamental operation in digital image processing (Quackenbush, 2004; Mendonca and Campilho, 2006; Papari and Petkov, 2011), relatively few studies have explored the application of automatic approaches for the mapping of geomorphologically relevant linear features (Baruch and Filin, 2011; Shruthi et al., 2011).

Considering the increasingly widespread availability of sub-decimetre resolution images from UAVs and other airborne platforms, this study targeted the development of a semi-automatic image analysis technique to support geomorphologists in the detection, mapping and characterization of landslide surface fissures from VHR aerial images. In this context, the term "semi-automatic" expresses that the technique requires user input to be adapted for different image types and environmental settings. The developed method is based on a combination of Gaussian directional filters, mathematical morphology and object-oriented image analysis (OOA) and was tested on a set of multi-temporal VHR images acquired at the Super-Sauze landslide (southeast French Alps). The obtained results were compared to manual mappings carried out by experts combining image interpretation and field surveys.

2. Types of surface fissure observed on landslides

Detailed observations of landslide surface fissures were provided by Krauskopf et al. (1939) who adapted analogies from structural geology for their interpretation and distinguished between strike–slip structures, normal faults, graben structures and compression structures. In addition, Ter-Stephanian (1946) noticed the mechanical significance of surface fissures and elaborated a classification scheme relating fissure morphology and location within the landslide mass to corresponding mechanical processes. This included a first-order differentiation between upper extension, side friction, central compression, and lower creep-on cracks. Although some authors used similar classification schemes (Bombard, 1968), the adopted terminology varies among different authors and affected lithologies (Fleming and Johnson, 1989; Cruden and Varnes, 1996; Fleming et al., 1999; Walter et al., 2009), and the terms crack and fissure are often used synonymously to refer to a variety of surface discontinuities.

Here, *fissure* is adopted as a generic term for open fractures on the topographic surface of a natural slope. At first instance, *transversal*, *longitudinal and diagonal fissures* are distinguished according to their main orientation axes relative to the dip of the slope. This terminology can be used ad hoc to classify fissures solely based on geometric properties observed in the field or in an image. A more refined mechanical classification such as provided in Ter-Stephanian (1946) will generally require considerations of the fissure patterns, the involved material and the local geometry of the slip surface. The term *crack* is used in this manuscript when referring to genetic processes described within classical fracture mechanics (Anderson, 2005). It should be noted that the term *crack* is also often adopted to refer to shrinking–swelling induced fractures (Malet et al., 2003) which are not the objective of this study.

Classical fracture mechanics postulates tensile opening, sliding and tearing as the three basic modes for crack propagation (Fig. 1a). The concept has been developed for brittle material but is also adopted to explain fracturing of plastic materials at high deformation rates (Schulson and Duval, 2009). Surface fissures may develop from a combination of all three modes, whereas in practice, considering the relatively low tensile fracture toughness of most geomaterials (Backers, 2004; Ke et al., 2008; Schulson and Duval, 2009), tensile fracturing can be expected to dominate the formation of fissures at the free surfaces of a landslide. However, interpreting tension cracks as a direct indicator for a purely tensile stress regime may often fall too short. In fact, tensile fracturing may also result from relaxation of tensile stresses that originate from deformation induced by shearing and compression as well (Wang and Shrive, 1995). A mechanical interpretation and classification of the fissures must therefore consider the fissure pattern, material and landslide geometry.

Fig. 1b–d illustrates three typical fissure patterns that are frequently used as geoindicators of specific deformation processes in the above-cited studies. One commonly observed example for such patterns is the formation of en-echelon fissure arrays (Fig. 1c), often also termed *Riedel* shears (Riedel, 1929). They accommodate tensile stress and shear stress typically resulting from shear in the bounding zone of blocks moving with different displacement rates. Certain patterns such as arrays of transversal fissures (Fig. 1b) are typically associated with tension in the steeper upper slopes, whereas fissures resulting from compression and lateral extension (Fig. 1d) are more typically associated with gentler slopes in the transit and accumulation zones of landsides (Sowers and Royster, 1978). For landslides with a complex geometry, the position of those fissure patterns may however deviate considerably from this simple scheme (Niethammer et al., 2011a).

3. Study site and data

The Super-Sauze mudslide is an active slow-moving landslide located in the Barcelonnette Basin in the Southern French Alps (Fig. 2) that developed in weathered black marls in the 1960s, and features highly variable displacement rates (from 0.01 to 0.40 $m \cdot day^{-1}$) controlled by the local hydrological conditions (Malet et al., 2005a). The landslide measures 950 m from the main scarp to the toe, and is up to 150 m wide. The moving mass has a clay-rich matrix containing up to 30% coarse gravel as well as larger boulders and blocks (Malet et al., 2005a). The surface displays the signs of deformation in the form of ridges, bulges, lobes and fissures but also markers of surface erosion such as rills and small gullies. Unlike surrounding stable areas the landslide surface is largely bare and only at a few locations, especially at its toe, cushion plants form small vegetation patches. Fissure widths of 0.01–0.40 m, lengths of more than 1.0 m and depths of up to 1.5 m (Espinosa, 2009) can be observed in the field (Fig. 4b). During the last 15 years, the landslide has been investigated through numerous monitoring campaigns including in-situ geophysical measurements, terrestrial and airborne LiDAR (light detection and ranging) and the acquisition of VHR optical imagery. In the VHR airborne optical images, the fissures can be recognized as dark curvilinear structures (Fig. 2c–e) as soon as their width approaches one pixel in size. Previous studies (Malet, 2003; Niethammer et al., 2011a; Walter et al., 2012) already discussed relationships between the observed fissure patterns (Fig. 2c–e) and strain resulting from a spatially heterogeneous displacement field and interactions between moving mass and the stable bedrock. However, a full reconstruction of the complex bedrock geometry that may allow for a more detailed characterisation of the underlying deformation mechanisms has been conducted only recently (Travelletti and Malet, 2012).

3.1. Airborne acquisitions of VHR optical imagery at the Super-Sauze landslide

Between April 2007 and October 2009, diverse imaging systems and airborne platforms were used to acquire VHR images of the landslide at five different dates (Fig. 3). In July 2008, October 2008, and October 2009, a low-cost UAV system equipped with compact camera was operated at flight heights between 100 and 250 m yielding images of the surface with a ground resolution between 0.03 and 0.10 m. The individual images were corrected for barrel lens distortion, rectified according to ground control points (GCPs) measured with differential GPS (DGPS), and finally merged into one large orthomosaic. Further details on the image acquisition and processing were provided by Niethammer et al. (2010, 2011b) who quantified

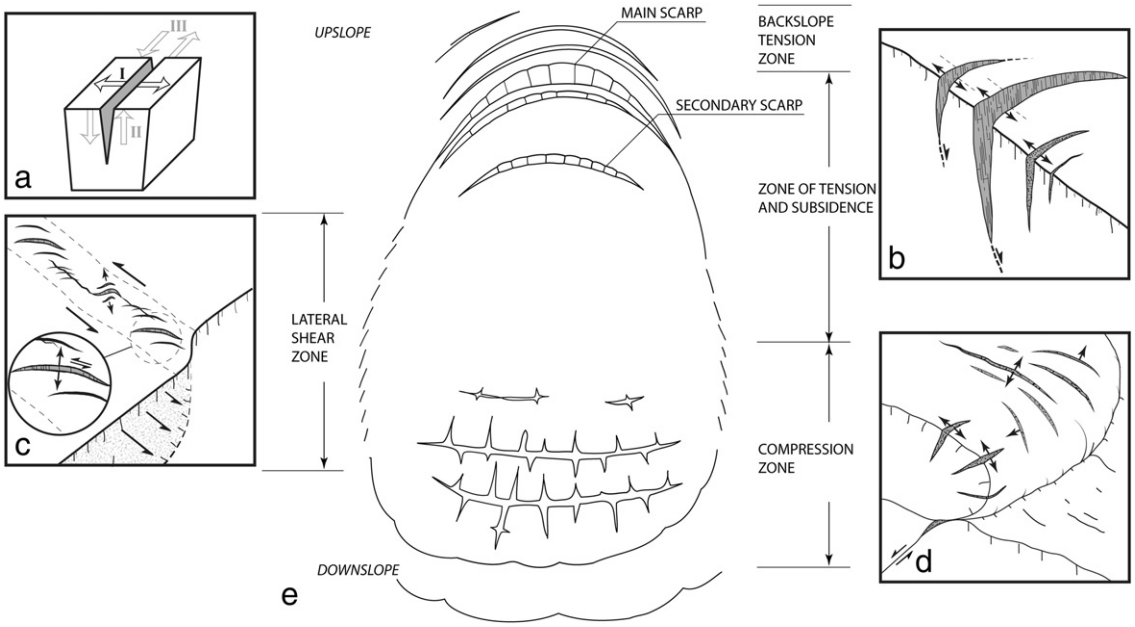


Fig. 1. Generic types of surface fissures and their typical spatial occurrence within a landslide mass. (a) Modes of fracture propagation: mode I (opening), mode II (sliding) and mode III (tearing). (b) Fissures developing predominately in mode I and resulting from tensile stress. (c) Fissures developing predominately in mode I and resulting from shear stress. (d) Fissures developing predominantly in mode I resulting from compressive stress and lateral expansion. (e) Division of a landslide mass (Sowers and Royster, 1978).

the residual positional error (x–y) for the October 2008 images with 0.5 ± 0.57 m within the boundaries of the sliding area. The UAV images for July 2008 and October 2009 are expected to provide better positional accuracies because they were orthorectified using elevation models that were generated from a photogrammetric analysis of the images.

During the two airborne LiDAR surveys in May 2007 and July 2009 (see Section 3.2), two orthomosaics of optical images with full coverage of the landslide were recorded using medium format cameras (Fig. 3) mounted on, respectively an airplane and a helicopter. The surveys used fully integrated systems for direct georeferencing and orthorectification with LiDAR surface models (see Section 3.2),

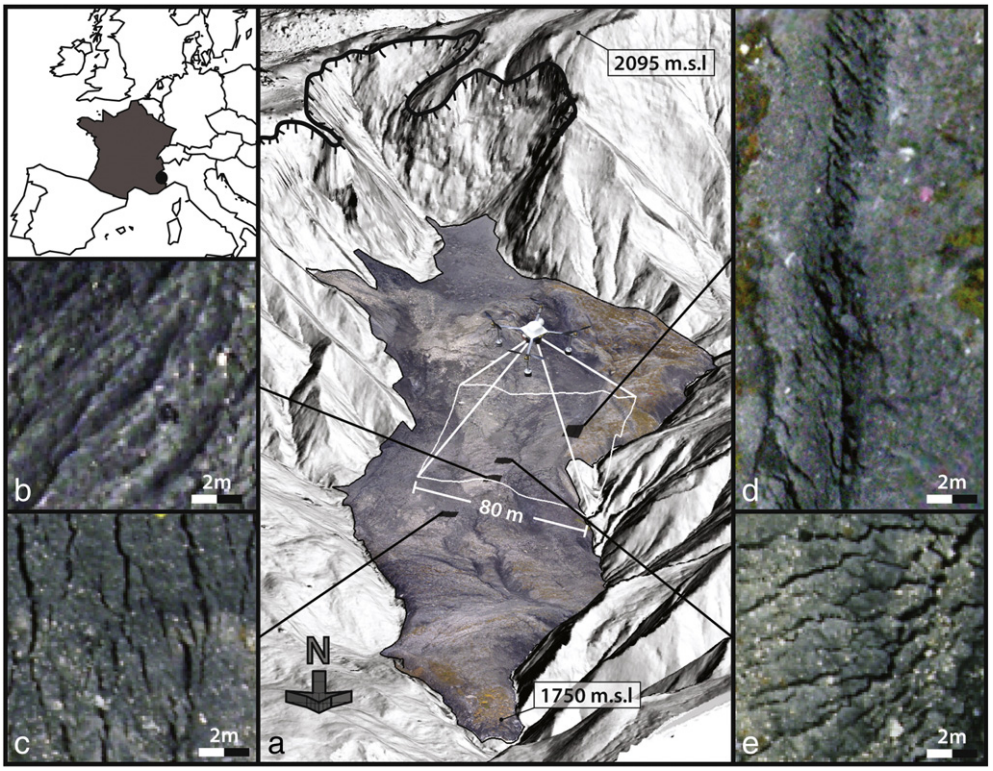


Fig. 2. Oblique view of the Super-Sauze landslide combining a hillshade image derived from an airborne LiDAR DTM (July 2009) and a UAV image (October 2008). (a) Main scarp (hashed black line), transport and accumulation zone (black outline), and area of interest for the multi-temporal analysis (white square). UAV image subsets show (b) compression ridges, (c) longitudinal fissures, (d) diagonal fissures at the boundary of the active part, and (e) and transversal fissures.

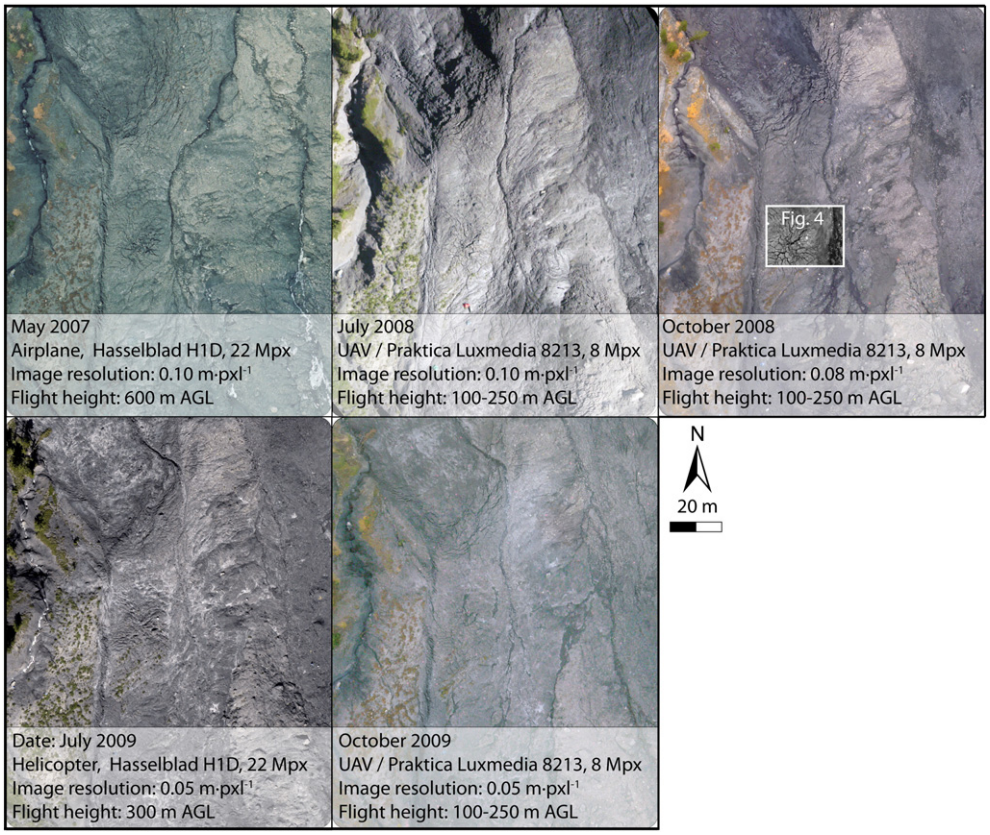


Fig. 3. Subsets of orthophotographs (see location in the white bounding box in Fig. 2a) acquired at five different dates with details of the acquisition systems and image ground resolutions.

which in general provide sub-decimetre positional accuracy in the *x*–*y* plane (Vallet, 2007).

For the study presented here, additionally 60 homologous tie points on stable areas were manually selected in the available images and showed a mean relative alignment error of 0.76 ± 0.82 m among the different acquisitions.

Further details on the adopted camera systems and the resolutions of the images resulting from the five surveys are summarized in Fig. 3. The figure also illustrates the considerable radiometric differences among the five images originating from illumination changes, seasonal variations and the distinct characteristics of the sensors. The scenes for May 2007, October 2008 and October 2009 were acquired under cloudy conditions with diffuse sky radiation and consequently show a more homogenous illumination of the surface. The scenes for July 2008 and July 2009 in contrast were recorded under sunny sky yielding strong contrast and many cast shadows. The latter are more prominent in the image for July 2008 which was recorded in the morning hours, at a relatively low sun angle. Although available methods for absolute and relative radiometric correction can be employed for the radiometric alignment of satellite images (Hong and Zhang, 2008; Vicente-Serrano et al., 2008), to the best of our knowledge, no approach exists to accurately align the radiometry of sub-decimetre images from different sensors, with substantial changes in illumination, a complex topography and changing surface characteristics. Initial test using histogram-matching, linear-regression (Schott et al., 1988) and iteratively re-weighted regression (Canty and Nielsen, 2008) did not provide satisfactory results. Consequently, no radiometric normalization was performed and the image analysis technique was designed and tested with radiometric diverse imagery.

In order to calibrate adjustable parameters of the detection algorithm to the targeted fissures and the variable scene characteristics, the processing was first tested on a subset of the terrain covering ~14,000 m^2 in the central part of the landslide (Figs. 2a and 3). This section was characterized by different fissure patterns and recorded during all surveys (including July 2008 and October 2009 which did not yield full coverage of the surface). Subsequently, the developed workflow was applied on the full scenes for a comprehensive mapping and analysis of the fissure distribution. Corresponding results for the full extent of the Super-Sauze landslide and their mechanical significance are discussed in Section 5.2.

3.2. LiDAR DTM

Two airborne LiDAR surveys were conducted in May 2007 and July 2009, respectively. The first survey used a Riegl LMS-Q560 laser scanner mounted on an airplane flying 600 m above the ground and resulted in a mean point density of 0.9 pts m^{-2} after vegetation filtering. The residual 3D positional error of the ground points was quantified as 0.12 m. The second survey was conducted with a Riegl Q240i laser scanner mounted on a helicopter and after vegetation filtering resulted in a mean point cloud density of 3.2 pts m^{-2}. The residual 3D positional error of the ground points was 0.07 m. Continuous surface rasters with a pixel size of 0.5 m were interpolated from the respective point clouds using Delaunay triangulation. The resulting surface was then adopted for the extraction of the principal hydrological drainage lines.

3.3. Reference datasets: expert maps of surface fissures

Reference mappings of the fissure characteristics (type and distribution) were elaborated by an expert geomorphologist familiar with the study site. The fissures were first identified on-site during a field survey carried out in October 2009 at the same time as the acquisition of the UAV images. The position of the fissures was mapped using a dGPS survey and terrestrial photographs. Then image interpretation rules were defined to identify and digitize the fissures on the images

as polyline vectors using a 2D view and at a scale of 1:250. The image interpretation rules were then applied to the four other images in order to elaborate an expert fissure map for each date. The resulting five maps were adopted as a reference to assess the performance of the semi-automatic method.

4. Image processing methods

While first generic edge detection operators were already proposed in the 1980s (Marr and Hildreth, 1980; Canny, 1986), the extraction of linear features from imagery remains a challenging task in many disciplines such as medical research (Mendonca and Campilho, 2006), earth science (Shao et al., 2011; Shruthi et al., 2011) or signal processing (Lampert and O'Keefe, 2011). For our focus, the specific challenges posed for an automation of fissure detection can be summarized as follows:

- The approach should be scalable to apply for variable fissure sizes and image resolutions, and as insensitive as possible to variable radiometric image characteristics;
- The technique should not respond to edges but enable the detection of dark curvilinear structures that may be oriented at any direction. Classical techniques such as Sobel operator and the Canny detectors (González and Woods, 2008) have been designed specifically for edge detection and are not directly applicable;
- The complex micro-topography, the presence of rock blocks and gravels as well as small patches of vegetation yield highly textured images. Consequently, the approach should enable us to smooth out spurious signals from the noisy background while still retaining small partially disconnected linear features of interest. Contextual scene information should be taken into account to resolve ambiguities of the local features.

Considering these challenges, a processing workflow including three main stages was developed. Firstly, a set of scalable Gaussian filters is applied to detect fissure candidates and suppress responses at edges. Secondly, a set of morphological filters is used to close small gaps along the extracted candidates. Thirdly, an object-oriented procedure is followed to eliminate some of the false positives exploiting higher-level scene information with contextual rules.

4.1. Stage 1: extraction of fissure candidates using a Gaussian matched filtering algorithm

A particularly well-studied example for the detection of dark curvilinear structures is the extraction of dark blood vessels in photographs of the human retina. Based on the observation that the cross-profiles of the vessels resembles a Gaussian distribution, Chaudhuri et al. (1989) proposed the use of a matched filter (MF) that is essentially a Gaussian convolution kernel subtracted by its own mean value. As illustrated in Fig. 4a, the cross-sections of surface fissures can be approximated with a Gaussian distribution and an MF scaled to the size of the fissure will give a peak response when crossing the fissure at an angle of approximately 90°. Because the MF still yields errors such as false detections at step edges (Fig. 5a, c) numerous extensions (Hoover et al., 2000; Sofka and Stewart, 2006) and alternative approaches (Mendonca and Campilho, 2006; Soares et al., 2006) have been developed. Recently, Zhang et al. (2010) proposed modification to the original MF filtering approach integrating a first order derivative of a Gaussian function (FDOG) to locally adapt the thresholds separating dark lines from non-target features. Compared to other state-of-the-art algorithms their approach provided competitive accuracies while being a computationally efficient and hence easier to apply on the large images resulting from VHR remote sensing.

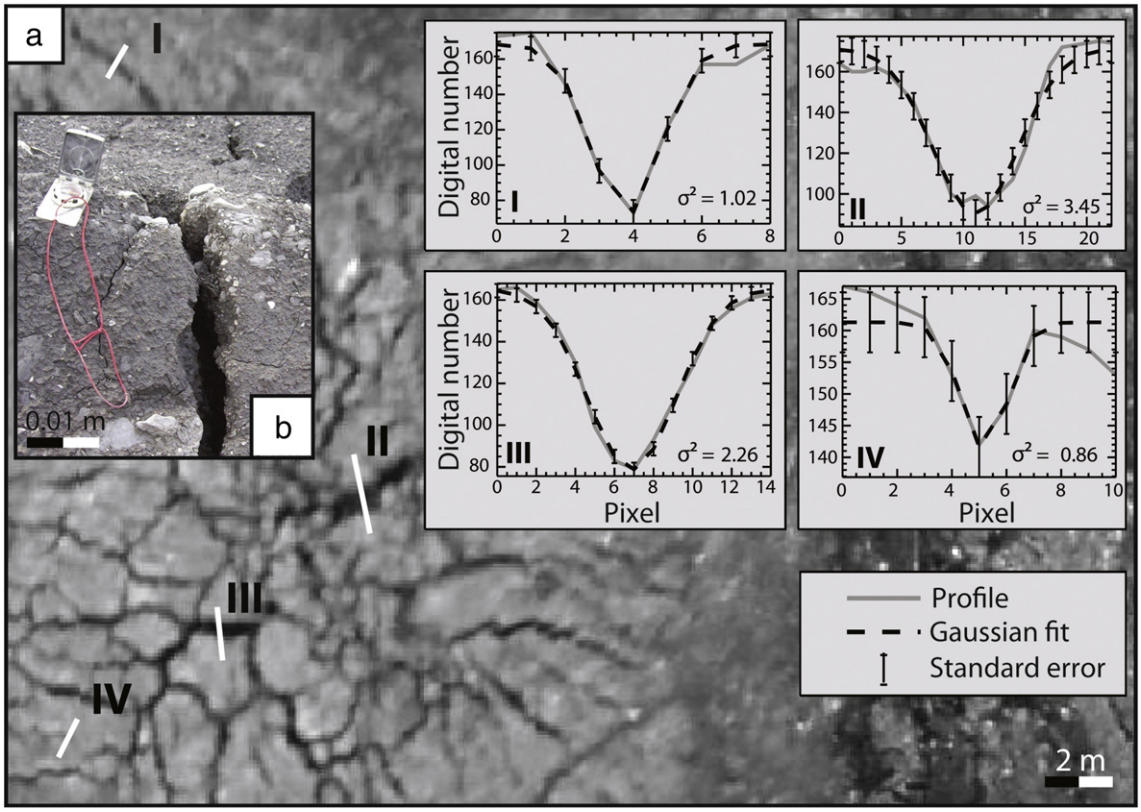


Fig. 4. Images of surface fissures. (a) Subset (see extent in Fig. 3) of the UAV image from October 2008 showing typical fissure patterns and (I–IV) grey-value profiles (green channel) approximated with Gaussian curves. (b) Field photograph taken in October 2009.

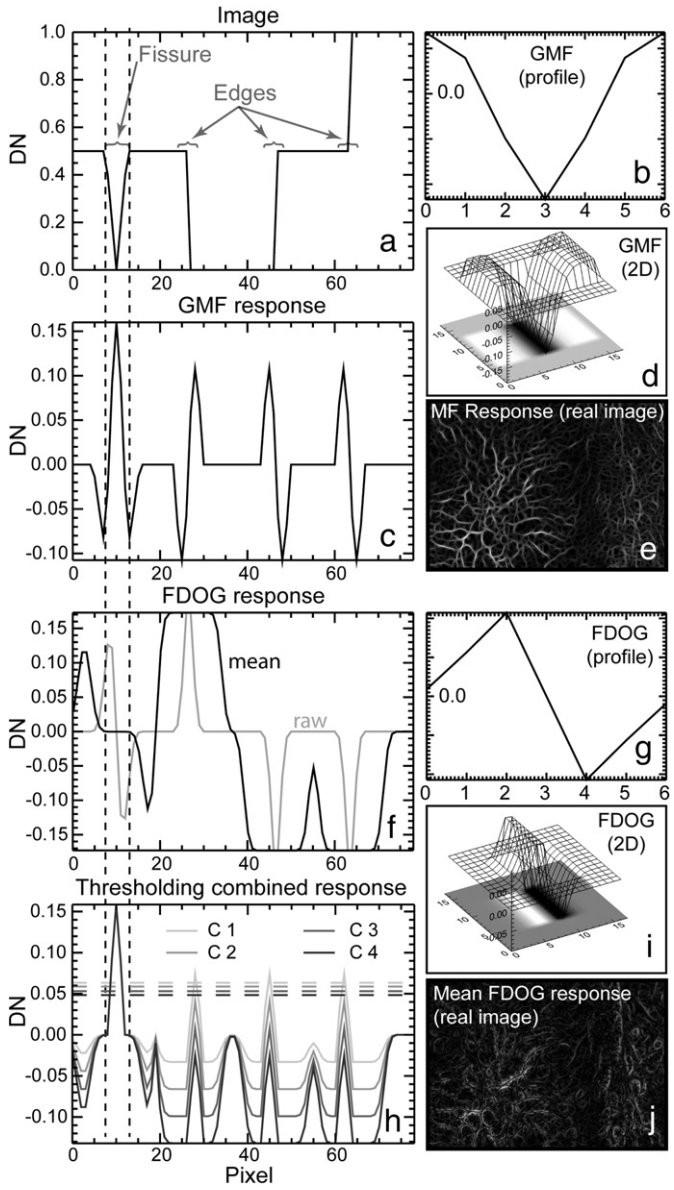


Fig. 5. Illustration of the principles of the Gaussian filtering for (a–c and f–h) a simplified 1-D case, (d and i) a 3D visualization of 2D filters and (e and j) the filter responses for the image subset in Fig. 4a. See text for details.

For this study, a similar approach was implemented in ENVI-IDL 4.8 (ITT Visual Information Solutions). The algorithm and its parameterization are detailed below.

The MF is a two dimensional kernel defined in the *x*-direction by an inverted Gaussian profile (Fig. 5b), and in the *y*-direction by replicates of the same profile (Fig. 5d). It may be denoted as:

$$\mathrm{MF} = g(x,y;\sigma) = -\frac{1}{\sqrt{2\pi}\sigma} e^{\left(\frac{x^2}{2\sigma^2}\right)} - m, \text{for } |x| \le 3\sigma, |y| \le L/2 \quad (1)$$

where σ denotes the standard deviation of the Gaussian functions and relates to the width of the targeted feature. To centre the kernel on zero, it is subtracted by its own mean m. The extent of the kernel in the *x*-direction is typically constrained to 3σ, whereas L defines the extent of the kernel in the *y*-direction and can be related to the length of the fissures. Because the matched filter still yields false responses at dark and bright step edges (Fig. 5c) Zhang et al. (2010) proposed to use the response of the FDOG to locally adjust the thresholds which are applied to classify the MF response into fissure and non-fissure structures. In analogy to Eq. (1), the first order derivative filter may be denoted as:

$$\mathrm{FDOG} = g'(x,y;\sigma) = -\frac{1}{\sqrt{2\pi}\sigma^3} e^{\left(\frac{x^2}{2\sigma^2}\right)}, \text{for } |x| \le 3\sigma, |y| \le L/2 \quad (2)$$

Fig. 5f illustrates that the FDOG responds with a single peak to edges but with a zero crossing at the centre of the idealized fissure. A simple mean filter can be applied to broaden the zero crossing into a plateau covering the whole width of the fissure (Fig. 5f). Subtracting the smoothed FDOG response from the MF response will attenuate the signal at edges while at the position of the fissure the full response is retained (Fig. 5h).

Since the orientation of the fissures is a priori unknown, multiple rotated versions of the Gaussian filters are applied to the image and for each pixel only the maximum response value is retained. This corresponds to finding the angle $\theta_{\max(x,y)}$ which maximizes the filter response at a given position in the image $I_{(x,y)}$ using:

$$\theta_{\max(x,y)} = \arg\max\left(I_{(x,y)} \otimes \mathrm{MF}_{\theta}\right), \text{for } 0 < \theta_i \le \pi \quad (3)$$

where $\otimes$ denotes the convolution operator and θ the orientation of the MF.

The calculation of the maximum response image R can then be obtained with:

$$R_{(x,y)} = \left[I_{(x,y)} \otimes \mathrm{MF}_{\theta\,\max(x,y)}\right] > 0 \quad (4)$$

where all negative response values are automatically set to zero and only values greater than zero are retained. The FDOG filter is rotated according to the determined $\theta_{\max(x,y)}$ and the corresponding response image D can be derived by:

$$D_{(x,y)} = \left|I_{(x,y)} \otimes \mathrm{FDOG}_{\theta_{\max(x,y)}} \otimes M\right| \quad (5)$$

where M denotes the above-mentioned mean filter used to broaden the zero crossing to the width of the fissures.

While Zhang et al. (2010) used a very broad mean filter with a fixed size, we suggest to use a kernel size that matches the width of the Gaussian kernel (6σ) and is thereby related to the width of the targeted features (Fig. 5a, f). In contrast to early studies where the FDOG response was used to locally adapt the threshold (Zhang et al., 2010) the final response image $\bar{R}$ is obtained by subtracting the FDOG from the GMF response using:

$$\overline{R_{(x,y)}} = R_{(x,y)} - Ct * D_{(x,y)} \quad (6)$$

where Ct denotes a user defined trade off parameter to adjust the sensitivity of the detection with typical range of values between 3 and 4. A threshold T is defined by:

$$T = \mu_{\bar{R}} + 2\,\sigma_{\bar{R}} \quad (7)$$

where $\mu_{\bar{R}}$ is the mean of the response image $\bar{R}$ and $\sigma_{\bar{R}}$ is the respective standard deviation.

A binary fissure candidate map F_{map} is obtained by applying the threshold T on the response image $\bar{R}$ using:

$$\overline{R_{(x,y)}} \ge T_{(x,y)} : F_{\mathrm{map}} = 1 \text{ and } \overline{R_{(x,y)}} \mathrm{b} T_{(x,y)} : F_{\mathrm{map}} = 0. \quad (8)$$

The thresholding after subtraction of the FDOG response was found to provide a generally more robust attenuation of undesired

Table 1
Parameter set of the Gaussian filters scaled according to the respective image resolution.

Image date	May 2007	July 2008	October 2008	July 2009	October 2009
Pixel size [m]	0.10	0.10	0.08	0.05	0.05
σ [pixel]	0.60	0.60	0.75	1.20	1.20
L [pixel]	10	10	12	20	20
n_θ	36	36	36	36	36
Ct	3.0	3.0	3.0	3.0	3.0

edge responses than the technique previously applied by Zhang et al. (2010).

In summary, the user needs to specify four simple parameters, namely (1) the scale of the filter kernels in terms of σ, (2) the length L of the kernel, (3) the constant Ct of the thresholding sensitivity and (4) the number of orientations n_θ at which the filters are calculated. In this study, n_θ was kept constant at 36 for all experiments, whereas, if computational time becomes an issue, the angular resolution may be reduced to 12 steps without major losses of accuracy. To determine σ a tool was created, which allows drawing profiles on the image and automatically estimates the fitting Gaussian function (Fig. 4). Cross-profiles of the smallest fissures visible in the image with the coarsest resolution (0.1 m pixel^{-1}) were best fitted by Gaussian curves with $\sigma \approx 0.6$. To ensure a homogenous scale of the detected features among all images, the kernel can be scaled by changing σ relative to the image resolution. If, for instance, the image resolution is increased to 0.08 m pixel^{-1}, a value of $\sigma \approx 0.75$ yields a kernel with the same physical size (Table 1). The same applies to the filter length L which was estimated at 1 m corresponding to the typical minimum length of the fissures. Resampling of the images can thereby be avoided. In our experience, σ establishes the lower bound for the width of the targeted features, whereas the filters still remain sensitive to features which are up to 5 times larger. For the choice of σ it is also helpful to note that the discrete kernel cannot represent FDOG functions with $\sigma \leq 0.5$.

To assess the sensitivity of the parameters and to determine a suitable threshold parameter Ct, a sensitivity analysis was carried out on a subset of the October 2008 image. Based on a visual assessment, values of $L = 1$ m and $\sigma = 0.75$ were found suitable for the detection of the fine fissure structures. The preliminary analysis also showed that increasing the parameters L and σ directs the detection towards more elongated and broader features, whereas in general the sensitivity of those parameters is rather low compared to the influence of the threshold Ct. Values of $Ct = \{0.0, 1.0, 2.0, 3.0, \text{and } 4.0\}$ were tested and based on a visual assessment of the outputs, a value of $Ct = 3$ was established for an optimal trade-off between detection rate and the amount of false positives. The final parameter set is summarized in Table 1.

Table 2
Summary of the thresholds adopted in the object-oriented post-processing routine. See text for details.

Feature	Thresholds
Shadow	Red < 100*, 40**
Shadow ratio	≤ 0.33
Vegetation	Ratio blue ≤ 0.33 + Otsu
Relative border to vegetation	≤ 0.15
Minimum angular difference	> 13°
Minimum length (clean up)	≥ 0.4 m
Minimum area (clean up)	> 0.1 m^2
Minimum fissure density (clean up)	> 1%.10 m^{-2}

* For May 2007, July 2008 and October 2008.
** For July 2009 and October 2009.

4.2. Stage 2: connection of broken lines using structuring elements

The highly textured surface of the landslide constitutes a noisy background that affects the detection especially at section where the fissures are very thin or partially occluded. While a human operator can easily interpolate broken lines through perceptual grouping (Metzger, 1975), this needs special attention for a semi-automated mapping technique.

To close small gaps between broken line segments of the detected candidates, a hit-or-miss transform algorithm (Serra, 1982) was used. The transform assigns a value of 1 to each pixel whose local neighbourhood fulfils the criteria defined by hit-and-miss structures (Fig. 6a), also known as structuring elements. They were defined to address all plausible 3-by-3 neighbourhoods representing small gaps in the detection starting from four prototype hit-structures shown in Fig. 6b. The respective miss-structures (Fig. 6c) are typically derived by simply inverting the prototype hit-structures, and both elements were rotated (Fig. 6d) to test for a total number of 24 possible neighbourhood arrangements. Exceptional cases were thereby the structuring elements for closing directly diagonal gaps, where an extended neighbourhood was used for the hit-and-miss structures (Fig. 6b, c) to prevent connections of lines running parallel to each other.

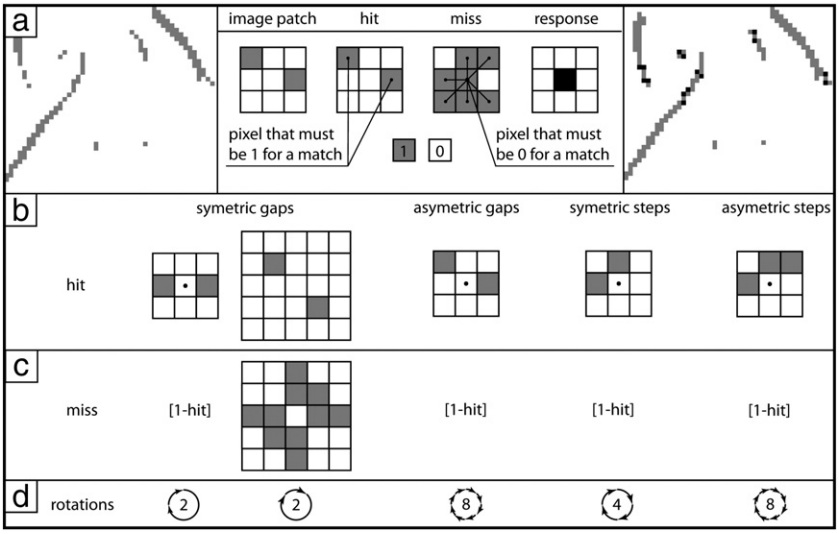


Fig. 6. Strategy used to connect broken line segments. (a) Working principle of the hit- and miss transform, (b) hit structures, (c) miss structures and (d) respective rotations used for the plausible pixel neighbourhoods.

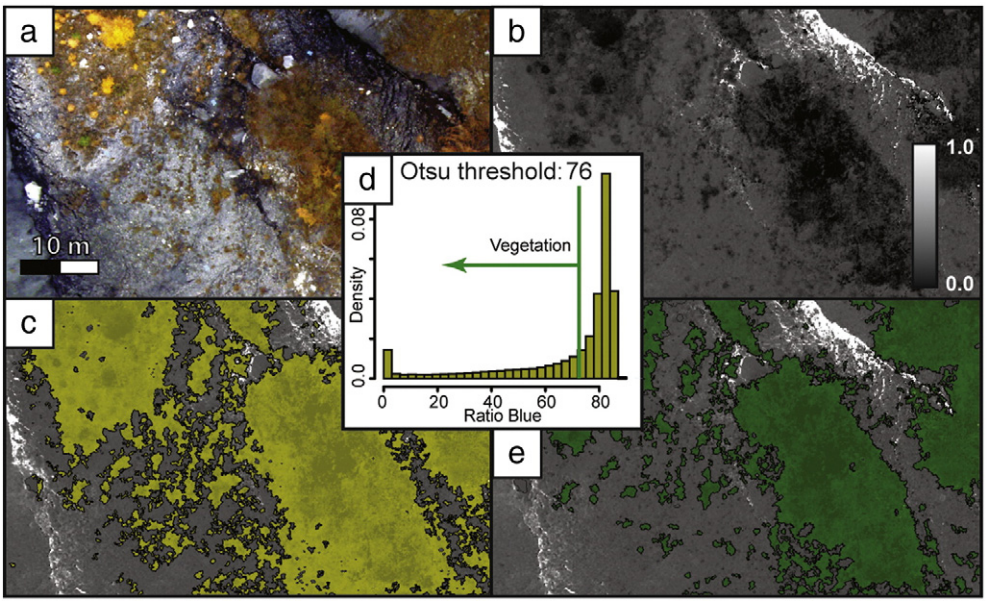


Fig. 7. Illustration of the automatic threshold detection for the intermediate mapping of vegetation. (a) Subset of the October 2008 image at the toe of the landslide. (b) Ratio blue. (c) Initial thresholding at ratio blue<0.33 to obtain vegetation candidates (yellow). (d) Histogram of the vegetation candidates with the automatically selected threshold. (e) Final map of the vegetation (green).

The connectivity of the line segments was also particularly important for the subsequent object-oriented post-processing, where objects constitute from pixels groups connected in a Von Neumann neighbourhood (four adjacent pixels at each side), and small isolated objects could be disregarded as noise.

4.3. Stage 3: object-oriented analysis for false positive removal

Due to visually similar objects, such as linear erosion features (rills and small gullies) and elongated shadows induced by the micro-topography, the fissure candidates resulting from the described filtering routine may still comprise numerous false positive detections. While a human interpreter can differentiate most of the false positives assessing the geospatial context of the scene, the efficient use of such information with automated systems is a challenge for object-oriented image analysis. To exploit the contextual scene information for an automatized refinement of the extracted fissure candidate maps, an object-oriented routine that integrates spatial reasoning into an explicit form was elaborated and implemented using eCognition 8.64 (Trimble, 2011). The routine included the following steps:

1) The ratio of shadow around the detections is evaluated and candidates with a ratio of shadow pixels in their smallest enclosing circle above 33% are regarded as false detections induced by shadings of the micro-topography. This ratio threshold was determined empirically through visual inspection of the candidate fissures, and selected to capture elongated false detections with one side lying fully in shaded zones. The threshold for shadow can thereby be adjusted according to the illumination conditions and the dynamic range of the image (Table 2).

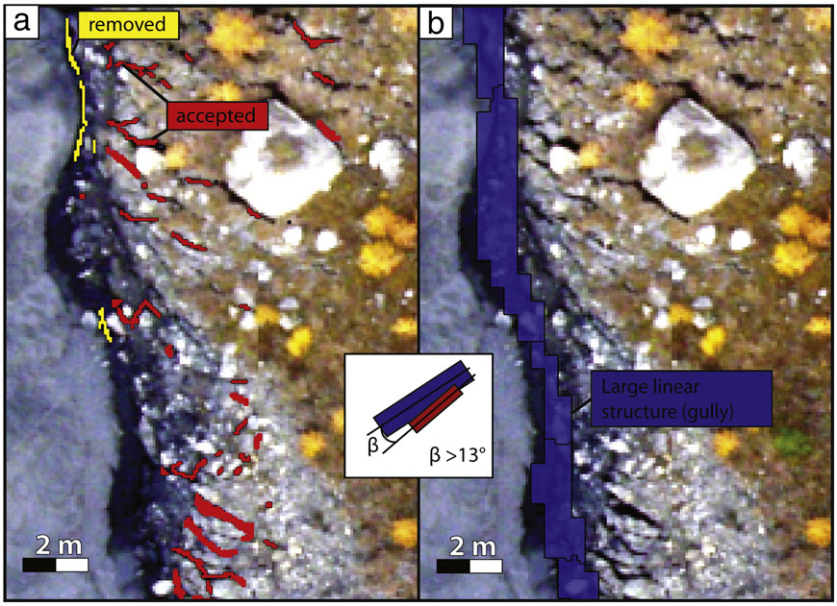


Fig. 8. Illustration of the object-oriented post-processing routine. (a) Fissure candidates that overlapped with linear structures. (b) Linear structure detected at a ten times greater filter scale. The fissure candidates aligned with the linear structures at angles below ±13° were removed.

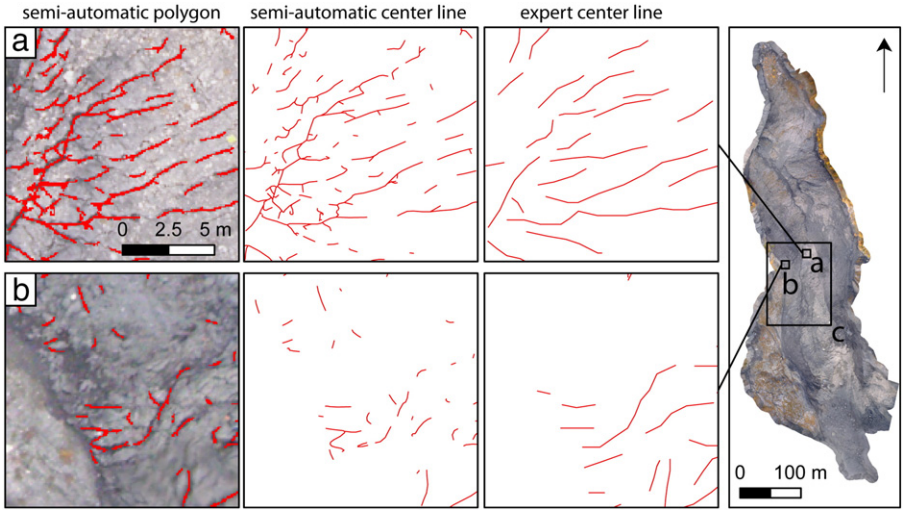


Fig. 9. Example of comparison of the obtained fissure maps with the expert mapping for October 2010 (fissures in red). (a) Area with relatively high agreement of the mapped fissure patterns. (b) Area with relatively high rate of false negatives and false positives. The scale of the representations corresponds approximately to the scale used for the expert mapping (1:250).

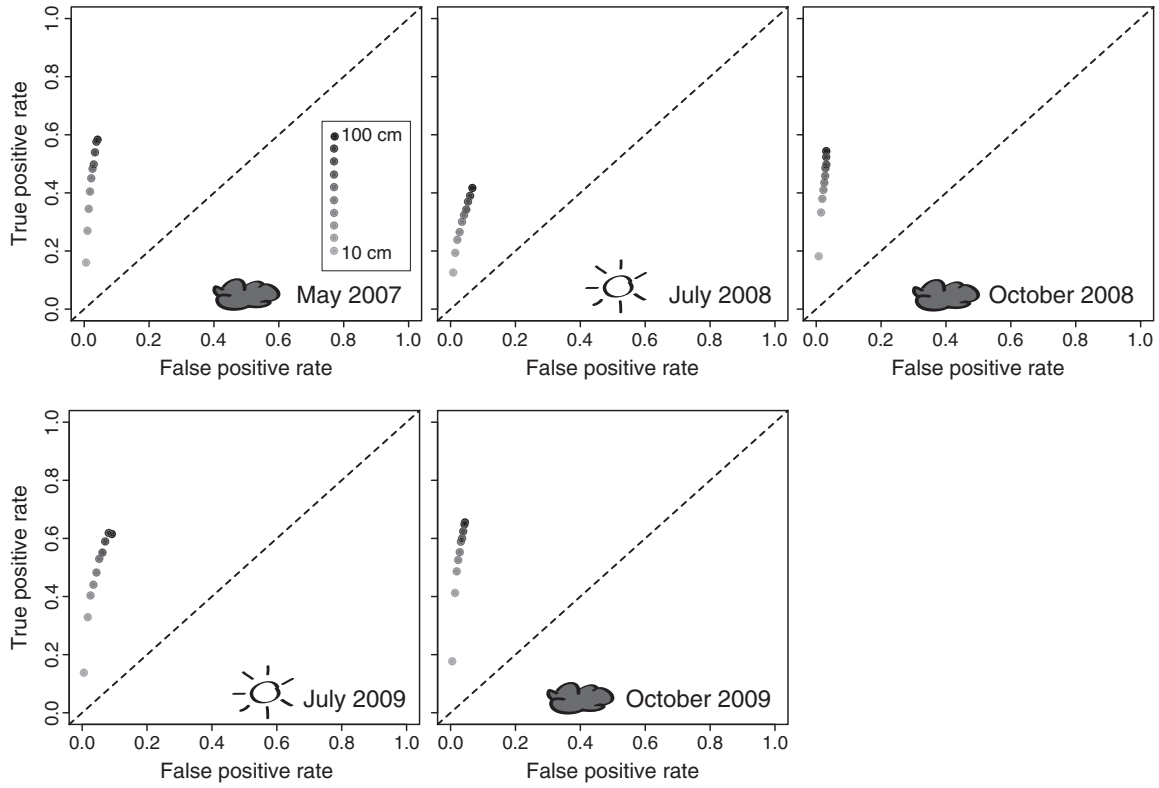
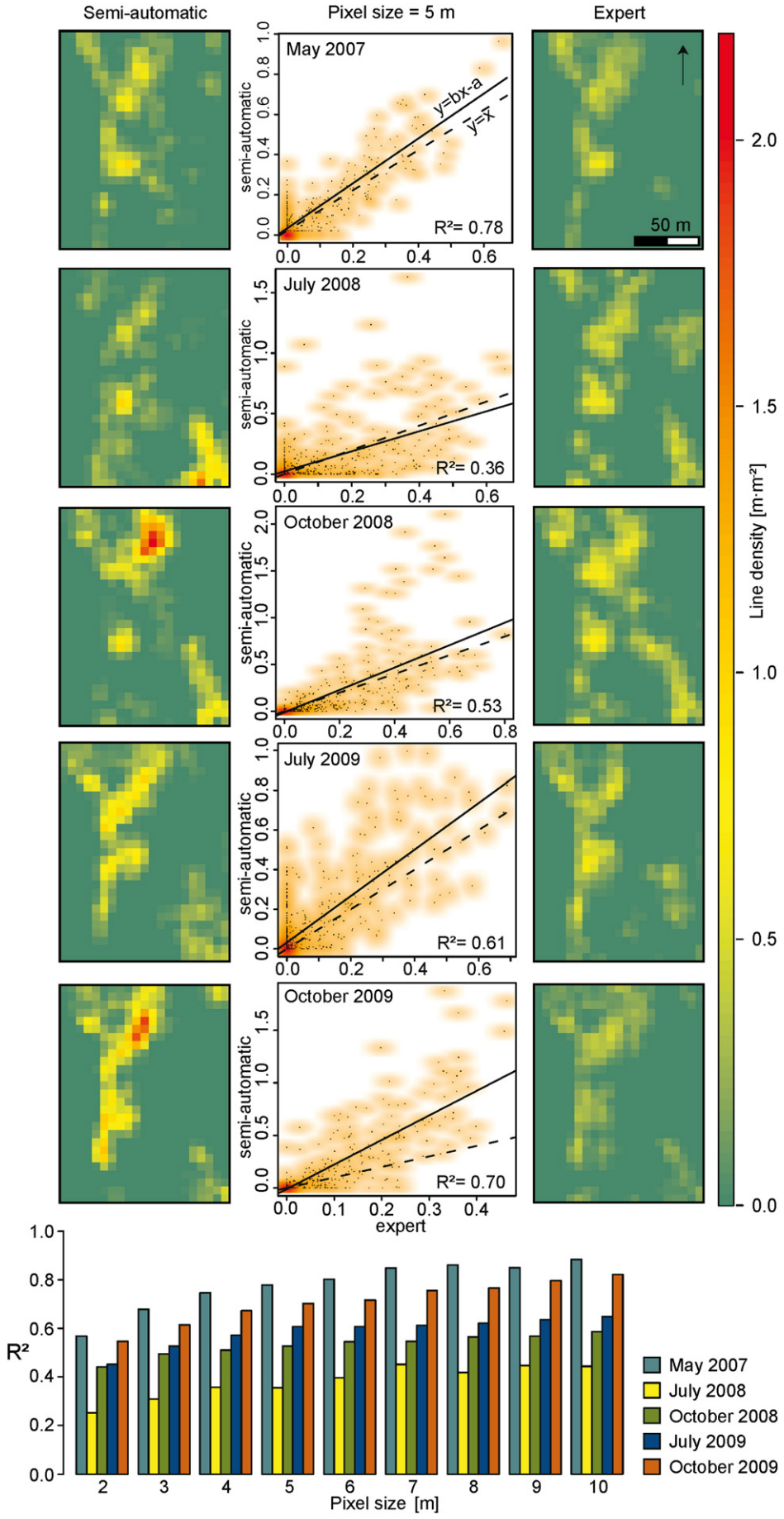


Fig. 10. Receiver operating characteristics (ROC) plots for the fissured area at different map resolutions. The sky conditions for the five different dates are indicated.

2) Further false detections may result from vegetation which typically shows a lower reflectance in the green and red channel compared to the blue. The blue ratio in the sum of all channels is consequently typically below one third for vegetated areas. The suitable value varies slightly with the illumination conditions and the season, and Otsu's method (Otsu, 1979) was employed to automatically adapt to such changes. Through an iterative testing of all possible values, Otsu's method determines threshold value that maximizes variance between two classes in an image. Hence, constraining the search space to all pixels with a ratio blue below 33%, the algorithm was used to determine the thresholds that maximize the contrast between vegetation and the background (Fig. 7). Fissure candidates covered by the resulting vegetation class, or having a relative border length larger than 0.15, were subsequently removed.
3) Another class of frequent false detections resulted from linear objects such as rills, gullies and nearly vertical steps at the landslide

Fig. 11. Correlation between fissure density estimates at 5 m raster resolution based on semi-automatic detections and expert mappings from the five images. The bar plots at the bottom display the R^2 coefficient at different raster resolutions.


Semi-automatic
Pixel size = 5 m
Expert
May 2007
y=bx-a
y=x
R²= 0.78
50 m
July 2008
R²= 0.36
October 2008
R²= 0.53
July 2009
R²= 0.61
October 2009
R²= 0.70
semi-automatic
expert
Line density [m·m⁻²]
R²
Pixel size [m]
May 2007
July 2008
October 2008
July 2009
October 2009

flanks, which may locally obtain similar characteristics as the targeted fissures. To test for the presence of larger linear features and evaluate their relationship with fissure candidates, a strategy to suppress additional false positives was required. For the mapping of the larger linear elements, two sources were adopted. First drainage lines were extracted from the LiDAR DTMs using hydrological standard tools (Tarboton et al., 1991) and enlarged with a surrounding buffer of 0.5 m. A second approach was to repeat the Gaussian filtering with the parameter set indicated in Table 1, but with a two times increased scale σ and a five times coarser image resolution (resampled with bilinear interpolation). This is equivalent to a search with a 10 times larger scale providing a sufficiently large scale difference to assure that none of the detected linear features would correspond to fissures. The linear objects extracted with both approaches were virtually overlaid with the fissure candidates, and the difference of the orientations of their respective centre lines was adopted as criteria to evaluate if the fissure candidate was in fact part of a larger linear object or constitutes an independent structure (Fig. 8). Image-based measurements of the angular offset of the fissured indicated a minimum offset of about $\pm 13°$. Considering that the lowest effective friction angle values measured for the landslide material are $\alpha' = 26°$ (Malet et al., 2005b), the thresholds are consistent with the orientation of $\alpha'/2$ that the Coulomb criterion predicts for the orientation of shear fissures at the landslide boundary (Tchalenko, 1970).

4) A last filtering step was implemented by removing all candidates with length not longer than 0.4 m and an area smaller than 0.1 m^2. Finally, all fissure candidates falling in areas with a fissure class density lower than 1% in a surrounding neighbourhood of 10 m^2 were considered as noise and also removed.

Table 2 displays that most adopted thresholds were kept the same among all the images and only the classification rule for the shadow areas was adapted in order to compensate radiometric differences in the input images.

5. Results and discussion

5.1. Comparison with multi-temporal manual mappings

The primary output of the developed processing routine is a map of the detected fissures represented by polygons. Applying a Delaunay triangulation that extracts the skeleton of those polygons (Trimble, 2011), a 2D line representation, which enables a more immediate comparison with expert mappings, can be obtained.

Fig. 9 displays an example of comparison between an expert map and the result of the semi-automatic detection. A first visual assessment of the obtained maps suggested better agreement of the fissure patterns in areas with high contrast and low texture (Fig. 9a), whereas false positives and false negatives concentrated in sections with low contrast and increased surface texture (Fig. 9b).

For a quantitative assessment of the mapping accuracy, the obtained results were compared with the expert mappings in the central part of the landslide (Fig. 9c) at all five dates. While several accuracy measures for geographic line datasets have been already proposed, there is still no consensus about one generally applicable technique and the metrics should be selected according to the problem at hand (Ariza-López and Mozas-Calvache, 2012). Here, we focus on three crucial aspects of the map accuracy that may have direct implications for their further use, namely the size of the affected (e.g. fissured) area, the length and density of the fissures, and their orientation.

5.1.1. Size of the fissured area

Tveite and Langaas (1999) suggested an accuracy measure for line datasets based on repeated buffering and overlay operations of detected and reference line datasets. A similar strategy was adopted in this study by repeatedly calculating true positive and false positives rates from two raster data representing the detections and the expert mapping at increasingly coarser resolutions. The raster data were calculated at 10 cm steps for resolutions between 0.1 and 1.0 m, and each pixel was assigned as fissured or non-fissured area according to the presence or absence of a fissure in the detections and the reference map, respectively. The resulting receiver operating characteristics (ROC) plots are presented (Fig. 10). The analysis showed a correspondence with the expert maps at true positive rates typically above 40% and up to 65%. The false positive rates were below 5% except for the scenes recorded with full sunlight where false positive rates up to 9% could be observed (Fig. 10).

5.1.2. Fissure length and density

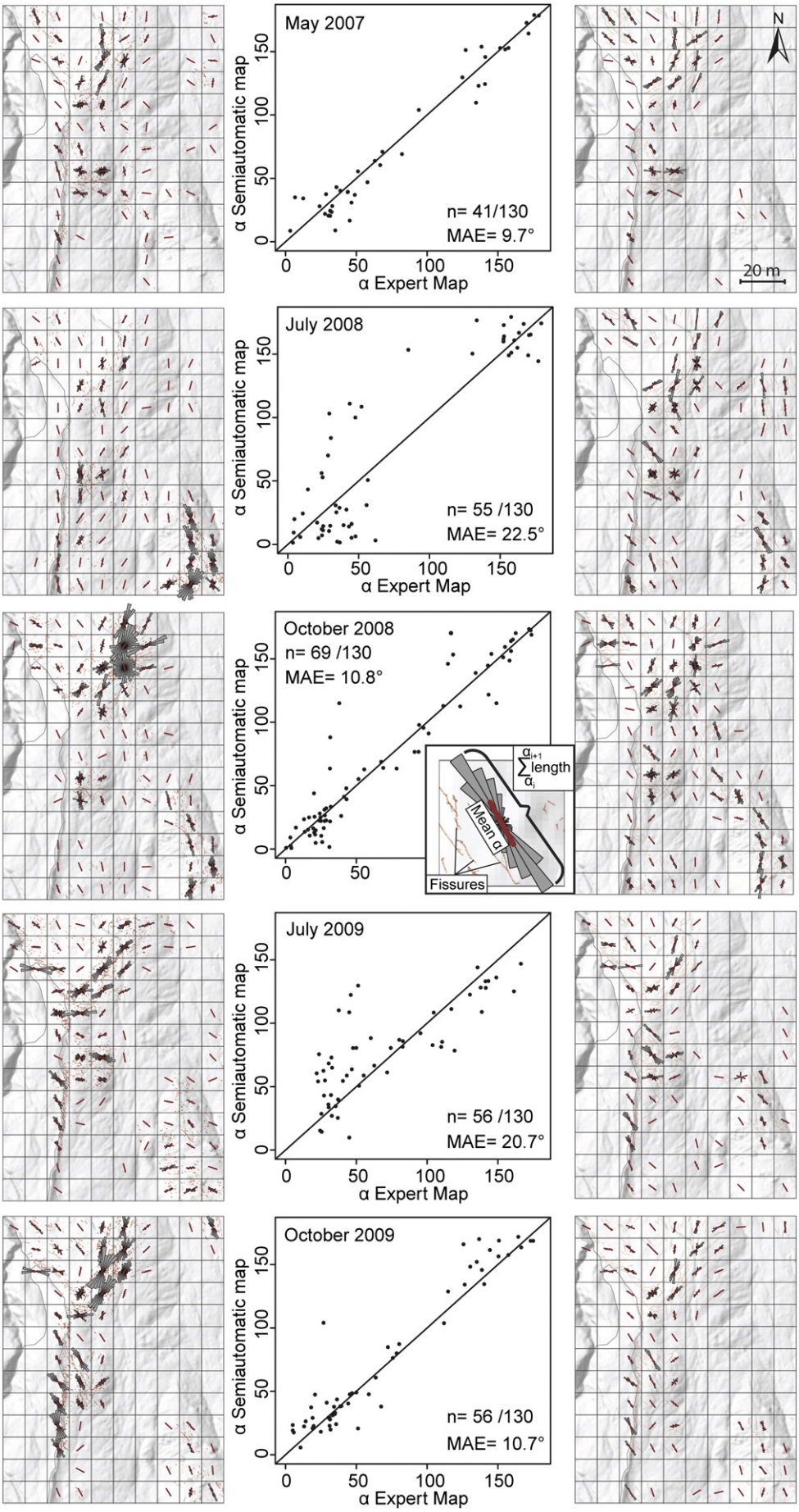
Hydrological models that integrate the influence of surface fissures on infiltration and preferential flow have demonstrated that the fraction of fissures per unit area is an important parameter with considerable influence on the modelled water storage (Malet et al., 2005a; Krzeminska et al., 2011). Such models are typically generated at slope scale with grid resolutions below 10 m. To assess the accuracy of the extracted maps with respect to this potential application, the fissure density was calculated as the line length in circular sliding windows with diameters between 2 and 10 m, and compared among automated detection and expert mappings.

The regression plots in Fig. 11 illustrate the correlation of the fissure density estimates with a 5 m circular sliding window yielding coefficient of determination (R^2) typically above 0.5. The regression analysis further indicated generally higher densities resulting from the semi-automatic detection originating from false positive detections but also from a stronger generalization of the fissure line drawings within the expert mapping. Exceptions from this general trend are the results obtained from the image of July 2008 which was recorded at a low sun incidence angle leading to a relatively low $R^2_{5m} = 0.36$. The bar plots in Fig. 11 display generally higher R^2 values at increasing resolutions of the density raster. This is a well-known effect of spatial aggregation on correlation statistics (Gotway and Young, 2002) but also reflects the contrast between stronger discrepancies of local details and a better correspondence of the global fissure pattern pictured in the respective maps. The highest correlation was observed among the mappings for May 2007 with $R^2_{10m} = 0.88$ indicating that the lower resolution of the corresponding input image was not an important factor for the accuracy of the detection.

5.1.3. Fissure orientation

As outlined in the Introduction section, different fissure patterns may signal respective mechanical processes, and statistics of the principal fracture orientation often allow us to estimate the directions of the principal stresses (Pollard and Fletcher, 2005). The fissure orientations were quantified as a third factor to assess the accuracy of the extracted maps using rose diagrams frequently employed for the analysis and interpretation of two dimensional orientation data (Jammalamadaka and SenGupta, 2001). Rose diagrams with a bin width of 10° were computed on a 10 m regular grid for the semi-automatic detections and the expert mappings at all five dates. Considering the length and direction of each bin expressed as a respective vector the preferred fissure orientation within a grid cell can be calculated by summing the vectors over all bins. Taking into account all cells containing fissures in both the

Fig. 12. Rose diagram plots with mean orientation (red line) and error statistics for the mean fissure orientation per 10 m grid cell for the test area at the five different dates. For visualization, the rose diagrams where plotted over a hillshade of the landslide surface and the scatterplot angles were centred at 90°.


May 2007
n= 41/130
MAE= 9.7°
July 2008
n= 55 /130
MAE= 22.5°
October 2008
n= 69 /130
MAE= 10.8°
Mean α
Fissures
July 2009
n= 56 /130
MAE= 20.7°
October 2009
n= 56 /130
MAE= 10.7°
α Semiautomatic map
α Expert Map
0
50
100
150
N
20 m

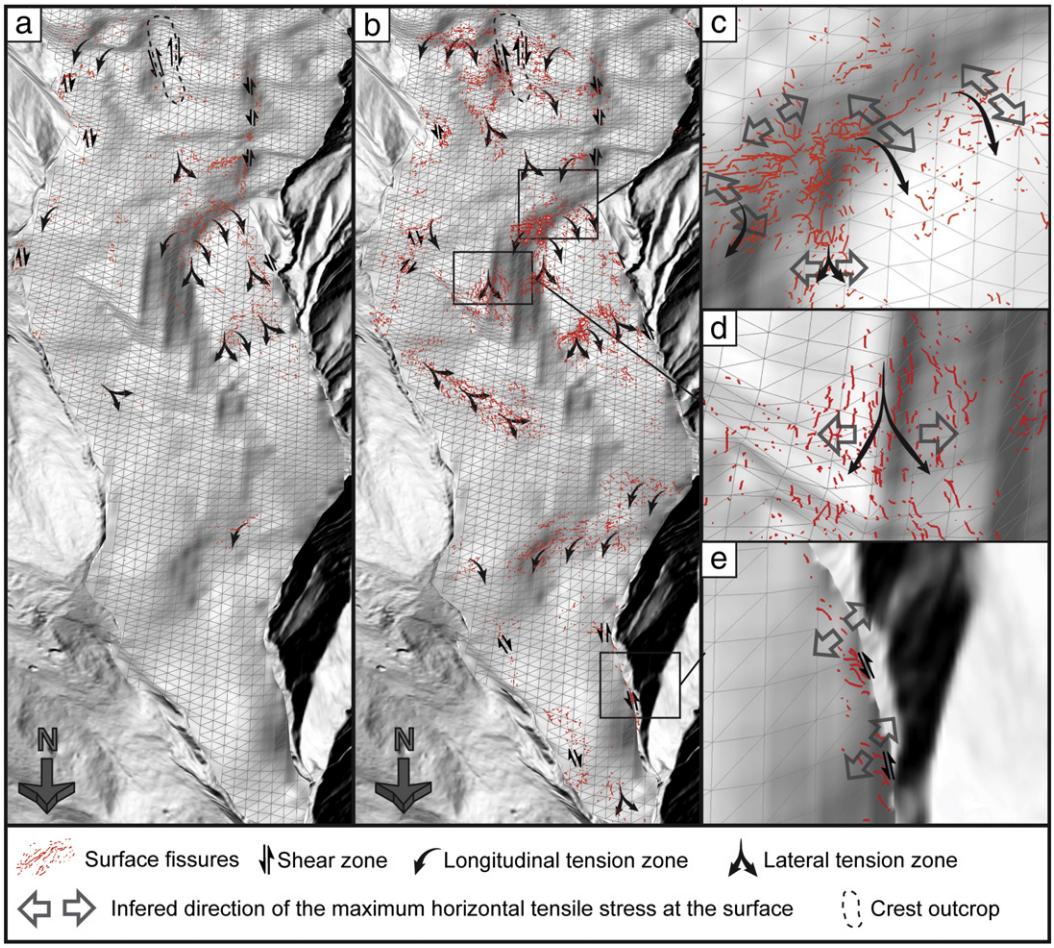


Fig. 13. Pseudo 3D view showing the landslide dynamics inferred from the fissure patterns detected in the aerial images of (a) May 2007 and (b) October 2008. (c, d, and e) Close up views for October 2008 showing inferred landslide dynamics and stress vectors. The results are overlaid on a hillshade model of the topography of the stable bedrock proposed by Travelletti and Malet (2012).

expert map and the semi-automatic detection, the mean absolute error (*MAE*) of the mean orientations provides a quantitative measure for the orientation accuracy.

The rose diagrams plots and error statistics in Fig. 12 depict *MAE* values between 9.7° and 22.5° for the five dates. The detections on the three scenes recorded under cloudy sky resulted in *MAE* not larger than 10.7°, whereas the error rate clearly exceeded 20° with the scenes of July 2008 and 2009 recorded with full sunlight at the surface. The lower orientation accuracies are largely consistent with the relatively low accuracies in terms of area (see Section 5.1.1) and density (see Section 5.1.2) resulting from the detection at the latter two dates.

5.2. Fissure patterns as possible geoindicators of deformation processes

For a comprehensive interpretation of the detected fissure patterns at the scale of the entire slope, the scenes of May 2007, October 2008 and July 2009 offering a full coverage of the landslide, have been analysed. However, considering the relatively low detection accuracy on the sunlit images of July 2009, the interpretation was focused essentially on the scenes of 2007 and 2008 spanning also over a period with displacement rates significantly above the average annual rates (Travelletti, 2011).

Comparing the detection results of May 2007 (Fig. 13a) and October 2008 (Fig. 13b), a significant increase in the abundance of fissures could be noted for the entire landslide. This can be attributed to a phase of strongly increased displacement rates (up to 3.5 m day^{-1}) in early June 2008 (Travelletti, 2011) preceding the UAV survey in October 2008. However, in October 2008, the displacement rates already consolidated again at average rates between 0.01 and 0.03 m day^{-1}, most of the fissures induced in June were preserved and evolved at the surface until October. This view is supported by the results obtained for the test area with the five scenes (Figs. 10 and 11) picturing rather a transient evolution than a complete reorganization of the fissure patterns. Despite partially strong disagreement in the absolute measured fissure density, both expert maps and semi-automatic mapping showed an increase in fissure density after May 2007, with higher values in October 2008 (Fig. 11) than directly after the peak displacement in spring. Pluviometric records for the area in 2008 show the relatively dry summer season with a cumulative rainfall of 110 mm for the month of July, August and September, suggesting that the increased fissure density in October is partially caused by an increased brittleness of the upper soil layer that dried out during summer.

Besides the general increase in the amount of fissures, it is intriguing to observe that at several local plots, similar fissure patterns can be observed at approximately the same positions through time (Figs. 12 and 13a,b), despite maximal displacements of up to 55 m between October 2008 and October 2009 (Niethammer et al., 2011a). This indicates the recurrent continuous in-situ formation where the fissures provide a close representation of the local strain field, similar as observed for the evolution of glacier crevasses (Harper et al., 1998).

Previous studies (Malet, 2003; Niethammer et al., 2011a; Walter et al., 2012) already observed close relationships between the occurrence of fissures and the geometry of the stable bedrock at the Super-Sauze landslide. They also noted a general contrast between higher water

content and rather ductile behaviour in the lower subsurface (<1 m) and typically lower water content of the topsoil yielding more brittle behaviour at the surface. The surface fissures can therefore be understood as the response to stresses induced in the topsoil through coupling with ductile strain in the deeper subsurface. A similar model was already described by Fleming and Johnson (1989) and adopted as a basis to qualitatively estimate the patterns of flow and stresses from a joint-interpretation (Fig. 13) of the detected fissure patterns and a geometrical model of the stable bedrock (Travelletti and Malet, 2012).

Considering the bedrock geometry and a formation of the open fissures normal to the direction of the least compressive stress (Pollard and Fletcher, 2005), three different flow field patterns leading to the fissure formation at the Super-Sauze landslide can be suggested. First, lateral shear at external and internal landslide boundaries aligned with the general flow field leads to the formation of diagonal shear fissure arrays (Fig. 13e). Second, longitudinal compressive and tensile strain related to abrupt changes in the slope of the sliding surface induces tensile stresses at the surface that results in transversal fissure arrays (Fig. 13c). Third, divergence of the flow field over topographic ridges and at the outlets of confining topographic channels induces lateral extension and tensile stresses resulting in longitudinal fissure arrays (Fig. 13d). At several locations, those processes overlap and lead to the formation of mixed structures such as a radial fissure patterns displayed in Figs. 4 and 13c, resulting from lateral shear and longitudinal strain, and from a divergent stress field, respectively.

5.3. Accuracy and related uses

Deformation patterns at the surface of landslides are important indicators for the mechanical processes, whereas the elaboration of detailed maps of such features remains a challenging and time-consuming task. While Sowers and Royster (1978) still argued that aerial photographs do not reach sufficient resolution for such mappings, modern digital sensors and new aerial platforms such as UAVs today provide the necessary level of detail. Furthermore, this study demonstrated the possible use of a semi-automatic image processing chain for the extraction of surface fissure maps.

The accuracy of the method was assessed by comparisons with expert maps and demonstrated heterogeneous areal accuracies with true positive rates of up to 65% and false positive rates generally below 10%. In addition, the orientation accuracy showed a variable quality of the resulting maps with mean deviations between 9.7° and 22.5°. The fissure densities derived from both maps have significant correlations ($R^2 = 0.36–0.78$), whereas the semi-automatic detections yield typically higher estimates. Interestingly, this difference is more pronounced with the images of 2009 (Fig. 10) reflecting the contrast between increased semi-automatic detection rates at higher resolutions and the fixed scale of the expert mapping. Contrariwise, the best agreement among detection and expert maps was measured for the scene of May 2007 showing that the lower resolution does not necessarily yield lower accuracies. Generally, lower accuracies were observed for the scenes recorded with full sunlight at the surface in July 2008 and 2009, and the worst results were obtained for July 2008 when images were recorded at a relatively low sun incidence angle. Since the direct sunlight induces shading that affects the local contrast and global image normalization methods cannot alleviate this problem, image acquisition with diffuse skylight appears to be the generally better option.

In the initial stage of the processing chain, a low-level linear feature detector is used. Similar techniques yield competitive results in medical image analysis (Zhang et al., 2010), whereas the accuracies achieved with aerial images in this study are still significantly lower. This must be attributed to the generally higher complexity of outdoor scenes and at the moment still requires additional steps and parameters to take the contextual scene information into account. The use of an OOA heuristic-based post-processing technique proved useful for the removal of false positives and helped to objectify the image analysis by transferring expert knowledge in an explicit form. The analysis still relies on a number of fixed thresholds which may hinder an easy transfer of the entire processing chain to a different geographic area. This concerns especially parameters that require knowledge of the local processes (e.g. minimum fissure length and effective friction angle), while thresholds that can be determined directly from the image (e.g. shadows and vegetation) may be adjusted more effortlessly.

The development of surface fissures precedes and accompanies especially slow- and very slow-moving landslides (Cruden and Varnes, 1996) making the developed technique particularly applicable to such types of landslides and to potentially unstable slopes. However, the spatial resolution of the acquired images must at least match, or should ideally exceed, the width of the targeted fissures, and the vegetation must be sufficiently sparse to permit direct view on the bare ground. The results of this study demonstrate that if those requirements are met, the obtained fissure maps can already provide sufficient accuracy to infer the landslide dynamics and mechanical processes at the slope scale (see Section 5.2). Density maps from both semi-automatic and expert mappings show a strong spatial and temporal variability of the fissure abundance pointing toward important local and temporal contrasts in the infiltration capacity which may need considerations in the design of hydro-mechanical models. An analysis of the evolution and mechanics of individual fissures will however require higher temporal resolution and terrestrial cameras have recently been installed at the surface of the landslide to record imagery for further research in this direction. It would also be desirable to test the developed technique for the investigation of other landslides with different characteristics in order to validate a more general applicability of the approach and the mechanical interpretation of the observed fissure dynamics.

The OOA heuristics already considers multi-scale information to some degree (see Section 4.3), whereas for further methodological improvements an explicit integration of an automatic scale selection technique at the low-level filtering stage appears as a promising approach to further reduce heuristics and tuneable parameters (Stumpf et al., 2012). The first and second stages of the proposed method are generic for the detection of dark linear features, and could in principle also be applied to detect other geomorphological and geological structures with similar characteristics. The proposed technique might be of interest for the mapping of gullies (Shruthi et al., 2011), geological lineaments (Mallast et al., 2011), ice-glacier crevasses (Vaughan, 1993) or tectonically induced fractures (Allmendinger and González, 2010), sufficiently larger to be depicted in sub-metre satellite images.

Considering the intrinsic disagreement in expert mappings of linear features, especially in the inter- and extrapolation of lines (Sander et al., 1997), further studies should also include an assessment of the uncertainties of reference maps since their quality can strongly bias the evaluation of different alternative approaches (Lampert et al., submitted for publication).

6. Conclusions

This study developed an image processing chain to extract surface fissures from heterogeneous sets of VHR aerial images and tested the approach with a challenging multi-temporal set of images recorded at the Super-Sauze landslide for five different dates. The first two stages of the developed workflow combine families of Gaussian matched filters and morphological filters, and are followed by an object-oriented analysis to reduce the amount of false positive detection using contextual information and auxiliary topographic information. The detection results can be represented in raster maps or optional by centre skeleton lines.

Under homogenous illumination conditions a comparison of the results with expert mapping demonstrated detection rates of up to 65% and orientation errors below 10°. Contrary, the technique is relatively sensitive to shading effects at full sunlight and prone to errors especially at low sun incidence angle. A joint-interpretation of obtained fissure

maps and of a 3D geometrical model of the stable bedrock demonstrated their complementary use for a better understanding of the geomorphological and geomechanical processes, such that the detected fissure pattern may be used for first approximation for mechanical processes in the recent deformation history of a slope. Possible directions for further research are the reduction of tuneable parameters and a more immediate exploitation of multi-scale information, as well as an adaption of the technique to other linear features with geomorphological and geological relevance.

Acknowledgements

The work was supported by the project SafeLand "Living with landslide risk in Europe: assessment, effects of global change, and risk management strategies" funded by the 7th Framework Programme of the European Commission (grant agreement no. 226479), the project SISCA 'Système Intégré de Surveillance de Crises de Glissements de Terrain' funded by the French Research Agency (contract ANR Risk-Nat, 2009–2012), the project FOSTER 'Spatio-temporal data mining: application to the understanding and monitoring of soil erosion' funded by the French Research Agency (contract ANR Cosinus, 2011–2013), and the project Grosshang "Coupling of Flow and Deformation Processes for Modeling the Movement of Natural Slopes" funded by the Deutsche Forschungsgemeinschaft (DFG). The authors would like to acknowledge Alexandre Mathieu (University of Strasbourg) for his assistance in the field and for the expert mapping of the different fissure patterns, and Julien Travelletti for discussion on the recent dynamics of the landslide. The IDL code implementing the first two detection stages and the eCognition rule set implementing the object-oriented analysis are available on demand.

References

Abramson, L.W., Lee, T.S., Sharma, S., Boyce, G.M., 2001. Slope Stability and Stabilization Methods, 2nd edition. Wiley, Chichester.
Allmendinger, R.W., González, G., 2010. Invited review paper: Neogene to Quaternary tectonics of the coastal Cordillera, northern Chile. Tectonophysics 495, 93–110.
Anderson, T.L., 2005. Fracture Mechanics: Fundamentals and Applications, 3rd edition. Taylor & Francis.
Ariza-López, F., Mozas-Calvache, A., 2012. Comparison of four line-based positional assessment methods by means of synthetic data. GeoInformatica 16, 221–243.
Backers, T., 2004. Fracture Toughness Determination and Micromechanics of Rock Under Mode I and Mode II Loading. PhD, Universität Potsdam, Potsdam.
Baruch, A., Filin, S., 2011. Detection of gullies in roughly textured terrain using airborne laser scanning data. ISPRS Journal of Photogrammetry and Remote Sensing 66, 564–578.
Bièvre, G., Jongmans, D., Winiarski, T., Zumbo, V., 2012. Application of geophysical measurements for assessing the role of fissures in water infiltration within a clay landslide (Trièves area, French Alps). Hydrological Processes 26, 2128–2142.
Bombard, J.-P., 1968. Une approche des problèmes posés par l'étude des mouvements de terrains. Thèse de Géologie, Faculté des Sciences de Grenoble, Grenoble, (in French).
Canny, J., 1986. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 8, 679–698.
Canty, M.J., Nielsen, A.N., 2008. Automatic radiometric normalization of multitemporal satellite imagery with the iteratively re-weighted MAD transformation. Remote Sensing of Environment 112, 1025–1036.
Chaudhuri, S., Chatterjee, S., Katz, N., Nelson, M., Goldbaum, M., 1989. Detection of blood vessels in retinal images using two-dimensional matched filters. IEEE Transactions on Medical Imaging 8, 263–269.
Chowdhury, R.N., Zhang, S., 1991. Tension cracks and slope failure. In: Chandler, R.J. (Ed.), Proceedings of the International Conference on Slope Stability Engineering: Developments and Applications. Institution of Civil Engineers, Isle of Wight, pp. 27–32.
Cruden, D.M., Varnes, D.J., 1996. Landslides types and processes. In: Turner, A.K., Schuster, R.L. (Eds.), Landslides: Investigation and Mitigation. Transportation Research Board, Special Report, 247. National Academy of Sciences, Washington, DC, pp. 36–75.
Davis, G.H., Reynolds, S.J., 1996. Structural Geology of Rocks and Regions. Wiley, Chichester.
Eisenbeiss, H., 2009. UAV Photogrammetry. PhD, ETH Zürich, Zürich.
Espinosa, A., 2009. Analysis and Quantification of Preferential Flow on the Super-Sauze Landslide. MSc, Delft University of Technology, Delft, Netherlands. (80 pp.).
Fleming, R.W., Johnson, A.M., 1989. Structures associated with strike–slip faults that bound landslide elements. Engineering Geology 27, 39–114.
Fleming, R.W., Baum, R.L., Giardino, M., 1999. Map and description of the active part of the slumgullion landslide, Hinsdale County, Colorado. Geologic Invstigations Series Map I-2672. U.S. Geological Survey, Denver, Colorado.
Frost & Sullivan Co., 2007. Study analysing the current activities in the field of UAV. European Comission Enterprise and Industry Directorate-General.
Glenn, N.F., Streutker, D.R., Chadwick, D.J., Thackray, G.D., Dorsch, S.J., 2006. Analysis of LiDAR-derived topographic information for characterizing and differentiating landslide morphology and activity. Geomorphology 73, 131–148.
González, R.C., Woods, R.E., 2008. Digital Image Processing, 3rd edition. Prentice Hall, Upper Saddle River.
Gotway, C.A., Young, L.J., 2002. Combining incompatible spatial data. Journal of the American Statistical Association 97, 632–648.
Graham, D.J., Rollet, A.-J., Piégay, H., Rice, S.P., 2010. Maximizing the accuracy of image-based surface sediment sampling techniques. Water Resources Research 46 http://dx.doi.org/10.1029/2008WR006940.
Grandjean, G., Bitri, A., Krzeminska, D.M., 2011. Characterisation of a landslide fissure pattern by integrating seismic azimuth tomography and geotechnical testing. Hydrological Processes 26, 2120–2127.
Günther, A., Carstensen, A., Pohl, W., 2004. Automated sliding susceptibility mapping of rock slopes. Natural Hazards and Earth System Sciences 4, 95–102.
Harper, J.T., Humphrey, N.F., Pfeffer, W.T., 1998. Crevasse patterns and the strain-rate tensor: a high-resolution comparison. Journal of Glaciology 44, 68–76.
Hoek, E., Bray, J.W., 1981. Rock Slope Engineering. The Institution of Mining and Metallurgy, London.
Hong, G., Zhang, Y., 2008. A comparative study on radiometric normalization using high resolution satellite images. International Journal of Remote Sensing 29, 425–438.
Hoover, A.D., Kouznetsova, V., Goldbaum, M., 2000. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Transactions on Medical Imaging 19, 203–210.
Jaboyedoff, M., Baillifard, F., Couture, R., Locat, J., Locat, P., 2004. Toward preliminary hazard assessment using DEM topographic analysis and simple mechanical modeling by means of sloping local base level. In: Lacerda, W.A., Ehrlich, M., Fontoura, A.B., Sayão, A. (Eds.), Landslides: Evaluation and Stabilization. Taylor & Francis Group, London, pp. 199–205.
Jammalamadaka, S.R., SenGupta, A., 2001. Topics in Circular Statistics. World Scientific, Singapore.
Ke, C.C., Chen, C.S., Tu, C.H., 2008. Determination of fracture toughness of anisotropic rocks by boundary element method. Rock Mechanics and Rock Engineering 41, 509–538.
Khattak, G.A., Owen, L.A., Kamp, U., Harp, E.L., 2010. Evolution of earthquake-triggered landslides in the Kashmir Himalaya, northern Pakistan. Geomorphology 115, 102–108.
Krauskopf, K.B., Feitler, S., Griggs, A.B., 1939. Structural features of a landslide near Gilroy, California. Journal of Geology 47, 630–648.
Krzeminska, D.M., Bogaard, T.A., van Asch, T.W.J., van Beek, L.P.H., 2011. A conceptual model of the hydrological influence of fissures on landslide activity. Hydrology and Earth System Sciences 8, 11039–11073.
Lampert, T.A., O'Keefe, S.E.M., 2011. A detailed investigation into low-level feature detection in spectrogram images. Pattern Recognition 44, 2076–2092.
Lampert, T.A., Stumpf, A., Gancarski, P., submitted for publication. An empirical study into expert agreement and ground truth estimation. IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA.
Malet, J.P., 2003. Les 'glissements de type écoulement' dans les marnes noires des Alpes du Sud. Morphologie, fonctionnement et modélisation hydro-mécanique. PhD, Université Louis Pasteur, Strasbourg (in French).
Malet, J.-P., Auzet, A.-V., Maquaire, O., Ambroise, B., Descroix, L., Esteves, M., Vandervaere, J.-P., Truchet, E., 2003. Soil surface characteristics influence on infiltration in black marls: application to the Super-Sauze earth flow (southern Alps, France). Earth Surface Processes and Landforms 28, 547–564.
Malet, J.P., van Asch, T.W.J., van Beek, R., Maquaire, O., 2005a. Forecasting the behaviour of complex landslides with a spatially distributed hydrological model. Natural Hazards and Earth System Sciences 5, 71–85.
Malet, J.P., Laigle, D., Remaître, A., Maquaire, O., 2005b. Triggering conditions and mobility of debris flows associated to complex earthflows. Geomorphology 66, 215–235.
Mallast, U., Gloaguen, R., Geyer, S., Rödiger, T., Siebert, C., 2011. Semi-automatic extraction of lineaments from remote sensing data and the derivation of groundwater flowpaths. Hydrology and Earth System Sciences Discussions 8, 1399–1431.
Marr, D., Hildreth, E., 1980. Theory of edge detection. Proceedings of the Royal Society of London, Series B: Biological Sciences 207, 187–217.
Martha, T., Kerle, N., van Westen, C.J., Kumar, K., 2010. Characterising spectral, spatial and morphometric properties of landslides for semi-automatic detection using object-oriented methods. Geomorphology 116, 24–36.
Matheson, G.D., 1983. Rock Stability Assessment in Preliminary Site Investigations — Graphical Methods. Transport and Road Research Laboratory, Crownthorne.
McCalpin, J., 1984. Preliminary age classification of landslides for inventory mapping. 21st Annual Engineering Geology and Soils Engineering Symposium, Moscow, Idaho, pp. 99–111.
Meisina, C., 2006. Characterisation of weathered clayey soils responsible for shallow landslides. Natural Hazards and Earth System Sciences 6, 825–838.
Melton, F.A., 1959. Aerial photographs and structural geomorphology. Journal of Geology 67, 351–370.
Mendonca, A.M., Campilho, A., 2006. Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction. IEEE Transactions on Medical Imaging 25, 1200–1213.
Metzger, W., 1975. Gesetze des Sehens, 3. Auflage. Waldemar Kramer, Frankfurt am Main.
Niethammer, U., Rothmund, S., James, M.R., Travelletti, J., Joswig, M., 2010. UAV-based remote sensing of landslides. ISPRS Commission V Mid Term Symposium XXXVIII, Part 5, pp. 496–501.
Niethammer, U., James, M.R., Rothmund, S., Travelletti, J., Joswig, M., 2011a. UAV-based remote sensing of the Super-Sauze landslide: evaluation and results. Engineering Geology 128, 2–11.

Niethammer, U., Rothmund, S., Schwaderer, U., Zeman, J., Joswig, M., 2011b. Open source image-processing tools for low cost UAV-based landslide investigations. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXVIII-1/C22.
Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics 9, 62–66.
Papari, G., Petkov, N., 2011. Edge and line oriented contour detection: state of the art. Image and Vision Computing 29, 79–103.
Parise, M., 2003. Observation of surface features on an active landslide, and implications for understanding its history of movement. Natural Hazards and Earth System Sciences 3, 569–580.
Passchier, C.W., Trouw, R.A.J., 2005. Microtectonics. Springer, Berlin Heidelberg.
Pollard, D.D., Fletcher, R.C., 2005. Fundamentals of Structural Geology. Cambridge University Press, New York. Third printing with corrections.
Prats, X., Ramírez, J., Delgado, L., Royo, P., 2012. Regulations and requirements. In: Angelov, P. (Ed.), In Sense and Avoid in UAS: Research and Applications. Wiley, Chichester, pp. 87–117.
Priest, S.D., 1993. Discontinuity Analysis for Rock Engineering. Chapman & Hall, London.
Quackenbush, L.J., 2004. A review of techniques for extracting linear features from imagery. Photogrammetric Engineering and Remote Sensing 70, 1383–1392.
Riedel, W., 1929. Zur Mechanik geologischer Brucherscheinungen. Zentralblatt für Mineralogie, Geologie und Paläontologie B 354–368.
Sander, P., Minor, T.B., Chesley, M.M., 1997. Ground-water exploration based on lineament analysis and reproducibility tests. Ground Water 35, 888–894.
Schott, J.R., Salvaggio, C., Volchok, W.J., 1988. Radiometric scene normalization using pseudoinvariant features. Remote Sensing of Environment 26, 1–16.
Schulson, E.M., Duval, P., 2009. Creep and Fracture of Ice. Cambridge University Press, New York.
Selby, M.J., 1993. Hillslope Materials and Processes, 2nd edition. Oxford University Press, Oxford.
Serra, J., 1982. Image Analysis and Mathematical Morpholgy. Academic Press, Orlando.
Shao, Y., Guo, B., Hu, X., Di, L., 2011. Application of a fast linear feature detector to road extraction from remotely sensed imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 4, 626–631.
Shreve, R.L., 1966. Sherman landslide, Alaska. Science 154, 1639–1643.
Shruthi, R.B.V., Kerle, N., Jetten, V., 2011. Object-based gully feature extraction using high spatial resolution imagery. Geomorphology 134, 260–268.
Soares, J.V.B., Leandro, J.J.G., Cesar, R.M., Jelinek, H.F., Cree, M.J., 2006. Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification. IEEE Transactions on Medical Imaging 25, 1214–1222.
Sofka, M., Stewart, C.V., 2006. Retinal vessel centerline extraction using multiscale matched filters, confidence and edge measures. IEEE Transactions on Medical Imaging 25, 1531–1546.
Sowers, G.F., Royster, D.L., 1978. Field investigation. In: Schuster, R.L., Krizek, R.J. (Eds.), Landslides: Analysis and Control. Transportation Research Board, Special Report, 176. National Academy of Science, Washington, DC, pp. 81–111.
Stumpf, A., Kerle, N., 2011. Object-oriented mapping of landslides using Random Forests. Remote Sensing of Environment 115, 2564–2577.
Stumpf, A., Lampert, T.A., Malet, J.-P., Kerle, N., 2012. Multi-scale line detection for landslide fissure mapping. International Geoscience and Remote Sensing Symposium, Munich, Germany, p. 4.
Tarboton, D.G., Bras, R.L., Rodriguez-Iturbe, I., 1991. On the extraction of channel networks from digital elevation data. Hydrological Processes 5, 81–100.
Tchalenko, J.S., 1970. Similarities between shear zones of different magnitudes. Geological Society of America Bulletin 81, 1625–1640.
Ter-Stephanian, G., 1946. On the landslide cracks classification. Bulletin of the Academy of Sciences of the Armenian SSR 10, 65–80 (in Russian with abstracts in Armenian and English).
Travelletti, J., 2011. Imagerie multi-paramètres et multi-résolutions pour l'observation et la caractérisation des mécanismes de glissements-coulées. PhD, Université de Caen Basse-Normandie, Strasbourg.
Travelletti, J., Malet, J.P., 2012. Characterization of the 3D geometry of flow-like landslides: a methodology based on the integration of heterogeneous multi-source data. Engineering Geology 128, 30–48.
Trimble, 2011. eCognition Developer 8.64.0. Reference Book. Trimble Documentation, München.
Tveite, H., Langaas, S., 1999. An accuracy assessment method for geographical line data sets based on buffering. International Journal of Geographical Information Science 13, 27–47.
Vallet, J., 2007. GPS/IMU and LiDAR integration to aerial photogrammetry: development and practical experiences with Helimap System®, 27. Wissenschaftlich-Technische Jahrestagung der DGPF, Dreiländertagung, Muttenz.
van Asch, T.W.J., van Beek, L.P.H., Bogaard, T.A., 2009. The diversity in hydrological triggering systems of landslides. The First Italian Workshop on Landslides, Napoli, Italy, pp. 151–156.
Vaughan, D.G., 1993. Relating the occurrence of crevasses to surface strain rates. Journal of Glaciology 39, 255–266.
Vicente-Serrano, S.M., Pérez-Cabello, F., Lasanta, T., 2008. Assessment of radiometric correction techniques in analyzing vegetation variability and change using time series of Landsat images. Remote Sensing of Environment 112, 3916–3934.
Walter, M., Niethammer, U., Rothmund, S., Joswig, M., 2009. Joint analysis of the Super-Sauze (French Alps) mudslide by nanoseismic monitoring and UAV-based remote sensing. First Break 27, 53–60.
Walter, M., Arnhardt, C., Joswig, M., 2012. Seismic monitoring of rockfalls, slide quakes, and fissure development at the Super-Sauze mudslide, French Alps. Engineering Geology 128, 12–22.
Wang, E.Z., Shrive, N.G., 1995. Brittle fracture in compression: mechanisms, models and criteria. Engineering Fracture Mechanics 52, 1107–1126.
Watts, A.C., Ambrosia, V.G., Hinkley, E.A., 2012. Unmanned aircraft systems in remote sensing and scientific research: classification and considerations of use. Remote Sensing 4, 1671–1692.
Youssef, A., Maerz, N., Hassan, A., 2009. Remote sensing applications to geological problems in Egypt: case study, slope instability investigation, Sharm El-Sheikh/Ras-Nasrani area, southern Sinai. Landslides 6, 353–360.
Zhang, B., Zhang, L., Zhang, L., Karray, F., 2010. Retinal vessel extraction by matched filter with first-order derivative of Gaussian. Computers in Biology and Medicine 40, 438–445.

3.4 *Measuring landslide horizontal surface displacements from VHR optical satellite imagery*

This section is based on: Stumpf, A., Malet, J.-P., Allemand, P., Submitted. Stereo-photogrammetry and displacement monitoring with Pleiades VHR satellite images, ISPRS Journal of Photogrammetry and Remote Sensing, 33p.

This section describes the elaboration of a processing chain combining spaceborne stereo-photogrammetry, orthorectification and sub-pixel image correlation to measure the horizontal surface displacement of three different landslides from VHR satellite images. To the best of our knowledge, it is the first study that explores the use of VHR satellite images for displacement measurements without external topographic information and ground control. The influence of the number of ground-control points on the accuracy of the image orientation, the extracted DSM and displacement rates is quantified with independent measurements. The accuracy of the generated stereo-photogrammetric DSM is assessed through comparison with ALS surveys and shows RMSEs between 1.34 m and 15.2 depending on the type of land cover and the topographic position. Surface displacements are measured through sub-pixel image correlation and the observed rates are compared with *in-situ* permanent GNSS measurements. The comparison shows a maximum error of 0.13 m which is one order of magnitude more accurate than what has been previously achieved with spaceborne optical images. It is also demonstrated that the technique can be applied without ground control points and no significant loss in accuracy which greatly facilitates its use for a broad range of applications.

3.4.1 Introduction

The monitoring of earth surface deformation is indispensable for the understanding of tectonic and geomorphological processes and the assessment of associated hazards. In particular, continuously active landslides are a major natural hazard threatening infrastructure and human settlements. Active landslides are controlled mainly by hydro-meteorological triggers and their behaviour is, therefore, sensitive to short- and long-term environmental changes.

During the last two decades, remote sensing has become an important tool to investigate landslide kinematics measuring 1D Line-Of-Sight (LOS) and 2D horizontal surface displacements (Delacourt et al., 2007; Stumpf et al., In Review-b). SAR interferometry has proven its ability to provide highly precise motion measurements (Ferretti et al., 2001; Berardino et al., 2002; Hooper, 2008) but yields only LOS displacement vectors with limited spatial coverage and range of measurable velocities. Image correlation of optical satellite and aerial images is often used to infer the 2D horizontal component of co-seismic and tectonic deformation (Leprince et al., 2007), glacier flow (Kääb, 2002; Heid and Kääb, 2012), landslides (Delacourt et al., 2007; Booth et al., 2013) and other earth surface processes (Vermeesch and Drake, 2008; de Michele et al., 2012). Theoretically DIC yields sub-pixel accuracy but image orientation, co-registration, georeferencing, decorrelation and especially the accurate modelling of topographic distortions are still challenging issues that have to be carefully addressed (Berthier et al., 2005; Scherler et al., 2008).

So far most studies have made use of medium and high-resolution satellite images and reported uncertainties (RMS or standard deviations) in the measured displacements of 1-4 m with Landsat and ASTER (Leprince et al., 2007; Scherler et al., 2008; Heid and Kääb, 2012; Redpath et al., 2013) and 0.3 m – 1 m with SPOT images (Van Puymbroeck et al., 2000; Michel and Avouac, 2002; Berthier et al., 2005; Binet and Bollinger, 2005; Taylor et al., 2008). The proposed methods are well adapted to measure displacements which are greater than one meter and coherent over large areas (e.g. glacier flow, coseismic slip).

Landslides, however, feature displacement fields with strong variability in space and time requiring observations with high temporal and spatial resolution. So far the use of spaceborne- and airborne DIC for landslide investigations has been constrained to historical reconstructions of cumulative displacement for time-intervals of several years limiting its operational use for monitoring and other applications. In Delacourt et al. (2004) and Casson et al. (2005) aerial stereo-pairs were exploited the remove the topographic component of the observed shifts. The analysis permitted to reconstruct the long-term dynamics of the La Clapière landslide but comprised the need of distributed ground-control points (GCPs) and errors of up to 2 m.

The latest generation optical satellites features shorter repeat-pass cycles and higher spatial resolutions (e.g. Pleiades, Spot 6-7, Geoeye-1, WorldView-2). It consequently provides enhanced capabilities for the acquisition of monoscopic and stereo images at shorter time intervals. Since both multi-temporal images and stereo-pairs can be derived from the same system stereo-photogrammetry, orthorectification and DIC could be applied using satellite images only, without any external topographic information or ground control.

This study investigates the use of VHR satellite images for measurements of landslide surface displacement with a particular focus on the impact of minimal or missing ground control. Pleiades satellite images acquired over the Ubaye valley (Southern French Alps) are analysed to measure the surface displacement of three active landslides over a period of two month. A processing chain comprising bundle adjustment, stereo-photogrammetric extraction of digital surface models (DSMs), and sub-pixel DIC is proposed and the accuracy of the

extracted motion fields is compared with permanent GNSS stations. A series of experiments is carried out to assess the impact of minimal or no ground control, and to quantify the accuracy of the sensor orientation and the extracted DSMs. To the best of our knowledge the study is the first to explore the use of Pleaides satellite images for displacement measurements and the first application of VHR satellite images only for the observation of landslide deformation.

In section 3.4.2, the adopted processing chain and analysed datasets are presented, the results are discussed in section 3.4.3 and conclusions are drawn in section 3.4.4.

3.4.2 Data and Methods

A prerequisite for accurate displacement measurements is the precise co-registration of multi-temporal images and the removal of the parallax resulting when the same topographic surface is observed from different view angles. This involves the compensation of biases in the original rational polynomial function (RPF) sensor model (section 3.4.2.1) and the orthorectification of the images with an accurate DSM (section 3.4.2.2). The method employed for sub-pixel correlation is explained in section in section 3.4.2.3 and an approach for the removal of noise and clutter based on the multi-spectral bands is detail in section 3.4.2.4. The accuracy assessment and datasets processed in this study are described in the sections 3.4.2.5 and 3.4.2.6., respectively.

3.4.2.1 RPF bias compensation through bundle block adjustment

During the last decade, RPFs have been widely accepted as a generic an accurate alternative to rigorous sensor models. Unlike rigorous sensor models, RPFs do not describe the interior and exterior orientation of the on-board camera but model the relationship between image coordinates and ground-coordinates with a rational polynomial that typically comprises 80 coefficients. The interior orientation of cameras used in VHR satellites can be considered as stable, whereas ground-positional errors result from uncertainties about exterior orientation parameters (pitch, raw, yaw, ephemeris) and a possible drift of those parameters over time (Grodecki and Dial, 2003). Depending on the satellite RPFs of modern VHR sensors yield geolocation accuracies in the range of 10-20 m (Hoja et al., 2008; Lussy et al., 2012) which is clearly not sufficient for a direct application of stereo-photogrammetry and sub-pixel image correlation. However, the errors can be greatly reduced using tie points and ground control points (GCPs) to estimate correction parameters. Several studies have demonstrated that least-square bundle-adjustment can yield geolocation accuracies of one meter and better (Grodecki and Dial, 2003; Fraser and Hanley, 2005).

Several commercial software solutions with end-to-end modules for satellite stereo-photogrammetry and bundle-adjustment are available and in this study the Leica Photogrammetric Suite (LPS, Intergraph, 2013) was employed for this purpose. All images (3 panchromatic, and 1 multispectral) were brought in to one single block and automatic tie point extraction was performed with an least-square image matching technique based on principles described in Gruen (1985). We used a 7x7 pixel correlation window and the minimum correlation coefficient was set 0.8. This resulted in a total number of 126 tie points. The residual errors of the tie points in the image space were checked and all points with residuals higher than 0.5 pixels were removed (~18% of the total number of points). Bias correction was performed through RPF bundle adjustment using an iteratively re-weighted least-square approach. The least-square approach allows defining prior weights for the observations (GCPs, and tie points). The RMSE of the GCPs was estimated with at 0.21 m (see section 3.4.2.6) and the uncertainty of the measurements in image space at 0.33 pixels.

The maximum number of least-square iterations was set to ten but convergence was usually reached after only three iterations.

The bias compensation can be performed with different error models being either a constant shift or two terms for shift and drift (Fraser and Hanley, 2005). Both error models were tested with a varying number of GCPs to assess the impact on DSM accuracy and displacement measurement. At least, 15 independent check points were reserved in all experiments to evaluate the residual errors after bundle adjustment.

3.4.2.2 DSM extraction and orthorectification

The quality of the RPF refinement can be judged by the residual errors of the tie and check points locations. For a precise description of the image-to image correspondence the residual error should generally not exceed 1 pixel. The optimal strategy for the subsequent extraction of the DSM generally depends on the characteristics of the terrain. We employed a hierarchical least square image matching algorithm (e.g. Zhang and Gruen, 2004) combined with techniques for the removal of outliers through principal component analysis (Xu et al., 2008). Area-based matching is performed in the image geometry considering epipolar constraints imposed by the refined sensor models.

In epipolar geometry, the topographic height relates directly to parallax shift in the x-direction and hence the search for correspondences can be reduced to a 1D search along the epipolar lines. Residual errors of the image orientation (i.e. the refined RPCs) can result in undesired offsets in the y-direction which can be partially compensated by extending the search with some pixels tolerance in the y-direction. The cost function for the matching is the normalized cross-correlation. Considering the rugged mountain topography of the study area are relatively small window size of 7 by 7 pixel was selected to avoid strong smoothing of the topographic surface. The search range in the y-parallax was to one pixel to compensate for residual errors in the image co-registration.

DSMs with a pixel spacing of 0.5 m (WGS84 UTM 32) were extracted and used together with the refined RPFs for the orthorectification of all panchromatic and multi-spectral bands.

3.4.2.3 Sub-pixel image correlation

In recent years several sub-pixel image correlation techniques have been developed and applied in many tectonic and co-seismic deformation studies (Leprince et al., 2007; Hollingsworth et al., 2013) and for the analysis of glacier (Kääb, 2002; Heid and Kääb, 2012) and landslide (Delacourt et al., 2004; Debella-Gilo and Kääb, 2011) motion. In a recent study Heid and Kääb (2012) compared the performance of state-of-the art image correlation techniques and found the correlator implemented in COSI-Corr (Leprince et al., 2007) to be the most robust approach. We used the implemented algorithm for sub-pixel image matching through phase correlation in the frequency domain. The algorithm is based on a robust coarse-to-fine scheme and a sub-pixel matching method with a theoretical precision of 1/50 pixel. A hierarchical scheme with iteratively decreasing windows sizes (64, 32, and 16) was used to measure the east-west (EW) and north-south (NS) components of the surface displacement from the orthorectified panchromatic satellite images. The values were used for the number of robustness iterations and masking thresholds were set to 2 and 0.8, respectively.

3.4.2.4 Post-processing

Displacement fields derived from DIC usually comprise important fractions of noise resulting from various factors such as instrumental sensor noise, changes in the illumination conditions, surface changes, random patterns of natural surfaces and atmospheric effects. This typically requires post-processing steps such as mean filtering and the removal of false matches over stable terrain. Glaciers respect well defined boundaries that persist over long time periods and allow to remove false matches outside the boundaries of the moving mass manually (Heid and Kääb, 2012). For landslides, however, this is not a reliable strategy since previously stable terrains at the boundaries of a moving mass frequently destabilize and new unstable zones can develop elsewhere. In order to provide objective post-processing steps adapted to landslides, a dedicated routine that considers multi-spectral information and terrain variables was developed.

Most VHR satellites do not only collect panchromatic images but also multispectral bands that convey rich information on the spectral characteristics of different land cover types. This information can be exploited to delineate areas where decorrelation and noise must be expected.

Optical data are generally not applicable to monitor displacements under dense vegetation and, since the repeating patterns of a dense canopy are likely to generate false matches, it is better to disregard such areas from the analysis. To this end, the near-infrared and red bands are used to compute the Normalized Difference Vegetation Index (NDVI). Further, a Gaussian mixture model (Benaglia et al., 2009) is employed to model the NDVI histogram with two normal distribution corresponding to sparse and dense vegetation, respectively. A threshold (t_{NDVI}) for the separation of the two classes can be defined with Eq. 1:

$$t_{NDVI} = \mu_{NDVI.dense} - 2\sigma_{NDVI.dense} \quad \text{Eq.1}$$

where $\mu_{NDVI.dense}$ and $\sigma_{NDVI.dense}$ are the mean and the standard deviation of the normal distribution corresponding to sparse and dense vegetation. A conservative threshold on the Digital Numbers (DN) of the panchromatic images (DN < 110) further allows the identification of areas with strong shading, which due to low contrast are another source of false matches. The two masks for dense vegetation and shaded areas are combined and refined with morphological filters to remove patches smaller than the size of the correlation window. The final mask is then applied to suppress corresponding pixels in the raster resulting from sub-pixel image correlation.

The generated DSMs can be exploited to extract topographic variables such as slope and aspect. Any displacement that deviates more than 135° from the local aspect of the slope can be considered as inconsistent and is consequently suppressed. In some cases also prior knowledge about the expected landslide displacement rates may exist and can be used to filter all measurements that exceed the range of likely displacement rates. As this study targets the monitoring of slow-moving landslides over a period of two months, a displacement rate of 2 m.month^{-1} was considered as an upper bound. The described post-processing chain (Figure 3-5) was implemented in R (R Core Team, 2013) to enable complete automation of the routine.

The described steps target the reduction of false matches, whereas a final filtering is still recommendable to account for noise in the remaining detections. The Non-Local Mean Filter (NLMF) (Buades et al., 2008) has proven good abilities for effective image denoising and preservation of local details at the same time. The filter takes into account the Signal-to-

Noise Ratio (SNR) computed during the image correlation and implemented as a module in COSI-Corr. The main parameters of the filter are the standard deviation of the Gaussian kernel (σ=1.6), the search radius (11 pixels) and the patch size (5 by 5 pixels).

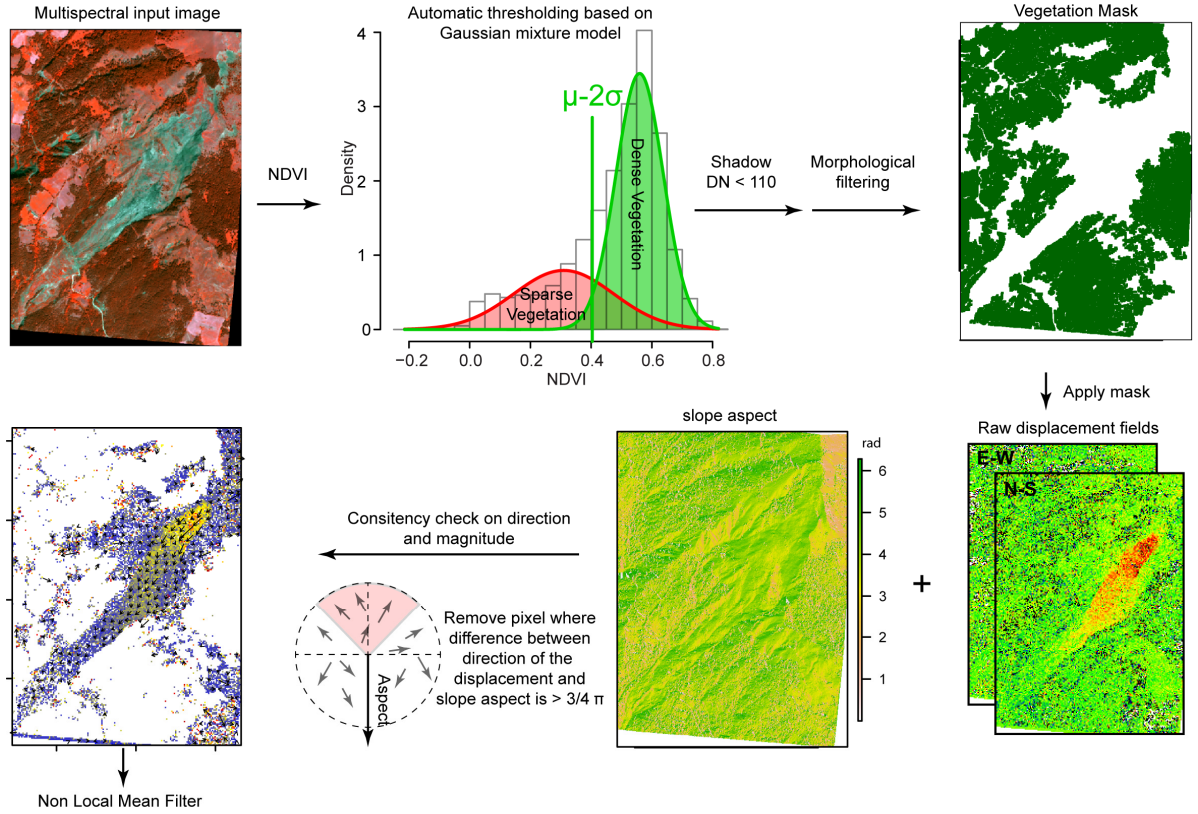


Figure 3-5: Schematic flow chart for the implemented post-processing routine (see section 3.4.2.4) for details.

3.4.2.5 Accuracy assessment

The accuracy of the processing chain is evaluated by considering three different aspects. (i) The quality of the image orientation is evaluated based on the residuals of the tie points and check points. (ii) The accuracy of the resulting DSM is quantified through comparison with airborne LiDAR surveys over stable terrain. (iii) The displacement fields are evaluated qualitatively for consistency with prior knowledge on the landslide processes and quantitatively through comparison with permanent on-site GNSS observations.

3.4.2.6 Processed datasets

The Pleiades satellite constellation comprises two identical satellites (Pleiades 1A launched 17/12/2011 and Pleiades-1B launched 02/12/2012). The two satellites have a phased sun-synchronous orbit with an orbital height of 694 km enabling short revisit times of up to 4 days. Panchromatic images are acquired with a ground sampling distance of approximately 0.7 m at Nadir and delivered with a nominal resolution of 0.5 m. Four multispectral bands (blue, green, red, and near infrared) are recorded simultaneously and delivered with a nominal resolution of 2 m. The images processed in this study (Figure 3-6) were acquired by Pleiades 1A over the Ubaye Valley (Barcelonnette, Southern French Alps) at two subsequent dates.

A monoscopic image was acquired in early August 2012 and a stereo-pair was recorded 59 days later. None of the images was recorded with the same incidence angle which must be considered as the default since satellite programming for Pleiades does not allow exact specifications of the view angle.

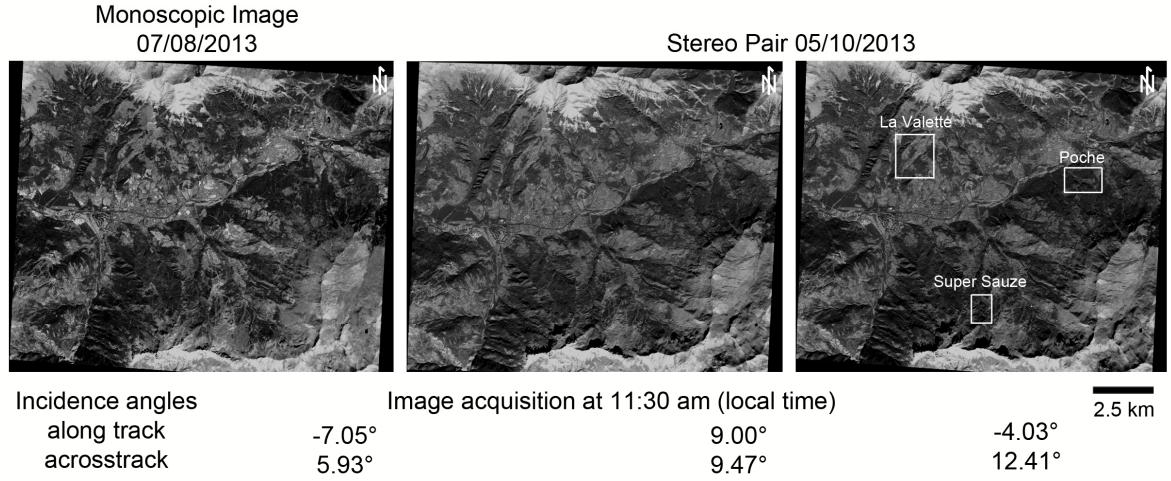


Figure 3-6: Overview of the panchromatic Pleiades images processed in this study.

Ground control points were extracted on stable terrain from LiDAR point clouds and associated orthophotographs from 2007, 2009 and 2012. A total number of 55 GCPs (Figure 3-7) was derived targeting stable natural objects (e.g. salient boulders) at proximity of the landslides and man-made objects such as cross roads and road markings. The datasets were available in the French Lambert 3 projection (Geoid heights IGN 69) and hence the GCPs had to be re-projected into ETRS89 TM32 (ellipsoid heights GRS80) to be consistent with the sensor model of the Pleiades satellite (WGS84 UTM 32).

The absolute error of the ground control points derived from the airborne LiDAR surveys is a composition of georeferencing errors (RMSE=0.12 m), scan errors (RMSE=0.16 m) and the uncertainty in the coordinate transformation (RMSE=0.01-0.05 m). Assuming that the terms are independent, the cumulative error ($RMSE_T$) can be computed with Eq. 2:

$$RMSE_T = \sqrt[2]{0.12^2 + 0.16^2 + 0.05^2} = 0.21 \text{ m} \qquad \text{Eq.2}$$

The latest aerial survey was carried out on 29/08/2012 and covered the Ubaye valley and the Super-Sauze landslide. The raw point cloud (comprising all returns) has an average point density of 90 pts.m^{-2} and was interpolated with a natural neighbour interpolator to a regular DSM raster (0.5 m pixel size) matching the grid of DSMs extracted from the Pleiades stereo-pairs.

The La Valette and the Super-Sauze landslides are monitored by the French Landslide Observatory (OMIV) with several permanent GNSS receivers providing measurements of the surface displacement with millimetre accuracy. In the time period between the image acquisitions three stations were operational and were used to quantify the accuracy of the correlation-based measurements.

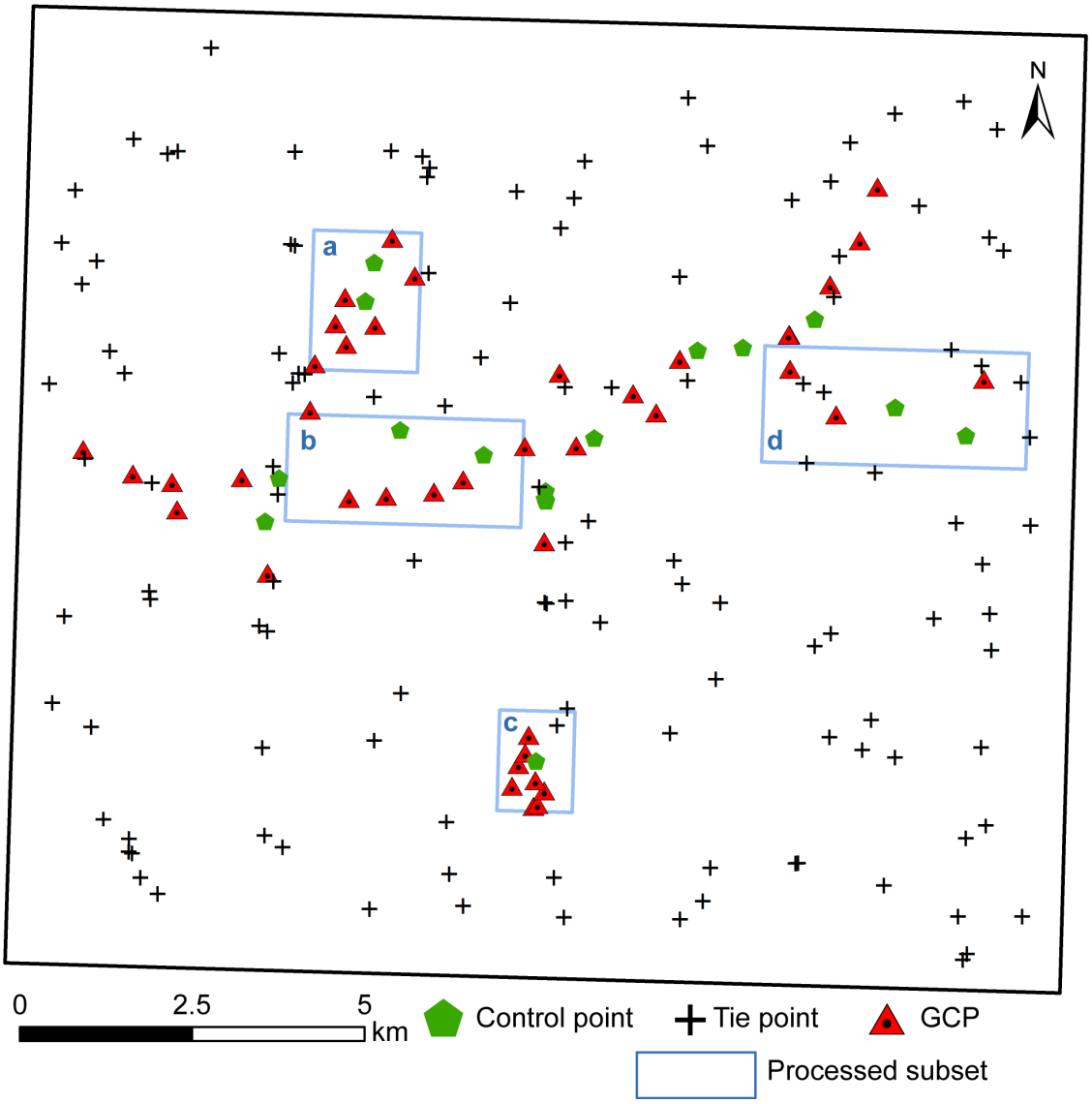


Figure 3-7: Distribution of control points, tie points and GCPs in the study area. The outlined subsets mark (a) the La Valette landslide, (b) the municipality of Barcelonnette, (c) the Super-Sauze landslide, and (d) the Poche landslide.

3.4.3 Results and Discussion

To evaluate the influence of the number of GCPs and different models employed for RPF refinement on the accuracy of the DSMs and motion measurements, four experiments were carried out using no GCPs, 5 GCPs and 40 GCPs with a simple translational bias correction, and 40 GCPs with a bias correction comprising two parameters for translation and drift. The results of those experiments and the derived displacement fields are presented and discussed in the subsequent sections.

3.4.3.1 Bundle adjustment

Residual errors in image and ground space after bundle adjustment are reported in Table 3-1. As can be expected, the residual errors in ground space reduced when more GCPs are used. The differences between using 5 and 40 GCPs are generally not larger than 0.1 m and the

enhancements that result from the use of and additional parameter to compensate for drift seem more important. The analysis suggest that the use of a small number of GCPs (5-10) and an error model that addresses translation and drift is close to optimal and only very small enhancements can be expected by collecting additional control points. This trend and the magnitude of the residual errors is consistent with several other studies on bias compensation with VHR satellite images (e.g. Fraser and Hanley, 2005; Aguilar et al., 2013). A correction without GCPs led to high residual errors especially in the z-axis. While the relative orientation among the images (indicated by the residuals in images space) is not significantly worse compared to the models with control points, the bias of the absolute ground coordinates is significantly higher than reported in other studies (Fraser and Hanley, 2005; Toutin et al., 2012; Aguilar et al., 2013).

Table 3-1: Residual errors in image and ground space after bundle adjustment for four different models. The highest accuracies are marked in green and red highlights the high ground residuals when no GCPs are used.

	No GCPs		5 GCPs		40 GCPs		40 GCPs translation +drift	
	Control RMSE	Check RMSE	Control RMSE	Check RMSE	Control RMSE	Check RMSE	Control RMSE	Check RMSE
Ground X [m]	N/A	3.80	0.06	0.46	0.24	0.44	0.21	**0.41**
Ground Y [m]	N/A	12.08	0.39	0.58	0.50	0.60	0.23	**0.27**
Ground Z [m]	N/A	34.19	0.04	0.60	0.06	**0.50**	0.06	**0.50**
Image X [pxl.]	N/A	0.27	0.27	0.25	0.28	**0.24**	0.27	**0.24**
Image Y [pxl.]	N/A	0.26	0.30	**0.16**	0.32	**0.16**	0.27	**0.16**
Total Image [pxl.]	0.27		**0.21**		0.32		0.23	

3.4.3.2 DSM accuracy

To analyse how the residual errors after bundle adjustment affect the quality of the DSM, stereo-reconstruction was performed with all four refined sensor models. While the general surface structure and level of detail depicted in the resulting DSM was very similar with all four models but Figure 3-8 shows that the residuals of the bundle adjustment without GCPs results in an offset of approximately 35 m in the z-axis. For the three reconstructions performed with GCPs, the mean elevation difference to the LiDAR ranges from -1.16 m (40 GCPs, translation) to -1.43 m (5 GCPs, translation). This bias is mainly caused by the stereophotogrammetric reconstruction yielding a relatively smooth surface in which many fine structures depicted in the LiDAR are truncated. Overall the RMSE of the DSMs shows only a minor influence of the number of GCPs ranging from 2.92 m (40 GCPs, translation) to 3.03 5 GCPs, translation).

Currently there are relatively few studies that analysed the stereophotogrammetric capabilities of Pleiades. Bernard et al. (2012) tested different imaging configurations over dense urban areas. Compared to ground points on bare earth, the RMSE of the models ranged

between 0.49 m and 1.17 m. Poli et al. (2013) quantified the accuracy of Pleiades stereo-models through comparison with a LiDAR DSM providing results that are more comparable to this study. The evaluation focused on an assessment over a dense urban area showing RMSE of 6.1-6.7 m. We observed that the errors are more important for dense urban areas (Figure 3-8) but the maximum RMSE was only 4.6 m. The analysis shows a good agreement of the stereophotogrammetric models and the LiDAR DSM with a minimum RMSE of 2.4 m over urban areas with low building density.

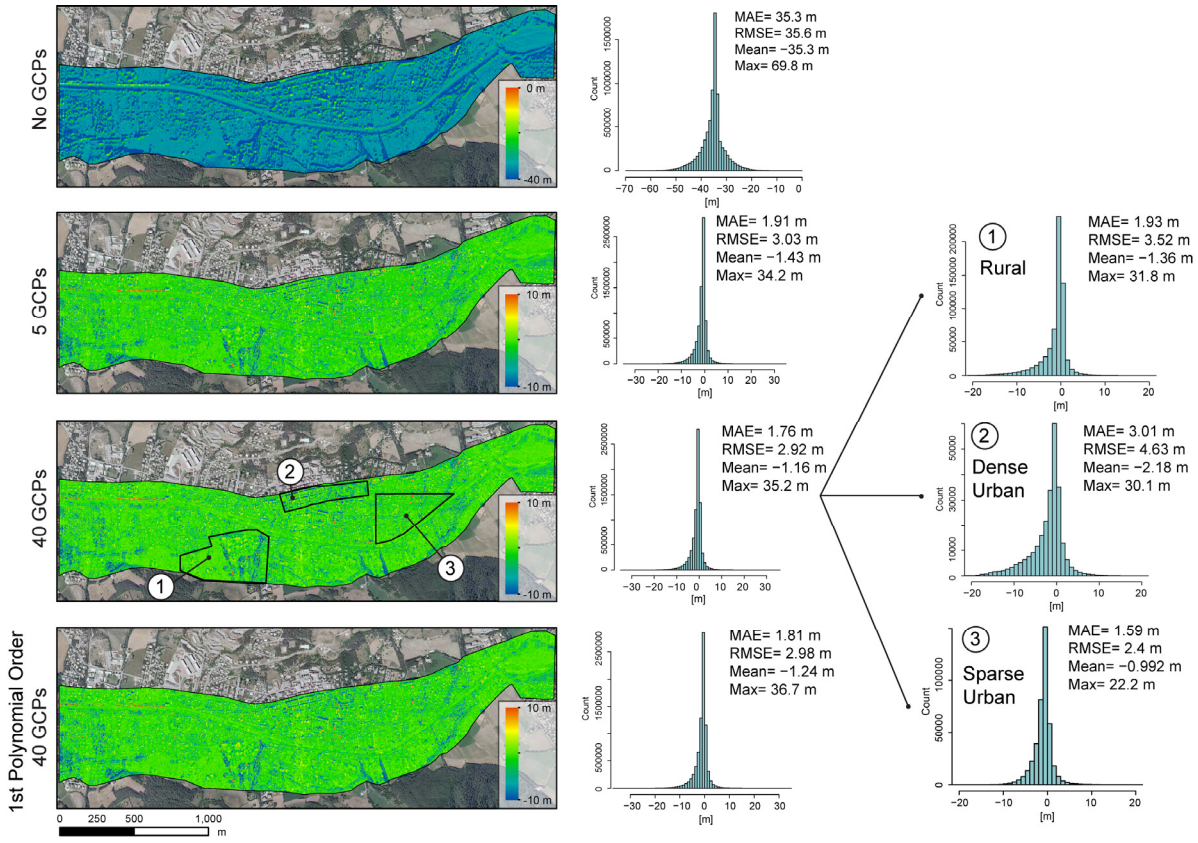


Figure 3-8: Accuracy assessment for the DSMs resulting from four different sensor models through subtraction from LiDAR DSM over the town of Barcelonnette. The analysis shows minor variations among the models generated with GCPs but important differences depending on the land cover.

A second analysis was carried out contrasting the LiDAR and the stereophotogrammetric DSMs at the Super-Sauze landslide (Figure 3-9). Compared to the results obtained for the municipality of Barcelonnette, the RMSEs were generally higher ranging between 5.6 m (40 GCPs, translation) and 5.9 m (5 GCPs, translation). Relative to the model based on 5 GCPs, the use of all 40 GCPs resulted only in minor enhancement of the total RMSE ranging from 0.03 m (translation and drift) to 0.28 m (translation). This corroborates what was already observed in the residual errors after bundle adjustment. Only small benefits can be expected from the collection of more than 5-10 GCPs.

The overall RMSEs for the DSM at the Super-Sauze landslide (Figure 3-9) are significantly higher than for the Barcelonnette test site (Figure 3-8) which can be explained by the steeper topography and the stronger vegetation cover on the stable parts around the landslide. Indeed for steep slopes, which were partially shaded at the time of the acquisition of the stereo-pair, the RMSE is 15.2 m and thereby significantly higher than for vegetated

areas with moderate slopes (RSME = 4.65 m) and for moderate slopes without vegetation (RSME = 1.34 m). The latter must be considered as representative for the accuracy achieved for the landslide body and indicates a satisfactory quality of the reconstruction for the zones targeted in this study.

The absolute heights of the DSMs without GCPs are strongly biased compared to the LiDAR DSMs resulting in RMSEs between 35.6 m (Figure 3-8) and 41.9 m (Figure 3-9). However, the relative orientation of the adjusted block shows very low residuals in the image space (Table 3-1) and could consequently still be used for relative displacement measurements (section 3.4.3.3).

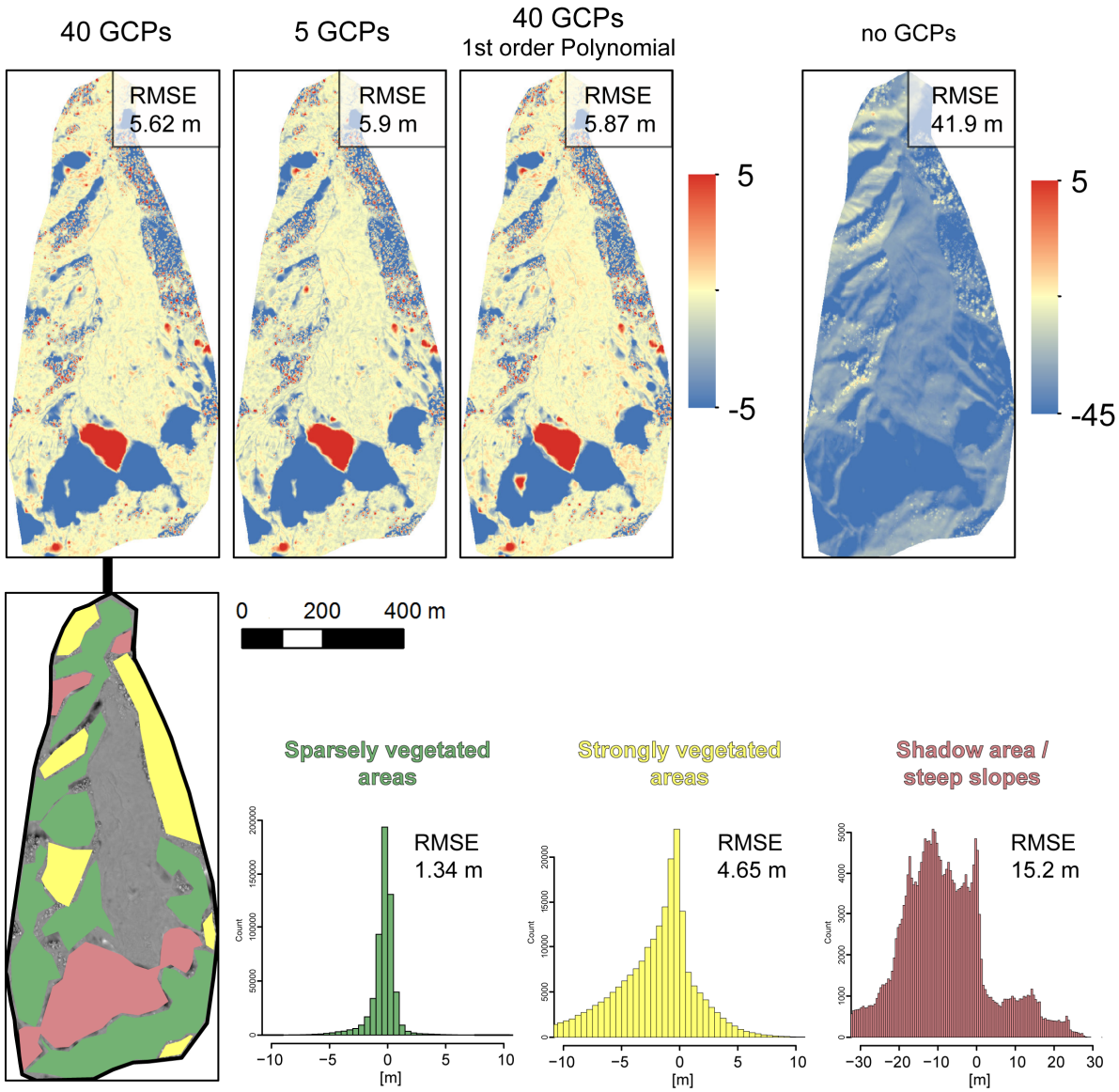


Figure 3-9: Accuracy assessment for the DSMs resulting from four different sensor models through subtraction from LiDAR DSM at stable areas around the Super-Sauze landslide. The analysis shows minor variations among the models generated with GCPs but important differences depending on the type of land cover and topographic position.

3.4.3.3 Displacement fields

Considering the low residual check point error of the model with 40 GCPs (translation only, Table 3-1) and the slightly higher accuracy of the resulting DSMs, it was considered as the primary option for displacement measurements. Although the complete omission of all GCPs yields a significant bias in the absolute coordinates of the resulting model, it could yield accurate motion measurements since the image block remains consistent in relative terms. Since measurements without the need for GCPs would greatly facilitate the automated processing of longer time series, it was considered as the second option in this study. Figure 3-10 shows that the motion fields for the La Valette landslide derived with and without GCPs. The general pattern of movement shows higher displacements at the scarp (maximum of 0.82 m) and gradually decreasing displacements downslope. This pattern is consistent with field observations and previous remote sensing studies at the La Valette landslide (Squarzoni et al., 2003; Delacourt et al., 2007; Leprince et al., 2008; Raucoules et al., 2013).Figure 3-10 indicates that the results obtained with or without GCPs are, though not identical, indeed very similar. Slightly higher displacements in the central part of the landslide are observed without GCPs (Figure 3-10b) but from visual comparison alone it is not obvious which of the two models provides the more accurate results.

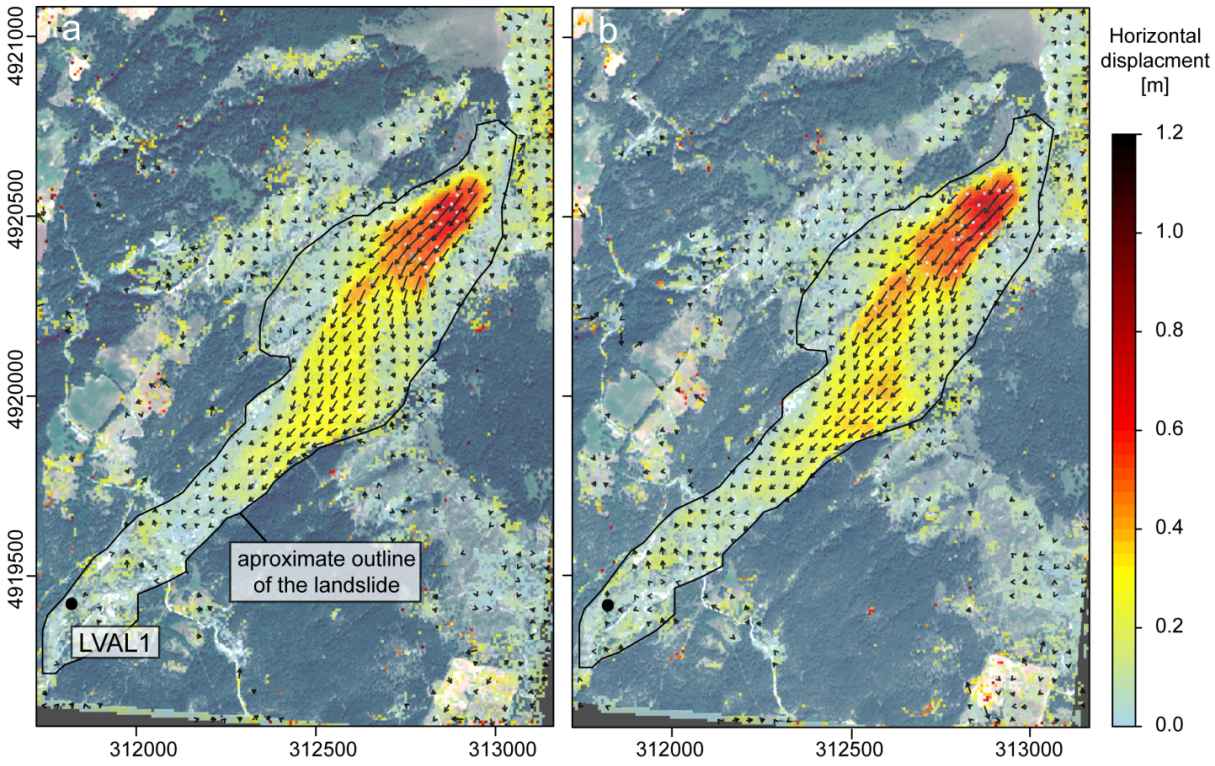


Figure 3-10: Displacement field of the La Valette landslide between 07/08/2012 and 05/10/2012 using (a) 40 GCPs and (b) no GCPs. The position of the permanent GNSS receivers operational in this period is indicated.

The displacement field obtained for the Super-Sauze landslide (Figure 3-11) shows a strong motion (maximum 2.15 m) at the central part of the landslide and a gradually decreasing rate further downslope. The fact that the pattern of movement changes significantly over time does not allow a direct comparison with previous studies but the general distribution of the motion could be confirmed by field observation a few days after the acquisition of the

Pleiades stereo-pairs. The stability of the upper most part of the landslide (Figure 3-11) is also consistent with terrestrial photogrammetric measurements in Stumpf et al. (In Review-a). Terrestrial photogrammetry indicates horizontal displacements at the central part of up to 4.0 m for the period 05/07/2013 to 09/10/2012 suggesting a decrease in displacement rates during the summer.

Similarly to the displacement fields at La Valette, the results for Super-Sauze show only slight differences between measurements with GCPs (Figure 3-11a) and without (Figure 3-11b). Overall the values are slightly lower without GCPs and an additional hotspot with locally high displacement rates (corresponding to a step in the topography of the sliding surfaces) is depicted (Figure 3-11b).

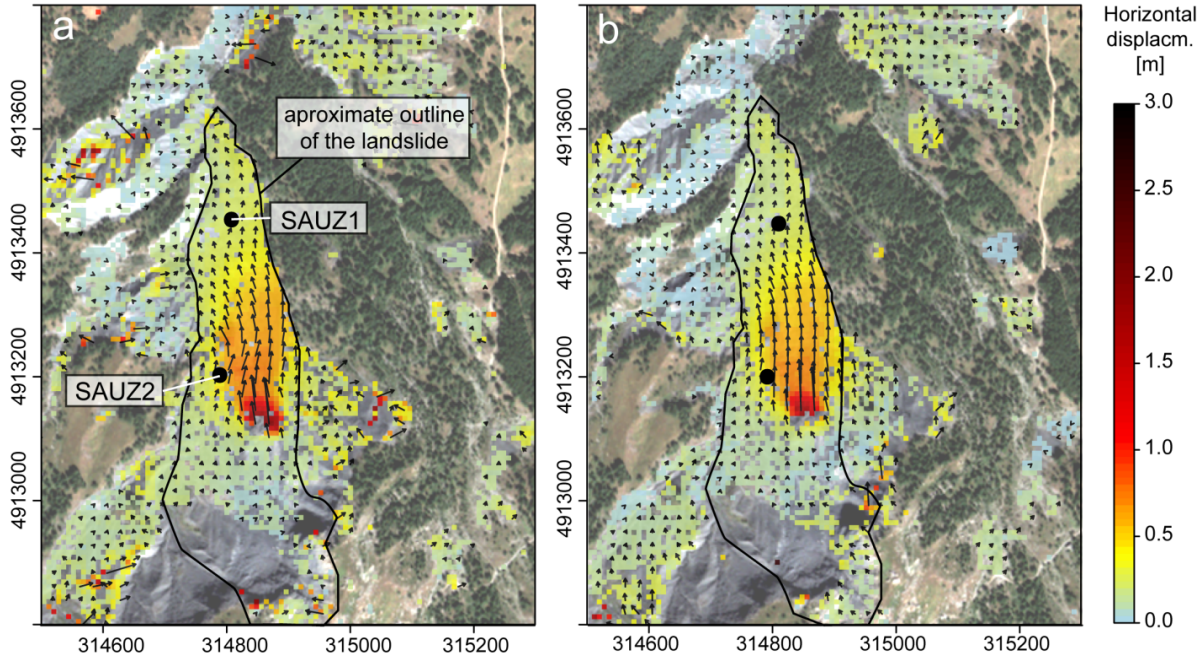


Figure 3-11: Displacement field of the Super-Sauze landslide between 07/08/2012 and 05/10/2012 using (a) 40 GCPs and (b) no GCPs. The positions of two permanent GNSS stations which were operational in this period are indicated.

Of the three investigated sites, the Poche landslide displayed the slowest movement during the observed period. The highest displacements (maximum 0.44 m) are observed in the central part (Figure 3-12). As for the other sites, the results obtained with GCPs and are very similar to those without ground control, while the measurement without GCPs contains more noise (Figure 3-12b). The magnitude and spatial pattern of the displacement field is consistent with the long-term kinematics observed during the last 50 years (< 1m.year^{-1} in the lower part of the slope) but displacement regime is highly variable over time and knowledge about the recent kinematics is limited.

La Valette and Super-Sauze are monitored with permanent GNSS receivers whose measurements were compared with the image-based velocities. Table 3-2 shows that the highest deviation of the 2D displacement with 40 GCPs is 0.13 m (LVAL1, N-component) and 0.13 m if no GCPs were used (SAUZ 2, N-component). The observed errors with 40 GCPs suggest RMSEs in the E-component ($RMSE_E$) of 0.04 m, in the N-component ($RMSE_N$) of 0.10 m and a total RMSE ($RMSE_T$) of 0.11 m. Surprisingly without GCPs the figures are slightly lower and correspond to $RMSE_E$ = 0.04 m, $RMSE_N$ = 0.08 m, $RMSE_T$ = 0.08 m.

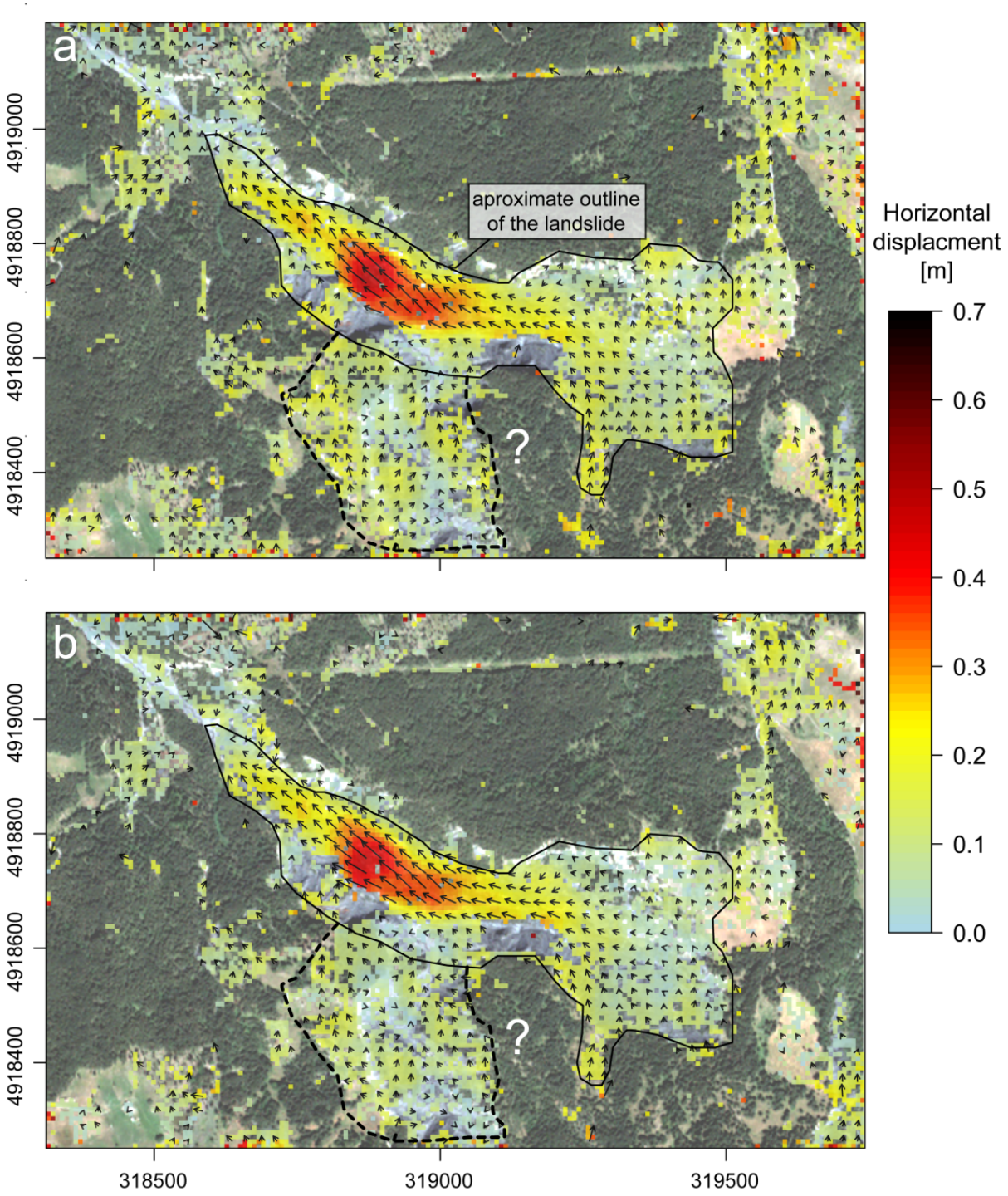


Figure 3-12: Displacement field of the Poche landslide between 07/08/2012 and 05/10/2012 using (a) 40 GCPs and (b) no GCPs. The area marked with a question sign indicates a coherent displacement pattern of a potential previously undiscovered instability.

Table 3-2: Comparison of the image-based motion measurements with three permanent GNSS stations at La Valette and Super-Sauze. The most accurate image-based measurements are highlighted in bold. The values in brackets correspond to the original pixel values before post-processing.

	LVAL1			SAUZ1			SAUZ2		
	GNSS	Image correlation		GNSS	Image correlation		GNSS	Image correlation	
		40 GCPs	no GCPs		40 GCPs	no GCPs		40 GCPs	no GCPs
E	0.00	**-0.02** (-0.05)	-0.06 (-0.03)	0.07	0.10 (-0.06)	**0.07** (2.88)	0.15	**0.13** (0.33)	**0.17** (0.13)
N	-0.01	0.12 (0.21)	**-0.06** (0.09)	0.23	0.23 (0.29)	**0.22** (-0.66)	0.58	**0.46** (0.43)	0.45 (0.21)
Total horizontal	0.01	0.12 (0.22)	**0.07** (0.09)	0.24	**0.25** (0.30)	**0.23** (2.96)	0.60	**0.48** (0.55)	**0.48** (0.25)

Considering that only three GNSS receivers were available, those figures are not statistically significant but, nevertheless, indicate that decimetre accuracy can be achieved with the described processing chain and that the availability of ground control is not essential for accurate measurements. Without GCPS a residual shift in x (3.80 m) and y (12.08 m) causes an offset of 1-2 pixel between the correlation images (8 m resolution) and the coordinates of the GNSS receivers. However, since the displacement of the observed landslides is relatively coherent over such distances this does not lead to an increased error. More in general, a lack of ground control seems to have no significant influence on the observed magnitude of the displacement, whereas, if a precise integration with in situ instruments is desired, the collection of a few precise GCPs should be considered.

Table 3-2 contains the measured displacements before the post-processing and indicating that the routine is not only useful for the removal of false positives but also improves the accuracy of the measured values. It should be recalled that the processed dataset comprised only one stereo-pair and consequently all images were orthorectified with the same DSM. Implicitly, this comprises the assumption that all movement which occurred between the two time steps is slope parallel (i.e. the surface relative height among different surface parts does not change). This is a reasonable assumption for glacier flow and the short time-period addressed in this study. However, multiple stereo-pairs will be required to account for significant changes in the terrain topography over longer time periods.

Compared to previous studies, using aerial images for the tracking of landslide motion (Delacourt et al., 2004; Casson et al., 2005) the residual errors are one order of magnitude smaller and in the same range than pixel-offset tracking techniques applied on SAR amplitude images (Raucoules et al., 2013). The possibility to reach decimetre accuracy with VHR satellite images using little or no ground control makes the elaborated processing chain potentially useful for seasonal monitoring of slow- and very-slow moving landslides or the reconstruction of the displacement measurement from image archives.

A bottleneck is still that processing chain involves two specialized commercial programs which increases the cost of the applications and hinders full automation. The use of an alternative image correlation technique implemented in open-source software (Deseilligny et al., 2013) could resolve this issue partially. However, there is currently no accurate and free tool for RPF bundle adjustment available and specialized commercial software is indispensable for the time being.

Further studies could evaluate the applicability of the proposed approach for other landslide types and surface displacement phenomena and the possibility to integrate displacement measurements derived from SAR interferometry and other sources to reconstruct all 3D components. An interesting option in this context is also the use of multiple stereo-pairs which theoretically enable the recovery of 3D surface displacement form disparity measurements directly in the image geometry.
As recently demonstrated in Booth et al. (2013) physical quantities such as the depth of the sliding surface and rheological parameters can be estimated from remotely sensed displacement fields and further research in this direction is needed to fully exploit the obtained measurements (e.g. assimilation in numerical models). A further possibility that arises from measurements without ground control is the application of correlation-based techniques for mapping purposes complementing services provided by SAR interferometry.

3.4.4 Conclusion

This study investigated the use of VHR satellite images for landslide displacement measurements with a particular focus on the reduction of ground control requirements. A processing chain comprising RPF bundle adjustment, DSM extraction, sub-pixel image correlation and post-processing was elaborated and used to quantify the surface displacement of three active slow-moving landslides.

Three Pleiades satellite images (1 monoscopic, one stereo-pair) were processed and a number of experiments allowed to quantify the accuracy of refined sensor models and extracted DSM for four scenarios differing mainly in the amount of available ground control. The analysis showed that GCPs are important for an accurate absolute georeferencing of the DSM of the displacement maps but have little impact on the absolute values of the measured displacement. A comparison of the extracted DSMs with airborne LiDAR surveys showed RMSE between 1.34 and 15.2 m depending mainly on the type of land cover at the surface. For the derived displacement fields a maximum error of 0.13 m was observed weather 40 GCPs were used or only relative orientation was performed without any GCPs.

To the best of our knowledge, study provides the first example for displacement measurements from VHR satellite images without ground control and external DEMs. Considering the limited number of processed datasets and validation measurements (i.e. three GNSS receivers), the results must be considered preliminary. We target in particular to exploit additional stereo-datasets that have been acquired for the same study areas and use an open-source correlator for better automation. Also the post-processing routine could be further enhanced taking into account additional topographic variables.

Acknowledgments
We thank the Centre national d'études spatiales for the acquisition of the Pléiades images and free access through the Recette Thématique Utilisateurs Pléiades (CNES-RTU). We are also grateful to the Observatoire Multidisciplinaire des Instabilités de Versants (SNO-INSU OMIV) for making the GNSS data publicly available

3.5 *3D landslide displacement monitoring and volume estimation using multi-view terrestrial photogrammetry*

This section is based on: Stumpf, A., Malet, J.-P., Allemand, P., Deseilligny, M.-P., Skupinski, G., In Review. Terrestrial multi-view photogrammetry for landslide monitoring, Journal of Geophysical Research., 48 p.

The research work presented in this section targets the elaboration of a photogrammetric processing chain for the monitoring of landslide dynamics in terms of surface changens, 3D displacement measurements and estimation of the transported volumes. It explores the use of SfM and MVS algorithms for the straightforward computation of accurate surface models (i.e. point clouds) and the derivation of relevant information through cloud-to-cloud change detection and point matching. Terrestrial photographs of a slow-moving landslide are acquired with a consumer-grade handheld camera for five different epochs. Based on publicly available and open-source software libraries, three different pipelines are elaborated and parameterized, and their accuracy is assessed through comparison with terrestrial and airborne LiDAR surveys. The most accurate of the three pipelines yields surface models that approach the quality of LiDAR. Change detections over a period of 2 years allow a detailed assessment of the seasonal dynamics of the landslide; the possibility to estimate volumes and 3D displacement estimates are illustrated for the most active parts of the landslide. Volumes and displacement rates derived from the photogrammetric processing chain are in close agreement with measurements based on terrestrial LiDAR and permanent GNSS observations.

3.5.1 Introduction

Digital Elevation Models (DEMs) are indispensable information sources in the geosciences. Modern remote sensing technologies have greatly facilitated their creation and frequent updating for applications in geomorphology, hydrology, geophysics and natural hazards research. Spaceborne observations are valuable sources for obtaining topographic information on global and regional scale whereas local scale studies typically require more precise measurements at higher spatial resolution, which can be derived with airborne/terrestrial photogrammetry or laser scanning. In particular Light-Detection and Ranging (LiDAR) is being employed in an increasingly large number of applications providing very accurate surface representations and the capability to penetrate vegetation (Heritage and Large, 2009; Jaboyedoff et al., 2010). However, the costs of the equipment and the logistics of a LiDAR surveys are currently still rather high and, therefore, acquisition at high temporal resolutions are not always feasible. Conventional photogrammetric techniques with metric and non-metric cameras are a frequently employed alternative for a wide range of applications (Fryer et al., 2007) but comprise high demands on the image acquisition geometry, ground control, processing software and the experience of the operator (Henry et al., 2002; Fryer et al., 2007).

Great advances of the photogrammetry and computer-vision communities in pose-estimation and bundle-adjustment (Triggs et al., 2000; Hartley and Zisserman, 2004), camera self-calibration (Fraser, 1997; Pollefeys et al., 1999) as well as feature-based and area-based image matching (Lowe, 2004; Pierrot-Deseilligny and Paparoditis, 2006; Hirschmuller, 2008; Furukawa and Ponce, 2010) have recently converged in a new class of photogrammetric algorithms and tools that enable more flexible 3D surface reconstruction from unordered non-metric image collections. They are summarized under the terms 'Structure-from-motion' (SfM, Ullman, 1979); being the process of estimating camera parameters and sparse point-clouds, and Multi-View Stereo (MVS), which is the process of deriving dense surface models once the correspondence among multiple cameras has been established. Many proposed approaches for SfM and MVS are now implemented in commercial software (AgiSoft PhotoScan, Pix4D, PhotoModeler Scanner, Trimble Inpho), web-based services (Microsoft Photosynth, Autodesk 123D, Arc3D, Cubify Capture) as well as in free and open-source software (Snavely et al., 2008; Furukawa and Ponce, 2010; Deseilligny and Clery, 2011; Rothermel et al., 2012; Wu, 2013).

The geoscience community has already started to take great interest in those new tools and began to explore their capabilities for a number of applications (Westoby et al., 2012; Fonstad et al., 2013). An overview of recent studies is provided in Table 3-3. Among many factors that condition the accuracy of SfM-MVS (e.g. camera, lens, acquisition geometry, quality of the ground control, illumination conditions, and processing software) the distance to the object is probably the most influential. Imaging distances between 1.8 m and 2400 m have been explored resulting in accuracies that are generally between 0.04 m and 1.68 m. An exception constitutes the application of SfM-MVS in vegetation studies where the errors can amount to several meters (Table 3-3).

Table 3-3: Overview of geoscience studies applying SfM and MVS.

Subject	Sensor platform	Distances	Employed tools	Reference data (accuracy)	Reported Accuracy	Reference
Coastal cliff	Boat	~20 m	Photosynth, Bundler(Snavely et al., 2008), Patch-based MVS algorithm (PMVS2, Furukawa and Ponce, 2010)	LiDAR (>0. 015 m)	RMSE, 0.07 m	(James and Robson, 2012; James and Varley, 2012)
Volcanic dome	Helicopter	500–2400 m		dGPS (0.1 m)	RMSE, 1.30 - 1.68m	
Forest canopy	Aerial kite	40 m	Agisoft Photoscan	Aerial LiDAR (not reported, typically 0.2-0.3 m)	0.80 - 5.69 (Mean ± RMSE, ground points), 3.9–10.9 m (RMSE, regression estimate of canopy heights)	(Dandois and Ellis, 2010, 2013)
Sparsely vegetated coastal area	UAV	40–50 m	Bundler, PMVS2	Total station and dGPS (0.02 m)	0.03-0.04 m	(Harwin and Lucieer, 2012)
Antarctic mossbeds	UAV	50 m	Agisoft Photoscan	dGPS (0.02 - 0.04 m)	0.04 m	(Lucieer et al.)
Alpine landslide	UAV	~200 m	Bundler, PMVS2	Terrestrial LiDAR (0.01-0.03 m)	0.25 m	(Niethammer et al., 2011b)
Glacier	UAV	300-400 m	Trimble Inpho	dGPS (0.05 m)	0.18 - 0.63 m	(Whitehead et al., 2013)
Volcanic lava flows	Terrestrial photographs	~1.8 m	Apero-MicMac	not available	Not available. (mm resolution)	(Bretar et al., 2013)

Though researchers have already explored the use of SfM-MVS tools in geoscience applications there are no studies that compare different pipelines making it problematic for interested users to select among the variety of different tools. This also applies to some extent for the choice of several algorithm parameters whose values are typically not reported. Several benchmark studies have evaluated MVS algorithms for toys (Seitz et al., 2006) or architectural outdoor scenes (Strecha et al., 2008). Remondino et al. (2012) also provided a comparative analysis of open-source tools for SfM on artificial objects and architectural environments. However, natural scenes yield fundamentally different image characteristics (Torralba and Oliva, 2003) and are typically more challenging in terms of surface characteristics, illumination and constraints on the view geometry. In Fallourd et al. (2010) and Travelletti et al. (2012) it has been demonstrated that terrestrial time-lapse photography is a valuable tool for the monitoring of landslides and glaciers but multi-view photogrammetry has not yet been tested in this context. Therefore further studies are required to explore the

potential and limitations of different SfM-MVS for multi-temporal geomorphological surveys.

Reliable accuracy assessments require reference measurements with known greater precision and dGPS and LiDAR scans are currently among the most accurate techniques. While long GPS observation sessions (typically > 3 hours) can yield millimetre accuracy (Malet et al., 2002), short sessions during photogrammetric campaigns typically comprise errors in the range of a few centimetres. Especially in rugged terrain it should further be considered that ground control points (GCPs) cannot be measured on very steep slopes and rugged surfaces introducing a potential bias. Terrestrial and aerial LiDAR scans provide spatially distributed point clouds at accuracies between 0.02 and 0.15 m, respectively (Jaboyedoff et al., 2010) and have been used for to assess the error of SfM-MVS techniques in several studies. In many cases the obtained points clouds are interpolated to continuous surfaces before comparison (Dandois and Ellis, 2010; Niethammer et al., 2011b; James and Robson, 2012; Westoby et al., 2012; Fonstad et al., 2013) which is not ideal since interpolation artefacts bias the accuracy estimates or the measured changes over time. Tools for more direct comparison of points clouds are widely used for co-registration and temporal change detection among LiDAR scans (Lague et al., 2013; EDF-R&D, TelecomParisTech, 2013) and might also be more appropriate in this context.

Therefore, the target of this work is to contribute to the on-going efforts to exploit SfM-MVS as a measurement technique in geoscience studies. The objective is to investigates the possibility to use freely available processing tools (Deseilligny and Clery, 2011; Wu et al., 2011; Deseilligny et al., 2013; Wu, 2013) for the monitoring of active landslides. Landslides occur frequently in high mountain terrain where the rugged topography, vegetation and fast changing weather conditions constitute a challenging environment for the application of photogrammetric techniques. In section 2 the study site (Super-Sauze landslide) is introduced together with the acquisition protocols of the terrestrial photographic surveys. The ground-control datasets obtained from LiDAR and dGPS surveys are explained in detail. Section 3 provides an overview of three SfM-MVS pipelines and details the implemented algorithms. Special attention is given to the significance and optimization of different algorithm parameters of the three SfM-MVS pipelines and methods for direct point cloud comparison. In section 4 the accuracy of photogrammetric models is assessed through comparison with LiDAR point clouds. In Section 5 change detection is performed to quantify the surface changes and dynamics of the landslide over a period of two years and Section 6 provides a discussion of current limitations, potentials and possible pitfalls of the survey design.

3.5.1.1 Study site and data acquisition

The Super-Sauze landslide (Figure 3-13a) is a clay-rich slow-moving slope movement located in the Southern French Alps. The landslide initially developed in the 1960s through retrogressive failures of the main scarp and is highly controlled by the local hydrology, climate and the accumulation of new material from successive failures at the main scarp. During the last decade several in-situ and remote sensing studies have contributed to a better understanding of its kinematics (e.g. Malet et al., 2002; Niethammer et al., 2011a; Travelletti et al., 2012; Stumpf et al., 2013b). Multi-technique displacement observations suggest a long-term average displacement rates of 0.01-0.03 m.day^{-1} (Malet et al., 2002) but regularly, daily cumulative displacements larger than 6 m are observed (Travelletti et al., 2012). The site is characterized by a rugged topography comprising vertical and overhanging cliffs as well as

quasi-horizontal surfaces. The eastward-adjacent slopes are largely forested, whereas badlands and several sub-parallel ridges border the landslide to the west. Constraints on possible view points, low incidence angles and occlusion make such type of terrain generally challenging for terrestrial photogrammetric measurements.

Regular acquisitions of terrestrial photographs in a MVS setup have been carried out since October 2011 in regular intervals (Table 3-4). Field campaigns are typically limited to the time between early May and late October since snow cover prohibits photogrammetric and most other measurements during the rest of the year. A Nikon D700 camera has been used, the focus has been set to infinity, and care has been taken to obtain a good trade-off between sufficiently short exposure time and large depth of field (narrow aperture) during all acquisitions.

Two target zones were monitored by photogrammetry. A first acquisition protocol was setup to reconstruct the evolution of the main scarp (Figure 3-13a) at five dates for the period October 2011-July 2013. The images were recorded in a surface-parallel linear array of panoramic shots yielding distances to the targeted surface between 20 m and 200 m. During the first survey images were recorded only at a reduced resolution (2128x1416) and in JPEG format, whereas for all subsequent surveys full resolution (4256x2832) images were stored in native Nikon (NEF) file-format to avoid information loss. A 60 mm lens was used at all dates except in July 2012 when a 35 mm lens was used to also investigate the influence of the focal length on the reconstruction. A second acquisition protocol was setup to obtain a full-scene model for the entire landslide in October 2012 and July 2013. Images were captured in a half circular array along the limits of the landslide (Fig. 1a) using a 35 mm lens. The distance to the surface varied between approximately 50 m and 1000 m.

The reference point clouds used in this study were acquired with a terrestrial (Optech ILRIS-3D) and aerial (Riegl LMS-Q560) laser scanner. To provide full coverage of the scarp, multiple terrestrial scans were carried out from different view angles aligned subsequently with the Iterative Closest Point (ICP) algorithm (InnovMetric, 2010). The scans were performed at an average distances between 3 and 800 m resulting in a ground-point density generally greater than 100 pts.m^2. While the theoretical accuracy of the terrestrial LiDAR scans is in the range of ± 0.01 m (Jaboyedoff et al., 2010), the alignment error generally amounted to an RMSE of 0.02-0.03 m (Travelletti, 2011). The aerial LiDAR survey was acquired on 29 August 2012 with an average flight height of 800 m above the surface resulting in an average ground-point density of approximately 90 pts.m^2. The accuracy of the airborne point-cloud is influenced by scanning, georeferencing and reprojection errors and typically between 0.20-0.30 m (Travelletti, 2011).

In addition, a set of 35 ground targets were used for georeferencing the point clouds. The target centres were measured with dGPS during the last photogrammetric acquisition on 18 July 2013. The 3D measurement error of the dGPS was reported with 2.5 cm in the 3 directions.

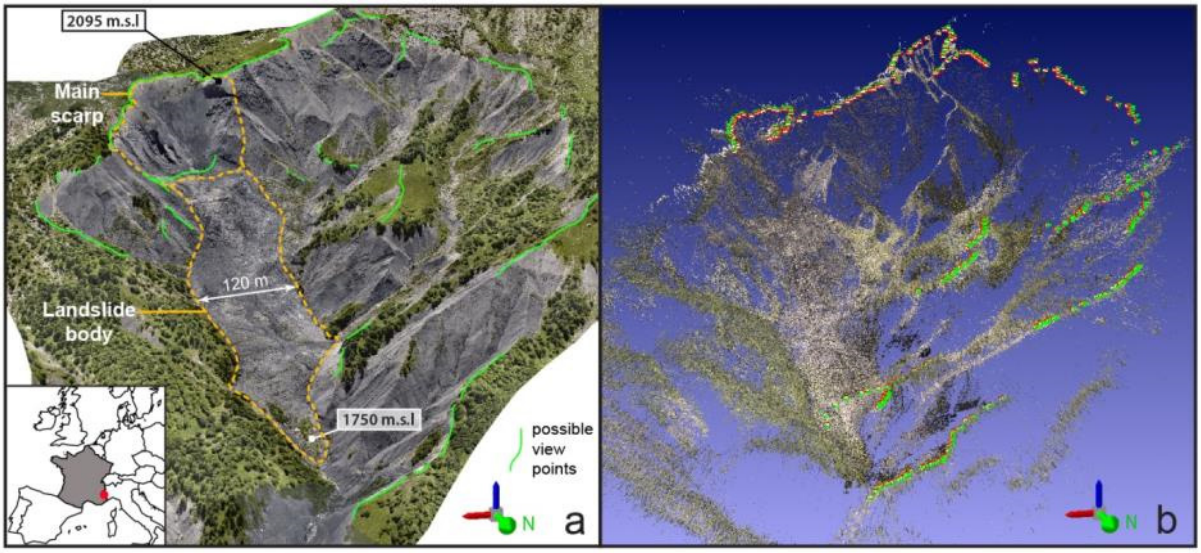


Figure 3-13: (a) Location and overview of the Super-Sauze landslide with possible view points for the image acquisition. (b) Exemplary reconstruction results (unscaled sparse point cloud, 19 July 2013) illustrating the camera acquisition protocol for the full scene reconstruction. Cameras positions are depicted by red/green icons.

Table 3-4: Overview of the collected data during the photographic, LiDAR and dGPS surveys.

Main scarp	**20-Oct-11**	**5-Jul-12**	**9-Oct-12**	**14-May-13**	**18-Jul-13**
# images	88	106	168	130	265
Focal length (mm)	60	35	60	60	60
Resolution	2128x1416	4256x2832	4256x2832	4256x2832	4256x2832
Terrestrial LiDAR	ILRIS-3D	ILRIS-3D	ILRIS-3D		
Distance (m)	3-800	3-800	3-800		
Point density (pts.m^{-2})	> 100	> 100	> 100		
RMSE$_{xyz}$ (m)	0.02-0.03	0.02-0.03	0.02-0.03		
Landslide body		**29-Aug-12**	**10-Oct-12**		**19-Jul-13**
# images			401		401
Focal length (mm)			35		35
Resolution			4256x2832		4256x2832
Airborne LiDAR		LMS-Q560			
Distance (m)		~ 800			
Point density (pts.m^{-2})		~ 90			
RMSE$_{xyz}$ (m)		0.20-0.30			

3.5.2 Data processing

The different surface reconstruction pipelines tested in this study are based on open-source libraries and follow the common SfM-MVS sequence starting from the extraction of homologous image points, estimation of camera parameters and bundle adjustment, and a final dense reconstruction.

VisualSfM (VSfM) is a freely-available software implementing algorithms for SIFT (Lowe, 2004) point extraction on graphical processing units (GPUs) (Wu, 2007), pose-estimation and multi-core bundle adjustment (Wu et al., 2011; Wu, 2013) for which source code is also partially available..

PMVS2 is a patch-based multi-view stereo algorithm (Furukawa and Ponce, 2010) targeting dense reconstruction after camera correspondence has already been established. The algorithm can be scaled to large reconstruction problems using a related cluster multi-view stereo (CMVS) algorithm (Furukawa et al., 2010).

Apero-MicMac is an open-source software for multi-view photogrammetry (Deseilligny et al., 2013). The project comprises tools for tie-point extraction (Tapioca), pose-estimation, camera-calibration, bundle-adjustment (Apero, Deseilligny and Clery, 2011), dense-matching (MicMac) and georeferencing as well as various further tools dedicated to miscellaneous tasks such as point cloud extraction, creation of masks, and orthorectification. The source code was compiled under Linux.

Three different pipelines combing VSfM with CMVS (VSfM+CMVS), Apero with CMVS (Apero+CMVS), and Apero with MicMac (Apero+MicMac) are compared. All processing was performed on a workstation with 8 cores, 36GB RAM and an NVIDIA Quadro 600 GPU. A general overview of all processing steps is provide in Figure 3-14 and detailed in the following sections.

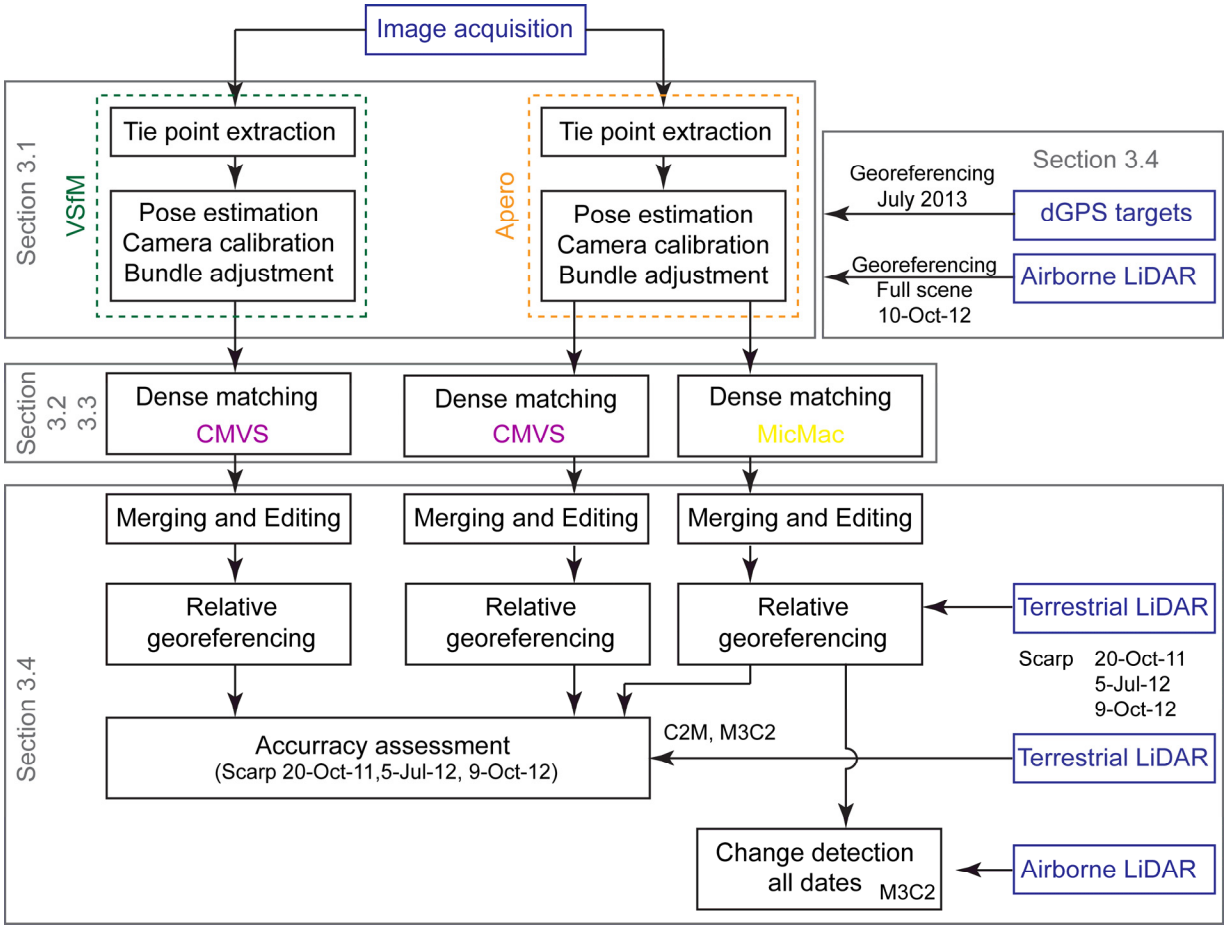


Figure 3-14: Generalized workflow for the reconstruction pipelines, accuracy assessment and change detection. See indicated text sections for details.

3.5.2.1 Pose estimation, camera calibration and bundle adjustment

For the robust estimation of all camera parameters, a large number of homologous points has to be identified among the different images.

For the tie point extraction with VSfM, it is necessary to convert the native NEF images to JPG format. For computational speed SIFT points are typically not extracted from the original very high resolution images and the maximum image dimension (width) was set to 2128 pixels when running the GPU SIFT (Wu, 2007) implemented in VSfM.

In order to provide a fair comparison, the tie point extraction in MicMac was also performed with a maximum image dimension of 2128 pixels. The corresponding module provides an interface for point-matching based on the open-source SIFT implementation SIFT++ (Vedaldi and Fulkerson, 2008). The code is parallelized for efficient computation and automatically filters outliers if there are less than 5 mutual matches among the 10 nearest neighbours (in the image space) of each matching point pair.

VSfM and Apero (the MicMac tool for the estimation of camera parameters) implement both an incremental reconstruction algorithm where the primary criteria for the selection of the initial pair are the maximum number of point matches. In VSfM, intrinsic camera parameters such as the focal length are initially retrieved form the EXIF tags and the five-point algorithm (Nister, 2004) is used to estimate the camera poses of the initial image pairs. The projection of the next camera (the image with the highest number of already projected tie points is selected) relative to initial pair is estimated using a direct linear transform inside a RANSAC routine (Hartley and Zisserman, 2004) and the focal length as recorded in the EXIF tags. After a defined number of images has been added, sparse bundle adjustment is used regularly to re-estimate a 8 parameter model for each camera (1 for focal length, 1 for radial distortion, 3 for rotation, and 3 for translation) minimizing the squared reprojection error of all matching points. The minimization problem is solved with a method adapted for large scale reconstruction problems (thousands of images) and combines a classical Levenberg-Marquardt algorithm with the Preconditioned Conjugate Gradients algorithm (Agarwal et al., 2010). For greater efficiency the algorithm is implemented in parallel and the number of full bundle adjustment steps decreases as the model size increases. An important aspect is that all camera parameters (including focal length and radial distortion) can vary among all reconstructed cameras. For a more detailed explanation of the implemented algorithms the reader is referred to (Agarwal et al., 2010; Wu et al., 2011; Wu, 2013).

In Apero the initial pair with the maximum number of mutual matching points is selected and both the eight-point algorithm and the direct linear transform are used within a RANSAC routine (Hartley and Zisserman, 2004) to estimate the relative camera poses (6 parameters derived from the essential matrix). Based on the L1 norms of the angular residual of the projection, the better solution is selected. The next image (the one with the highest number of already projected tie points is selected) is added and its pose is estimated relative to the already reconstructed views using the three-point algorithm (Haralick et al., 1994) within a RANSAC routine. The focal length at this stage is extracted from the EXIF tags and lens distortion parameters are set to zero. After a certain number of images has been added (default maximum 4), a bundle adjustment is performed through a Levenberg-Marquardt optimization minimizing the squared reprojection error of all matching points. During this step only the pose parameters are refined. When all images have been added in the reconstruction, a subsequent sequence of bundle adjustment iterations is launched to estimate the lens parameters according to a lens model selected by the user. Apero implements a number of different lens models and for the presented study a model with ten degrees of

freedom (1 for the focal length, 2 for the distortion centre, 3 for the radial distortion, 2 for the decentric distortion, and 2 for affine parameters) was used. A few final iterations are used to re-estimate the camera poses with the determined lens model and apply stricter criteria for the removal of outliers in the sparse point cloud. It is generally recommendable to perform auto-calibration only on a subset of the image sequence to avoid that cumulating matching errors propagate into the lens model. With Apero this issue can be addressed by performing the explained calibration step initially with a subset of suitable images (typically 10-15), and adding all further views incrementally starting from the previous estimated poses and a fixed lens model. During the reconstruction, the global and per-image residual errors and the used number of point-matches are reported and can serve as indicators to identify the overall quality of the sparse reconstruction and unstable views. For further details the reader is referred to (Deseilligny and Clery, 2011; Pierrot-Deseilligny and Clery, 2012; Deseilligny et al., 2013).

3.5.2.2 CMVS-PMVS

PMVS is a dense matching algorithm comprising three steps of matching, patch expansion and filtering. The matching step is performed by point-wise matching of Harris and Difference-of-Gaussians feature points. A patch of a certain size (in pixels in the reference image) is than defined around each matching point and the depth and orientation of the surface are estimated minimizing a discrepancy measure under epipolar constraints. The discrepancy measure is based on the normalized cross-correlation (NCC) of the patch intensity values among all images used for its reconstruction. Only cameras whose view direction is not too different from the direction of the estimated surface normal and whose photo-discrepancy measure is not too high are used.

In the patch-expansion step, neighbouring cells of already reconstructed patches are probed. The surface normals are initialized from the already determined neighbour and the depth is initialized according to the ray intersection of the included images (images in which the patch is visible). Orientation and depth of the patch are re-estimated minimizing the photo-discrepancy measure and further views in which the new patch is visible according to the estimated depth map are added.

The final step concerns the filtering with criteria combining visibility constraints, the photo-discrepancy score, and a form of regularization where the projection of a patch in all images should at least contain 25% of its projected neighbours. The expansion and filtering steps are typically iterated three times. A more thorough description of the algorithm is provided in Furukawa and Ponce (2010).

Furukawa et al. (2010) proposed CMVS to cluster the output of SfM reconstructions before dense matching. The algorithm forms clusters of well-connected views (sharing many matching points) and removes redundant views while preserving good coverage of the scene. The maximum size of the cluster can be specified by the user in order to split the reconstruction in subsets that can be handled by the hardware at hand.

PMVS requires undistorted images and both VSfM and MicMac include tools to undistort the input images with the determined lens model. Preliminary test were performed after the first image acquisition testing different parameter settings. Considering hardware constraints, model coverage and the amount of visually apparent errors the set of parameters summarized in Table 3-5 was found to provide the best results possible with the hardware available for this study.

Table 3-5: Overview of the main parameter settings used for CMVS-PMVS.

Parameter	Value (default)	Explanation
CMVS cluster size	40 (-)	Larger view clusters require more than 36 GB RAM
Level	0 (1)	Do computation with full resolution images
csize	2 (2)	At least one patch in every csize x csize pixel square region is reconstructed. Smaller csize can increase the point density but required more than 36 GB RAM.
Patch size	14 (7)	Slower but more robust reconstruction
n_{Imin}	3 (3)	Point which are visible in less n_{Imin} images are removed
t	0.7 (0.7)	Patches with a normalized cross correlation score below this value are rejected

3.5.2.3 Multi-Image Correlation MicMac

The core library of MicMac is a hierarchical MVS image correlation algorithm for dense reconstruction. Several aspects of the algorithm are described in Deseilligny et al. (2013), whereas here only a high-level overview of the main steps and parameters when working in "ground-image geometry" is provided. One base image I_b and at least two matching images I_m must be selected and the 3D object surface will be estimated as the depth relative to I_b. Matching is performed projecting squared patches W_s of a size s * s pixel between all images involved according to the known epipolar geometry constraints (Figure 3-15a) and the minimum and maximum depth Z_{min} and Z_{max}, respectively. The algorithm is initialized with downsampled version of the original images (e.g. 1/32 of the original dimension) and Z_{min} and Z_{max} are estimated from the minimum and maximum distances in the sparse point cloud. The similarity of the projected patches is evaluated via the NCC and the mean correlation coefficient among all image combination is computed for each possible disparity value (inversely related to the depth). Based on the mean correlation coefficient a matching cost C_{match} can be computed, which takes the value 0 if the image patches are perfectly similar and increases with dissimilarity. The range of probed disparity values is discretized according to a parameter Z_{step}.

The goal of this process is to determine for each pixel in I_b a corresponding disparity value D for which the global matching cost C_{glob} is minimal. The resulting disparity map is prone to comprise numerous outliers if C_{glob} is based on the NCC alone and that is why a second term (C_{trans}) is added to the cost function penalizing different disparities among neighboring pixels and promoting the smoothness of the surface. The global cost function is indicated in Eq. 3:

$$C_{glob} = \sum_{k=1}^{N} C_{match} + C_{trans}$$
$$C_{trans} = \alpha * \nabla D$$
$$\nabla D > \nabla D_{max} \Rightarrow C_{trans} = +\infty \qquad \text{Eq. 3}$$

, where C_{trans} is a regularization term that depends on the disparity gradient among neighbouring pixels (∇D). The strength of the regularization is controlled by the factor α.

Surface discontinuities translate into strong changes in the disparity of neighboring pixels and to preserve such features C_{trans} increases to infinity if ∇D exceeds the threshold ∇D_{max}.

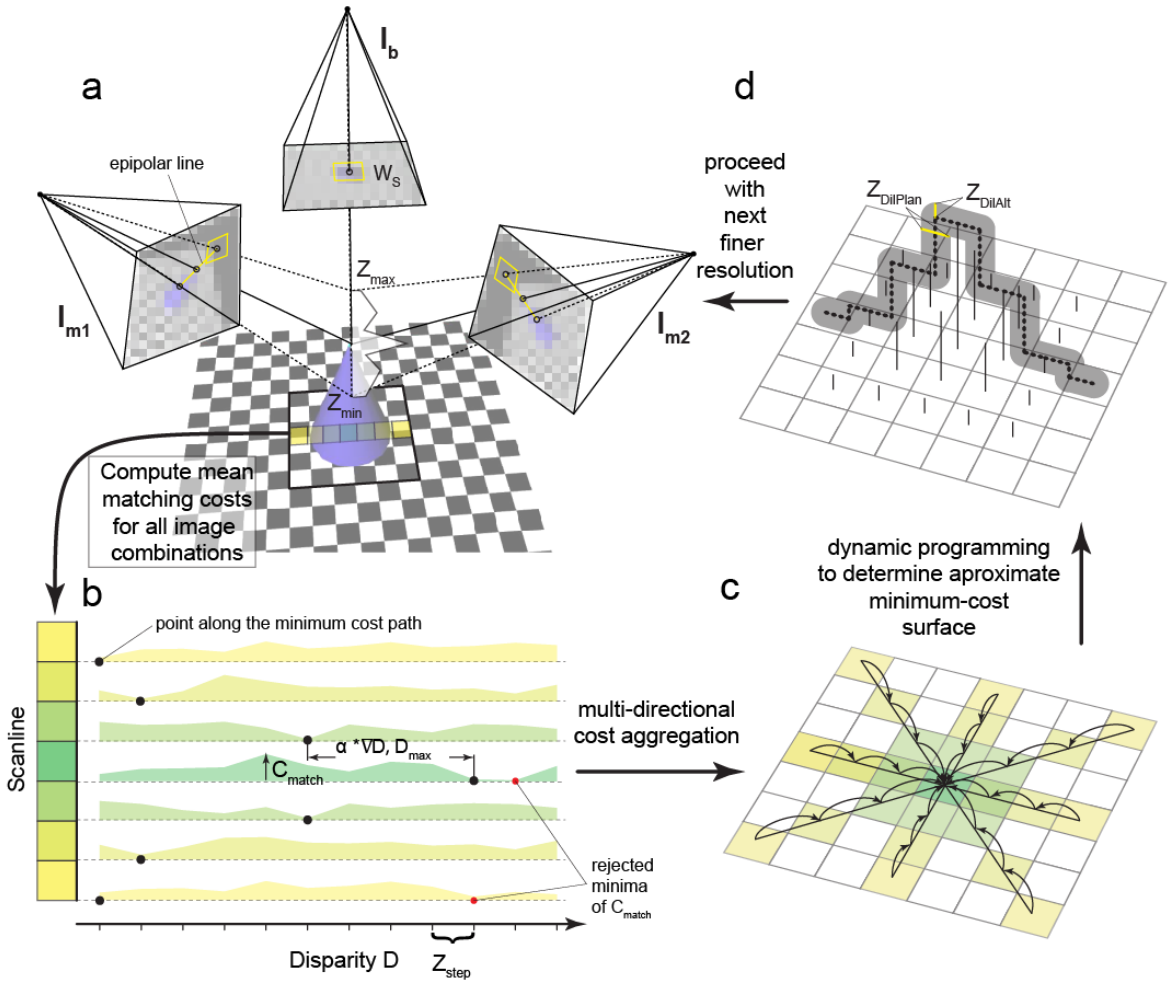


Figure 3-15: Schematic representation of the different steps and parameters of MiMac's dense matching algorithm. (a) Epipolar constraints on the matching according to the determined image orientation, (b) aggregation of matching costs and smoothness term to determine the minimum-cost path along one scanline, (c) multi-directional aggregation of the costs and approximation of the minimum cost surface via dynamic programming, and (d) buffering operation with uncertainties to constrain the search space for the next iteration.

In traditional scanline optimization (Scharstein and Szeliski, 2002), the minimum of C_{glob} is determined as the minimum-cost path in the disparity map along one scanline (Figure 3-15b). While this enables very efficient computation, it does not take into account information from neighbouring pixels outside the scanline and yields streaking effects. This issue can be resolved using the semi-global matching (SGM) technique proposed in Hirschmuller (2008) where C_{glob} is determined through the aggregation of minimum-cost paths from several directions on the disparity map (Hirschmuller, 2008; Deseilligny et al., 2013). The number of different directions should provide good coverage of the image and is typically set to 16 (8 in Figure 3-15c). After the multi-directional aggregation a minimum-cost cost surface can be determined efficiently via dynamic programming.

Starting from downsampled versions of the original image the algorithm determines a first approximation of the surface at a coarse resolution. This approximation serves as a prior to constrain the search during the subsequent step at the next finer resolution. A similar approach that combines SGM with a hierarchical scheme has recently been proposed in Rothermel et al. (2012).

The uncertainties on this prior are expressed by the two parameters Z_{DilAlt} (in multiples of Z_{step}) and $Z_{DilPlan}$ (in pixel of the disparity map). The minimum/maximum disparities are computed in a neighbourhood of $Z_{DilPlan}$ pixel and Z_{DilAlt} is subtracted/added to it, respectively. The resulting two disparity surfaces serve as a bound on the search in the next finer resolution step (Figure 3-15d).

The final iteration is typically carried out on the full resolution images and since the disparity is estimated in integer multiples of Z_{step} a final interpolation step is added to obtain a floating point depth map.

MicMac provides standard procedures for dense matching but enhanced results might be obtained if the parameters are tuned for the reconstruction at hand. Table 3-6 contrasts the parameters of three different schemes that were tested when processing the first acquisition in 20-Oct-11. In the customized routine the matching step on full resolution is repeated three times with a reduced Z_{step} for greater sub-pixel precision, decreasing uncertainty parameters and an adaptive median filter to remove outliers. Also ∇D_{max} is set to a rather low value in order to preserve depth discontinuities that are abundant in the rugged topography. Figure 3-16 shows exemplary results of the dense matching with three different settings and demonstrates that the additional steps in the customized scheme significantly reduce the amount of noise and justify the additional computation time (Table 3-6). It should be noted that the routine 'Standard 2' provides better coverage since the threshold on C_{match} is more conservative. Consequently fewer areas are masked out and the increased runtime can partially be attributed to this. However, the customized scheme was used for all further processing to not compromise the accuracy of the resulting models.

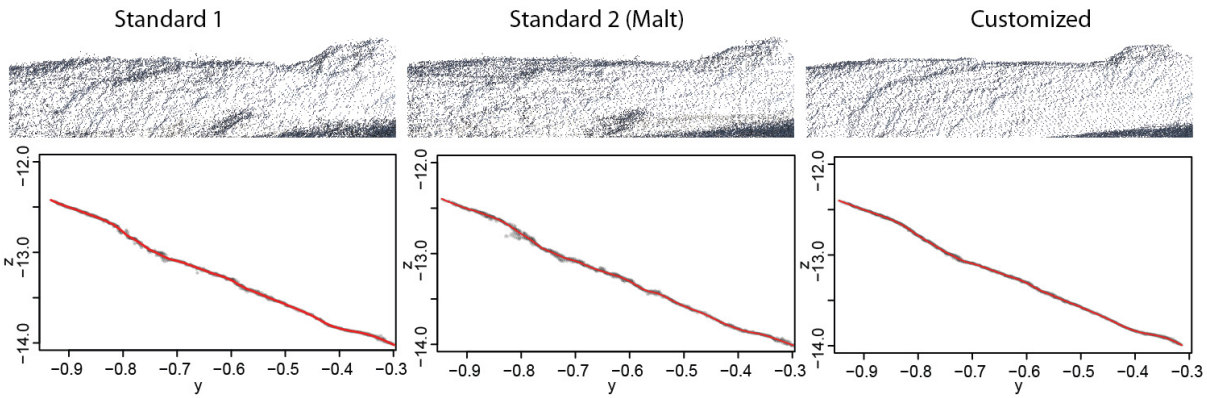


Figure 3-16: Comparison of exemplary point-clouds sections obtained with the three tested matching schemes (Table 3-6) of MicMac. A very smooth section of a ridge at the foot of the scarp is shown (unscaled model for an acquisition at the scarp in October 2011). The red line marks the average surface along the displayed ridge. Note that the scaling of x- and z-axes are in coordinates arbitrarily fixed during the SfM step.

Table 3-6: Overview of parameter settings for three tested MicMac matching schemes (see text for details).

General parameters	**Standard 1**	**Standard 2 (Malt)**	**Customized**
Parameter name	Value	Value	Value
Z_{min}	0.3 * the minimum distance in the sparse point cloud	0.5 * the minimum distance in the sparse point cloud	0.3 * the minimum distance in the sparse point cloud
Z_{max}	3 * the minimum distance in the sparse point cloud	3 * the minimum distance in the sparse point cloud	3 * the minimum distance in the sparse point cloud
NbDir	8	7	8
C_{match} threshold	0.4	0.2	0.4
α	0.05	0.02	0.05
∇D_{max}	3.0	2.0	1.0
Min. number of images	3	3	3
Changing parameters	Z_{step}, $Z_{DilPlan}$, Z_{DilAlt}, s	Z_{step}, $Z_{DilPlan}$, Z_{DilAlt}, s	Z_{step}, $Z_{DilPlan}$, Z_{DilAlt}, s
Resolution			
32	0.5, 3, 4, 3	0.8, 3, 4, 3	0.5, 3, 4, 3
16	0.5, 3, 4, 3	0.4, 3, 4, 3	0.5, 3, 4, 3
8	0.5, 3, 4, 3	0.4, 3, 4, 3 (∇D_{max}=1)	0.5, 3, 4, 3
4	0.5, 3, 4, 3	0.4, 3, 4, 3 (∇D_{max}=1)	0.5, 3, 4, 3
2	0.5, 3, 4, 3	0.4, 3, 4, 3 (∇D_{max}=1)	0.5, 3, 4, 3
1	0.5, 3, 4, 3	0.2, 1, 1, 3 (∇D_{max}=1)	0.3, 3, 4, 3 + adaptive median filter with s=5
1	Interpolation step	Interpolation step	0.3, 1, 2, 5 + adaptive median filter with s=5
1	-		0.3, 1, 1, 3
1	-		Interpolation step
Runtime for the model of the subset in Figure 3-16 (1 base and 6 matching images)	303.5 s	435.4 s	559.3 s

Currently the base images still have to be selected manually. The MicMac library comprises optional tools to create masks constraining the area that should be modelled. Ideally, the selected base images should provide a normal view of the surface, have sufficient neighbouring views with support from multiple directions and provide good coverage of the scene. For the selection of the matching images the library provides a tool which, based on the view angle and coverage relative to the base image, automatically determines a ranking of the most suitable matching images. In general the six highest ranked images were selected, whereas in few cases it was necessary to adjust the selection to account for occlusion and large differences in the object distance.

3.5.2.4 Georeferencing, accuracy assessment and change detection

The dense point clouds from all three pipelines (VSfM+CMVS, Apero+CMVS, Apero+MicMac) were visually inspected (Cignoni et al., 2008) and obvious outliers (point patches far from the average surface, unnatural correlation artefacts) were manually removed. Both CMVS and MicMac typically yield overlapping point clouds originating from multiple view clusters and base images, respectively. Redundant points in those overlapping areas were removed and preference was given to the point clouds with higher density.

The models of the scarp for which no GCPs were available were referenced directly after the generation of the dense-point clouds. The models for 20-Oct-2011, 5-Jul-2012, 9-Oct-2012 were georeferenced with respect to the corresponding terrestrial LiDAR acquisitions (using Polyworks, (InnovMetric, 2010)). For a coarse alignment (rotation, translation, scaling), four homologous points were selected manually and the ICP algorithm was used to refine the alignment (rotation, translation). The model of the scarp for 14-May-2013 was georeferenced by the same procedure but the SfM-MVS model of the previous date was used as a reference.

The georeferencing of the remaining models was performed with GCPS using the Campari tool which is part of the MicMac library. Campari is an interface to the Apero algorithm and allows integrating both tie points and GCPs in the bundle adjustment. A related graphical interface is used to mark the position of the GCPs directly in the images. The uncertainty of the ground and image measurements is provided by the users and constitutes weights for the least-square optimization. Table 3-9 (Section 3.5.3) summarizes the number of involved GCPs and their respective uncertainties which were set according to the errors of the respective reference measurements.

To assess the accuracy of the point clouds resulting from the three SfM-MVS pipelines, the datasets were compared against the terrestrial LiDAR scans taken at the same date. One of the most commonly used approaches for point-cloud comparison is to interpolate a surface from one of the two point clouds and compute the distance along the surface normal (cloud-to-mesh, C2M). This approach has proven to provide very accurate measurements but poses problems in areas where the point clouds do not overlap (due to inaccurate interpolation of the surface). To obtain accurate measurements is it hence necessary to remove all non-overlapping areas during time-consuming manual editing.

Lague et al. (2013) proposed a technique (M3C2) that does not require an explicit surface interpolation but estimates the surface normals and distances directly from the point clouds. In a first step the surface normals at each point are computed considering all points in neighbourhood of the size D/2 (the scale factor). The normals can be computed from one reference cloud (typically the earlier date) or averaging the normals of the two point clouds. Subsequently the distances are calculated at each point as the mean distance between all points in a neighbourhood of the size d/2 (the projection scale). An additional parameter p_{max} defines the length of the projection cylinder and thereby the maximum distance of points considered in the computation. The scale D should be chosen sufficiently large to make the distance measurement independent of the surface roughness and a range of 0.3 m < d < 2m for the projection scale is recommended (Barnhart and Crosby, 2013; Lague et al., 2013).

For the accuracy assessment, we focused on the first three acquisitions at the scarp and tested both comparison methods (C2M, M3C2). In order to obtain accurate and comparable results non-overlapping parts of the point clouds were removed manually beforehand. The C2M implementation in CloudCompare (EDF-R&D, TelecomParisTech, 2013) was used and the LiDAR point clouds were interpolated to form the reference mesh using a Delaunay triangulation with a local least-square plane approximation. The parameters for the M3C2 algorithm were adapted according to the surface roughness and the maximum

expected changes. For the accuracy assessment the surfaces are relatively similar and hence a rather small scale D and p_{max} are sufficient. Since the LiDAR point cloud is assumed to provide the more accurate surface representation, it is reasonable to use it for the estimation of the normals. As the general shape of the surface varies considerably over time larger scale parameters were selected to promote a greater robustness of the distance calculation. For the same reason the normals were estimated as the average from both point-clouds. The calculation of the confidence interval is based on the local surface roughness and the registration error which can be set to account for both referencing and measurement errors (see Lague et al., 2013 for further details). Table 3-7 provides an overview of the parameter settings. An interesting feature of the M3C2 algorithm is the possibility to estimate a 95%-confidence interval for significant changes between two point clouds.

Table 3-7: Parameter settings for the M3C2 algorithm for the accuracy assessment and the change detection.

M3C2 parameter	Accuracy assessment: target scarp	Change detection: target scarp	Change detection target: landslide body
D	5 m	10 m	15 m
d	0.5 m	1.0 m	1.5 m
p_{max}	5 m	7 m	7 m
Normal computation	LiDAR	Average normal from both point clouds	Average normal from both point clouds
Registration error	-	0.1 m	0.2 m

3.5.3 Accuracy assessment and comparison of different open-source SfM-MVS pipelines

An overview of some quality indicators for the reconstructed dense point clouds is provided in Table 3-8. As a general strategy for the use of Apero, 10-15 images were selected for lens self-calibration and the bundle adjustment was repeated several times adding images successively, while keeping the lens model fixed and using the estimated poses from the previous run for initialization. It was targeted to integrate all collected images, however, views with high residuals, blur or unfavourable illumination conditions were removed during a final bundle adjustment aiming at a reduction of the residuals below 0.5. For the full scene acquisition of 9-Oct-12 a block of 69 images remained unconnected and could not be used for the modelling. Also with VSfM it was possible to integrate most of the images for the first three surveys of the landslide scarp. Since no residuals are reported by the tool no refinement was carried out. An attempt to reconstruct the full scene of 9-Oct-12 resulted in several disconnected blocks. Since none of them provided satisfactory coverage of the landslide dense matching was not pursued. In Table 3-8 also the number of points is compared illustrating that MicMac provides a significantly denser point clouds (and coverage, see section 3.5.4.1).

To evaluate the accuracy of the three tested SfM-MVS workflows the results were compared against the corresponding terrestrial LiDAR scans and the mean-absolute error (MAE), the root-mean-squared error (RMSE) and the mean distance (MD) between the two clouds are reported. While MD does not reflect the accuracy of the 3D model it was found to be a useful indicator for a potential bias after scaling and co-registration via ICP. Figure 3-17 shows that the MD is at most 1.3 cm and generally below 0.07 cm which is in all cases only a minor fraction of the RMSE. The results of the direct georeferencing with Apero are reported

in Table 3-8 and the residuals were found to be within the uncertainty of the GCP sources (Table 3-4). A visual inspection of stable areas in the full scene photogrammetric models of 19-Jul-13 and the airborne LiDAR point clouds revealed a small systematic residual misalignment. This could be attributed to inaccuracies in the georeferencing of the airborne LiDAR data which generally dominates the absolute error in the point location (Ussyshkin and Smith, 2006). To address this issue the ICP algorithm was employed to align all point clouds to the full scene photogrammetric reconstruction of 19-Jul-13. In general we found direct georeferencing during bundle adjustment to be the more convenient strategy since it is easier to locate GCPs on the photos than within the point clouds and further processing within additional software can be avoided.

Table 3-8: Quality indicators for the sparse reconstructions and dense point clouds.

Model	Scarp 20-Oct-11	Scarp 5-Jul-12	Scarp 9-Oct-12	Full scene 9-Oct-12	Scarp 14-May-13	Scarp 18-Jul-13	Full scene 19-Jul-13
Reconstructed views (total) with Apero	84 (88)	106 (106)	142 (168)	332 (401)	130 (130)	244 (265)	401 (401)
Global residuals (pixel)	0.458	0.325	0.497	0.460	0.474	0.496	0.433
Reconstructed views (total) with VSfM # points	82 (88)	106 (106)	164 (168)	209 (401)	-	-	-
VSfM+CMVS	1,646 k	12,679 k	12,645 k	-	-	-	-
Apero+CMVS	1,693 k	15,563 k	15,843 k	-	-	-	-
Apero+MicMac	14,233 k	27,613, k	23,628 k	44,395 k	49,971 k	67,042 k	48,378 k

Table 3-9: Summary of the direct georeferencing with GCPs.

	Full Scene October 2012	Full Scene July 2013	Scarp July 2013
GCP source	Orthophoto+Aerial LiDAR (Salient objects on stable areas)	dGPS measured targets	dGPS measured targets
Number	31	35	11
Ground uncertainty	0.25 m	0.05 m	0.05 m
Image uncertainty	1 pixel	0.5 pixel	0.5 pixel
Residuals after compensation	0.32 ± 0.20 m	0.05 ± 0.04 m	0.10 ± 0.05 m

The two tested methods for the distance computation (C2M, M3C2) provide very similar results but M3C2 results in slightly lower errors estimates (the maximum RMSE difference between both was 1.6 cm). Manual masking of non-overlapping areas is a pre-requisite for the application of C2M and higher error estimates are likely if small non-overlapping areas have not been recognized correctly. Consequently the M3C2 was considered as the more reliable estimator and used for change detection throughout the rest of the study.

The RMSEs of the obtained models (Figure 3-17) were between 2.7 cm and 7.5 cm. Among the different workflows Apero+MicMac provided the most accurate point clouds (RMSE 2.7 cm - 5.6 cm) and VSfM+CMVS yielded slightly higher errors (RMSE 5.7 – 7.5 cm). In this context it is worth noting that the uncertainty of the terrestrial LiDAR is within a range of 2-3 cm and at an RMSE of 2.7 cm it is, therefore, not possible to state with certainty if the LiDAR or the photogrammetric point cloud is more accurate. Apero+CMVS provided accuracies similar to VSfM+CMVS except for 5-Jul-12 where it improved the RMSE by 1.3 cm. This is probably related to the fact that a 35 mm lens with greater distortion was used for this acquisition and the more complex lens model used in Apero generally compensate the distortion better than the simple lens model used in VSfM. The acquisition with the 35 mm lens yielded the most accurate results indicating that the reduced ground sampling distance with a wider angle lens was not a significant factor. The greater overlap among the different views provided a better redundancy of the tie-point network and probably promoted the more accurate pose estimations (cf. residuals in Table 3-8).

In general all three pipelines provided accurate models for the scarp and considering the high degree of automation of the VSfM+CMVS the pipeline still provides very competitive results. However, since the full scene reconstruction did not succeed with VSfM+CMVS and Apero+MicMac provided the more accurate results it was used for the processing of all remaining models.

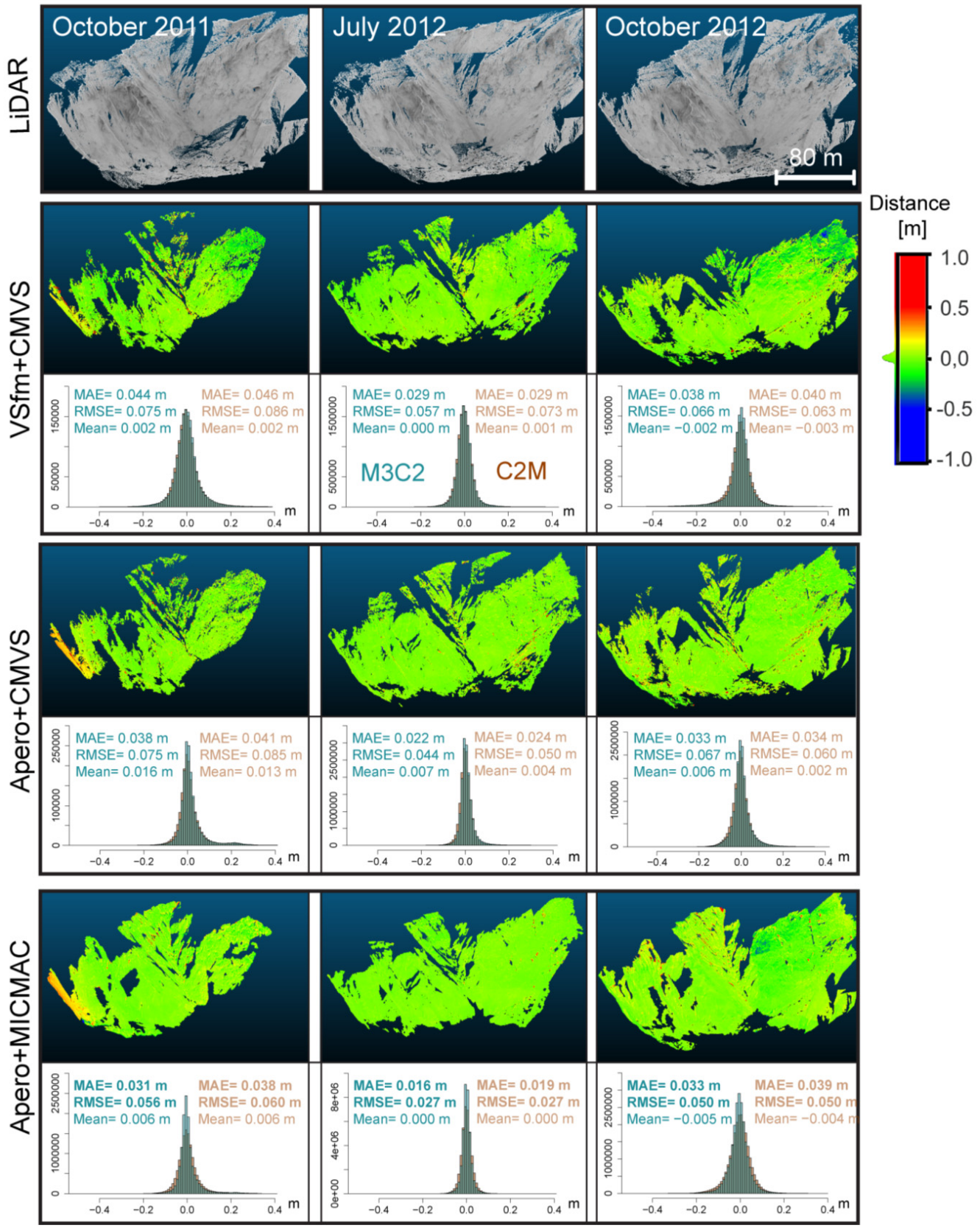


Figure 3-17: Comparison of three SfM-MVS pipelines (VSfM+CMVS, Apero+CMVS, Apero+MicMac) against terrestrial LiDAR scans. MAEs, RMSEs, and mean differences from two different point-cloud comparison methods are reported.

3.5.4 Temporal evolution of the Super-Sauze landslide

In this section, the results of the change detection (M3C2) are reported and relationships with geomorphological processes during the monitored period are discussed. The possibility to obtain volume estimates and 3D displacement vectors are illustrated exemplarily.

3.5.4.1 Quantification and interpretation of retrogression at the main scarp

Over the monitored period, significant chances were observed for all time intervals and the maximum detected differences amounted to between -5.20 m (ablation, Figure 3-21a) and 7.13 m (accumulation, Figure 3-20a). According to the co-registration error set to 0.1 m (Table 3-7), the minimum significant change at 95% confidence was 0.2 m (Figure 3-18b, Figure 3-19b,Figure 3-20b, Figure 3-21b).

Figure 3-22a provides a synoptic view of the scarp indicating areas of major changes that will be discussed in further detail.

The first period (8.5 months, autumn-early summer) was dominated by ablation resulting from rockslides, rockfalls and runoff on the scree slopes at the base of the scarp. The source of a medium size rockslide could be detected at the eastern end of the scarp and its volume was estimated at approximately 1090 m^3 (compared to 1050 m^3 from terrestrial LiDAR, Figure 3-18a).

During the following period (~3 months, summer to autumn), only minor rockfalls and rockslides occurred. Some minor ablation that was registered at the central part could be traced back to errors in the photogrammetric model of 05-07-2012 since it appeared as a positive change during the subsequent period (compare Figure 3-19a, Figure 3-20a).

The subsequent period (~5 months, autumn to spring) was characterized by important activity constituted by major rockfalls, rockslides, runoff and accumulation of the resulting sediments at the base of the scarp (Figure 3-20a). The major rockslide that occurred at the western part of the scarp was only partially covered by the photogrammetric models but a lower bound for the released volume was estimated at approximately 2530 m^3. The high magnitude of changes is consistent with the prolonged winter and high snow accumulation resulting in extensive thaw-freeze cycles and an abundance of melt-water in early spring.

The last period (~2 months, summer-autumn) was again dominated by ablation resulting from runoff induced erosion of the scree slopes. The rockslide zone at the western part of the scarp, which was activated during the previous time interval, preserved a strong activity (1350 m^3; Figure 3-21a) and some minor rockfall events could be detected.

In summary the change detection provided a detailed picture of the surface evolution and main processes at the scarp (Figure 3-22). Residual false detections could be excluded through the interpretation of successive periods and the overall accuracy of the method was found satisfactory for process monitoring and volume estimation.

It is noteworthy that owing the complex surface topography the coverage obtained from the initial surveys was not fully satisfactory but could be successively increased building on the experience of previous surveys (Table 3-10).

Table 3-10: Coverage of the photogrammetric models of the main scarp relative to the terrestrial LiDAR surveys

Date	20-Oct-11	5-Jul-12	9-Oct-12	14-May-13	18-Jul-13
Coverage (% of projected LiDAR points)	42.9%	63.0%	76.0 %	72.7%	80.7%

The coverage has been computed as the percentage of terrestrial LiDAR points that could be projected on the photogrammetric point clouds using the M3C2 algorithm. The last survey indeed resulted in a point cloud that closely approaches the coverage of the terrestrial LiDAR surveys and encourages further use of SfM-MVS for the monitoring of complex surfaces in high-mountain environments.

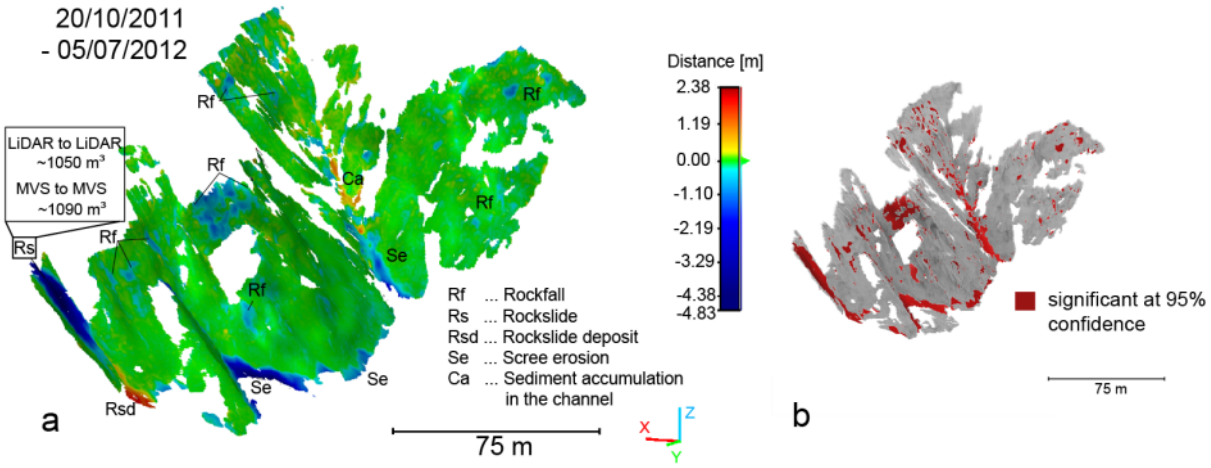


Figure 3-18: (a) Photogrammetry-based change detection and geomorphological interpretation for the period 20-10-2011 till 05-07-2012. (b) SfM-MVS reconstruction for the 20-10-2011 and significant changes.

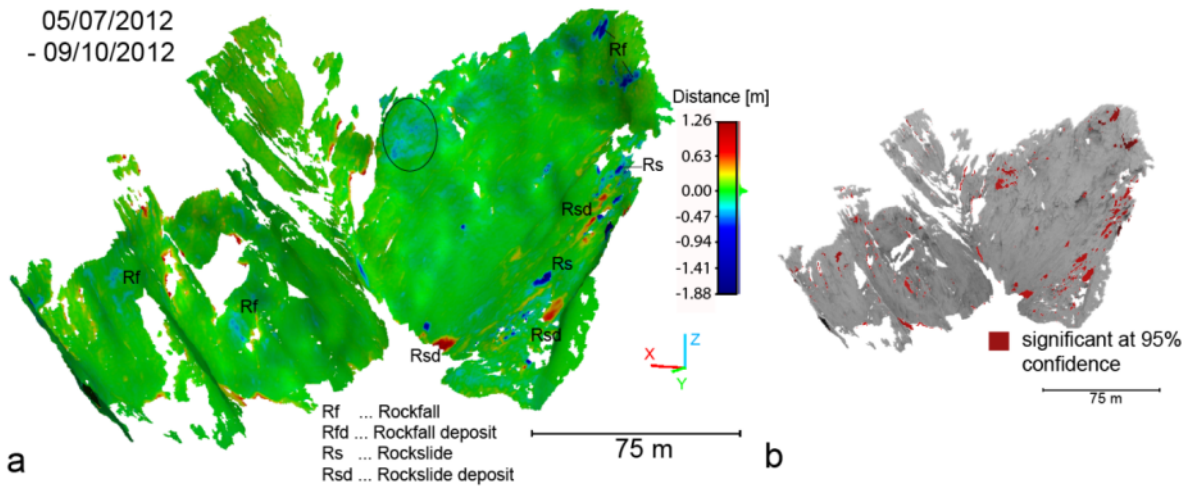


Figure 3-19: (a) Photogrammetry-based change detection and geomorphological interpretation for the period 05-07-2012 till 09-10-2012. Some false detections resulting from errors in the model of 05-07-2012 are indicated (black ellipse). (b) SfM-MVS reconstruction for the 05-07-2012 and significant changes.

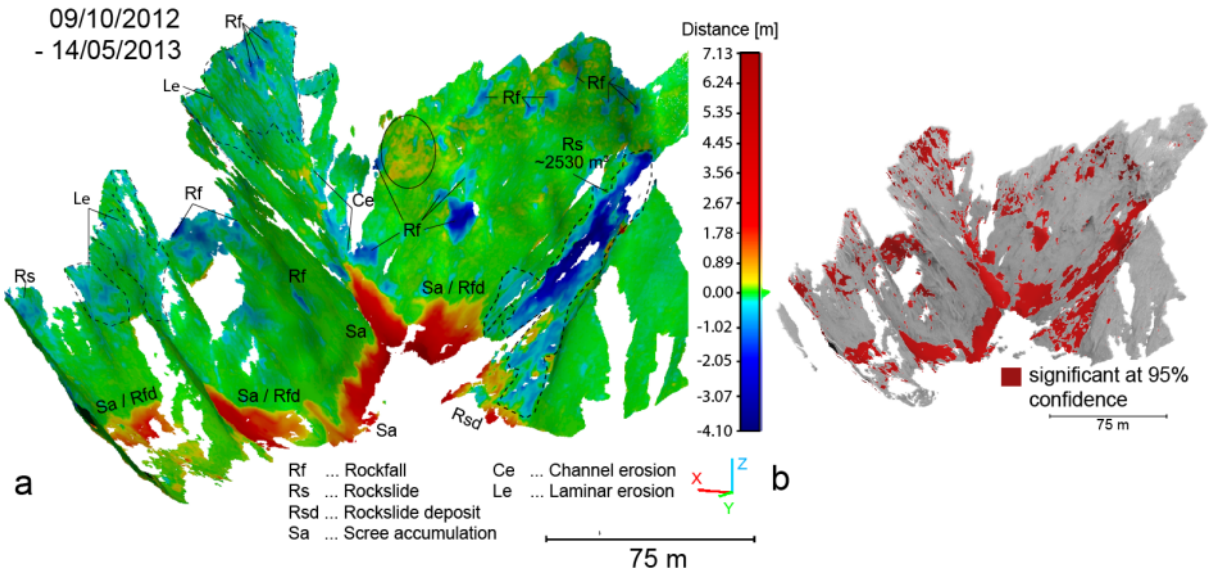


Figure 3-20: (a) Photogrammetry-based change detection and geomorphological interpretation for the period 09-10-2012 till 14-05-2013. Some false detections resulting from errors in the model of 05-07-2012 are indicated (black ellipse). (b) SfM-MVS reconstruction for the 05-07-2012 and significant changes.

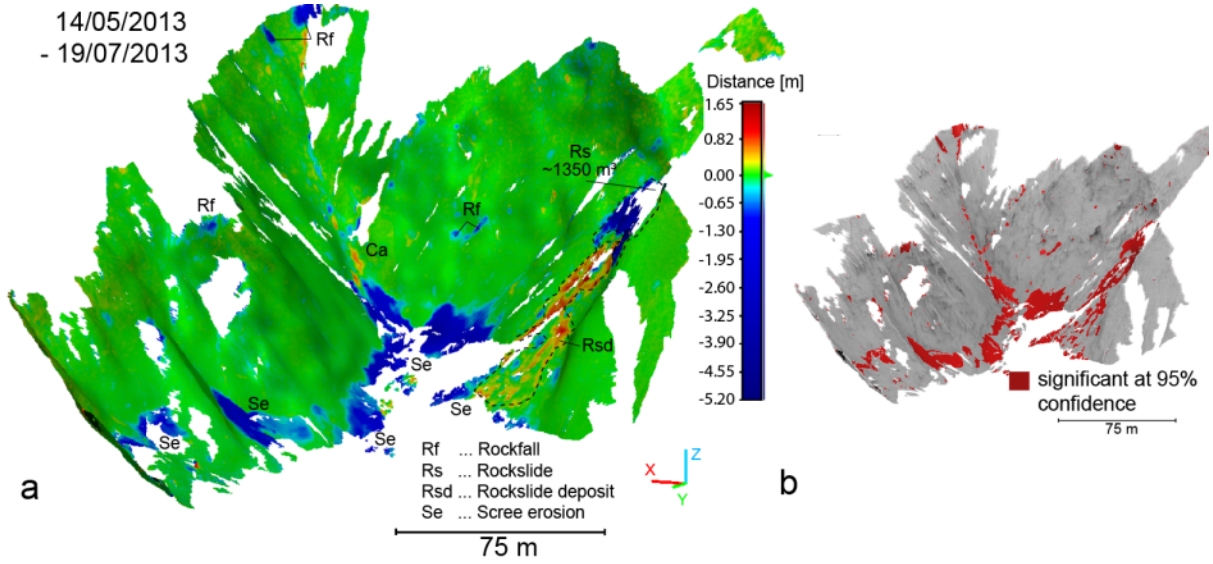


Figure 3-21: (a) Photogrammetry-based change detection and geomorphological interpretation for the period 14-05-2013 till 19-07-2013. (b) SfM-MVS reconstruction for the 14-05-2013 and significant changes.

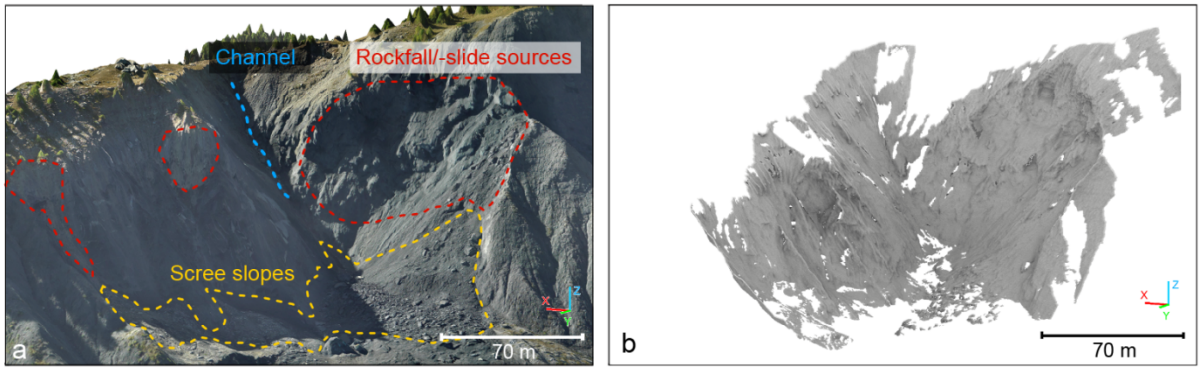


Figure 3-22: (a) Synthetic view (Aerial LiDAR and Orthophotograph 29-08-2012) of the scarp. Areas with major detected changes are indicated. (b) SfM-MVS reconstruction for the 19-07-2013.

3.5.4.2 Quantification and interpretation of the landslide dynamics

Preliminary change detections between the photogrammetric models and the aerial LiDAR over stable zones close to the landslide showed an RMSE between 0.08 m and 0.17 m for the full scene models (Figure 3-23, Figure 3-25). Consequently, a co-registration error of 0.20 m was used for the computation of the confidence interval during the change detection with the full point clouds (see also Table 3-7). Three change detections were carried out. One between the aerial LiDAR and the MVS point cloud of 10-Oct-2012, a second between the two MVS point clouds of 10-Oct-2012 and 19-Jul-2013 and a third one to capture the cumulative displacement between the date of the aerial LiDAR acquisition (29-Aug-2012) and the 19-Jul-2013. Their results are presented in Figure 3-23, Figure 3-24 and Figure 3-25, respectively. At an estimated co-registration error of 0.2 m the minimum displacement that can be detected at 95% confidence is 0.4 m. The maximum measured change over the full monitored period (324 days) varied between -6.88 m and 4.77 m and shows a strong activation of almost the entire landslide as well as a major rockfall at the scarp and corresponding accumulation (Figure 3-25). While the area directly below the main scarp remained relatively stable, a secondary scarp formed several meters further downslope and currently constitutes the most active part of the landslide. This secondary scarp was already active during the first 42 days (29-Aug-2012 till 10-Oct-2012) and downslope displacement induced surface changes between -3.52 m and 2.50 m. Minor detected changes include rockfalls and scree accumulation at the main scarp as well as traces of a translational failure in the central part of the landslide body (Figure 3-23). Several further areas were classified with significant changes but a closer inspection of the photogrammetric point cloud (10-Oct-2012) revealed errors in the reconstruction as their origin. Three respective areas are indicated in Figure 3-23. They have in common that they were all imaged from large distances and relatively oblique incidence angles (<30˚) at which fine structures such as small thalwegs and gullies could not be carved out correctly. To address the issue dedicated images for those error prone areas were recorded during the second acquisition (19-Jul-2013) and enabled to reduce the amount of artefacts and resulting false positive changes (Figure 3-25).

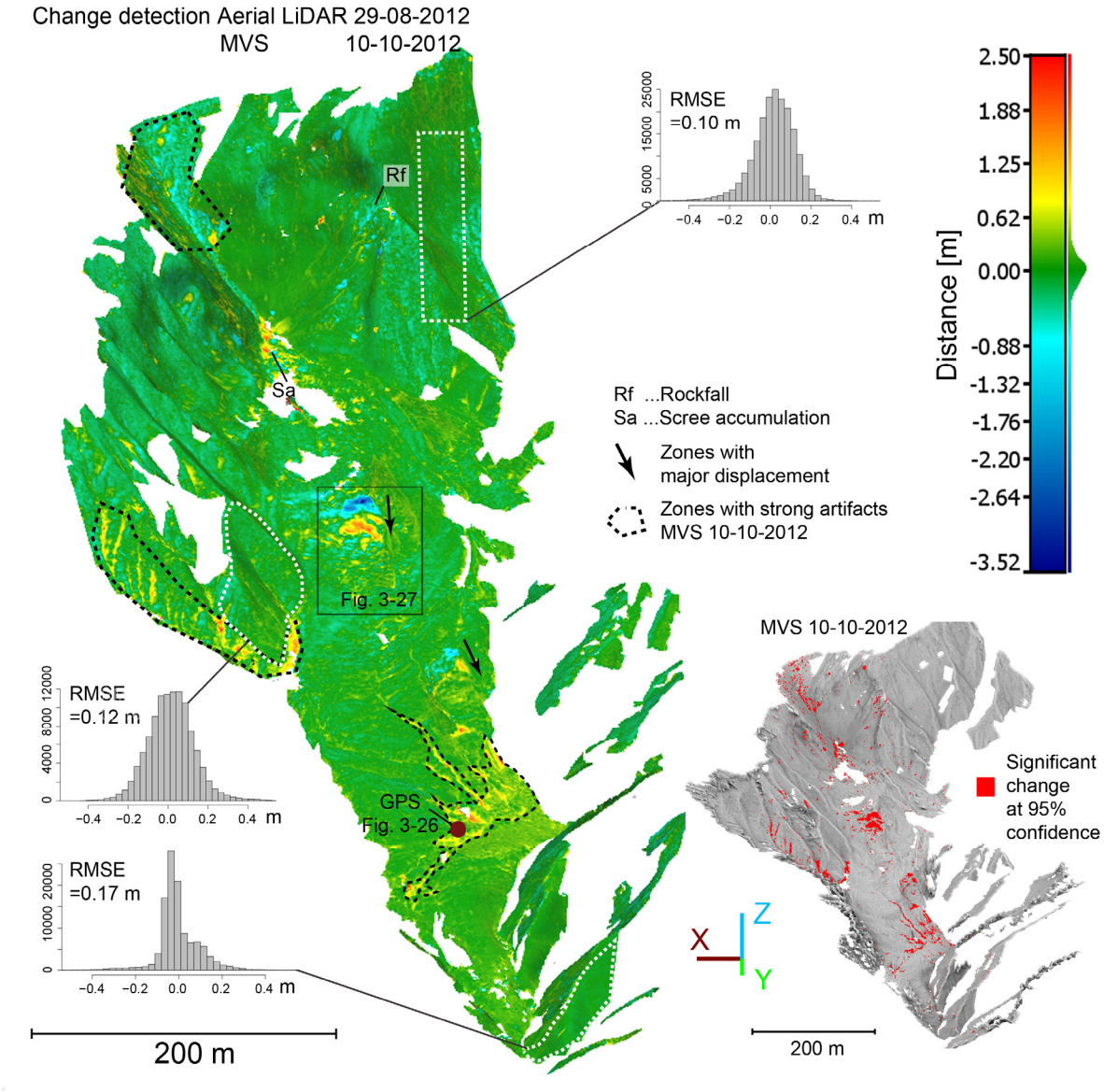


Figure 3-23: Change detection results (29-08-2012 till 10-10-2012) for the full landslide and their geomorphological interpretation. The inserted histograms show the residual error between the photogrammetric point cloud and aerial LiDAR. Vegetated areas were masked out for better visualization.

The activity at the secondary scarp increased in the 282 days period between 10-Oct-2012 and 19-Jul-2013. The general change pattern indicates a strong displacement from the central part of the landslide (mainly negative distances) towards the lower part (mainly positive distances, Figure 3-24a). Notable features include the progression of the previously initialized translational failure (Figure 3-24b), lobes that displayed coherent downslope movement (Figure 3-24 c) and a significant advancement of the landslide toe. The depicted rockfalls and accumulations resemble the results of the change detection at the main scarp (compare Figure 3-20, Figure 3-21). Volumes released from two source areas were estimated at 260 m^3 and 3760 m^3 (Figure 3-24a), which proved to be consistent with estimates from the scarp models for the same areas and time periods (230 m^3 and 3710 m^3, respectively). Errors in MVS model of 10-Oct-2012 hindered to track the evolution of deeply incised gullies at the

central part of the landslide for the two sub-periods (Figure 3-23, Figure 3-24). However, reconstruction of the respective area could be significantly enhanced with the photogrammetric model of 19-Jul-2013 and the cumulative change detection indicates an incision of the central gully of up to 0.8 m (Figure 3-25a).

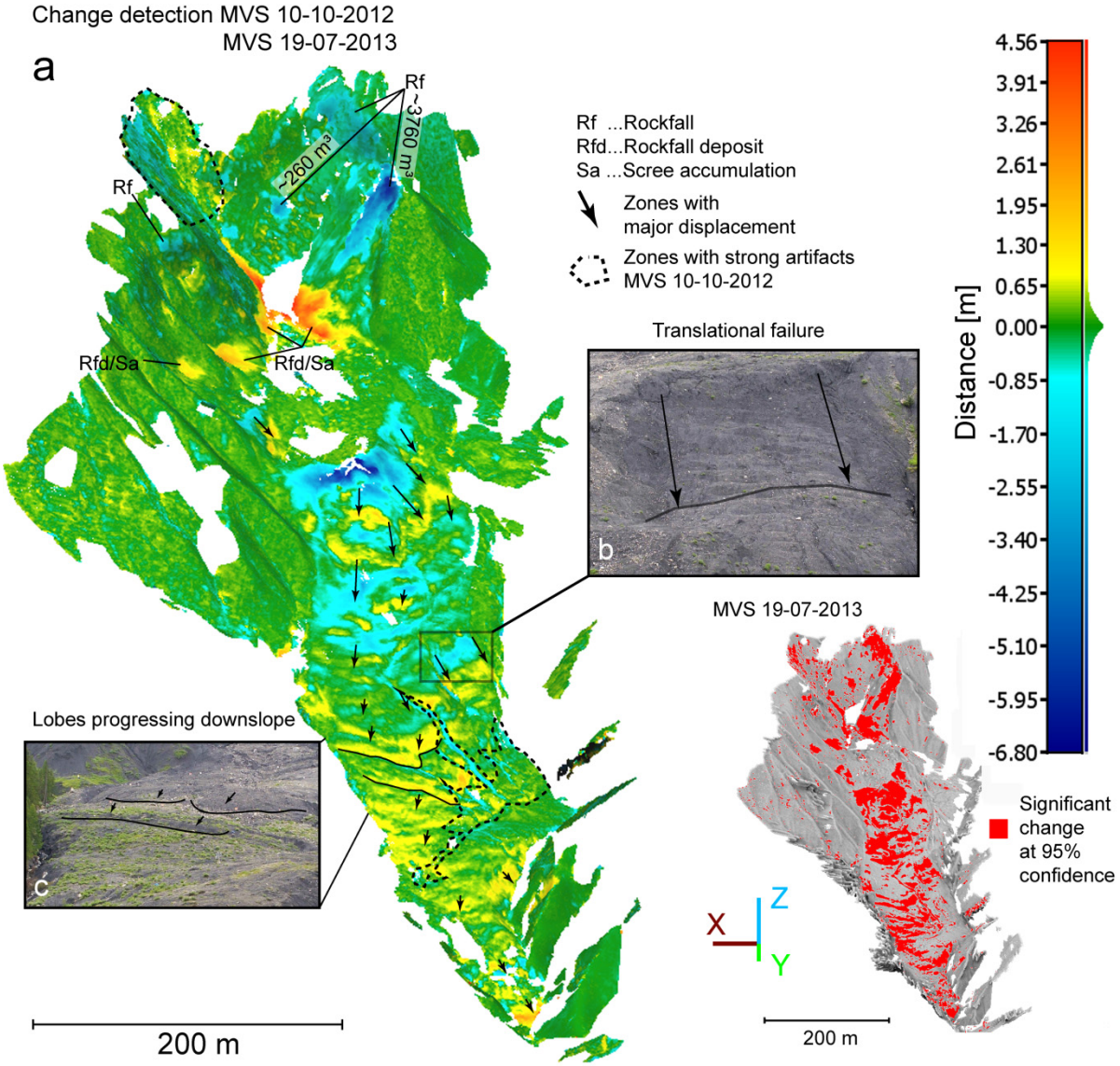


Figure 3-24 (a) Change detection results (10-Oct-2012 till 19-Jul-2013) for the full landslide, and their geomorphological interpretation. Vegetated areas were masked out for better visualization. Subsets of a terrestrial photograph show (b) the final state of a detected translational failure and (c) lobes that displayed coherent downslope movement.

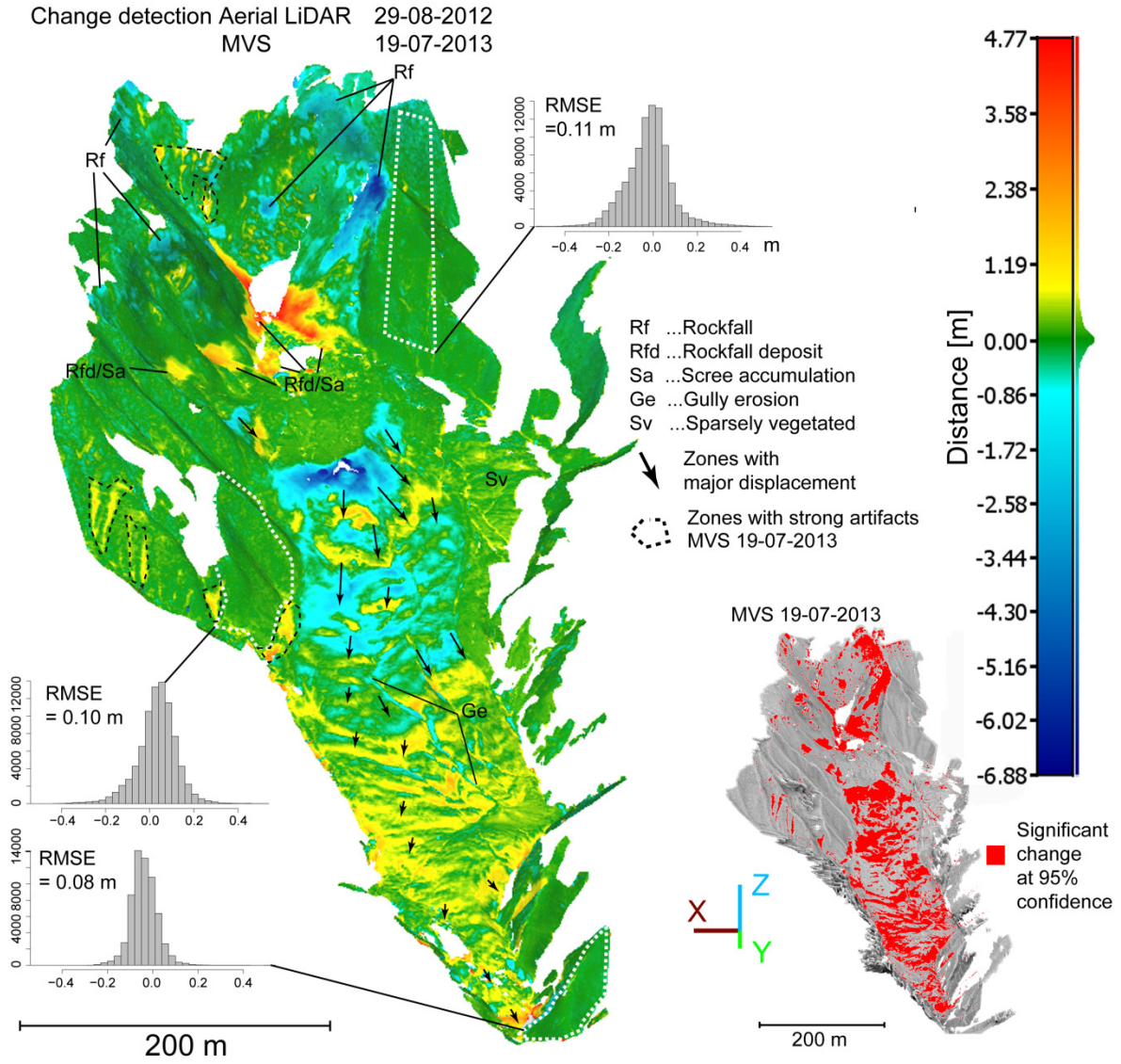


Figure 3-25: Cumulative change detection results (29-Oct-2012 till 19-Jul-2013) for the full landslide, and their geomorphological interpretation. The inserted histograms show the residual error between the photogrammetric point cloud and aerial LiDAR. Vegetated areas were masked out for better visualization.

The general pattern of the observed movement is consistent with previous studies of the landslide dynamics (Malet et al., 2002; Travelletti et al., 2012). However, due to the flow-like behaviour of the landslide, the largest component of the 3D displacement is typically parallel to the slope and hence an important component of the displacement is not comprised in the distances measured normal to the surfaces. To illustrate the possibility to quantify full 3D displacement vectors two sub-areas where analysed in greater detail. One analysis focuses on the most active part at the secondary scarp and a second analysis targets to validate the obtained displacement vectors through a comparison with a permanent GPS stations installed on the landslide. The rather strong displacement during the monitored period affected two out three GPS station and consistent measurements thus were only available from one of the stations. The respective subsets are indicated in Figure 3-23.

To measure 3D displacement at the exact position of the GPS the corresponding points in the aerial LiDAR point clouds were aligned manually (simple translation) to the position of the GPS as depicted in the MVS point clouds. The same procedure was performed for points on a pillar with solar panels next to the GPS. Displacement vectors are obtained from the point coordinates before and after the transforms. The results of this analysis are shown in Figure 3-26 indicating a total displacement of 3.62 m measured among the different point clouds, whereas the permanent GPS indicates a total displacement of 3.15 m. The 0.47 m difference accumulates errors in the absolute geolocation, the co-registration, the point clouds and errors induced by the matching process. The residual georeferencing error of the MVS model for 19-Jul-2013 was only 0.05 ± 0.04 m and the relative RMSE over stable areas after alignment only 0.17 m maximum. To understand the comparatively high error of the displacement measurement it is important to note that fine structures such as the GPS station are generally over-generalized due to the regularization in the photogrammetric modelling. This was especially true for the model of 10-Oct-2012 in which the GPS position was imaged from a relatively remote position (~ 700 m). It explains the greater deviation for the period 29-Aug-2012 till 10-Oct-2012. Thus the relative error of 0.47 m can be considered as an upper bound for the displacement estimates for the central most active part where matching was performed on well-defined boulders using the ICP algorithm for fine registration.

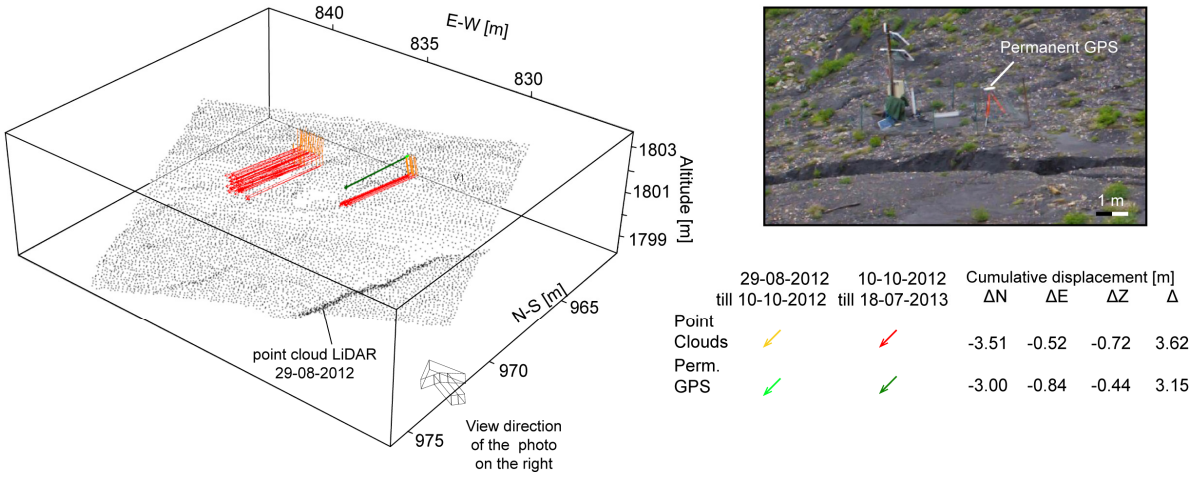


Figure 3-26: 3D displacement vectors measured via point matching among the different point clouds and a comparison with permanent GPS measurements.

The 3D displacement measurements for the central part of the landslide were carried out through piecewise alignment (translation and rotation) of rigid blocks that could be identified in at least two of the three point clouds. The ICP algorithm was used for fine registration, whereas, if the distances were larger than 1 m, an initial guess for the translational component was provided manually. The measurements for the period 29-Aug-2012 till 10-Oct-2012 suggest a maximum displacement of 4.69 m with at most 3.83 in the z-component and 2.70 m in the horizontal component (Figure 3-27a). During the second period (10-Oct-2012 till 19-Jul-2013) the displacement increased up to 25.55 m with a maximum 22.76 m in the x-y component and a maximum of 11.82 m in z direction (Figure 3-27b). In relative terms this corresponds to average maxima of 1.12 $cm.d^{-1}$ for the time between late August and early September, and 0.90 $cm.d^{-1}$ for the following months until mid-July of the next year. A

meaningful interpretation those figures, however, must considered that displacement typically decreases significantly during the winter month and only takes up again in spring.

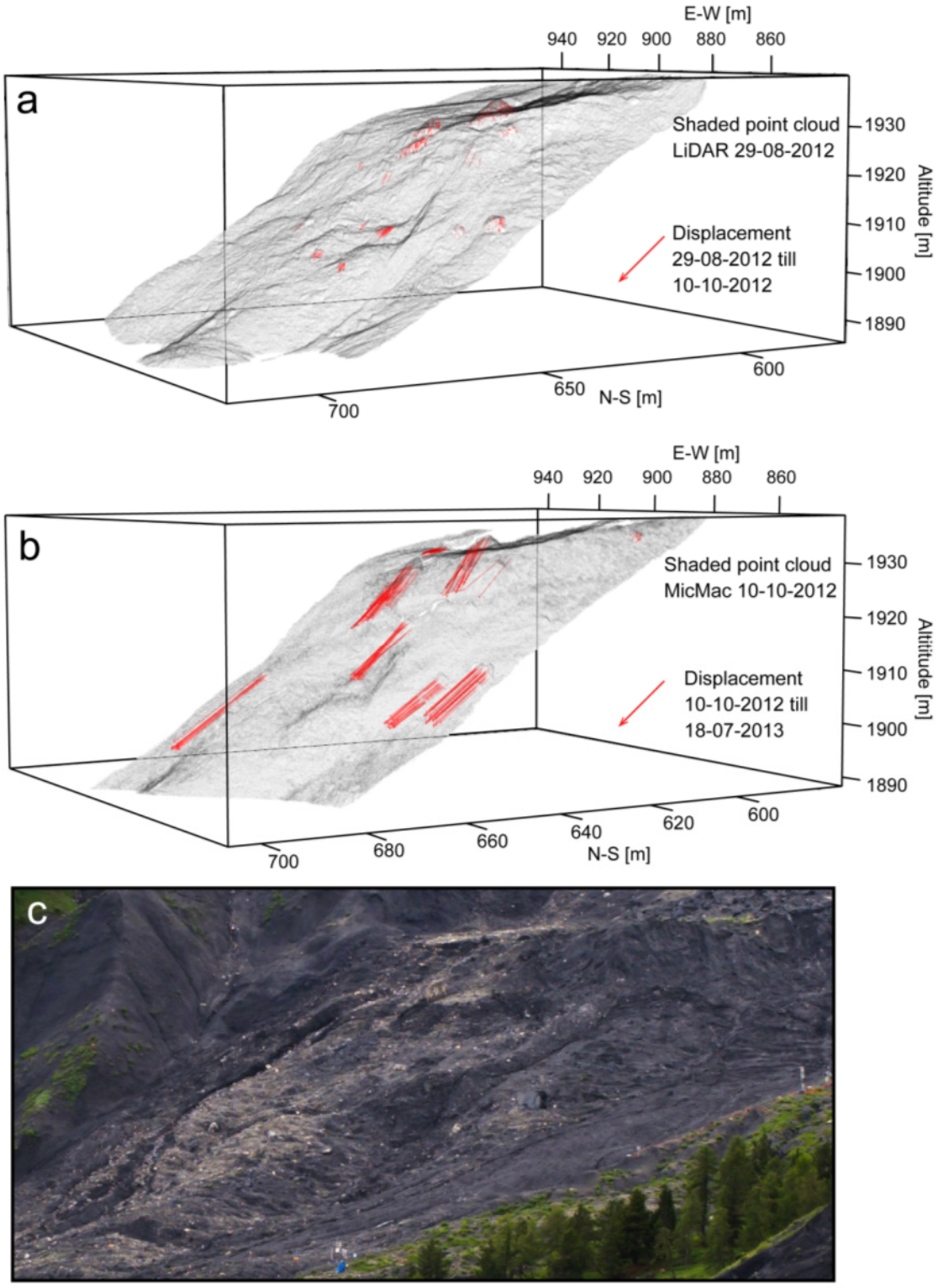


Figure 3-27: 3D displacement vectors measured on rigid structures via ICP between (a) the aerial LiDAR of 29-Aug-2012 and MicMac point cloud of 10-Oct-12, and (b) the MicMac point cloud of 10-Oct-12 and 19-Jul-2013. (c) Terrestrial photograph showing the subset of analysed area.

3.5.5 Potential and limitations of terrestrial SfM-MVS for landslide monitoring

Similar to several previous studies (see Section 3.5.1) we found that SfM-MVS should be considered as an accurate and cost-effective technique to obtain 3D information for geomorphological studies. The use of different open-source SfM-MVS pipelines was evaluated for terrestrial landslide monitoring and it was demonstrated that a MicMac-based pipeline can provide accuracies and coverage which approach the quality of terrestrial and aerial LiDAR if the geometry of the image acquisition is well adapted to the surveyed object. Using the M3C2 algorithm for change detection allowed taking into account the quantified errors and establishing a 95% confidence for the reliable detection of changes above 20 cm at the landslide scarp and above 40 cm at the slope scale. Remaining outliers could be traced back to local artefacts in the 3D reconstruction and excluded through multi-temporal analyses and considerations of the spatial context of erosion, accumulation and displacement. A displacement measurement at the location of a permanent GPS station suggested a maximum error of 0.47 m and rockfall volume estimates differed by 4% from the respective volumes calculated from terrestrial LiDAR.

Despite those encouraging results and a high degree of automation, several potential pitfalls deserve further consideration since they may increase the need for manual intervention in the pipeline and compromise the accuracy. A non-trivial aspect of terrestrial MVS in complex natural landscapes is the planning of the acquisition geometry. Practical guidelines for acquisition planning with buildings, indoor spaces and cultural artefacts can be found in Luca et al. (2013) and Wenzel et al. (2013) but it remains difficult to provide general rules for natural terrain. Smaller base-to-height ratios (i.e. short distances between neighbouring views) diminish the accuracy of the surface reconstruction, while larger baselines complicate the matching. The MicMac library includes a tool for image selection and an overly dense linear network is preferable to not risk unconnected image clusters. The angle between neighbouring views must not exceed 15˚ and their overlap should be greater 80%. Motion blurs and changes in the lens parameters (e.g. auto-focus) should be avoided. The surveys were conducted under variable weather conditions ranging from bright sunshine to light rain. However, diffused lighting (i.e. clouded sky) and no rain are ideal to avoid effects from shadows and noise. Each surface should be visible in 3-6 well separated views and include ideally one view close to nadir. For the landslide terrain investigated in this study nadir views could not be obtained for all positions and many of the remaining artefacts must be attributed to this issues. At sites where the terrain is even more unfavourable for terrestrial imaging at nadir alternative platforms such as UAVs should be considered. To fully understand the constraints of a specific site it is generally recommendable to conduct at least one preliminary survey to optimize the protocol for operational monitoring. For future operational monitoring at the investigated landslide, it is recommended to integrate the acquisition protocols for the scarp and the landslide body (~500 images) to avoid multiple processing chains. Tools for the optimization of the camera network based on preliminary video acquisitions have been recently proposed (Alsadik et al., 2013) and should be tested for the optimization of acquisition geometry in natural terrain. A further option could be near-real-time user interaction to determine the next best viewing position (e.g. Hoppe et al., 2012), which however requires a direct downlink from the camera to a computer and might not be easy to implement in the field.

A general limitation of landslide monitoring with passive optical sensors is the inability to penetrate vegetation. The reconstructed surfaces over vegetated areas represent an average canopy surface and due to high surface roughness and random pattern generally incur

large errors. The Super-Sauze landslide is very sparsely vegetated and the manual removal of few vegetated areas was straightforward. For more complex settings an automated point cloud classification with dedicated tools (Brodu and Lague, 2012) might be desirable. In SfM-MVS point clouds the vertical structure of the vegetation is represented much poorer than in LiDAR point clouds, whereas RGB colour information that can be mapped automatically to each 3D point provides useful features not available with LiDAR.

Regarding the georeferencing we found direct referencing with photogrammetric targets during the SfM phase to be a more convenient strategy than establishing point to point correspondence after dense matching. In cases of limited GPS accuracy the ICP algorithm proofed to be a valuable tool to resolve residual alignment errors (translation and rotation). An intermediate solution that was not explored in this study would be to resolve the model scale with a scale bar that must be visible in at least two images and recover translation and rotation through matching of the dense point clouds of stable terrain. In any case it is indispensable for the alignment to include a sufficiently large proportion of stable and vegetation-free terrain within the reconstruction, which should be kept in mind during the image acquisition. For the validation of displacement rates we recommend measurements at clearly visible photogrammetric targets rather than GPS antenna.

3.5.6 Conclusion

This study investigated the use of structure-from-motion and multi-view stereo pipelines for the terrestrial monitoring of landslides. In a comparative assessment of different open-source solutions the MicMac library yielded the more accurate results, whereas VisualSfM and PMVS permitted a higher degree of automation. Compared to LiDAR the RMSE did generally not exceed 0.2 m for the reconstruction of the entire landslide and 0.06 m for the reconstruction of the main scarp. The M3C2 algorithm was found to be a versatile and accurate tool for the reliable detection of changes and the possibility to obtain volume and 3D displacement estimates was illustrated for especially active zones. It could be demonstrated that at the slope scale terrestrial multi-view photogrammetry is sufficiently accurate to detect surface changes in the range of decimetres. Thus, the technique currently remains less precise than terrestrial LiDAR or dGPS but provides spatially distributed information at significant lower costs and is, therefore, valuable for many practical landslide investigations.

An option that has not been explored in this study is the possibility to measure 3D displacement from temporal sequence of stereo-pairs which is a technique frequently used experimental mechanics (Sutton et al., 2009). The determination of 3D vectors through matching in the image space can be expected to provide more precise results than matching of the resulting point clouds. More frequent acquisitions might be required in this context to avoid temporal decorrelation. The planning and optimization of the image acquisition protocols still remains challenging in complex natural terrain and further research is needed to establish tools that enable an a more straightforward optimization of camera network for operational monitoring.

3.6 *Discussion and perspectives*

The research presented in this chapter was dedicated to the development, elaboration and validation of image processing techniques and workflows for the detailed characterization and monitoring of landslides at the local scale.

In **section 3.3** it was demonstrated that the adaptation of medical image processing algorithms and OBIA permits the **detection of surface fissures** from low-altitude surveys. The interpretation of resulting maps provided valuable insights into the patterns of stress and strain acting on the surface and allowed to detail relationships between displacement rates, the geometry of the sliding surfaces and the formation of the fissures. The obtained maps are also useful as an input for the parameterization of hydro-mechanical models that take into account preferential flows and mechanical feedbacks (Krzeminska et al., 2012).

The developed image processing chain provides one example how modern image processing techniques can valorise the resulting images for landslide investigation. It also suggests that further qualitative and quantitative information could be extracted by adapting pattern recognition, texture analysis or image-based grain-size analysis (e.g. for the analysis of rockfall deposits). While the developed OBIA routine is rather application specific, the Gaussian-matched filters are generically applicable for the detection of linear features and could be interesting for further geoscientific application such as the mapping of linear erosion features (e.g. gullies), tectonic fractures or glacier crevasses.

The main limitations of the employed processing techniques are (i) the sensitivity of the Gaussian-matched filters to illumination conditions resulting in higher errors with direct sun light at low sun elevation angles and (ii) a significant number of free parameters that must be calibrated by the users, and (iii) relatively low true positive rates. In subsequent studies, we addressed some of those issues and implemented an extension of the Gaussian-matched filters that follows a multi-scale framework for automatic scale selection. This eliminates to adjust the scale parameter manually and provides results better to manually-tuned single scale detectors (Figure 3-28). As part of further studies on the influence of ground truth uncertainties on the comparison between different detectors, we also benchmarked several low-level line detectors for the detection of surface fissures. Though the Gaussian matched filter was ranked second among nine state-of-the-art detectors, it was also found that an algorithm based on supervised learning and Gabor wavelets (Soares et al., 2006) provided more accurate results (Figure 3-29). The algorithm requires no manual tuning and could serve as a more robust low-level detector enhancing the developed processing chain.

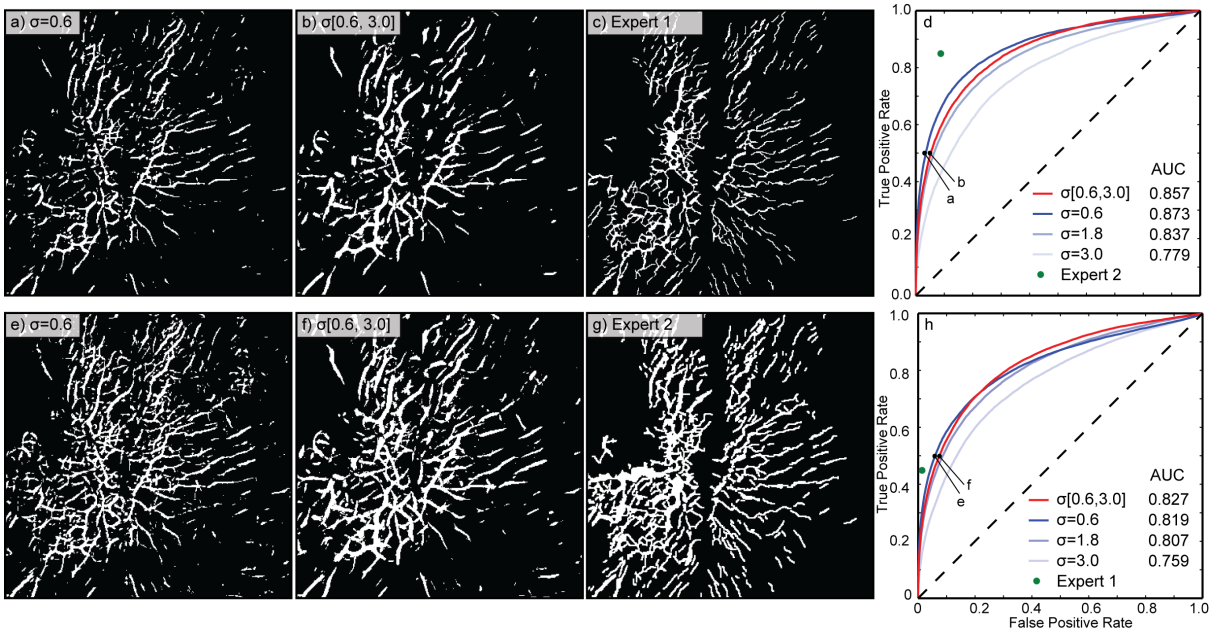


Figure 3-28: Comparative accuracy assessment of (a, e) single scale detection and (b, f) multi-scale detection against (c, g) two expert mappings. (d, h) The ROC analysis shows that the multi-scale detector yields accuracies (area under the curve – AUC) comparable or better to single scale detection.

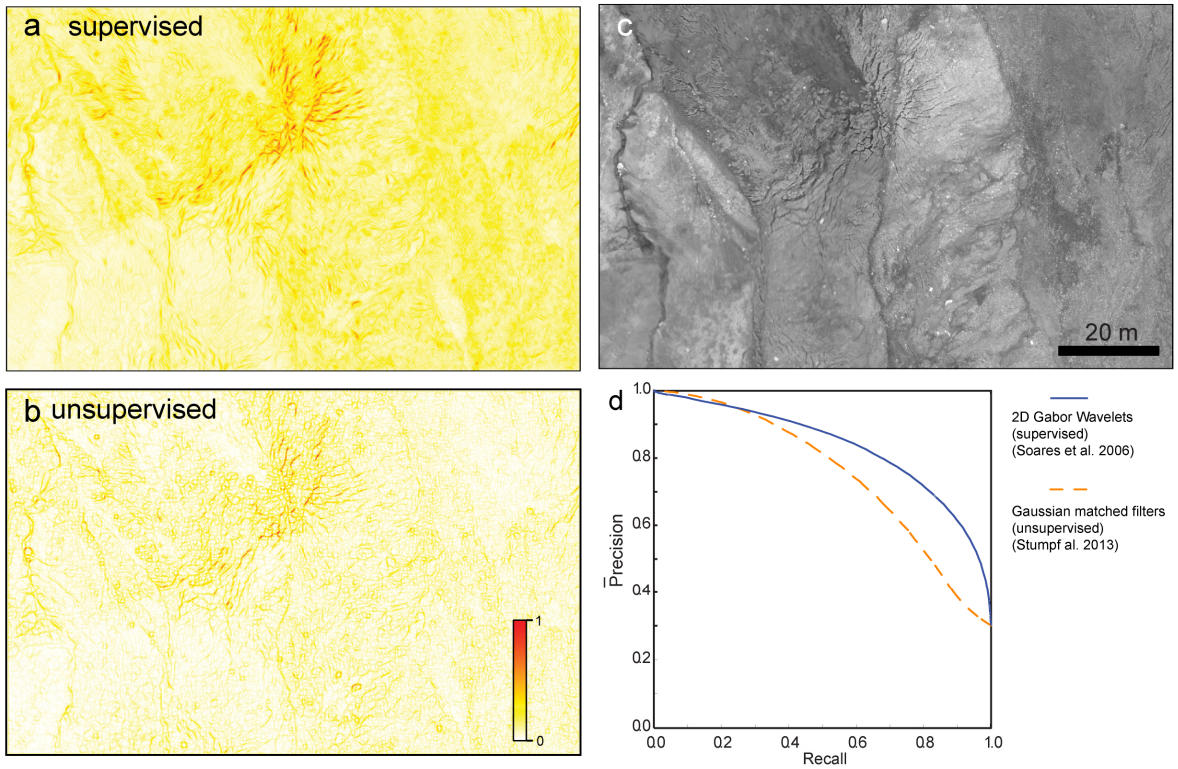


Figure 3-29: Comparison between two low-level feature detectors for the detection of landslide surface fissures. (a) Probabilistic outputs of supervised approach (Soares et al., 2006), Response of a Gaussian filter as proposed in Stumpf et al. (2013b) , (c) the input image and (d) the average precision-recall curves (Lampert and Gancarski, In Review) with respect to ten different expert maps.

In Section 3.4, a processing chain for **displacement measurements from VHR satellites images** was developed and applied to derive the displacement fields of three active slow-moving landslides. Measured displacement rates and surface heights were compared with GNSS measurements and airborne LiDAR surveys respectively. The analysis revealed that the resulting DSMs comprise metric errors that vary strongly depending on the land cover type, whereas the displacement measurements showed an unprecedented accuracy with maximum errors of 13 cm. Furthermore, it was demonstrated that this accuracy can be achieved without ground control points. The analysed images covered a time period of two month and the obtain displacement fields provided important insights in the magnitude of the seasonal displacement rates and changes in the spatial pattern of the activity.

The achieved accuracy of the horizontal displacement measurement is approximately one order of magnitude higher than results from previous studies using medium- and high resolution satellites but also aerial images. The study suggests the possibility to measure horizontal displacement accurately without ground control which enables greater automation of the processing chain and opens the door for more continuous monitoring of landslides and other surface deformation processes. While in this research only a bi-temporal dataset was available, an extension to multi-date measurements could be easily achieved through integration of images from multiple time steps into one single bundle adjustment block. Further case studies with longer-time series and applications to other landslide-types and environmental conditions are needed to explore full potential and limitations of the proposed processing chain.

Besides, general limitations of optical imaging (e.g. occlusion through vegetation, shadows and clouds), the proposed processing chain is only limited by the availability of suitable datasets and salient objects that remain stable over the targeted monitoring period. Spaceborne stereo-pair image acquisition is currently possible at a minimum price of 3250 € (100 km^2) while archived stereo-data can be assessed at a minimum of 500 € (25 km^2) and prices are likely to decrease in the near future. The requirement of stable objects may complicate the applicability to deformation phenomena that extend over entire VHR images (e.g. coseismic displacement) and then requires some prior knowledge about the extent of the motion in order to exclude moving points from the RPF bias compensation.

Another limitation is currently the distribution of the processing chain over several specialized commercial and open-source environments which implies a long learning curve for the use of the technique and hinders full automation. Open source solutions for precise image correlation are available (Deseilligny et al., 2013) and could resolve this issue partially. However, there is currently a lack of open software solutions for precise RPF bundle adjustment. The Orfeo toolbox developed by CNES (http://blog.orfeo-toolbox.org/) provides some promising features for RPF bundle adjustment that, with further development, could fill this gap.

A further promising research direction would be the integration of multi-sensor satellite data, which would significantly enhance the possibilities to exploit historical archives and help to reduce the time-intervals between subsequent observations. RPFs delivered with the latest generation of VHR satellite images provide a generic sensor model that makes multi-sensor adjustment possible. However, differences in the spatial resolution and radiometric characteristics of different sensors pose high demands on the robustness of automatic tie point detectors and image matching algorithms. The availability of multiple stereo-pairs in principle could also enable to recover the full 3D surface displacements from disparity measurements in the imaging geometry which would avoid intermediate steps of

DSM generation and orthorectification and possibly make the processing chain more efficient and precise.

On the long-term, further efforts are also needed to better exploit the derived displacement fields to infer other physical landslide parameters such as strain, subsurface geometry and material parameters. This will require the development of new inversion techniques and a better integration of the measurements with conceptual and numerical landslide models.

In **Section 3.5** the potential of **open-source terrestrial photogrammetry for landslide surface reconstruction and 3D monitoring** was investigated at the Super-Sauze landslide over a period of 2 years. Three different processing chains were established and their main algorithm parameters were detailed and adapted for the generation of high-resolution surface models in a high mountain environment with complex topography. A comparison of the resulting 3D point clouds with terrestrial LiDAR surveys revealed that a customized pipeline implemented with the MicMac library yielded higher accuracies approaching those of terrestrial LiDAR point clouds (RMSE 0.027–0.060 m) and airborne LiDAR point clouds (RMSE 0.08-0.17 m). The spatial coverage relative to the LiDAR surveys was 43% in the worst case but could be improved to more than 80% trough successive optimization of the imaging geometry. In all cases, the point density was generally significant higher when compared to the LiDAR scans.

While those results are comparable or better than figures presented in earlier studies, it must be noted that the accuracies depend strongly on the application domain and the imaging distance to the object, and are, thus, not directly comparable. Volumes estimates for rockslides that occurred at the main scarp of the landslide and 3D displacement measurement were found to be in line with the quantities obtained from LiDAR surveys and permanent GNSS observations. An innovative cloud-to-cloud change detection method was employed and permitted to differentiate significant changes with high confidence from residual errors. The detected changes and measured 3D displacement vectors revealed an initial activation of in the central part of the landslide that intensified and gradually progressed downslope at the end of a prolonged winter with abundant snowfall. Seasonal cycles of strong ablation and sediment runoff could be observed at the main scarp of the landslide.

The main contributions of this work lie in establishing and validating an accurate SfM-MVS pipeline and demonstrating the feasibility to accurately monitor seasonal landslide dynamics with terrestrial photogrammetry at the slope scale. Once a suitable processing chain is established the operational costs for data acquisition and SfM-MVS processing are minimal and render photogrammetric surveys a tool that should be always considered in landslide investigations.

Current limitations and potential pitfalls have been discussed in detailed at the end of the manuscript in section 3.5 and are summarized again here. Though the underlying algorithms are relatively robust to variable baselines and illumination the planning of the imaging geometry in natural terrain is not a trivial task and at least basic knowledge on photogrammetric principles and the site characteristics are indispensable. Complex topography might also constrain the possibility to obtain nadir and at sites where the terrain is unfavourable for terrestrial imaging at nadir alternative platforms such as UAVs should be considered. Again it should be stressed that optical imaging has very limited capabilities to penetrate the vegetation and decorrelation is likely over densely areas. This hinders reliable measurements of surface displacement and deformation over densely vegetated surfaces. For georeferencing we found direct referencing with photogrammetric targets to be a more

reliable strategy than establishing point to point correspondence after dense matching. For a sustainable use of the ground control points during long-term monitoring they should be installed at fixed positions on stable terrain and be evenly distributed on the slope.

Enhancing the accuracy and automation of SfM-MVS algorithm remains an active field of research. The improving processing chains and decreasing costs of SLR digital cameras are likely to render the use of this technique for landslide investigations more and more attractive in the future. To further exploit the capabilities of SfM-MVS for landslide monitoring several possible research directions are prominent.

To step forward to more continuous monitoring and integration with *in situ* measurement systems, the fixed installation of stereo setups is a promising option. The MicMac library tested in this study provides the option to resolve the scaling of the model from a scale bar visible within at least two views and could be adapted to derive metric displacement vectors without further ground control. Stereo-camera rigs with known baselines could provide similar capabilities but have yet not provided accurate field measurements over long distances.

In this study, 3D displacement measurements have been realized with a classical point-matching algorithm that is frequently employed for the analysis of terrestrial LiDAR point clouds. This involves several steps namely the computation of disparity maps, depth maps and the subsequent point matching. More accurate measurements could be obtained by computing first 2D displacements in the image geometry and inversion of the 3D displacement with a least-square approach.

On the long-term, obtained measurements should be exploited to derive physical quantities such as the depth of the movement or the rheological parameters. This will require closer integration with conceptual and numerical models of the landslide and time series derived from *in situ* measurements.

Synthesis of the research findings on landslide monitoring at local scales:

Main results: The research presented in this chapter set out to develop new image processing chains for the retrieval of surface displacements, transported volumes and surface fissure maps that enable to complement historical and point-wise observations, and to infer the slope dynamics. An image analysis workflow for the mapping of surface fissures was established combining Gaussian matched filterers, mathematical morphology and OBIA. The joint interpretation of the resulting maps, displacement rates and bedrock geometry permitted to correlate increased fissure density and increasing displacement rates, patterns of stress with abrupt changes of the underlying bedrock geometry. A processing chain for the computation of high-resolution surface models and displacement fields from VHR satellite images was elaborated and enabled measurements of the horizontal displacement rates with unprecedented decimeter accuracy. Furthermore, it was demonstrated that this accuracy can be achieved without ground control points which will greatly facilitate the automation for operational monitoring in the future. The analyzed bi-temporal datasets covered a time period of two month and the obtain displacement fields provided important insights in the magnitude of the seasonal displacement rates and changes in the spatial pattern of the activity compared to historical. Three different pipelines for high resolution surface reconstructions from multi-view terrestrial photographs were elaborated and compared regarding their capability for multi-temporal reconstruction and 3D monitoring of landslides. Comparisons with terrestrial and airborne LiDAR surveys showed that open-source stereo-photogrammetry can achieve similar accuracies and higher point-cloud density at significantly lower costs. Cloud-to-cloud change detection and matching allowed the quantification of the dominant geomorphological processes (e.g. rockslides, sediment runoff) and insights into deformation patterns emerging during acceleration phases.

Perspectives: UAVs are becoming increasingly frequently used in landslide investigations and the application of the proposed image processing method for fissures mapping with more frequent surveys could help to better characterize the seasonal dynamics of the fissure formation and to establish quantitative relationships with displacement rates, rainfall and pore water pressure. Multi-scale and supervised extensions investigated in this study should be considered to improve the mapping accuracy and would help to also measure the fissure width as an additional parameter. The possible integration of fissure maps in hydro-mechanical models has been described recently and the developed image processing chain could facilitate the elaboration of such maps. The mapping of linear features investigated in other geoscientific fields (e.g. gully erosion, seismo-tectonic fractures) could also benefit from proposed feature detectors.
The possibility to measure horizontal surface displacement from VHR satellite images at decimeter accuracy and without ground control opens the door for a number of possible applications. Further efforts are required for a full automation of the processing chain and especially the development of an open-source library for precise bundle adjustment would close an important gap in the chain of publicly available tools. The integration of multi-sensor datasets will be an important step for the construction of long-term continuous time series and could benefit from the high geolocation accuracy of SAR X-band satellites. Ongoing research is targeting the derivation of physical parameters such as strain, landslide geometry and material properties from the derived displacement fields.

Enhancing the accuracy and automation of SfM-MVS algorithm remains an active field of research and landslide monitoring with terrestrial cameras will certainly benefit from those efforts. To step forward to more continuous monitoring and integration with in situ measurement systems the fixed installation of stereo setups is a promising option. More accurate measurements could be obtained by computing first 2D displacements in the image geometry and inversion of the 3D displacement with a least-square approach. On the long-term obtained measurements should be combined with conceptual and numerical landslide models to derive physical quantities such as the depth of the movement or rheological parameters.

General Conclusion

Detailed investigations of landslides at regional and local scales is indispensable for a better understanding of the governing geomechanical processes, the assessment of the associated hazards, and their forecast for early warning. Those investigations comprise tasks of landslide detection and mapping, characterization of the landslide type, geometry and intensity as well as continuous monitoring to construct parameter time series and issue warnings.

The immense progress and diversification of remote sensing technologies has significantly enhanced our capabilities to map and monitor slope movements with increasing accuracy and spatial and temporal resolutions. SAR and LiDAR have become commonly adopted technologies in landslide investigations. Especially SAR interferometry can be considered as a mature technique for the use in operational services. Still, both techniques comprise several limitations such as limited spatio-temporal coverage or high costs. **Optical imaging is a cost-efficient and versatile remote sensing technology** that provides complementary information and overcomes several current limitations.

The doctoral research described in this thesis was dedicated to the **development of automated and robust image analyses techniques** to better exploit the potential value of optical data (from different platforms) for rapid **landslide mapping at the regional scale**, and for **detailed characterization and monitoring at the local scale**.

Considering the great diversity of modern remote sensing systems and processing techniques a comprehensive review of recent innovations in the application of remote sensing for landslide monitoring was undertaken and presented in **Chapter 1.** The review provides a detailed overview of currently available systems and processing methods and elaborates criteria that help to contrast different possible solutions for landslide monitoring. The capabilities and limitations of all techniques are compared based on criteria such as spatial resolution and coverage, accuracy, data and processing costs, and maturity of the method. Rule sets for the selection of the most suitable remote sensing technologies according to different landslide types, displacement rates and observational scales are proposed. The rule sets target scientists and end-users in charge of landslide research and risk management.

Chapter 2 addressed the development, implementation and validation of generic supervised image processing chains for landslide mapping from VHR satellite and aerial images in combination with topographic data. Algorithms for image segmentation, feature extraction, feature selection and supervised learning were adapted and combined in successive processing chains. The developed techniques were applied and validated on five remote sensing datasets at different study sites and achieved satisfactory accuracies in all test cases.

A Random Forest (RF) based feature selection method allowed selecting the most relevant features and permitted to reduce the complexity of the machine learning model while yielding comparable or even higher classification accuracies. New object-based features particularly engineered for the recognition of landslides were found to be generically applicable and provided significant improvements of the classification accuracy in all tested cases. Relationships between the segmentation scale, the importance of different object features and the accuracy of the resulting maps were studied in detailed and showed that over-segmentation generally provided higher accuracies. The implementation of stratified resampling schemes allowed compensating the biases resulting from inherent class-imbalances. A new Active Learning (AL) heuristic was proposed which, contrariwise to previous approaches, considers spatial constraints on the samples selection, and costs in terms

of labeling time instead of number of samples. The developed region-based sampling scheme significantly reduced the labeling time required by the user. An in-depth analysis of the remaining errors and comparison with the mapping uncertainties among multiple experts revealed that the achieved accuracy is within the range of the variance among multiple experts. This indicates that the proposed processing chain could be used for operational applications providing map products whose quality is comparable to manual mappings.

Chapter 3 described the development and application of remote sensing techniques to acquire spatially distributed information on surface structures, displacements and transported volumes at the local scale. Three dedicated image processing chains were developed for the extraction of spatio-temporal information on the landslide dynamics from VHR satellites images, UAV images and terrestrial photographs. First, an image processing workflow that combines Gaussian matched filters, mathematical morphology and OBIA methods for the generation of surface fissure maps from multi-date UAV images. The resulting maps were validated with expert mappings and interpreted regarding the evolution of the fissures in relation to displacement rates and bedrock geometry. Second, a complete processing chain for the computation of high-resolution surface models and displacement fields from VHR satellite images was elaborated and enabled measurements of the horizontal displacement rates with unprecedented decimetre accuracy. Furthermore, it was demonstrated that this accuracy can be achieved without ground control points which will greatly facilitate the automation for operational monitoring in the future. The analysed bi-temporal datasets covered a time period of two months and the obtained displacement fields provided important insights in the magnitude of the seasonal displacement rates. Third, three different pipelines for high resolution surface reconstructions from multi-view terrestrial photographs were elaborated and compared regarding their capability for multi-temporal reconstruction and 3D monitoring of landslides. Comparisons with terrestrial and aerial LiDAR scans showed that open-source stereo-photogrammetry can achieve similar accuracies and higher point-cloud density at significantly lower costs. Estimated volumes and 3D displacement rates were found to be in line with LiDAR and GNSS-based measurements, respectively.

While the use of optical image analysis as an operational tool in landslide investigations will still require additional efforts, we believe that this thesis provides several contributions which bring this goal within closer reach and encourage further research on this topic. Future studies could consider the following key directions.

To enable the **operational use** of the developed AL algorithms for landslide mapping and possibly other applications in disaster response (e.g. damage mapping), important aspects will be to optimize and parallelize the implemented algorithms and complement them with a user-friendly interface enabling real-time interaction. Experiments in simulated or real-world rapid mapping scenarios should be envisaged at the prototyping stage ensuring that it is robust and will be accepted by users during image interpretation and field work. To handle the **large volumes of data** resulting from current satellite missions such as Sentinel-2, it seems recommendable to implement such a system on dedicated computational infrastructure. Segmentation and feature extraction can be carried out offline and a web-interface might be an interesting option to facilitate access to the service.

Additional series of Pleiades series stereo-pairs have been acquired to construct longer displacement time series for the landslides investigated in the Ubaye valley. Further collaborations with the *OTB development team (CS Company)* and the *Laboratoire Méthodes d'Analyses pour le Traitement d'Images et la Stéréorestitution* are envisaged to arrive at an **integrated open-source based framework for stereo reconstruction and displacement**

measurements. An important aspect for future research will be the development of and open-source tool for RPF bundle adjustment which would fill an important gap in the chain of available open-source tools. Multi-temporal stereo-pairs in theory enable the least-square inversion of full 3D displacement from directly from disparity measurements in the image geometry and would help to better track the complex deformation patterns of landslides.

The experiences in this research with terrestrial photogrammetry for landslide monitoring support the continuation and optimization of the acquisition protocols for the Super-Sauze landslide and encouraged the installation of new **fixed stereo camera setups for continuous 3D measurements**. The high frequency displacement measurements obtained with this system will help to establish quantitative relationships with other physical parameters such as pore-water pressure, height of the ground-water table and seismic signals emitted by the landslide. The latter are being recorded continuously with broadband seismic antenna and will be correlated with the observed displacement, rock falls and fissure formation to better pinpoint the sources of micro-seismicity and investigate the potential role of the signals for monitoring and early warning.

To fully exploit displacement fields and time series measured from satellites, aerial platforms, and terrestrial cameras, it will be of great importance to develop better approaches for the derivation of physical parameter such as strain, depth and material properties. This will require a closer **integration of remote sensing and data mining techniques with conceptual and numerical landslide models**, which should be envisaged as interdisciplinary research project involving geomorphologist, geophysicist, geographers and computer scientists.

Bibliography

Abed, R.S., Abdullah, G.N., 2013. Measurements in Alhadba Minaret Using Robotic Total Station. Engineering & Technology Journal 31, 910-920.

Abellán, A., Jaboyedoff, M., Oppikofer, T., Vilaplana, J.M., 2009. Detection of millimetric deformation using a terrestrial laser scanner: experiment and application to a rockfall event. Natural Hazards and Earth System Sciences 9, 365-372.

Abellán, A., Oppikofer, T., Jaboyedoff, M., Rosser, N.J., Lim, M., Lato, M.J., 2013. Terrestrial laser scanning of rock slope instabilities. Earth Surface Processes and Landforms, 48.

Abellán, A., Vilaplana, J.M., Calvet, J., García-Sellés, D., Asensio, E., 2011. Rockfall monitoring by Terrestrial Laser Scanning – case study of the basaltic rock face at Castellfollit de la Roca (Catalonia, Spain). Nat. Hazards Earth Syst. Sci. 11, 829-841.

Abramson, L.W., Lee, T.S., Sharma, S., Boyce, G.M., 2001. Slope stability and stabilization methods. Wiley.

Adrian, R.J., 2005. Twenty years of particle image velocimetry. Experiments in Fluids 39, 159-169.

Agarwal, S., Snavely, N., Seitz, S., Szeliski, R., 2010. Bundle Adjustment in the Large, in: Daniilidis, K., Maragos, P., Paragios, N. (Eds.), Computer Vision – ECCV 2010. Springer Berlin Heidelberg, 29-42.

Aguilar, M.A., Saldaña, M.d.M., Aguilar, F.J., 2013. Assessing geometric accuracy of the orthorectification process from GeoEye-1 and WorldView-2 panchromatic images. International Journal of Applied Earth Observation and Geoinformation 21, 427-435.

Alexander, D.E., 2002. Principles of Emergency Planning and Management. Oxford University Press, New York.

Allasia, P., Manconi, A., Giordan, D., Baldo, M., Lollino, G., 2013. ADVICE: A New Approach for Near-Real-Time Monitoring of Surface Displacements in Landslide Hazard Scenarios. Sensors 13, 8285-8302.

Alsadik, B., Gerke, M., Vosselman, G., 2013. Automated camera network design for 3D modeling of cultural heritage objects. Journal of Cultural Heritage 14.

Antonello, G., Casagli, N., Farina, P., Leva, D., Nico, G., Sieber, A.J., Tarchi, D., 2004. Ground-based SAR interferometry for monitoring mass movements. Landslides 1, 21-28.

Arattano, M., Marchi, L., 2000. Video-derived velocity disribution along a debris flow surge. Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere 25, 781-784.

Arattano, M., Marchi, L., 2008. Systems and Sensors for Debris-flow Monitoring and Warning. Sensors 8, 2436-2452.

Ardizzone, F., Cardinali, M., Carrara, A., Guzzetti, F., Reichenbach, P., 2002. Impact of mapping errors on the reliability of landslide hazard maps. Nat. Hazards Earth Syst. Sci. 2, 3-14.

Aryal, A., Brooks, B.A., Reid, M.E., Bawden, G.W., Pawlak, G.R., 2012. Displacement fields from point cloud data: Application of particle imaging velocimetry to landslide geodesy. Journal of Geophysical Research: Earth Surface 117, F01029.

Baldi, P., Cenni, N., Fabris, M., Zanutta, A., 2008. Kinematics of a landslide derived from archival photogrammetry and GPS data. Geomorphology 102, 435-444.

Baldo, M., Bicocchi, C., Chiocchini, U., Giordan, D., Lollino, G., 2009. LIDAR monitoring of mass wasting processes: The Radicofani landslide, Province of Siena, Central Italy. Geomorphology 105, 193-201.

Balz, T., Zhang, L., Liao, M., 2013. Direct stereo radargrammetric processing using massively parallel processing. ISPRS Journal of Photogrammetry and Remote Sensing 79, 137-146.

Bamler, R., Hartl, P., 1998. Synthetic aperture radar interferometry. Inverse problems 14, R1.

Barazzetti, L., Remondino, F., Scaioni, M., 2009. Combined use of photogrammetric and computer vision techniques for fully automated and accurate 3D modeling of terrestrial objects.

Barlow, J., Franklin, S., Martin, Y., 2006. High spatial resolution satellite imagery, DEM derivatives, and image segmentation for the detection of mass wasting processes. Photogrammetric Engineering and Remote Sensing 72, 687-692.

Barlow, J., Lim, M., Rosser, N., Petley, D., Brain, M., Norman, E., Geer, M., 2012. Modeling cliff erosion using negative power law scaling of rockfalls. Geomorphology 139, 416-424.

Barnhart, T., Crosby, B., 2013. Comparing Two Methods of Surface Change Detection on an Evolving Thermokarst Using High-Temporal-Frequency Terrestrial Laser Scanning, Selawik River, Alaska. Remote Sensing 5, 2813-2837.

Baroň, I., Supper, R., Winkler, E., Motschka, K., Ahl, A., Čarman, M., Kumelj, Š., 2013. Airborne geophysical survey of the catastrophic landslide at Stože, Log pod Mangrtom, as a test of an innovative approach for landslide mapping in steep alpine terrains. Nat. Hazards Earth Syst. Sci. 13, 2543-2550.

Benaglia, T., Chauveau, D., Hunter, D.R., Young, D., 2009. mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software 32, 1--29.

Berardino, P., Fornaro, G., Lanari, R., Sansosti, E., 2002. A new algorithm for surface deformation monitoring based on small baseline differential SAR interferograms. IEEE Transactions on Geosciece and Remote Sensing 40, 2375-2383.

Bernard, M., Decluseau, D., Gabet, L., Nonin, P., 2012. 3D capabilities of Pleaides satellite. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXIX-B3, 553-557.

Berthier, E., Vadon, H., Baratoux, D., Arnaud, Y., Vincent, C., Feigl, K.L., Rémy, F., Legrésy, B., 2005. Surface motion of mountain glaciers derived from satellite optical imagery. Remote Sensing of Environment 95, 14-28.

Binet, R., Bollinger, L., 2005. Horizontal coseismic deformation of the 2003 Bam (Iran) earthquake measured from SPOT-5 THR satellite imagery. Geophysical Research Letters 32.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 65, 2-16.

Bonnard, C., 2006. Technical and human aspects of historic rockslide dammed lakes and landslide dam breaches. A chapter of the book Italian Journal of Engineering Geology and Environment, Special Issue on Security of natural and artificial Rockslide Dams, 21-31.

Booth, A.M., Lamb, M.P., Avouac, J.-P., Delacourt, C., 2013. Landslide velocity, thickness, and rheology from remote sensing: La Clapière landslide, France. Geophysical Research Letters 40, 4299-4304.

Bovenga, F., Refice, A., Pasquariello, G., 2012a. Using corner reflectors and X-band SAR interferometry for slope instability monitoring, Tyrrhenian Workshop on Advances in Radar and Remote Sensing (TyWRRS), 2012 114-120.

Bovenga, F., Wasowski, J., Nitti, D.O., Nutricato, R., Chiaradia, M.T., 2012b. Using COSMO/SkyMed X-band and ENVISAT C-band SAR interferometry for landslides analysis. Remote Sensing of Environment 119, 272-285.

Breiman, L., 2001. Random forests. Machine Learning 45, 5-32.

Bretar, F., Arab-Sedze, M., Champion, J., Pierrot-Deseilligny, M., Heggy, E., Jacquemoud, S., 2013. An advanced photogrammetric method to measure surface roughness: Application to volcanic terrains in the Piton de la Fournaise, Reunion Island. Remote Sensing of Environment 135, 1-11.

Brodu, N., Lague, D., 2012. 3D terrestrial lidar data classification of complex natural scenes using a multi-scale dimensionality criterion: Applications in geomorphology. ISPRS Journal of Photogrammetry and Remote Sensing 68, 121-134.

Brückl, E., Brunner, F.K., Kraus, K., 2006. Kinematics of a deep-seated landslide derived from photogrammetric, GPS and geophysical data. Engineering Geology 88, 149-159.

Bruzzone, L., Carlin, L., 2006. A multilevel context-based system for classification of very high spatial resolution images. IEEE Transactions on Geoscience and Remote Sensing 44, 2587-2600.

Buades, A., Coll, B., Morel, J.-M., 2008. Nonlocal image and movie denoising. International Journal of Computer Vision 76, 123-139.

Bulmer, M.H., Petley, D.N., Murphy, W., Mantovani, F., 2006. Detecting slope deformation using two-pass differential interferometry: Implications for landslide studies on Earth and other planetary bodies. Journal of Geophysical Research: Planets 111, E06S16.

Canals, M., Lastras, G., Urgeles, R., Casamor, J.L., Mienert, J., Cattaneo, A., De Batist, M., Haflidason, H., Imbo, Y., Laberg, J.S., Locat, J., Long, D., Longva, O., Masson, D.G., Sultan, N., Trincardi, F., Bryn, P., 2004. Slope failure dynamics and impacts from seafloor and shallow sub-seafloor geophysical data: case studies from the COSTA project. Marine Geology 213, 9-72.

Canny, J., 1986. A Computational Approach to Edge Detection. IEEE Transaction on Pattern Analysis and Machine Intelligence 8, 679–698.

Capaldo, P., 2013. High Resolution Radargrammetry with Cosmo-Skymed, Terrasar-X And Radarsat-2 Imagery: Development And Implementation of an Image Orientation Model for Digital Surface Model Generation. Department of Civil, Constructional and Environmenal Engineering, Geodesy and Geomatic Division, Sapienza University of Rome, 189 p.

Cardellini, S., Osimani, P., 2013. The Ancona Early Warning Centre, Instrumentation and Continuous Monitoring of the Landslides, in: Margottini, C., Canuti, P., Sassa, K. (Eds.), Landslide Science and Practice. Springer Berlin Heidelberg, 57-65.

Cardenal, J., Mataa, E., Perez-Garcia, J.L., Delgado, J., Hernandez, M.A., Gonzalez, A., Diaz-de-Teran, J., 2008. Close range digital photogrammetry techniques applied to landslides monitoring, ISPRS Congress 2008. Commission VIII, Beijing, 235-240.

Carrara, A., 1993. Uncertainty in evaluating landslide hazard and risk. Kluwer Academic Publishers: Dordrecht, 101-109.

Casagli, N., Catani, F., Del Ventisette, C., Luzi, G., 2010. Monitoring, prediction, and early warning using ground-based radar interferometry. Landslides 7, 291-301.

Casagli, N., Farina, P., Leva, D., Tarchi, D., 2006. Application of ground-based radar interferometry for monitoring active rockslides and implications for emergency management, in: Evans, S., Mugnozza, G., Strom, A., Hermanns, R. (Eds.), Landslides from Massive Rock Slope Failure. Springer Netherlands, 157-173.

Cascini, L., Fornaro, G., Peduto, D., 2009. Analysis at medium scale of low-resolution DInSAR data in slow-moving landslide-affected areas. ISPRS Journal of Photogrammetry and Remote Sensing 64, 598-611.

Cascini, L., Fornaro, G., Peduto, D., 2010. Advanced low- and full-resolution DInSAR map generation for slow-moving landslide analysis at different scales. Engineering Geology 112, 29-42.

Casson, B., Delacourt, C., Allemand, P., 2005. Contribution of multi-temporal remote sensing images to characterize landslide slip surface. Application to the La Clapière landslide (France). Nat. Hazards Earth Syst. Sci. 5, 425-437.

Casson, B., Delacourt, C., Baratoux, D., Allemand, P., 2003. Seventeen years of the "La Clapière" landslide evolution analysed from ortho-rectified aerial photographs. Engineering Geology 68, 123-139.

Castagnetti, C., Bertacchini, E., Corsini, A., Capra, A., 2013. Multi-sensors integrated system for landslide monitoring: critical issues in system setup and data management. European Journal of Remote Sensing 46, 104-124.

Chandler, J.H., Moore, R., 1989 Analytical photogrammetry: a method for monitoring slope instability. Quarterly Journal of Engineering Geology and Hydrogeology 22, 97-110.

Chen, R.-F., Chang, K.-J., Angelier, J., Chan, Y.-C., Deffontaines, B., Lee, C.-T., Lin, M.-L., 2006. Topographical changes revealed by high-resolution airborne LiDAR data: The 1999 Tsaoling landslide induced by the Chi-Chi earthquake. Engineering Geology 88, 160-172.

Chowdhury, R.N., Zhang, S., 1991. Tension cracks and slope failure, in: Chandler, R.J. (Ed.), Proceedings of the International Conference on Slope Stability Engineering: developments and applications. Institution of Civil Engineers, Isle of Wight, 27–32.

Cigna, F., Bateson, L., Jordan, C. and Dashwood, C., 2012. Feasibility of InSAR technologies for nationwide monitoring of geohazards in Great Britain, The remote sensing and

photogrammetry society conference. Remote Sensing and Photogrammetry Society, London, UK.
Cignoni, P., Callieri, M., Corsini, M., Dellepiane, M., Ganovelli, F., Ranzuglia, G., 2008. Meshlab: an open-source mesh processing tool, Eurographics Italian Chapter Conference. The Eurographics Association, 129-136.
Clague, J.J., Stead, D., 2012. Landslides - types, mechanisms and modelling Cambridge University Press, Cambridge, UK.
Cleaves, A., 1961. Landslide investigations: a field handbook for use in highway location and design. United States Bureau of Public Roads, Washington, DC.
Coates, A., Huval, B., Wang, T., Wu, D., Catanzaro, B., Andrew, N., 2013. Deep learning with cots hpc systems, Proceedings of the 30th International Conference on Machine Learning (ICML-13), 1337-1345.
Coates, A., Ng, A.Y., 2012. Learning feature representations with k-means, Neural Networks: Tricks of the Trade. Springer, 561-580.
Colesanti, C., Wasowski, J., 2006. Investigating landslides with space-borne Synthetic Aperture Radar (SAR) interferometry. Engineering Geology 88, 173-199.
Cramer, M., Haala, N., 2010. DGPF Project: Evaluation of Digital Photogrammetric Aerial-based Imaging Systems- Overview and Results from the Pilot Center. Photogrammetric Engineering and Remote Sensing 76, 1019-1029.
Crosetto, M., Gili, J.A., Monserrat, O., Cuevas-González, M., Corominas, J., Serral, D., 2013a. Interferometric SAR monitoring of the Vallcebre landslide (Spain) using corner reflectors. Nat. Hazards Earth Syst. Sci. 13, 923-933.
Crosetto, M., Monserrat, O., Luzi, G., Cuevas-Gonzalez, M., Devanthery, N., 2013b. A Noninterferometric Procedure for Deformation Measurement Using GB-SAR Imagery. Geoscience and Remote Sensing Letters, IEEE PP, 1-1.
Cruden, D.M., 1991. A simple definition of a landslide. Bulletin of the International Association of Engineering Geology - Bulletin de l'Association Internationale de Géologie de l'Ingénieur 43, 27-29.
Cruden, D.M., Varnes, D.J., 1996. Landslides Types and Processes, in: Turner, A.K., Schuster, R.L. (Eds.), Landslides: Investigation and Mitigation. Transportation Research Board, Special Report 247, National Academy of Sciences, Washington D.C, 36–75.
Cuenca, M.C., Hanssen, R., Hooper, A., Arıkan, M., 2011. Surface deformation of the whole Netherlands after PSI analysis, ESA Fringe Workshop 2011, Frascati, Italy.
Daehne, A., Corsini, A., Ronchetti, F., 2011. Correlation image velocimetry applied to multitemporal LiDAR data for automated landslide movement assessment, EGU General Assembly 2011. Copernicus, Vienna, Austria.
Dandois, J.P., Ellis, E.C., 2010. Remote Sensing of Vegetation Structure Using Computer Vision. Remote Sensing 2, 1157-1176.
Dandois, J.P., Ellis, E.C., 2013. High spatial resolution three-dimensional mapping of vegetation spectral dynamics using computer vision. Remote Sensing of Environment 136, 259-276.
de Michele, M., Leprince, S., Thiébot, J., Raucoules, D., Binet, R., 2012. Direct measurement of ocean waves velocity field from a single SPOT-5 dataset. Remote Sensing of Environment 119, 266-271.
De Michele, M., Raucoules, D., De Sigoyer, J., Pubellier, M., Chamot-Rooke, N., 2010. Three-dimensional surface displacement of the 2008 May 12 Sichuan earthquake (China) derived from Synthetic Aperture Radar: evidence for rupture on a blind thrust. Geophysical Journal International 183, 1097-1103.
Debella-Gilo, M., Kääb, A., 2011. Sub-pixel precision image matching for measuring surface displacements on mass movements using normalized cross-correlation. Remote Sensing of Environment 115, 130-142.
Debella-Gilo, M., Kääb, A., 2012. Measurement of Surface Displacement and Deformation of Mass Movements Using Least Squares Matching of Repeat High Resolution Satellite and Aerial Images. Remote Sensing 4, 43-67.

Delacourt, C., Allemand, P., Berthier, E., Raucoules, D., Casson, B., Grandjean, P., Pambrun, C., Varel, E., 2007. Remote-sensing techniques for analysing landslide kinematics: a review. Bulletin de la Societe Geologique de France 178, 89-100.

Delacourt, C., Allemand, P., Casson, B., Vadon, H., 2004. Velocity field of the "La Clapière" landslide measured by the correlation of aerial and QuickBird satellite images. Geophysical Research Letters 31, L15619.

Delacourt, C., Raucoules, D., Le Mouélic, S., Carnec, C., Feurer, D., Allemand, P., Cruchet, M., 2009a. Observation of a Large Landslide on La Reunion Island Using Differential Sar Interferometry (JERS and Radarsat) and Correlation of Optical (Spot5 and Aerial) Images. Sensors 9, 616-630.

Delacourt, C., Raucoules, D., Mouélic, S.L., Carnec, C., Feurer, D., Allemand, P., Cruchet, M., 2009b. Observation of a Large Landslide on La Reunion Island Using Differential Sar Interferometry (JERS and Radarsat) and Correlation of Optical (Spot5 and Aerial) Images. Sensors 9, 616-630.

Demir, B., Minello, L., Bruzzone, L., In Press. Definition of Effective Training Sets for Supervised Classification of Remote Sensing Images by a Novel Cost-Sensitive Active Learning Method. IEEE Transactions onGeoscience and Remote Sensing, 13.

Deseilligny, M.-P., Belveaux, J., Choqueux, G., Deveau, M., Girod, L., 2013. MicMac, Apero and Other Beverages in a Nutshell, ENSG - Marne-la-Vallée, 275.

Deseilligny, M.P., Clery, I., 2011. Apero, an open source bundle adjusment software for automatic calibration and orientation of set of images, in: Remondino, F., El-Hakim, S. (Eds.), 4th ISPRS International Workshop 3D-ARCH 2011: ″3D Virtual Reconstruction and Visualization of Complex Architectures″, Trento, Italy, 269-276.

Dewitte, O., Jasselette, J.C., Cornet, Y., Van Den Eeckhaut, M., Collignon, A., Poesen, J., Demoulin, A., 2008. Tracking landslide displacements by multi-temporal DTMs: A combined aerial stereophotogrammetric and LIDAR approach in western Belgium. Engineering Geology 99, 11-22.

DORIS, 2007-2013. http://www.doris-project.eu [last access 31/10/2013].

Drăguţ, L., Tiede, D., Levick, S.R., 2010. ESP: a tool to estimate scale parameter for multiresolution image segmentation of remotely sensed data. International Journal of Geographical Information Science 24.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sensing of Environment 120, 25-36.

EDF-R&D, TelecomParisTech, 2013. CloudCompare (version 2.5) [GPL software]. Retrieved from http://www.danielgm.net/cc/, [last access: 21/10/2013].

Eisenbeiss, H., 2009. UAV Photogrammetry. Institute of Geodesy and Photogrammetry, ETH Zürich.

Erten, E., Reigber, A., Hellwich, O., Prats, P., 2009. Glacier Velocity Monitoring by Maximum Likelihood Texture Tracking. IEEE Transactions on Geoscience and Remote Sensing 47, 394-405.

ESA-Copernicus, 2013. https://sentinel.esa.int/web/sentinel/missions/sentinel-1/data-distribution-policy, [last access: 21/10/2013].

ESA, 2013. http://www.esa.int/Our_Activities/Observing_the_Earth/Oil_and_gas_sector_gears_up_for_Sentinels, [last access: 21/10/2013].

Fallourd, R., Vernier, F., Friedt, J.-M., Martinc, G., Trouvé, E., Moreau, L., Nicolas, J.-M., 2010. Monitoring temperate glacier with high resolution automated digital cameras - Application to the Argentière Glacier. International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences 38, 1-23.

Favalli, M., Fornaciai, A., Mazzarini, F., Harris, A., Neri, M., Behncke, B., Pareschi, M.T., Tarquini, S., Boschi, E., 2010. Evolution of an active lava flow field using a multitemporal LIDAR acquisition. J. Geophys. Res. 115, B11203.

Fell, R., Corominas, J., Bonnard, C., Cascini, L., Leroi, E., Savage, W.Z., 2008. Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. Engineering Geology 102, 85-98.
Fernández, T., Pérez, J.L., Colomo, C., Mata, E., Delgado, J., Cardenal, J., Irigaray, C., Chacón, J., 2012. Digital Photogrammetry and LiDAR Techniques to study the Evolution of a Landslide. on Geo-information for Disaster Management–Best Practices, 95.
Ferretti, A., Fumagalli, A., Novali, F., Prati, C., Rocca, F., Rucci, A., 2011. A New Algorithm for Processing Interferometric Data-Stacks: SqueeSAR. IEEE Transactions on Geosciece and Remote Sensing 49, 3460-3470.
Ferretti, A., Prati, C., Rocca, F., 1999. Permanent scatterers in SAR interferometry, IEEE International Geoscience and Remote Sensing Symposium, IGARSS 1999, 1528-1530.
Ferretti, A., Prati, C., Rocca, F., 2001. Permanent scatterers in SAR interferometry. IEEE Transactions on Geosciece and Remote Sensing 39, 8-20.
Finsterwalder, R., 1931. Geschwindigkeitsmessungen an Gletschern mittels Photogrammetrie. Zeitschrift für Gletscherkunde und Glazialgeologie 19, 251-262.
Fiorucci, F., Cardinali, M., Carlà, R., Rossi, M., Mondini, A.C., Santurri, L., Ardizzone, F., Guzzetti, F., 2011. Seasonal landslide mapping and estimation of landslide mobilization rates using aerial and satellite images. Geomorphology 129, 59-70.
Fleming, R.W., Baum, R.L., Giardino, M., 1999. Map and description of the active part of the slumgullion landslide, Hinsdale County, Colorado, Geologic Invstigations Series Map I-2672. U.S. Geologcial Survey, Denver, Colorado.
Fleming, R.W., Johnson, A.M., 1989. Structures associated with strike-slip faults that bound landslide elements. Engineering Geology 27, 39–114.
Fonstad, M.A., Dietrich, J.T., Courville, B.C., Jensen, J.L., Carbonneau, P.E., 2013. Topographic structure from motion: a new development in photogrammetric measurement. Earth Surface Processes and Landforms 38, 421-430.
FOSTER, 2011-2014. Spatio-temporal data mining: application to the understanding and monitoring of soil erosion [last access 31/10/2013].
Fraser, C.S., 1997. Digital camera self-calibration. ISPRS Journal of Photogrammetry and Remote Sensing 52, 149-159.
Fraser, C.S., Edmundson, K.L., 2000. Design and implementation of a computational processing system for off-line digital close-range photogrammetry. ISPRS Journal of Photogrammetry and Remote Sensing 55, 94-104.
Fraser, C.S., Hanley, H.B., 2005. Bias-compensated RPCs for sensor orientation of high-resolution satellite imagery. Photogrammetric Engineering and Remote Sensing 71, 909-915.
Fraz, M.M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A.R., Owen, C.G., Barman, S.A., 2012. Blood vessel segmentation methodologies in retinal images – A survey. Computer Methods and Programs in Biomedicine 108, 407-433.
Fruneau, B., Achache, J., Delacourt, C., 1996. Observation and modelling of the Saint-Etienne-de-Tinee landslide using SAR interferometry. Tectonophysics 265.
Fryer, J., Mitchell, H., Chandler, J., 2007. Applications of 3D Measurement from Images. Whittles Publishing, Caithness, Scotland.
Fu, W., Guo, H., Tian, Q., Guo, X., 2010. Landslide monitoring by corner reflectors differential interferometry SAR. International Journal of Remote Sensing 31, 6387-6400.
Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R., 2010. Towards Internet-scale multi-view stereo, IEEE Conference on Computer Vision and Pattern Recognition, 1434-1441.
Furukawa, Y., Ponce, J., 2010. Accurate, Dense, and Robust Multiview Stereopsis. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32, 1362-1376.
Galli, M., Ardizzone, F., Cardinali, M., Guzzetti, F., Reichenbach, P., 2008. Comparing landslide inventory maps. Geomorphology 94, 268-289.
Gance, J., Malet, J.-P., Dewez, T., Travelletti, J., In Press. Detection and continuous tracking of moving objects for characterizing landslide displacements from terrestrial optical images. Engineering Geology, 20.

Ganz, J., 1914. Die Gipfelbewegung der Rosablanche, SAC-Jahrbuch, 50, cited in Zölly, H., 1948. Geschichte der geodätischen Grundlagen für Karten und Vermessungen in der Schweiz. Stiftung Dr. Joachim de Giacomi, Schweizerischen Naturforschenden Gesellschaft, Eidgenössischen Landestopographie.

García-Davalillo, J., Herrera, G., Notti, D., Strozzi, T., Álvarez-Fernández, I., 2013. DInSAR analysis of ALOS PALSAR images for the assessment of very slow landslides: the Tena Valley case study. Landslides, 1-22.

George, R., Woodgate, P., 2002. Critical factors affecting the adoption of airborne geophysics for management of dryland salinity. Exploration Geophysics 33, 84-89.

Ghuffar, S., Székely, B., Roncat, A., Pfeifer, N., 2013. Landslide Displacement Monitoring Using 3D Range Flow on Airborne and Terrestrial LiDAR Data. Remote Sensing 5, 2720-2745.

Giordan, D., Allasia, P., Manconi, A., Baldo, M., Santangelo, M., Cardinali, M., Corazza, A., Albanese, V., Lollino, G., Guzzetti, F., 2013. Morphological and kinematic evolution of a large earthflow: The Montaguto landslide, southern Italy. Geomorphology 187, 61-79.

Glade, T., Anderson, M., Crozier, M., 2005. Landslide Hazard and Risk. Wiley, Chichester, England, 834.

González-Díez, A., Fernández-Maroto, G., Doughty, M.W., Díaz de Terán, J.R., Bruschi, V., Cardenal, J., Pérez, J.L., Mata, E., Delgado, J., 2013. Development of a methodological approach for the accurate measurement of slope changes due to landslides, using digital photogrammetry. Landslides, 1-14.

Gorum, T., van Westen, C.J., Korup, O., van der Meijde, M., Fan, X., van der Meer, F.D., 2013. Complex rupture mechanism and topography control symmetry of mass-wasting pattern, 2010 Haiti earthquake. Geomorphology 184, 127-138.

Grandin, R., Doin, M.-P., Bollinger, L., Pinel-Puysségur, B., Ducret, G., Jolivet, R., Sapkota, S.N., 2012. Long-term growth of the Himalaya inferred from interseismic InSAR measurement. Geology 40, 1059-1062.

Grandjean, G., Bitri, A., Krzeminska, D.M., 2011. Characterisation of a landslide fissure pattern by integrating seismic azimuth tomography and geotechnical testing. Hydrological Processes 26, 2120–2127.

Greslou, D., de Lussy, F., Amberg, V., Dechoz, C., Lenoir, F., Delvit, J.-M., Lebègue, L., 2013. PLEIADES-HR 1A & 1B image quality commissioning: innovative geometric calibration methods and results, SPIE Earth Observing Systems XVIII, 886611-886611-886612.

Grodecki, J., Dial, G., 2003. Block adjustment of high-resolution satellite images described by rational polynomials. Photgrammetric Engineering & Remote Sensing 69, 59-68.

Gruen, A., 1985. Adaptive least squares correlation: a powerful image matching technique. South African Journal of Photogrammetry, Remote Sensing and Cartography 14, 175-187.

Guzzetti, F., Malamud, B.D., Turcotte, D.L., Reichenbach, P., 2002. Power-law correlations of landslide areas in central Italy. Earth and Planetary Science Letters 195, 169-183.

Guzzetti, F., Mondini, A.C., Cardinali, M., Fiorucci, F., Santangelo, M., Chang, K.-T., 2012. Landslide inventory maps: New tools for an old problem. Earth-Science Reviews 112, 42-66.

Haralick, B., Lee, C.-N., Ottenberg, K., Nölle, M., 1994. Review and analysis of solutions of the three point perspective pose estimation problem. International Journal of Computer Vision 13, 331-356.

Harp, E.L., Keefer, D.K., Sato, H.P., Yagi, H., 2010. Landslide inventories: The essential part of seismic landslide hazard analyses. Engineering Geology.

Hartley, R.I., Zisserman, A., 2004 Multiple View Geometry in Computer Vision. Cambridge University Press.

Harwin, S., Lucieer, A., 2012. Assessing the Accuracy of Georeferenced Point Clouds Produced via Multi-View Stereopsis from Unmanned Aerial Vehicle (UAV) Imagery. Remote Sensing 4, 1573-1599.

He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21, 1263-1284.

Heid, T., Kääb, A., 2012. Evaluation of existing image matching methods for deriving glacier surface displacements globally from optical satellite imagery. Remote Sensing of Environment 118, 339-355.
Heim, A., 1882. Der Bergsturz von Elm. Zeitschrift der Deutschen Geologischen Gesellschaft 34, 74-115.
Heim, A., 1932. Bergsturz und Menschenleben. Fretz and Wasmuth, Zurich.
Henry, J.-B., Malet, J.-P., Maquaire, O., Grussenmeyer, P., 2002. The Use of Small-Format and Low-Altitude Aerial Photos for the Realization of High-Resolution DEMs in Mountainous Areas: Application to the Super-Sauze Earthflow (Alpes-De-Haute-Provence, France). Earth Surface Processes and Landforms 27, 1339–1350.
Heritage, G.L., Large, A.R.G., 2009. Laser scanning for the environmental sciences. Wiley-Blackwell, London.
Herrera, G., Fernández-Merodo, J.A., Mulas, J., Pastor, M., Luzi, G., Monserrat, O., 2009. A landslide forecasting model using ground based SAR data: The Portalet case study. Engineering Geology 105, 220-230.
Herrera, G., Gutiérrez, F., García-Davalillo, J.C., Guerrero, J., Notti, D., Galve, J.P., Fernández-Merodo, J.A., Cooksley, G., 2013. Multi-sensor advanced DInSAR monitoring of very slow landslides: The Tena Valley case study (Central Spanish Pyrenees). Remote Sensing of Environment 128, 31-43.
Hervás, J., Barredo, J.I., Rosin, P.L., Pasuto, A., Mantovani, F., Silvano, S., 2003. Monitoring landslides from optical remotely sensed imagery: the case history of Tessina landslide, Italy. Geomorphology 54, 63-75.
Hirschmuller, H., 2008. Stereo Processing by Semiglobal Matching and Mutual Information. IEEE Transaction on Pattern Analysis and Machine Intelligence 30, 328-341.
Hoja, D., Schneider, M., Müller, R., Lehner, M., Reinartz, P., 2008. Comparison of orthorectification methods suitable for rapid mapping using direct georeferencing and RPC for optical satellite data. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXVII-B4, 1617-1624.
Hollingsworth, J., Leprince, S., Ayoub, F., Avouac, J.-P., 2013. New constraints on dike injection and fault slip during the 1975–1984 Krafla rift crisis, NE Iceland. Journal of Geophysical Research: Solid Earth 118, 3707-3727.
Hooper, A., 2008. A multi-temporal InSAR method incorporating both persistent scatterer and small baseline approaches. Geophys. Res. Lett. 35, L16302.
Hooper, A., Bekaert, D., Spaans, K., Arıkan, M., 2012. Recent advances in SAR interferometry time series analysis for measuring crustal deformation. Tectonophysics 514–517, 1-13.
Hoppe, C., Klopschitz, M., Rumpler, M., Wendel, A., Kluckner, S., Bischof, H., Reitmayr, G., 2012. Online Feedback for Structure-from-Motion Image Acquisition, Vritish Machine Vision Conference, Guildford, UK, 1-12.
Hough, G., Green, J., Fish, P., Mills, A., Moore, R., 2011. A geomorphological mapping approach for the assessment of seabed geohazards and risk. Marine Geophysical Research 32, 151-162.
Iasio, C., Novali, F., Corsini, A., Mulas, M., Branzanti, M., Benedetti, E., Giannico, C., Tamburini, A., Mair, V., 2012. COSMO SkyMed high frequency - high resolution monitoring of an alpine slow landslide, corvara in Badia, Northern Italy, IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2012, 7577-7580.
Inaba, H., Itakura, Y., Kasahara, M., 2000. Surface velocity computation of debris flows by vector field measurements. Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere 25, 741-744.
InnovMetric, 2010. PolyWorks User's Manual - 3-D scanner and 3-D digitizer software from InnovMetric Software Inc. http://www.innovmetric.com/, [last access: 21/10/2013].
Intergraph, 2013. http://geospatial.intergraph.com/products/LPS/LPS/Details.aspx, [last access: 21/10/2013].
Intrieri, E., Gigli, G., Casagli, N., Nadim, F., 2013. Brief communication "Landslide Early Warning System: toolbox and general concepts". Nat. Hazards Earth Syst. Sci. 13, 85-90.

Intrieri, E., Gigli, G., Mugnai, F., Fanti, R., Casagli, N., 2012. Design and implementation of a landslide early warning system. Engineering Geology 147–148, 124-136.
IPCC, 2012. Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation, in: Field, C.B., V. Barros, T.F. Stocker, D. Qin, D.J. Dokken, K.L. Ebi, M.D. Mastrandrea, K.J. Mach,, G.-K. Plattner, S.K.A., M. Tignor, P.M. Midgley (Eds.), Cambridge, UK, 582.
Itakura, Y., Inaba, H., Sawada, T., 2005. A debris-flow monitoring devices and methods bibliography. Nat. Hazards Earth Syst. Sci. 5, 971-977.
Itakura, Y., Suwa, H., 1989. Measurement of surface velocity of debris flows by spatial filtering velocimetry, in: Tochiki, S. (Ed.), Landslides and Debris-Flows, The Japan-China Symposium. Japan Landslide Society and The Japan Society of Erosion Control Engineering, Niigata & Tokyo, Japan, 199–203.
Iverson, R.M., Reid, M.E., Logan, M., LaHusen, R.G., Godt, J.W., Griswold, J.P., 2010. Positive feedback and momentum growth during debris-flow entrainment of wet bed sediment. Nature Geoscience 4, 116-121.
Iwe, H., 2012. Ground based interferometric synthetic aperture radar for monitoring slowly moving surfaces. Faculty of Mathematics and Natural Sciences, University of Oslo, 184 p.
Jaboyedoff, M., Couture, R., Locat, P., 2009. Structural analysis of Turtle Mountain (Alberta) using digital elevation model: Toward a progressive failure. Geomorphology 103, 5-16.
Jaboyedoff, M., Oppikofer, T., Abellán, A., Derron, M.-H., Loye, A., Metzger, R., Pedrazzini, A., 2010. Use of LIDAR in landslide investigations: a review. Natural Hazards, 1-24.
James, M.R., Robson, S., 2012. Straightforward reconstruction of 3D surfaces and topography with a camera: Accuracy and geoscience application. Journal of Geophysical Research, in press.
James, M.R., Varley, N., 2012. Identification of structural controls in an active lava dome with high resolution DEMs: Volcán de Colima, Mexico. Geophysical Research Letters 39, L22303.
Joerg, P.C., Morsdorf, F., Zemp, M., 2012. Uncertainty assessment of multi-temporal airborne laser scanning data: A case study on an Alpine glacier. Remote Sensing of Environment 127, 118-129.
Johnstone, J.J., Mintzer, O.W., Parvis, M., Montano, P., Mile, R.D., Shepard, J.R., Frost, R.E., 1953. A manual on the airphoto interpretation of soils and rocks for engineering purposes. Purdue University, School of Civil Engineering and Engineering Mechanics, West Lafayette, USA.
Jolivet, R., Grandin, R., Lasserre, C., Doin, M.P., Peltzer, G., 2011. Systematic InSAR tropospheric phase delay corrections from global meteorological reanalysis data. Geophysical Research Letters 38.
Joyce, K.E., Belliss, S.E., Samsonov, S.V., McNeill, S.J., Glassey, P.J., 2009. A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters. Progress in Physical Geography 33, 183-207.
Julea, A., Méger, N., Bolon, P., Rigotti, C., Doin, M.P., Lasserre, C., Trouve, E., Lăzărescu, V.N., 2011. Unsupervised Spatiotemporal Mining of Satellite Image Time Series Using Grouped Frequent Sequential Patterns. IEEE Transactions on Geosciece and Remote Sensing 49, 1417-1430.
Jung, H.-S., Lu, Z., Zhang, L., 2013. Feasibility of Along-Track Displacement Measurement From Sentinel-1 Interferometric Wide-Swath Mode. IEEE Transactions on Geoscience and Remote Sensing 51, 573-578.
Kääb, A., 2002. Monitoring high-mountain terrain deformation from repeated air- and spaceborne optical data: examples using digital aerial imagery and ASTER data. ISPRS Journal of Photogrammetry and Remote Sensing 57, 39-52.
Kasperski, J., Delacourt, C., Allemand, P., Pothérat, P., 2010. Evolution of the Sedrun landslide (Graubünden, Switzerland) with ortho-rectified air images. Bulletin of Engineering Geology and the Environment 69, 421-430.
Kaufmann, V., 2012. The evolution of rock glacier monitoring using terrestrial photogrammetry: the example of Äußeres Hoch-ebenkar rock glacier (Austria). Austrian Journal of Earth Sciences 105, 63-77.

Kean, J.W., Staley, D.M., Cannon, S.H., 2011. In situ measurements of post-fire debris flows in southern California: Comparisons of the timing and magnitude of 24 debris-flow events with rainfall and soil moisture conditions. Journal of Geophysical Research 116, F04019.

Kerle, N., 2002. Volume estimation of the 1998 flank collapse at Casita volcano, Nicaragua: a comparison of photogrammetric and conventional techniques. Earth Surface Processes and Landforms 27, 759-772.

Kerle, N., Heuel, S., Pfeifer, N., 2008. Real-time data collection and information generation using airborne sensors, in: Zlatanova, S., Li, C. (Eds.), Geospatial Information Technology for Emergency Response. Taylor & Francis Group, London, 43-73.

Kettig, R.L., Landgrebe, D.A., 1976. Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects. IEEE Transactions on Geocience Electronics GE-14, 19-26.

Khattak, G.A., Owen, L.A., Kamp, U., Harp, E.L., 2010. Evolution of earthquake-triggered landslides in the Kashmir Himalaya, northern Pakistan. Geomorphology 115, 102–108.

Kirschbaum, D., Adler, R., Hong, Y., Kumar, S., Peters-Lidard, C., Lerner-Lam, A., 2012. Advances in landslide nowcasting: evaluation of a global and regional modeling approach. Environmental Earth Sciences 66, 1683-1696.

Kjekstad, O., Highland, L., 2009. Economic and Social Impacts of Landslides, in: Sassa, K., Canuti, P. (Eds.), Landslides: Disaster Risk Reduction. Springer, 573-587.

Kobayashi, T., Takada, Y., Furuya, M., Murakami, M., 2009. Locations and types of ruptures involved in the 2008 Sichuan earthquake inferred from SAR image matching. Geophysical Research Letters 36, L07302.

Krauskopf, K.B., Feitler, S., Griggs, A.B., 1939. Structural features of a landslide near Gilroy, California. The Journal of Geology 47, 630–648.

Krzeminska, D.M., Bogaard, T.A., Debieche, T.-H., Marc, V., Ponton, J., Malet, J.-P., 2009. Quantitative analysis of preferential flow during small scale infiltration tests on an active mudslide, French Alps, in: Malet, J.-P., Remaître, A., Boogard, T. (Eds.), International Conference on Landslide Processes: from geomorpholgic mapping to dynamic modelling. CERG Editions, Strasbourg, 151–156.

Krzeminska, D.M., Bogaard, T.A., Malet, J.P., H van Beek, L.P., 2012. A model of hydrological and mechanical feedbacks of preferential fissure flow in a slow-moving landslide. Hydrol. Earth Syst. Sci. Discuss. 9, 11161-11197.

Kumari, P., Carter, W.E., Shrestha, R.L., 2011. Adjustment of systematic errors in ALS data through surface matching. Advances in Space Research 47, 1851-1864.

Kurtz, C., Stumpf, A., Malet, J.-P., Puissant, A., Gançarski, P., Passat, N., In Press. Hierarchical extraction of landslides from multiresolution remotely sensed optical images. ISPRS Journal of Photogrammetry and Remote Sensing.

Lacroix, P., Zavala, B., Berthier, E., Audin, L., 2013. Supervised Method of Landslide Inventory Using Panchromatic SPOT5 Images and Application to the Earthquake-Triggered Landslides of Pisco (Peru, 2007, Mw8.0). Remote Sensing 5, 2590-2616.

Lague, D., Brodu, N., Leroux, J., 2013. Accurate 3D comparison of complex topography with terrestrial laser scanner: Application to the Rangitikei canyon (N-Z). ISPRS Journal of Photogrammetry and Remote Sensing 82, 10-26.

Lampert, T.A., Gancarski, P., In Review. The Bane of Skew: Uncertain Ranks and Unrepresentative Precision. Machine Learning.

Lampert, T.A., O'Keefe, S.E., 2010. A survey of spectrogram track detection algorithms. Applied acoustics 71, 87-100.

Lampert, T.A., Stumpf, A., Gançarski, P., In Review. An Empirical Study into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation. International Journal of Computer Vision.

LAMPRE, 2013 - 2015. http://www.lampre-project.eu/ [last access 31/10/2013].

Lane, S.N., Richards, K.S., Chandler, J.H., 1993. Developments in photogrammetry; the geomorphological potential. Progress in Physical Geography 17, 306-328.

Larsen, I.J., Montgomery, D.R., 2012. Landslide erosion coupled to tectonics and river incision. Nature Geosci 5, 468-473.
Lastras, G., Canals, M., Amblas, D., Ivanov, M., Dennielou, B., Droz, L., Akhmetzhanov, A., 2006. Eivissa slides, western Mediterranean Sea: morphology and processes. Geo-Marine Letters 26, 225-233.
Leberl, F., Irschara, A., Pock, T., Meixner, P., Gruber, M., Scholz, S., Wiechert, A., 2010. Point Clouds: Lidar versus 3 D Vision. Photogrammetric Engineering and Remote Sensing 76, 1123-1134.
Leprince, S., 2008. Monitoring Earth Surface Dynamics With Optical Imagery. California Institute of Technology, California Institute of Technology, 252 p.
Leprince, S., Barbot, S., Ayoub, F., Avouac, J.P., 2007. Automatic and Precise Orthorectification, Coregistration, and Subpixel Correlation of Satellite Images, Application to Ground Deformation Measurements. Geoscience and Remote Sensing, IEEE Transactions on 45, 1529-1558.
Leprince, S., Berthier, E., Ayoub, F., Delacourt, C., Avouac, J.P., 2008. Monitoring Earth Surface Dynamics With Optical Imagery. EOS, Transaction American Geophysical Union 89.
Ling, L., Yinqing, Z., Jingwen, L., Burgmann, R., 2011. Persistent scatterer SAR interferometry application on berkeley hills landslides, IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2011 4285-4287.
Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60, 91-110.
Lu, P., Stumpf, A., Kerle, N., Casagli, N., 2011. Object-Oriented Change Detection for Landslide Rapid Mapping. IEEE Geoscience and Remote Sensing Letters 8, 701-705.
Luca, L.D., Pierrot-Deseilligny, M., Remondino, F., Beninstant, P., Cléry, I., Lombardo, J., Nex, F., Nony, N., 2013. Tapenade - Tools and Acquisition Protocols for Enancing Artifacts Documentation, http://www.tapenade.gamsau.archi.fr/TAPEnADe/Protocols.html, [last access: 05-10-2013].
Lucieer, A., Turner, D., King, D.H., Robinson, S.A., Using an Unmanned Aerial Vehicle (UAV) to capture micro-topography of Antarctic moss beds. International Journal of Applied Earth Observation and Geoinformation.
Lussy, F., Greslou, D., Dechoz, C., Amberg, V., Delvit, J.M., Lebegue, L., Blanchet, G., Fourest, S., 2012. Pleiades HR in flight geometrical calibration : Location and mapping of the focal plane. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXIX-B1, 519-523.
Luzi, G., 2010. Ground Based SAR Interferometry: a Novel Tool for Geoscience, in: Imperatore, P., Riccio, D. (Eds.), Geoscience and Remote Sensing New Achievements. Intech.
Mackey, B., Roering, J., McKean, J., 2009. Long-term kinematics and sediment flux of an active earthflow, Eel River, California. Geology 37, 803-806.
Mackey, B.H., Roering, J.J., 2011. Sediment yield, spatial characteristics, and the long-term evolution of active earthflows determined from airborne LiDAR and historical aerial photographs, Eel River, California. Geological Society of America Bulletin 123, 1560-1576.
Maddalon, J.M., Hayhurst, K.J., Koppen, D.M., Upchurch, J.M., Morris, A.T., Verstynen, H.A., 2013. Perspectives on Unmanned Aircraft Classification for Civil Airworthiness Standards. National Aeronautics and Space Administration, Langley Research Center, Hampton, Virginia, 55.
Malberg, H., 1973. Comparison of mean cloud cover obtained by satellite photographs and ground-based observations over Europe and the Atlantic. Monthly Weather Review 101, 893-897.
Malet, J.P., Maquaire, O., Calais, E., 2002. The use of Global Positioning System techniques for the continuous monitoring of landslides: application to the Super-Sauze earthflow (Alpes-de-Haute-Provence, France). Geomorphology 43, 33-54.
Malet, J.P., van Asch, T.W.J., van Beek, R., Maquaire, O., 2005. Forecasting the behaviour of complex landslides with a spatially distributed hydrological model. Natural Hazards and Earth System Sciences 5, 71–85.

Manetti, L., Steinmann, G., 2007. 3DeMoN ROBOVEC—integration of a new measuring instrument in an existing generic remote monitoring platform, 7th international symposium on field measurements in geomechanics, 24-27.
Mantovani, F., Soeters, R., Van Westen, C.J., 1996. Remote sensing techniques for landslide studies and hazard zonation in Europe. Geomorphology 15, 213-225.
Marr, D., Hildreth, E., 1980. Theory of edge detection. Proceedings of the Royal Society of London Series B, Biological Sciences 207, 187–217.
Martha, T., Kerle, N., van Westen, C.J., Kumar, K., 2010a. Characterising spectral, spatial and morphometric properties of landslides for semi-automatic detection using object-oriented methods. Geomorphology 116, 24–36
Martha, T.R., Kerle, N., Jetten, V., van Westen, C.J., Vinod Kumar, K., 2010b. Landslide Volumetric Analysis Using Cartosat-1-Derived DEMs. IEEE Geoscience and Remote Sensing Letters 7, 582 - 586
Martha, T.R., Kerle, N., van Westen, C.J., Jetten, V., Kumar, K.V., 2011. Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis. IEEE Transactions on Geoscience and Remote Sensing 49, 4928-4943.
Martha, T.R., Kerle, N., van Westen, C.J., Jetten, V., Vinod Kumar, K., 2012. Object-oriented analysis of multi-temporal panchromatic images for creation of historical landslide inventories. ISPRS Journal of Photogrammetry and Remote Sensing 67, 105-119.
Massonnet, D., Rossi, M., Carmona, C., Adragna, F., Peltzer, G., Feigl, K., Rabaute, T., 1993. The displacement field of the Landers earthquake mapped by radar interferometry. Nature 364, 138-142.
Matsuura, S., Asano, S., Okamoto, T., 2008. Relationship between rain and/or meltwater, pore-water pressure and displacement of a reactivated landslide. Engineering Geology 101, 49-59.
McCalpin, J., 1984. Preliminary age classification of landslides for inventory mapping, 21st Annual Engineering Geology and Soils Engineering Symposium, Moscow, Idaho, 99–111.
McCoy, S.W., Kean, J.W., Coe, J.A., Staley, D.M., Wasklewicz, T.A., Tucker, G.E., 2010. Evolution of a natural debris flow: In situ measurements of flow dynamics, video imagery, and terrestrial laser scanning. Geology 38, 735-738.
Meisina, C., 2006. Characterisation of weathered clayey soils responsible for shallow landslides. Natural Hazards and Earth System Sciences 6, 825–838.
Metternicht, G., Hurni, L., Gogu, R., 2005. Remote sensing of landslides: An analysis of the potential contribution to geo-spatial systems for hazard assessment in mountainous environments. Remote Sensing of Environment 98, 284-303.
Michel, R., Avouac, J.-P., 2002. Deformation due to the 17 August 1999 Izmit, Turkey, earthquake measured from SPOT images. Journal of Geophysical Research 107, 2062.
Michel, R., Avouac, J.P., Taboury, J., 1999. Measuring ground displacements from SAR amplitude images : Application to the Landers earthquake. Geophysical Research Letters 26, 875-878.
Michoud, C., Bazin, S., Blikra, L.H., Derron, M.H., Jaboyedoff, M., 2013. Experiences from site-specific landslide early warning systems. Nat. Hazards Earth Syst. Sci. 13, 2659-2673.
Moine, M., Puissant, A., Malet, J.-P., 2009. Detection of landslides from aerial and satellite images with a semi-automatic method. Application to the Barcelonnette basin (Alpes-de-Haute-Provence, France), in: Malet, J.-P., Remaître, A., Bogaard, T. (Eds.), International Conference 'Landslide Processes', Strasbourg, France, 63-68.
Mondini, A.C., Marchesini, I., Rossi, M., Chang, K.-T., Pasquariello, G., Guzzetti, F., 2013. Bayesian framework for mapping and classifying shallow landslides exploiting remote sensing and topographic data. Geomorphology 201, 135-147.
Monserrat, O., Crosetto, M., 2008. Deformation measurement using terrestrial laser scanning data and least squares 3D surface matching. ISPRS J Photogramm Remote Sens 63, 142–154.
Motagh, M., Wetzel, H.-U., Roessner, S., Kaufmann, H., 2013. A TerraSAR-X InSAR study of landslides in southern Kyrgyzstan, Central Asia. Remote Sensing Letters 4, 657-666.
Motta, M., Gabrieli, F., Corsini, A., Manzi, V., Ronchetti, F., Cola, S., 2013. Landslide Displacement Monitoring from Multi-Temporal Terrestrial Digital Images: Case of the Valoria Landslide

Site, in: Margottini, C., Canuti, P., Sassa, K. (Eds.), Landslide Science and Practice. Springer Berlin Heidelberg, 73-78.
Nadim, F., Kjekstad, O., Peduzzi, P., Herold, C., Jaedicke, C., 2006. Global landslide and avalanche hotspots. Landslides 3, 159-173.
Nagai, M., Chen, T., Shibasaki, R., Kumagai, H., Ahmed, A., 2009. UAV-Borne 3-D Mapping System by Multisensor Integration. IEEE Transactions on Geoscience and Remote Sensing 47, 701-708.
Nascetti, A., 2013. High resolution radargrammetry: development and implementation of an innovative image matching strategy. Department of Civil, Constructional and Environmenal Engineering, Geodesy and Geomatic Division, Sapienza University of Rome, 189 p.
Nichol, J., Wong, M.S., 2005. Satellite remote sensing for detailed landslide inventories using change detection and image fusion. International Journal of Remote Sensing 26, 1913 - 1926.
Niethammer, U., James, M.R., Rothmund, S., Travelletti, J., Joswig, M., 2011a. UAV-based remote sensing of the Super-Sauze landslide: Evaluation and results. Engineering Geology 128, 2–11.
Niethammer, U., Rothmund, S., James, M.R., Travelletti, J., Joswig, M., 2010. UAV-based remote sensing of landslides. ISPRS Commission V Mid Term Symposium XXXVIII, Part 5, 496-501.
Niethammer, U., Rothmund, S., Schwaderer, U., Zeman, J., Joswig, M., 2011b. Open source image-processing tools for low cost UAV-based landslide investigations. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences Vol XXXVIII-1/C22.
Nikolaeva, E., Walter, T.R., Shirzaei, M., Zschau, J., 2013. Landslide dynamics and coupling revealed by L-band InSAR in central Georgia. Nat. Hazards Earth Syst. Sci. Discuss. 1, 4925-4962.
Nister, D., 2004. An efficient solution to the five-point relative pose problem. Pattern Analysis and Machine Intelligence, IEEE Transactions on 26, 756-770.
Noferini, L., Pieraccini, M., Mecatti, D., Luzi, G., Atzeni, C., Tamburini, A., Broccolato, M., 2005. Permanent scatterers analysis for atmospheric correction in ground-based SAR interferometry. Geoscience and Remote Sensing, IEEE Transactions on 43, 1459-1471.
Notti, D., Davalillo, J.C., Herrera, G., Mora, O., 2010. Assessment of the performance of X-band satellite radar data for landslide mapping and monitoring: Upper Tena Valley case study. Nat. Hazards Earth Syst. Sci. 10, 1865-1875.
Oppikofer, T., Bunkholt, H.S.S., Fischer, L., Saintot, A., Hermanns, R.L., Carrea, D., Longchamp, C., Derron, M.-H., Michoud, C., Jaboyedoff, M., 2012. Investigation and monitoring of rock slope instabilities in Norway by terrestrial laser scanning, in: al., E.e. (Ed.), Landslides and Engineered Slopes: Protecting Society through Improved Understanding, Proceedings of the 11th International and 2nd North American Symposium on Landslides and Engineered Slopes. Taylor & Francis Group, London, Banff, Canada, 1235-1241.
Oppikofer, T., Jaboyedoff, M., Blikra, L., Derron, M.H., Metzger, R., 2009. Characterization and monitoring of the Åknes rockslide using terrestrial laser scanning. Nat. Hazards Earth Syst. Sci. 9, 1003-1019.
Oppikofer, T., Jaboyedoff, M., Keusen, H.-R., 2008. Collapse at the eastern Eiger flank in the Swiss Alps. Nature Geosci 1, 531-535.
Optech, 2013. Gemini summary specification sheet. http://www.optech.ca/pdf/ALTM_Gemini_SpecSheet_110309_Web.pdf, [last access 19/10/2013].
Oswald, D., 2003. Analyse de l'activité de glissements de terrain et relation avec les conditions climatiques: Exemples dans les Préalpes fribourgeoises (Suisse). Faculté des Sciences, Université de Fribourg, 151 p.
OTB Development Team, 2013. http://orfeo-toolbox.org [last access 31/10/2013].
Paar, G., Huber, N.B., Bauer, A., Avian, M., Reiterer, A., 2012. Vision-Based Terrestrial Surface Monitoring, Terrigenous Mass Movements. Springer, 283-348.
Paine, J., Minty, B.S., 2005. Airborne Hydrogeophysics, in: Rubin, Y., Hubbard, S. (Eds.), Hydrogeophysics. Springer Netherlands, 333-357.
Pan, B., 2011. Recent Progress in Digital Image Correlation. Experimental Mechanics 51, 1223-1235.

Parise, M., 2003. Observation of surface features on an active landslide, and implications for understanding its history of movement. Natural Hazards and Earth System Sciences 3, 569–580.
Parker, R.N., Densmore, A.L., Rosser, N.J., de Michele, M., Li, Y., Huang, R., Whadcoat, S., Petley, D.N., 2011. Mass wasting triggered by the 2008 Wenchuan earthquake is greater than orogenic growth. Nature Geosci 4, 449-452.
Pasolli, E., Melgani, F., Alajlan, N., Conci, N., 2013. Optical Image Classification: A Ground-Truth Design Framework. IEEE Transactions on Geoscience and Remote Sensing 51, 3580-3597.
Petitjean, F., Inglada, J., Gançarski, P., 2012. Satellite image time series analysis under time warping. IEEE Transactions on Geoscience and Remote Sensing 50, 3081-3095.
Petley, D., 2012. Global patterns of loss of life from landslides. Geology 40, 927-930.
Pfaffhuber, A., Grimstad, E., Domaas, U., Auken, E., Foged, N., Halkjær, M., 2010. Airborne EM mapping of rockslides and tunneling hazards. The Leading Edge 29, 956-959.
Pieraccini, M., 2013. Real Beam vs. Synthetic Aperture Radar for Slope Monitoring, Progress In Electromagnetics Research Symposium, Stockholm, Sweden, 1629-1632.
Pieraccini, M., Casagli, N., Luzi, G., Tarchi, D., Mecatti, D., Noferini, L., Atzeni, C., 2003. Landslide monitoring by ground-based radar interferometry: A field test in Valdarno (Italy). International Journal of Remote Sensing 24, 1385-1391.
Pierrot-Deseilligny, M., Clery, I., 2012. Some possible protocols of acquisition for optimal use of the 'Apero' Open Source software in automatic orientation and calibration, Tutorial for EuroCow, Castelldefels, Espagne.
Pierrot-Deseilligny, M., Paparoditis, N., 2006. A Multiresolution and Optimization-Based Image Matching Approach : An Application to Surface Reconstruction from Spot 5- HRS Stereo Imagery, WG I/5 & I/6 Workshop on Topographic Mapping from Space (with Special Emphasis on Small Satellites), Ankara, Turkey.
Pinel-Puyssegur, B., Michel, R., Avouac, J.P., 2012. Multi-Link InSAR Time Series: Enhancement of a Wrapped Interferometric Database. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 5, 784-794.
Pipia, L., Fabregas, X., Aguasca, A., Lopez-Martinez, C., 2008. Atmospheric Artifact Compensation in Ground-Based DInSAR Applications. Geoscience and Remote Sensing Letters, IEEE 5, 88-92.
Plank, S., Singer, J., Minet, C., Thuro, K., 2012. Pre-survey suitability evaluation of the differential synthetic aperture radar interferometry method for landslide monitoring. International Journal of Remote Sensing 33, 6623-6637.
Plank, S., Singer, J., Thuro, K., 2013. Assessment of number and distribution of persistent scatterers prior to radar acquisition using open access land cover and topographical data. ISPRS Journal of Photogrammetry and Remote Sensing 85, 132-147.
Poli, D., Remondino, F., Angiuli, E., Agugiaro, G., 2013. Evaluation of Pleiades-1A triplet on Trento testfield. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL-1/W1, 287-292.
Poli, D., Toutin, T., 2012. Review of developments in geometric modelling for high resolution satellite pushbroom sensors. The Photogrammetric Record 27, 58-73.
Pollefeys, M., Koch, R., Gool, L., 1999. Self-Calibration and Metric Reconstruction Inspite of Varying and Unknown Intrinsic Camera Parameters. International Journal of Computer Vision 32, 7-25.
Pomeroy, J.S., 1982. Landslides in the greater Pittsburgh region, Pennsylvania. USGS Professional Paper, United States Gouvernment Printing Office, 66.
Powers, P.S., Chiarle, M., Savage, W.Z., 1996. A digital photogrammetric method for measuring horizontal surficial movements on the Slumgullion earthflow, Hinsdale County, Colorado. Computers & Geosciences 22, 651-663.
Prasad, S., Bruce, L.M., Chanussot, J., 2011. Optical Remote Sensing Advances in Signal Processing and Exploitation Technique, in: Hammoud, R., Wolff, L.B. (Eds.), Augmented Vision and Reality Springer, Berlin - Heidelberg, Germany, 351.

Prats, P., Scheiber, R., Reigber, A., Andres, C., Horn, R., 2009. Estimation of the Surface Velocity Field of the Aletsch Glacier Using Multibaseline Airborne SAR Interferometry. IEEE Transactions on Geoscience and Remote Sensing 47, 419-430.
PREVIEW, 2005 - 2008. http://www.preview-risk.com, [last access: 31/10/2013].
Prior, D.B., Bornhold, B.D., Coleman, J.M., Bryant, W.R., 1982. Morphology of a submarine slide, Kitimat Arm, British Columbia. Geology 10, 588-592.
Prochaska, A.B., Santi, P.M., Higgins, J.D., 2008. Relationships between size and velocity for particles within debris flows. Canadian Geotechnical Journal 45, 1778-1783.
Prokešová, R., Kardoš, M., Medveďová, A., 2010. Landslide dynamics from high-resolution aerial photographs: A case study from the Western Carpathians, Slovakia. Geomorphology 115, 90-101.
PST-A, 2008. Mappe di misure di accelerazioni derivate. Geoportale Nazionale, 6.
R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Raucoules, D., Bourgine, B., de Michele, M., Le Cozannet, G., Closset, L., Bremmer, C., Veldkamp, H., Tragheim, D., Bateson, L., Crosetto, M., Agudo, M., Engdahl, M., 2009. Validation and intercomparison of Persistent Scatterers Interferometry: PSIC4 project results. Journal of Applied Geophysics 68, 335-347.
Raucoules, D., de Michele, M., Malet, J.P., Ulrich, P., 2013. Time-variable 3D ground displacements from high-resolution synthetic aperture radar (SAR). application to La Valette landslide (South French Alps). Remote Sensing of Environment 139, 198-204.
Razak, K.A., Straatsma, M.W., van Westen, C.J., Malet, J.P., de Jong, S.M., 2011. Airborne laser scanning of forested landslides characterization: Terrain model quality and visualization. Geomorphology 126, 186-200.
Redpath, T.A.N., Sirguey, P., Fitzsimons, S.J., Kääb, A., 2013. Accuracy assessment for mapping glacier flow velocity and detecting flow dynamics from ASTER satellite imagery: Tasman Glacier, New Zealand. Remote Sensing of Environment 133, 90-101.
Reeh, N., Mohr, J.J., Madsen, S.N., Oerter, H., Gundestrup, N.S., 2003. Threedimensional surface velocities of Storstrømmen glacier, Greenland, derived from radar interferometry and ice-sounding radar measurements. Journal of Glaciology 49, 201−209.
Reigber, A., Scheiber, R., Jager, M., Prats-Iraola, P., Hajnsek, I., Jagdhuber, T., Papathanassiou, K.P., Nannini, M., Aguilera, E., Baumgartner, S., Horn, R., Nottensteiner, A., Moreira, A., 2013. Very-High-Resolution Airborne Synthetic Aperture Radar Imaging: Signal Processing and Applications. Proceedings of the IEEE 101, 759-783.
Reinartz, P., Müller, R., Lehner, M., Schroeder, M., 2006. Accuracy analysis for DSM and orthoimages derived from SPOT HRS stereo data using direct georeferencing. ISPRS Journal of Photogrammetry and Remote Sensing 60, 160-169.
Reiterer, A., Lehmann, M., Miljanovic, M., Ali, H., Paar, G., Egly, U., Eiter, T., Kahmen, H., 2009. A 3D optical deformation measurement system supported by knowledge-based and learning techniques. Journal of Applied Geodesy 3, 1-13.
Remondino, F., Fraser, C., 2006. Digital camera calibration methods: considerations and comparisons. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 36, 266-272.
Remondino, F., Pizzo, S., Kersten, T., Troisi, S., 2012. Low-Cost and Open-Source Solutions for Automated Image Orientation – A Critical Overview, in: Ioannides, M., Fritsch, D., Leissner, J., Davies, R., Remondino, F., Caffo, R. (Eds.), Progress in Cultural Heritage Preservation. Springer Berlin Heidelberg, 40-54.
Rib, H.T., Liang, T., 1978. Recognition and identification, in: Schuster, R.L., Krizek, R.J. (Eds.), Landslide Analysis and Control National Academy of Sciences, Washington DC, 34–80.
Richards, J.A., 2009. Remote sensing with imaging Radar. Springer.
Riedel, B., Walther, A., 2008. InSAR processing for the recognition of landslides. Adv. Geosci. 14, 189-194.

Righini, G., Pancioli, V., Casagli, N., 2011. Updating landslide inventory maps using Persistent Scatterer Interferometry (PSI). International Journal of Remote Sensing 33, 2068-2096.
Rivera, A., Corripio, J., Bravo, C., Cisternas, S., 2012. Glaciar Jorge Montt (Chilean Patagonia) dynamics derived from photos obtained by fixed cameras and satellite image feature tracking. Annals of Glaciology 53, 147.
Rocca, F., 2003. 3D Motion recovery with multi-angle and/or left right Interferometry., 3rd InternationalWorkshop on ERS SAR Interferometry (FRINGE 2003), Frascati, Italy.
Roering, J.J., Stimely, L.L., Mackey, B.H., Schmidt, D.A., 2009. Using DInSAR, airborne LiDAR, and archival air photos to quantify landsliding and sediment transport. Geophys. Res. Lett. 36, L19402.
Rohmer, J., Raucoules, D., 2012. On the applicability of Persistent Scatterers Interferometry (PSI) analysis for long term CO2 storage monitoring. Engineering Geology 147–148, 137-148.
Rosser, N., Lim, M., Petley, D., Dunning, S., Allison, R., 2007. Patterns of precursory rockfall prior to slope failure. J. Geophys. Res. 112, F04014.
Rosser, N.J., Brain, M.J., Petley, D.N., Lim, M., Norman, E.C., 2013. Coastline retreat via progressive failure of rocky coastal cliffs. Geology 41, 939-942.
Rosser, N.J., Petley, D.N., Lim, M., Dunning, S., Allison, R.J., 2005. Terrestrial laser scanning for monitoring the process of hard rock coastal cliff erosion. Quarterly Journal of Engineering Geology and Hydrogeology 38, 363-375.
Rothermel, M., Wenzel, K., Fritsch, D., Haala, N., 2012. SURE: Photogrammetric Surface Reconstruction from Imagery, LC3D Workshop, Berlin, Germany.
Rothmund, S., Niethammer, U., Malet, J.-P., Joswig, M., 2013. Landslide surface monitoring based on UAV- and ground-based images and terrestrial laser scanning: accuracy analysis and morphological interpretation. First Break, Special topic: Near Surface Geoscience 31.
Rucci, A., Ferretti, A., Monti Guarnieri, A., Rocca, F., 2012. Sentinel 1 SAR interferometry applications: The outlook for sub millimeter measurements. Remote Sensing of Environment 120, 156-163.
Rüeger, J.M., 1990. History, Electronic Distance Measurement. Springer Berlin Heidelberg, 1-2.
Russell, S., Norvig, P., 2009. Artificial Intelligence: A Modern Approach. Pearson.
SafeLand, 2009-2011. Living with landslide risk in Europe. http://www.safeland-fp7.eu, [last access: 31/10/2013].
SAFER, 2007-2013. http://safer.emergencyresponse.eu, [last access: 31/10/2013].
Santangelo, M., Cardinali, M., Rossi, M., Mondini, A.C., Guzzetti, F., 2010. Remote landslide mapping using a laser rangefinder binocular and GPS. Nat. Hazards Earth Syst. Sci. 10, 2539-2546.
Scambos, T.A., Dutkiewicz, M.J., Wilson, J.C., Bindschadler, R.A., 1992. Application of image cross-correlation to the measurement of glacier velocity using satellite image data. Remote Sensing of Environment 42, 177-186.
Scharstein, D., Szeliski, R., 2002. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. International Journal of Computer Vision 47, 7-42.
Scherer, M., Lerma, J.L., 2009. From the Conventional Total Station to the Prospective Image Assisted Photogrammetric Scanning Total Station: Comprehensive Review. Journal of Surveying Engineering 135, 173-178.
Scherler, D., Leprince, S., Strecker, M.R., 2008. Glacier-surface velocities in alpine terrain from optical satellite imagery - Accuracy improvement and quality assessment. Remote Sensing of Environment 112, 3806-3819.
Schlögel, R., Doubre, C., Malet, J.-P., T., L., in elaboration. ALOS/PALSAR DInSAR processing for the monitoring of landslides characterized by large displacements.
Schwab, M., Rieke-Zapp, D., Schneider, H., Liniger, M., Schlunegger, F., 2008. Landsliding and sediment flux in the Central Swiss Alps: A photogrammetric study of the Schimbrig landslide, Entlebuch. Geomorphology 97, 392-406.
Schwarz, D.F., König, I.R., Ziegler, A., 2010. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics 26, 1752-1758.

Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R., 2006. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms, Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, 519-528.
Settles, B., 2010. Active learning literature survey. Computer Sciences Technical Report 1648, 67.
Shan, J., Toth, C.K., 2009. Topographic Laser Ranging and Scanning, Principles and Processing. CRC Press, Taylor and Francis Group, 590.
Shao, Y., Guo, B., Hu, X., Di, L., 2011. Application of a fast linear feature detector to road extraction from remotely sensed imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 4, 626–631.
Shreve, R.L., 1966. Sherman landslide, Alaska. Science 154, 1639–1643.
Singhroy, V., Molch, K., 2004. Characterizing and monitoring rockslides from SAR techniques. Advances in Space Research 30, 290-295.
SLAM, 2003-2005. http://dup.esrin.esa.it, [last access: 31/10/2013].
Smith, A., 2010. Image segmentation scale parameter optimization and land cover classification using the Random Forest algorithm. Journal of Spatial Science 55, 69-79.
Snavely, N., Seitz, S., Szeliski, R., 2008. Modeling the World from Internet Photo Collections. International Journal of Computer Vision 80, 189-210.
Soares, J.V.B., Leandro, J.J.G., Cesar, R.M., Jelinek, H.F., Cree, M.J., 2006. Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification. IEEE Transactions on Medical Imaging 25, 1214–1222.
Sowers, G.F., Royster, D.L., 1978. Field Investigation, in: Schuster, R.L., Krizek, R.J. (Eds.), Landslides: Analysis and Control. Transportation Research Board, Special Report 176, National Academy of Science, Washington, D.C., 81–111.
Squarzoni, C., Delacourt, C., Allemand, P., 2003. Nine years of spatial and temporal evolution of the La Valette landslide observed by SAR interferometry. Engineering Geology 68, 53-66.
Sterzai, P., Vellico, M., Berti, M., Coren, F., Corsini, A., Rosi, A., Mora, P., Zambonelli, F., Ronchetti, F., 2010. LiDAR and hyperspectral data integration for landslide monitoring: The Test Case Of Valoria Landslide. Italian Journal of Remote Sensing 42, 89-99.
Strecha, C., Von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U., 2008. On benchmarking camera calibration and multi-view stereo for high resolution imagery, Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, 1-8.
Strozzi, T., Farina, P., Corsini, A., Ambrosi, C., Thüring, M., Zilger, J., Wiesmann, A., Wegmüller, U., Werner, C., 2005. Survey and monitoring of landslide displacements by means of L-band satellite SAR interferometry. Landslides 2, 193-201.
Strozzi, T., Luckman, A., Murray, T., Wegmuller, U., Werner, C.L., 2002. Glacier motion estimation using SAR offset-tracking procedures. Geoscience and Remote Sensing, IEEE Transactions on 40, 2384-2391.
Stumpf, A., Kerle, N., 2011. Object-oriented mapping of landslides using Random Forests. Remote Sensing of Environment 115, 2564–2577.
Stumpf, A., Lachiche, N., Malet, J.-P., Kerle, N., Puissant, A., 2013a. Active learning in the spatial domain for remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 16.
Stumpf, A., Malet, J.-P., Allemand, P., Submitted. Stereo-photogrammetry and displacement monitoring with Pleiades VHR satellite images. ISPRS Journal of Photogrammetry and Remote Sensing, 33 p.
Stumpf, A., Malet, J.-P., Allemand, P., Deseilligny, M.-P., Skupinski, G., In Review-a. Terrestrial multi-view photogrammetry for landslide monitoring Journal of Geophysical Research.
Stumpf, A., Malet, J.-P., Kerle, N., Michoud, C., Tofani, V., Segoni, S., Jaboyedoff, M., Casagli, N., In Review-b. Criteria for the selection of appropriate remote sensing technologies for landslide monitoring. Earth Sciences Review. Earth Science Reviews.
Stumpf, A., Malet, J.-P., Kerle, N., Niethammer, U., Rothmund, S., 2013b. Image-based mapping of surface fissures for the investigation of landslide dynamics. Geomorphology 186, 12-27.

Sturzenegger, M., Stead, D., 2009a. Close-range terrestrial digital photogrammetry and terrestrial laser scanning for discontinuity characterization on rock cuts. Engineering Geology 106, 163-182.

Sturzenegger, M., Stead, D., 2009b. Quantifying discontinuity orientation and persistence on high mountain rock slopes and large landslides using terrestrial remote sensing techniques. Natural Hazards and Earth System Sciences 9, 267–287.

Supper, R., Baroň, I., Ottowitz, D., Motschka, K., Gruber, S., Winkler, E., Jochum, B., Römer, A., 2013. Airborne geophysical mapping as an innovative methodology for landslide investigation: evaluation of results from the Gschliefgraben landslide, Austria. Nat. Hazards Earth Syst. Sci. Discuss. 1, 2281-2318.

Supper, R., Römer, A., Jochum, B., Bieber, G., Jaritz, W., 2008. A complex geo-scientific strategy for landslide hazard mitigation – from airborne mapping to ground monitoring. Advances in Geosciences 14, 195–200.

Sutton, M.A., Orteu, J.-J., Schreier, H.W., 2009. Image correlation for shape, motion and deformation measurements: basic concepts, theory and applications. Springer.

Talling, P.J., Wynn, R.B., Masson, D.G., Frenz, M., Cronin, B.T., Schiebel, R., Akhmetzhanov, A.M., Dallmeier-Tiessen, S., Benetti, S., Weaver, P.P.E., Georgiopoulou, A., Zuhlsdorff, C., Amy, L.A., 2007. Onset of submarine debris flow deposition far from original giant landslide. Nature 450, 541-544.

Tamburini, A., Del Conte, S., Larini, G., Lopardo, L., Malaguti, C., Vescovi, P., 2011. Application of SqueeSAR TM to the characterization of deep seated gravitational slope deformations: the Berceto case study (Parma, Italy), Proceedings of the Second World Landslide Forum, 7.

Tarchi, D., Casagli, N., Fanti, R., Leva, D.D., Luzi, G., Pasuto, A., Pieraccini, M., Silvano, S., 2003. Landslide monitoring by using ground-based SAR interferometry: an example of application to the Tessina landslide in Italy. Engineering Geology 68, 15-30.

Tarchi, D., Oliveri, F., Sammartino, P.F., 2013. MIMO Radar and Ground-Based SAR Imaging Systems: Equivalent Approaches for Remote Sensing. IEEE Transactions on Geoscience and Remote Sensing 51, 425-435.

Taylor, M.H., Leprince, S., Avouac, J.-P., Sieh, K., 2008. Detecting co-seismic displacements in glaciated regions: An example from the great November 2002 Denali earthquake using SPOT horizontal offsets. Earth and Planetary Science Letters 270, 209-220.

TERRAFIRMA, 2003-2012. http://www.terrafirma.eu.com, [last access: 31/10/2013].

Teza, G., Galgaro, A., Zaltron, N., Genevois, R., 2007. Terrestrial laser scanner to detect landslide displacement fields: a new approach. International Journal of Remote Sensing 28, 3425 - 3446.

Thomson, S., Fountain, D., Watts, T., 2007. Airborne geophysics—Evolution and revolution, Proceedings of Exploration, 19-37.

Thuro, K., Wunderlich, T., Heunecke, O., Singer, J., Schuhbäck, S., Wasmeier, P., Glabsch, J., Festl, J., 2009. Low cost 3D early warning system for instable alpine slopes – the Aggenalm Landslide monitoring system. Kostengünstiges 3D Frühwarnsystem für alpine instabile Hänge – Das Überwachungssystem der Aggenalm-Hangbewegung. Geomechanik und Tunnelbau 2, 221-237.

Tinel, C., Fontannaz, D., De Boissezon, H., Grizonnet, M., Michel, J., 2012. The ORFEO accompaniment program and ORFEO ToolBox, IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2012, , 7102-7105.

Tofani, V., Raspini, F., Catani, F., Casagli, N., 2013a. Persistent Scatterer Interferometry (PSI) Technique for Landslide Characterization and Monitoring. Remote Sensing 5, 1045-1065.

Tofani, V., Segoni, S., Agostini, A., Catani, F., Casagli, N., 2013b. Technical Note: Use of remote sensing for landslide studies in Europe. Nat. Hazards Earth Syst. Sci. 13, 299-309.

Torralba, A., Oliva, A., 2003. Statistics of natural image categories. Network: Computation in Neural Systems 14, 391-412.

Toutin, T., Schmitt, C.V., Wang, H., 2012. Impact of no GCP on elevation extraction from WorldView stereo data. ISPRS Journal of Photogrammetry and Remote Sensing 72, 73-79.

Travelletti, J., 2011. Imagerie multi-paramètres et multi-résolutions pour l'observation et la caractérisation des mécanismes de glissements-coulées. Laboratoire de Géographie, Physique et Environnement, Université de Caen Basse-Normandie.

Travelletti, J., Delacourt, C., Allemand, P., Malet, J.P., Schmittbuhl, J., Toussaint, R., Bastard, M., 2012. Correlation of multi-temporal ground-based optical images for landslide monitoring: Application, potential and limitations. ISPRS Journal of Photogrammetry and Remote Sensing 70, 39-55.

Travelletti, J., Delacourt, C., Malet, J.-P., In Review. Multi-date correlation of Terrestrial Laser Scanning data for the characterization of landslide kinematics. International Journal of Applied Earth Observation and Geoinformation, 15.

Travelletti, J., Malet, J.P., Samyn, K., Grandjean, G., Jaboyedoff, M., 2013. Control of landslide retrogression by discontinuities: evidence by the integration of airborne- and ground-based geophysical information. Landslides 10, 37-54.

Travelletti, J., Oppikofer, T., Delacourt, C., Malet, J., Jaboyedoff, M., 2008. Monitoring landslide displacements during a controlled rain experiment using a long-range terrestrial laser scanning (TLS). Int Arch Photogramm Remote Sens Spatial Inf Sci 37, 485-490.

Triggs, B., McLauchlan, P., Hartley, R., Fitzgibbon, A., 2000. Bundle Adjustment — A Modern Synthesis, in: Triggs, B., Zisserman, A., Szeliski, R. (Eds.), Vision Algorithms: Theory and Practice. Springer Berlin Heidelberg, 298-372.

Tsai, Z.-X., You, G.J.Y., Lee, H.-Y., Chiu, Y.-J., 2012. Use of a total station to monitor post-failure sediment yields in landslide sites of the Shihmen reservoir watershed, Taiwan. Geomorphology 139–140, 438-451.

Tsutsui, K., Rokugawa, S., Nakagawa, H., Miyazaki, S., Chin-Tung Cheng Shiraishi, T., Yang, S.-D., 2007. Detection and Volume Estimation of Large-Scale Landslides Based on Elevation-Change Analysis Using DEMs Extracted From High-Resolution Satellite Stereo Imagery IEEE Transactions on Geoscience and Remote Sensing Letters 45, 1681 - 1696

Tuia, D., Muñoz-Marí, J., Camps-Valls, G., 2012. Remote sensing image segmentation by active queries. Pattern Recognition 45, 2180-2192.

Tuia, D., Ratle, F., Pacifici, F., Kanevski, M.F., Emery, W.J., 2009. Active learning methods for remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing 47, 2218-2232.

Turner, D., Lucieer, A., Wallace, L., In Press. Direct Georeferencing of Ultrahigh-Resolution UAV Imagery. IEEE Transactions on Geosciece and Remote Sensing, 8.

Ullman, S., 1979. The interpretation of visual motion. MIT Press, Cambridge, Massachusetts, and London, England.

Ussyshkin, R.V., Smith, B., 2006. Performance analysis of ALTM 3100EA: Instrument specification and accuracy of LiDAR data. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 34, Part XXX.

Ussyshkin, V., Boba, M., Sitar, M., 2008. Performance Characterization of an Airborne Lidar System: Bridging System Specifications and Expected Performance. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 37.

van Asch, T., Malet, J.P., 2009. Flow-type failures in fine-grained soils: an important aspect in landslide hazard analysis. Nat. Hazards Earth Syst. Sci. 9, 1703-1711.

van Asch, T.W.J., van Beek, L.P.H., Bogaard, T.A., 2009. The diversity in hydrological triggering systems of landslides, The First Italian Workshop on Landslides, Napoli, Italy 151–156.

Van Den Eeckhaut, M., Hervás, J., 2012. State of the art of national landslide databases in Europe and their potential for assessing landslide susceptibility, hazard and risk. Geomorphology 139–140, 545-558.

Van Den Eeckhaut, M., Kerle, N., Poesen, J., Hervás, J., 2012. Object-oriented identification of forested landslides with derivatives of single pulse LiDAR data. Geomorphology 173–174, 30-42.

Van Puymbroeck, N., Michel, R., Binet, R., Avouac, J.-P., Taboury, J., 2000. Measuring Earthquakes from Optical Satellite Images. Appl. Opt. 39, 3486-3494.

van Westen, C.J., Castellanos, E., Kuriakose, S.L., 2008. Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview. Engineering Geology 102, 112-131.
van Westen, C.J., van Asch, T.W.J., Soeters, R., 2006. Landslide hazard and risk zonation—why is it still so difficult? Bulletin of Engineering Geology and the Environment 65, 167-184.
Vaunat, J., Leroueil, S., 2002. Analysis of Post-Failure Slope Movements within the Framework of Hazard and Risk Analysis. Natural Hazards 26, 81-107.
Vedaldi, A., Fulkerson, B., 2008. VLFeat: An Open and Portable Library of Computer Vision Algorithms. http://www.vlfeat.org/, [last access: 21/10/2013].
Ventura, G., Vilardo, G., Terranova, C., Sessa, E.B., 2011. Tracking and evolution of complex active landslides by multi-temporal airborne LiDAR data: The Montaguto landslide (Southern Italy). Remote Sensing of Environment 115, 3237-3248.
Vermeesch, P., Drake, N., 2008. Remotely sensed dune celerity and sand flux measurements of the world's fastest barchans (Bodélé, Chad). Geophysical Research Letters 35.
Wallace, L., Lucieer, A., Watson, C., Turner, D., 2012. Development of a UAV-LiDAR System with Application to Forest Inventory. Remote Sensing 4, 1519-1543.
Walstra, J., Chandler, J.H., Dixon, N., Dijkstra, T.A., 2007. Aerial photography and digital photogrammetry for landslide monitoring. Geological Society, London, Special Publications 283, 53-63.
Walstra, J., J.H. Chandler, N. Dixon, Dijkstra, T.A., 2004. Time for Change - Quantifying Land-slide Evolution Using Historical Aerial Photographs and Modern Photogrammetric Methods. International Archives Photogrammetry, Remote Sensing and Spatial Information Science XXXV, 475-480.
Walter, M., Niethammer, U., Rothmund, S., Joswig, M., 2009. Joint analysis of the Super-Sauze (French Alps) mudslide by nanoseismic monitoring and UAV-based remote sensing. first break 27, 53–60.
Wenzel, K., Rothermel, M., Fritsch, D., Haala, N., 2013. Image Acquisition and model selection for multi-view stereo, 3D-ARCH 2013 - 3D Virtual Reconstruction and Visualization of Complex Architectures, Trento, Italy.
Westoby, M.J., Brasington, J., Glasser, N.F., Hambrey, M.J., Reynolds, J.M., 2012. 'Structure-from-Motion' photogrammetry: A low-cost, effective tool for geoscience applications. Geomorphology 179, 300-314.
Whitehead, K., Moorman, B.J., Hugenholtz, C.H., 2013. Low-cost, on-demand aerial photogrammetry for glaciological measurement. The Cryosphere Discuss. 7, 3043-3057.
Wills, C.J., McCrink, T.P., 2002. Comparing Landslide Inventories: The Map Depends on the Method. Environmental and Engineering Geoscience 8, 279-293.
Wu, C., 2007. A GPU Implementation of Scale Invariant Feature Transform SIFT. http://cs.unc.edu/~ccwu/siftgpu, [last access: 31/10/2013].
Wu, C., 2013. Towards Linear-time Incremental Structure from Motion, International Conference on 3D Vision, University of Washington, Seattle, USA, 8.
Wu, C., Agarwal, S., Curless, B., Seitz, S.M., 2011. Multicore bundle adjustment, Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 3057-3064.
Wu, J., Gilliéron, P.-Y., Merminod, B., 2012. Cell-based Automatic Deformation Computation by Analyzing Terrestrial Lidar Point Clouds. Photogrammetric Engineering and Remote Sensing 78, 317-329.
Xia, Y., Kaufmann, H., Guo, X.F., 2004. Landslide Monitoring in the Three Gorges Area Using D-INSAR and Corner Reflectors. PHOTOGRAMMETRIC ENGINEERING & REMOTE SENSING 70, 1167–1172.
Xu, F., Woodhouse, N., Xu, Z., Marr, D., Yang, X., Wang, Y., 2008. Blunder elimination techniques in adaptive automatic terrain extraction. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 29, 1139-1144.
Yamaguchi, Y., Tanaka, S., Odajima, T., Kamai, T., Tsuchida, S., 2003. Detection of a landslide movement as geometric misregistration in image matching of SPOT HRV data of two different dates. International Journal of Remote Sensing 24, 3523-3534.

Yang, I., Park, J., Kim, D., 2007. Monitoring the symptoms of landslide using the non-prism total station. KSCE Journal of Civil Engineering 11, 293-301.

Zhang, B.T., Veenker, G., 1991. Neural networks that teach themselves through genetic discovery of novel examples, Neural Networks, 1991. 1991 IEEE International Joint Conference on, 690-695 vol.691.

Zhang, L., Gruen, A., 2004. Automatic DSM Generation from Linear Array Imagery Data. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXV-B3, 128-133.

Zölly, H., 1948. Geschichte der geodätischen Grundlagen für Karten und Vermessungen in der Schweiz. Stiftung Dr. Joachim de Giacomi, Schweizerischen Naturforschenden Gesellschaft, Eidgenössischen Landestopographie.

Appendices

Appendix 1: Lu et al. 2011. Object-Oriented Change Detection for Landslide Rapid Mapping.

Lu, P., Stumpf, A., Kerle, N., Casagli, N., 2011. Object-Oriented Change Detection for Landslide Rapid Mapping. Geoscience and Remote Sensing Letters, IEEE 8, 701-705.

Object-Oriented Change Detection for Landslide Rapid Mapping

Ping Lu, André Stumpf, Norman Kerle, and Nicola Casagli

***Abstract*—A complete multitemporal landslide inventory, ideally updated after each major event, is essential for quantitative landslide hazard assessment. However, traditional mapping methods, which rely on manual interpretation of aerial photographs and intensive field surveys, are time consuming and not efficient for generating such event-based inventories. In this letter, a semiautomatic approach based on object-oriented change detection for landslide rapid mapping and using very high resolution optical images is introduced. The usefulness of this methodology is demonstrated on the Messina landslide event in southern Italy that occurred on October 1, 2009. The algorithm was first developed in a training area of Altolia and subsequently tested without modifications in an independent area of Itala. Correctly detected were 198 newly triggered landslides, with user accuracies of 81.8% for the number of landslides and 75.9% for the extent of landslides. The principal novelties of this letter are as follows: 1) a fully automatic problem-specified multiscale optimization for image segmentation and 2) a multitemporal analysis at object level with several systemized spectral and textural measurements.**

***Index Terms*—Change detection, landslide, object-oriented analysis (OOA), rapid mapping.**

I. Introduction

LANDSLIDING is a major natural hazard with growing global concern. A recent study has indicated landslide occurrences to be geomorphological indicators of global climate change [1], similar to Jakob and Lambert [2], who reported that the potential influence of climate change is reflected in an increase of landslide frequency. For a successful quantitative landslide hazard assessment, a compilation of historical event-based landslide inventory is crucial [3].

Traditionally, landslide mapping has relied on visual interpretation of aerial photographs and intensive field surveys. However, for mapping of large areas, those methods are time consuming, creating a gap that remote sensing has been increasingly filling. Due to restrictions in spatial resolution, traditional optical satellite imagery, such as acquired by the Landsat Thematic Mapper (TM), has limited utility for landslide studies [4]. More recently, high-resolution images and Light Detection and Ranging (LiDAR) elevation derivatives have started to offer an alternative way for effective landslide mapping. Most research works, however, have been focusing on pixel-based analysis. For example, Borghuis *et al.* [5] employed unsupervised image classification in automated landslide mapping using Satellite Pour l'Observation de la Terre 5 (SPOT-5) imagery. McKean and Roering [6] also successfully delineated landslide features using measures of surface roughness from LiDAR digital terrain model (DTM). With increasing spatial resolution, however, pixel-based methods have fundamental limitations in addressing particular landslide characteristics due to finite spatial extent. Only those object characteristics allow landslides to be further assigned to different type classes and other features of similar appearance to be discarded. Such methods focusing on features instead of pixels are the basis of object-oriented analysis (OOA).

OOA, which is based on image segmentation and subsequent classification of derived image primitives, represents a more advantageous approach for analyzing high-resolution data because image pixels can be meaningfully grouped into networked homogeneous objects and noise consequently reduced [7], [8]. Moreover, OOA offers a potentially automated approach for landslide mapping, with a consideration of spectral, morphological, and contextual landslide features supported by expert knowledge [9], thus allowing a cognitive approach that is comparable to visual image analysis. Nonetheless, so far, very few studies have used OOA for landslide mapping. Preliminary efforts by Barlow *et al.* [10] and Martin and Franklin [11] focused on automatic landslide detection using low-resolution Landsat Enhanced TM Plus (ETM+) images. The methodology was further improved by Barlow *et al.* [12] through the use of higher resolution SPOT-5 data, as well as an inclusion of more robust geomorphic variables. Moreover, Martha *et al.* [9] integrated spectral, spatial, and morphometric features to successfully recognize and classify five different types of landslides in difficult terrain in the High Himalayas. These studies show the increasing utility and potential of OOA in detecting and mapping landslides rapidly. However, all of the proposed approaches tend to fail in situations where both fresh and older landslides are present and prevent an accurate event-related landslide inventory.

A potential solution could be the integration of pre-event image data. Change detection from satellite imagery before and after a landslide event has already been proven useful for identification of newly triggered landslides at pixel-based level. Most frequently, change detection has been based on image ratios and image differencing with a defined threshold [4]. In addition, image subtraction and postclassification comparison

Manuscript received September 28, 2010; revised November 17, 2010; accepted December 13, 2010. Date of publication February 9, 2011; date of current version June 24, 2011.

P. Lu and N. Casagli are with the Department of Earth Sciences, University of Firenze, 50121 Florence, Italy (e-mail: ping.lu@unifi.it; nicola.casagli@unifi.it).

A. Stumpf is with the Faculty of Geo-Information Science and Earth Observation, University of Twente, 7500 AE Enschede, The Netherlands, and also with the Laboratoire Image, Ville et Environnement, University of Strasbourg, 67000 Strasbourg, France (e-mail: stumpf24883@itc.nl).

N. Kerle is with the Faculty of Geo-Information Science and Earth Observation, University of Twente, 7500 AE Enschede, The Netherlands (e-mail: kerle@itc.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2010.2101045

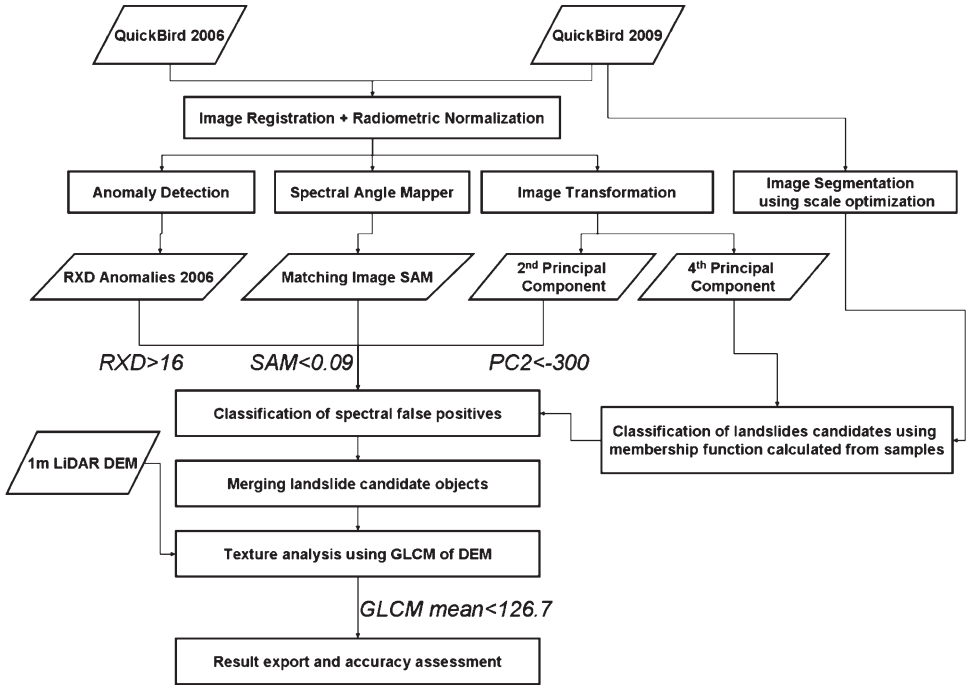


Fig. 1. General flowchart of landslide mapping by OOA change detection. (RXD) Reed–Xiaoli detector. (SAM) Spectral angle mapper. (PC) Principal component. (GLCM) Gray-level co-occurrence matrix.

have been attempted. For example, Nichol and Wong [13] reported that a postclassification comparison using a maximum likelihood classifier produced a detection rate of 70%. Park and Chi [14] introduced the concept of change detection into OOA, using very high resolution (VHR) images before and after landslide occurrence. However, their aim was only to recognize change/nonchange objects, without further efforts to remove those false positives from the "changed" objects. The approach is apparently only suited for situations where all major changes are induced by landslides and all landslides occurred in forested terrain.

The purpose of this letter is to introduce a new approach for a rapid mapping of newly triggered landslides using an objected-oriented change detection technique. The methodology aims at a semiautomatic and rapid analysis with a minimum of operator involvement and manual analysis steps. Compared to conventional approaches for landslide mapping, this approach benefits from the following: 1) an image segmentation with problem-specified scale optimization and 2) a multitemporal analysis at object level with several systemized spectral and textural metrics.

II. Methodology

The adopted methodology includes two parts: 1) image segmentation with multiscale optimization and 2) classification of landslide objects. The general methodology is shown in Fig. 1. Two QuickBird images were used in the study, acquired on September 6, 2006 and October 8, 2009, with 0.3% and zero cloud cover, respectively. For each image, only four multispectral bands [blue, green, red, and near infrared (NIR)] with a spatial resolution of 2.4 m were used, i.e., without pansharpening with the 0.6-m panchromatic band, to avoid artefacts introduced by image fusion and to increase the efficiency of computation time. Moreover, the 2.4-m resolution is sufficient for landslide mapping. In addition, a 1-m DTM was created from airborne LiDAR data acquired on October 6–19, 2009, shortly after the event, with a maximum point density of 8 points/m^2 (vertical and horizontal accuracies: 15 and 40 cm, 1-σ). The spectral analysis was performed with ENVI 4.7 software. The OOA and textural analysis were implemented in eCognition Developer 8.

The application of this approach is demonstrated by a case study in Messina Province of Sicily, southern Italy. During the night of October 1, 2009, intensive prolonged rainfall (ca. 223 mm in 7 h) affected several catchments south of Messina city. Numerous debris flows and shallow landslides were triggered, and 31 people were reported dead. Two of the most damaged areas are studied, including a training area of Altolia (ca. 1.8 km^2) for algorithm development and a larger independent testing area of Itala (ca. 8.1 km^2). The latter allows the robustness and transferability of the algorithm (without any change of rule set and parameter thresholds) and the corresponding accuracy to be assessed by comparison with a manually mapped landslide inventory.

A. Image Segmentation With Scale Optimization

Image segmentation defines the building blocks for object-oriented image analysis and, to ease further analysis, should aim at meaningful delineation of targeted real-world objects. However, considering the complex characteristics of landslides, including land cover variance, illumination difference, diversity of spectral behavior, and size variability, it is difficult to delineate each individual landslide as a single object [9]. Notwithstanding this difficulty, over- and undersegmentation can be reduced by means of a multiscale optimization approach.

The multiresolution segmentation based on fractal net evolution approach (FNEA) implemented in Definiens eCognition [7] was employed for the initial segmentation, parameterized according to the specific needs of event-based rapid mapping

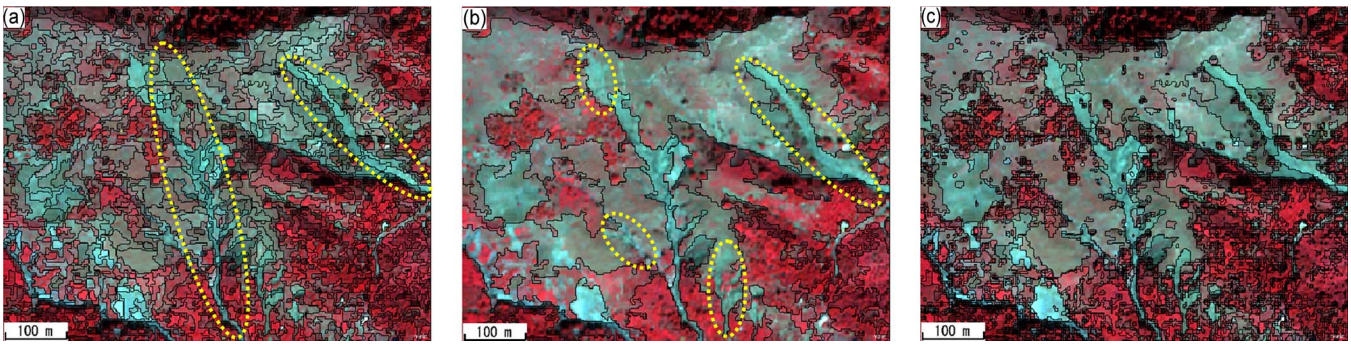


Fig. 2. Detailed view of the image segmentation at (a) a fixed scale of 30, (b) a specified scale of 200, and (c) a described multiscale optimization. The dotted ellipses are used to highlight (a) over- and (b) undersegmentation of landslides.

of landslides, and incorporated in a multiscale optimization routine. FNEA requires the user to define weights for input layers (bands), as well as a scale parameter that defines the maximum allowed heterogeneity within individual segments. Catastrophic slope failures typically remove the vegetation and result in high ratios between the red and NIR bands. These bands are also the least affected by atmospheric effects and were assigned equal weights w_c. The scale parameter is consequently defined as

$$f = \sum_c w_c(n_{\mathrm{mrg}} \cdot \sigma_{c,\mathrm{mrg}}) - (n_{o1} \cdot \sigma_{c,o1} + n_{o2} \cdot \sigma_{c,o2}) \quad (1)$$

with n corresponding to the number of pixels within an object and σ_c to the standard deviation of pixel values within the band c. Since FNEA performs the most homogeneous merge in the local vicinity [15], the subscripts indicate objects prior to merge ($_{o1}$ and $_{o2}$) and the respective resulting object after merging ($_{mrg}$). The fact that suitable values for f usually need to be determined by the user in time-consuming "trial-and-error" procedures has previously been identified as one of its major limitations [15]. Statistical optimization methods (e.g., [16] and [17]) allowed the choice of the scale parameter to be made more objective, provided that targeted elements exhibit one operational scale. However, slope failures feature several orders of magnitudes in volume and area, which prohibit the definition of one single scale parameter. To overcome this difficulty, Esch *et al.* [18] proposed a multilevel segmentation optimization procedure, which iteratively compares the spectral characteristics of image objects generated at multiple scales. A simplified version of this approach, which uses less spectral information and an automatically derived threshold, was used in this study.

In an initial step, the image was segmented with two hierarchical scales ($f_1 = 5$ and $f_2 = 10$). The mean Percentage Difference (mPD) between subobject level $(L1)$ and superobject level $(L2)$ was calculated as

$$mPD = \frac{|\nu_{L1} - \nu_{L2}|}{\nu_{L2}} \quad (2)$$

where ν is the ratio of the intensities in the NIR and red bands of the respective sub- and superobjects. Each subobject whose mPD exceeds the mean mPD of all subobjects by more than 2σ was consequently classified as a "real" subobject and transferred to the superobject level

$$real\ object = \begin{cases} 1, & mPD > 2\sigma_{mPD} \\ 0, & \text{else.} \end{cases} \quad (3)$$

In this sense, $2\sigma_{mPD}$ replaced the user-defined thresholds introduced by Esch *et al.* [18]. In the next step, the similarity of transferred adjacent subobjects ($ob1$ and $ob2$) was evaluated by their intensity difference in the NIR and red bands. Similar objects were merged according to the following condition:

$$sim_{ob1,ob2} = \begin{cases} 1, & (0.5 * |RED_{ob1} - RED_{ob2}| \\ & \quad + 0.5 * |NIR_{ob1} - NIR_{ob2}|) < 10 \\ 0, & \text{else.} \end{cases} \quad (4)$$

The procedure was repeated for a total of 11 scales (15, 20, 30, 50, 70, 100, 150, 200, 300, 500, and 700), where, in each step, the result of the previous cycle became the subobject level, and according to the next larger scale factor, a number of objects were merged to create a superobject level above. With each iteration, further objects exceeding the initially derived $2\sigma_{mPD}$ were transferred to the next level. The complete procedure aims to provide a segmentation that represents sufficiently distinct objects independent of their particular scale.

Fig. 2 shows the segmentation result of the multiscale optimization on the postevent imagery. Compared to the original FNEA with only one segmentation scale [Fig. 2(a) and (b)], image segmentation using multiscale optimization [Fig. 2(c)], although still facing some difficulties to delineate every individual landslide, decreases the degree of over- and undersegmentation and is able to capture better landslides as image objects among a number of different scales. Furthermore, the optimization runs fully automatically and liberates the user from a time-consuming trial-and-error evaluation of the optimal parameterization for the image segmentation.

B. Classification of Landslide Objects

Landslide classification in previous studies has become increasingly complex. While initial works were largely restricted to the digital number values of multispectral bands, later indices such as Normalized Difference Vegetation Index, different texture measures, digital elevation model derivatives, and externally prepared vector layers (e.g., of flow accumulation and stream networks) or shadow masks were employed [9]. The landslide detection approach presented here makes use of additional spectral and textural measurements: change detection using temporal principal component analysis (PCA), image matching through spectral angle mapper (SAM), anomaly detection by the Reed–Xiaoli detector (RXD), and textural analysis with gray-level co-occurrence matrix (GLCM). The

derivatives of PCA, SAM, and RXD were calculated as separate layers and incorporated in OOA as features of each object that were derived during the hierarchical segmentation.

The change detection was first carried out using temporal PCA, an image transformation of stacked pre- and postevent images based on eigenvector analysis of their image covariance matrix [19]. Temporal PCA combined all eight bands of the pre- and postevent QuickBird images (four bands each) and transformed these bands into eight uncorrelated components. The components which concentrate changes have relatively smaller eigenvalues and can be determined by visual inspection [20]. An assessment of the principal components in the training area revealed that the signatures of newly triggered landslides were primarily concentrated in the fourth principal component (PC4). The minor components beyond PC4 were mainly composed of residuals of the transformation, in most cases noise. In the training area, landslide candidates were preliminarily chosen from PC4 using a membership function calculated from ten selected samples of landslide objects. This membership function was then incorporated in the algorithm of classification and later employed without modification in the testing area. A further inspection of remaining false positives, such as roads, deforestation areas, and water, revealed relatively low values of those objects in the second principal component (PC2), and values < -300 were found to be suitable threshold for their removal.

Since shadows were also possibly recorded as changes in PCA, a spectral matching image between the pre- and postimages was created using SAM [21] and then imported in OOA. The purpose of SAM is to remove the influence of these subtle spectral changes due to illumination differences and viewing angle variation. The matching image derived from SAM estimated spectral similarity by comparing spectral angle differences in terms of image space between the pre- and postevent QuickBird images. For both images, each pixel was represented by a spectrum identified as a 4-D vector with specified length and direction. As SAM only considers the angle between the spectral vectors but not the vector length, it is less sensitive to changes due to illumination and shadowing [21]. Excluding objects with low SAM values $(SAM < 0.09)$ allowed a removal of spectral false positives that resulted from subtle spectral changes in illumination, as well as shadow, which cannot always be excluded from the change component of PCA.

In addition, in order to remove false positives, such as urban areas as well as existing outcrops and clear-cuts, the RXD anomaly detector [22] was used to estimate spectral anomalies based on the pre-event image. Assuming that urban areas, deforestation, roads, and other infrastructure demonstrate spectral signatures that are significantly different from the background, RXD can be used to highlight those areas. In this letter, RXD was applied on the pre-event imagery to detect spectral anomalies that existed already before the event, which are consequently excluded as newly triggered landslides. A created anomaly image was then derived and employed as an additional layer in eCognition. Objects with large RXD values were considered as spectral anomalies, and a threshold of $RXD > 16$ was defined to exclude these anomalous false positives.

Following the spectral processing that identified landslide candidate objects, a texture analysis of a 1-m LiDAR DTM was performed after merging those candidates. The texture analysis was performed on elevation data for the purpose of analyzing topographic variability, using second-order statistics of the widely applied GLCM [23]. The objective is to remove false positives with low-frequency elevation variation, such as undisturbed or unfractured areas, homogeneous flat surfaces, and objects with low height variation (i.e., roads and water bodies). Texture features calculated from $GLCM_{\text{mean}}$ were used in our study. Neighboring pixels in all directions (0°, 45°, 90°, and 135°) were considered for the GLCM generation, accounting for the potential different aspects of landslide objects. Objects with low $GLCM_{\text{mean}}$ values were considered to be false positives, and a threshold of $GLCM_{\text{mean}} < 126.7$ was defined. The remaining landslide candidates were then classified as final output of newly triggered landslides.

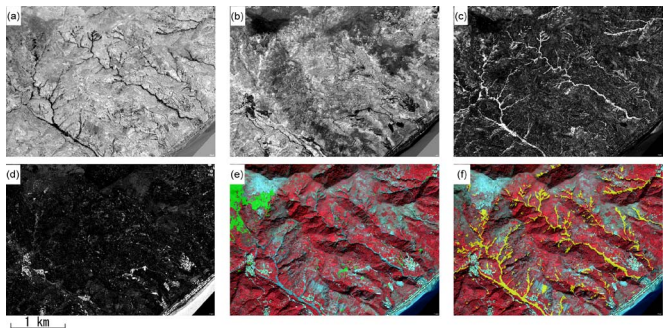


Fig. 3. Testing area of Itala. (a) Fourth component of PCA. (b) Second component of PCA. (c) Matching image of SAM. (d) Result of RXD anomaly detection on pre-event image. (e) (Green) False positives detected using $GLCM_{\text{mean}}$. (f) (Yellow) Final result of detected landslides.

TABLE I
ACCURACY ASSESSMENT FOR OOA-MAPPED LANDSLIDES

	Manual mapping	Auto. mapping	True positive	User Acc.	Prod. Acc.
Number	285	242	198	81.8%	69.5%
Extent (km^2)	0.602	0.555	0.421	75.9%	69.9%

III. RESULT AND ACCURACY ASSESSMENT

The algorithm developed based on the training area of Altolia was directly applied in the testing area of Itala, without any change in membership function values and defined thresholds. The intermediate derivatives and final outputs for the testing area are shown in Fig. 3. To evaluate the accuracy of this approach, OOA-derived landslides were compared with a manually mapped landslide inventory. The accuracy assessment was carried out for the number and the spatial extent of mapped landslides (Table I), both of which are considered critical in a subsequent quantitative landslide hazard and risk assessment. The number of landslides is useful for a quantitative estimate of the temporal probability of landslide occurrence, whereas the spatial extent of landslide is beneficial for the estimate of probability of landslide size through the landslide frequency-area statistics [3].

The accuracy assessment calculates the commission and omission errors, which are measures of the user's and producer's accuracies of the mapped landslides, respectively. For the spatial extent of landslides, a user's accuracy of 75.9% and a producer's accuracy of 69.9% were achieved. In terms of the number of landslides, user's and producer's accuracies of 81.8% and 69.5%, respectively, were reached. For both the

number and spatial extent of landslides, the results show a lower producer's than user's accuracy: Specifically, ca. 31% of manually mapped landslides were omitted in the OOA-based detection. This indicates an overestimation of false positives during their classification, accompanied with an underestimation of true positives obtained from the membership function of the selected samples. Further improvements should include a more accurate definition of these thresholds for classifying false positives and a more careful selection of representative samples.

IV. Conclusion

This letter has described a novel approach of object-oriented change detection for rapid mapping of newly triggered landslides after major events, using VHR satellite images and LiDAR data. The approach used a transparent semiautomatic mapping technique that reduces the user involvement to the determination of a few thresholds for a systemized predesigned OOA work process. First, a problem-specific multiscale optimization of FNEA was proposed to reduce the degree of over- and undersegmentation of landslides among a number of different scales, avoiding a time-consuming trial-and-error evaluation of the optimal segmentation parameters that has characterized most OOA research works in the past. Second, change detection using image transformation of PCA was not only found to be useful for a preliminary selection of landslide candidates from PC4 but also enabled a removal of false positives directly from PC2. Third, the matching image derived from SAM allowed the detection of subtle spectral changes from the change of spectrum vector direction. Fourth, spectral anomalies detected by RXD in the pre-event image allowed the removal of false positives, such as landslides that already existed before the landslide event. Finally, surface texture measures based on a 1-m LiDAR DTM were incorporated to remove false positives with low-frequency elevation variation.

For the case study in Messina, the approach achieved user's and producer's accuracies of 75.9% and 69.9%, respectively, for the extent of landslides, and 81.8% and 69.5%, respectively, for the number of landslides. Although the accuracy of the automatic approach does not entirely match what can be achieved in manual mapping, it provides an efficient supplement for traditional methods. The chosen spectral object features are expected to be useful to accommodate multispectral information from a great variety of different sensors. The proposed thresholds typically need further adjustment for the application in other cases, whereas in the presented example, the visual inspection of one-fifth of the study area was sufficient for this purpose. Moreover, it should not be forgotten that considerable time can be saved for landslide mapping because the manual drawing of landslides boundaries is replaced by image segmentation. Hence, for an effective landslide hazard assessment, the approach provides an efficient tool to retrieve lacking temporal data for an event-based landslide inventory, thus allowing the assessment of temporal probability and magnitude of landslide events for a quantitative hazard assessment.

References

[1] L. Borgatti and M. Soldati, "Landslides as a geomorphological proxy for climate change: A record from the Dolomites (northern Italy)," *Geomorphology*, vol. 120, no. 1/2, pp. 56–64, Aug. 2010.

[2] M. Jakob and S. Lambert, "Climate change effects on landslides along the southwest coast of British Columbia," *Geomorphology*, vol. 107, no. 3/4, pp. 275–284, Jun. 2009.

[3] F. Guzzetti, P. Reichenbach, M. Cardinali, M. Galli, and F. Ardizzone, "Probabilistic landslide hazard assessment at the basin scale," *Geomorphology*, vol. 72, no. 1–4, pp. 272–299, Dec. 2005.

[4] J. Hervas, J. I. Barredo, P. L. Rosin, A. Pasuto, F. Mantovani, and S. Silvano, "Monitoring landslides from optical remotely sensed imagery: The case history of Tessina landslide, Italy," *Geomorphology*, vol. 54, no. 1/2, pp. 63–75, Aug. 2003.

[5] A. M. Borghuis, K. Chang, and H. Y. Lee, "Comparison between automated and manual mapping of typhoon-triggered landslides from SPOT-5 imagery," *Int. J. Remote Sens.*, vol. 28, no. 8, pp. 1843–1856, Jan. 2007.

[6] J. McKean and J. Roering, "Objective landslide detection and surface morphology mapping using high-resolution airborne laser altimetry," *Geomorphology*, vol. 57, no. 3/4, pp. 331–351, Feb. 2004.

[7] U. C. Benz, P. Hofmann, G. Willhauck, I. Lingenfelder, and M. Heynen, "Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information," *ISPRS J. Photogramm. Remote Sens.*, vol. 58, no. 3/4, pp. 239–258, Jan. 2004.

[8] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 65, no. 1, pp. 2–16, Jan. 2010.

[9] T. R. Martha, N. Kerle, V. Jetten, C. J. van Westen, and K. V. Kumar, "Characterising spectral, spatial and morphometric properties of landslides for semi-automatic detection using object-oriented methods," *Geomorphology*, vol. 116, no. 1/2, pp. 24–36, Mar. 2010.

[10] J. Barlow, Y. Martin, and S. E. Franklin, "Detecting translational landslide scars using segmentation of Landsat ETM + and DEM data in the northern Cascade Mountains, British Columbia," *Can. J. Remote Sen.*, vol. 29, no. 4, pp. 510–517, 2003.

[11] Y. E. Martin and S. E. Franklin, "Classification of soil- and bedrock-dominated landslides in British Columbia using segmentation of satellite imagery and DEM data," *Int. J. Remote Sens.*, vol. 26, no. 7, pp. 1505–1509, Apr. 2005.

[12] J. Barlow, S. Franklin, and Y. Martin, "High spatial resolution satellite imagery, DEM derivatives, and image segmentation for the detection of mass wasting processes," *Photogramm. Eng. Remote Sens.*, vol. 72, no. 6, pp. 687–692, 2006.

[13] J. Nichol and M. S. Wong, "Satellite remote sensing for detailed landslide inventories using change detection and image fusion," *Int. J. Remote Sens.*, vol. 26, no. 9, pp. 1913–1926, May 2005.

[14] N. W. Park and K. H. Chi, "Quantitative assessment of landslide susceptibility using high-resolution remote sensing data and a generalized additive model," *Int. J. Remote Sens.*, vol. 29, no. 1, pp. 247–264, Jan. 2008.

[15] G. J. Hay, T. Blaschke, D. J. Marceau, and A. Bouchard, "A comparison of three image-object methods for the multiscale analysis of landscape structure," *ISPRS J. Photogramm. Remote Sens.*, vol. 57, no. 5/6, pp. 327–345, Apr. 2003.

[16] G. M. Espindola, G. Camara, I. A. Reis, L. S. Bins, and A. M. Monteiro, "Parameter selection for region-growing image segmentation algorithms using spatial autocorrelation," *Int. J. Remote Sens.*, vol. 27, no. 14, pp. 3035–3040, Jul. 2006.

[17] L. Drăguţ, T. Schauppenlehner, A. Muhar, J. Strobl, and T. Blaschke, "Optimization of scale and parametrization for terrain segmentation: An application to soil-landscape modelling," *Comput. Geosci.*, vol. 35, no. 9, pp. 1875–1883, Sep. 2009.

[18] T. Esch, M. Thiel, M. Bock, A. Roth, and S. Dech, "Improvement of image segmentation accuracy based on multiscale optimization procedure," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 3, pp. 463–467, Jul. 2008.

[19] J. S. Deng, K. Wang, Y. H. Deng, and G. J. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *Int. J. Remote Sens.*, vol. 29, no. 16, pp. 4823–4838, Aug. 2008.

[20] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: A systematic survey," *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 294–307, Mar. 2005.

[21] F. A. Kruse, A. B. Lefkoff, J. W. Boardman, K. B. Heidebrecht, A. T. Shaprio, P. J. Barloon, and A. F. H. Goetz, "The spectral image processing system (SIPS)—Interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, no. 2/3, pp. 145–163, May/Jun. 1993.

[22] I. S. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 10, pp. 1760–1770, Oct. 1990.

[23] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

Appendix 2: Kurtz et al. In Press. Hierarchical extraction of landslides from multiresolution remotely sensed optical images.

Kurtz, C., Stumpf, A., Malet, J.-P., Puissant, A., Gançarski, P., Passat, N., In Press. Hierarchical extraction of landslides from multiresolution remotely sensed optical images. ISPRS Journal of Photogrammetry and Remote Sensing, 26 pages.

Hierarchical extraction of landslides from multiresolution remotely sensed optical images

Camille Kurtz[a,*], André Stumpf[b,c,*], Jean-Philippe Malet[b], Pierre Gançarski[d], Anne Puissant[c], Nicolas Passat[e]

[a]*LIPADE, Université Paris Descartes, France*
[b]*IPGS, Université de Strasbourg, CNRS, France*
[c]*LIVE, Université de Strasbourg, CNRS, France*
[d]*ICube, Université de Strasbourg, CNRS, France*
[e]*CReSTIC, Université de Reims Champagne-Ardenne, France*

Abstract

The automated detection and mapping of landslides from Very High Resolution (VHR) images present several challenges related to the heterogeneity of landslide sizes, shapes and soil surface characteristics. However, a common geomorphological characteristic of landslides is to be organized with a series of embedded and scaled features. These properties motivated the use of a multiresolution image analysis approach for their detection. In this work, we propose a hybrid segmentation/classification region-based method, devoted to this specific issue. The method, which uses images of the same area at various spatial resolutions (Medium to Very High Resolution), relies on a recently introduced top-down hierarchical framework. In the specific context of landslide analysis, two main novelties are introduced to enrich this framework. The first novelty consists of using non-spectral information, obtained from Digital Terrain Model (DTM), as *a priori* knowledge for the guidance of the segmentation/classification process. The second novelty consists of using a new domain adaptation strategy, that allows to reduce the expert's interaction when handling large image datasets. Experiments performed on satellite images acquired over terrains affected by landslides demonstrate the efficiency of the proposed method with different hierarchical levels of detail addressing various operational needs.

Keywords: landslide mapping, VHR images, multiresolution region-based analysis, hierarchical approach, binary partition tree, domain adaptation

1. Introduction

1.1. Context

In the field of Earth observation, a new generation of sensors with meter and sub-meter resolution has led to an increased production of Very High Resolution (VHR) optical images (Benediktsson et al., 2012), and to improved operational capabilities for monitoring geohazards. Especially, several studies demonstrated that such kind of imagery enables to inventory and delineate landslide-affected areas (Nichol and Wong, 2005; Barlow et al., 2006; Martha et al.,

*Camille Kurtz and André Stumpf are considered as equal first authors of this manuscript. Corresponding author: Camille Kurtz. Tel.: +33 (0)1.83.94.58.07, Fax: +33 (0)1.83.94.58.07.

Email addresses: camille.kurtz@parisdescartes.fr (Camille Kurtz), andre.stumpf@unistra.fr (André Stumpf), jeanphilippe.malet@unistra.fr (Jean-Philippe Malet), gancarski@unistra.fr (Pierre Gançarski), anne.puissant@unistra.fr (Anne Puissant), nicolas.passat@univ-reims.fr (Nicolas Passat)

2010; Mondini et al., 2011; Stumpf and Kerle, 2011), providing valuable information for the estimation of potential risks to infrastructures and human lives. Comprehensive landslide inventory maps should ideally also provide information about the respective sub-parts of each single landslide which are often characterized by different kinematic patterns. Such type of spatial information on landslide sub-units is of paramount importance for quantitative hazard assessments (Thiery et al., 2007) and the identification of landslide that are more prone to imminent acceleration or fluidization (Raucoules et al., 2013). VHR satellite images contain sufficient spatial details to depict small geomorphological surface features such as faults, scarps, fissures, rock blocks and lobes, and should in principle also enable an analysis at the level of sub-parts composing a landslide. However, established concepts of landslide geomorphology, such as a first order differentiation between source area, transport area and toe (Figure 1), do not directly correspond to spectrally homogeneous pixels or regions in VHR images. Indeed, the different sub-parts appear heterogeneous as they generally contain different kinds of basic elements with a specific spatial organization (*e.g.*, different kinds of fissures, different sizes and shapes of rock blocks, etc.). Consequently, by opposition to lower resolution images (Yilmaz, 2009; Kayastha et al., 2013), landslides can be considered as hierarchies of complex patterns composed by sub-objects of interest. One may notice that the conceptual model in Figure 1 allows us to identify potentially useful features (*e.g.*, slope gradient, slope curvature, differences in altitude) to distinguish the different landslide sub-parts.

Due to the expanding fleet of VHR optical satellites such as the French PLEIADES constellation (de Lussy et al., 2005), the challenges associated with VHR images are gaining increasingly greater importance in the context of landslide mapping (Guzzetti et al., 2012). On the one hand, the size and the complexity of the images make the visual analysis a time consuming and error prone task (Galli et al., 2008; Fiorucci et al., 2011). On the other hand, state-of-the-art image analysis tools, which are usually considered for the mapping of landslides from lower resolution images, rely on radiometric homogeneous hypotheses of the landslides represented in the images. Consequently, these tools cannot handle the new levels of spatial details provided by VHR images (Blaschke, 2010). Therefore, new image analysis methodologies have to be proposed for the hierarchical mapping of landslides from VHR optical satellite images.

1.2. Related works

In this context, various region-based approaches have been proposed to automate the extraction of landslides from VHR images (Barlow et al., 2006; Martha et al., 2010; Lu et al., 2011; Stumpf and Kerle, 2011; Lahousse et al., 2011; Hölbling et al., 2012; Stumpf et al., 2013). In opposition to pixel-based approaches that mainly use spectral and textural information (Townshend et al., 2000; Mallinis et al., 2008), region-based approaches enable to consider high-level (*e.g.*, contextual, geometrical) features to describe the objects to be classified. Indeed, region-based approaches enable to transfer high-level knowledge in computer-accessible features leading to discriminatory decision sets. Such decision sets have been employed for the mapping of landslides from HR (Martha et al., 2010) and multi-temporal VHR images (Lu et al., 2011). Stumpf et al. (2013) have also proposed a supervised framework to automatically select discriminative features among a multitude of potentially useful ones. Nevertheless, most of the proposed approaches do not consider the hierarchical organization of the objects of interest (Benz et al., 2004), that is a serious drawback when dealing with VHR optical satellite images.

To tackle this issue, two key-concepts can be considered: using multiple images, and using multiple spatial resolutions. On the one hand, multiple images provide complementary information, that can enrich each other. This is

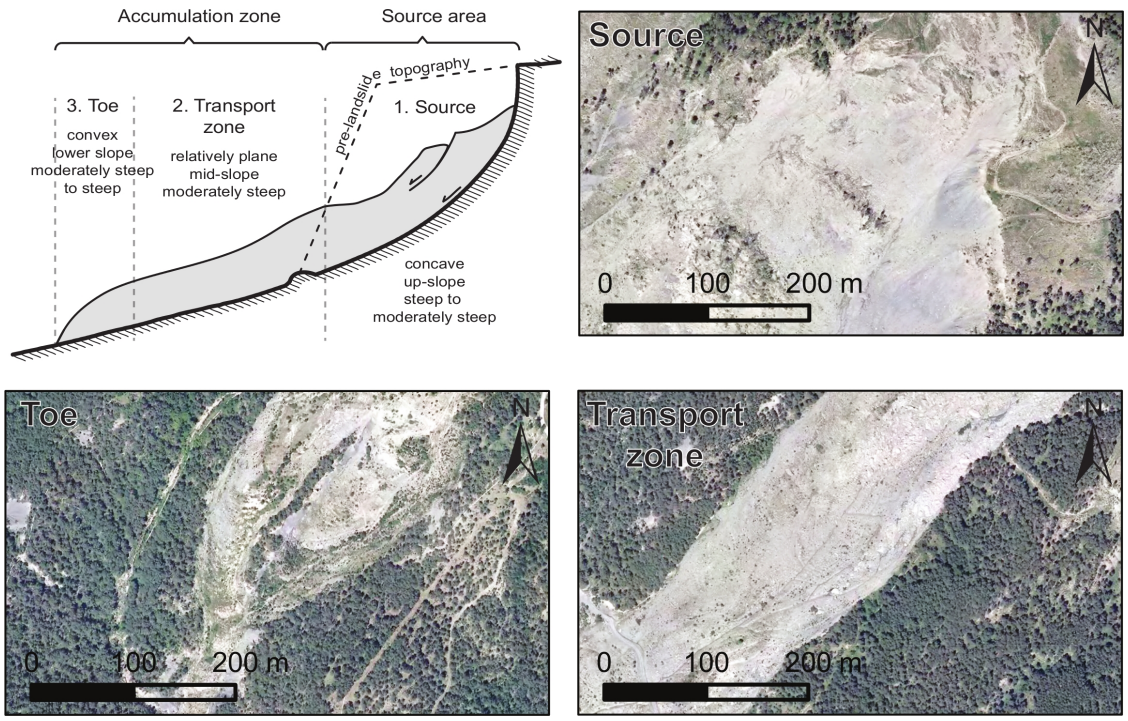


Figure 1: Schematic partition of a landslide into sub-units (source area, transport area and toe) and their typical geomorphological features. The image depicts the distinguished sub-units on an aerial photograph (50 cm, © IGN, 2008) and shows the heterogeneous surfaces including deformation features such as fissures and different sizes of rock blocks. This example illustrates that geomorphological definitions do not correspond directly to spectrally homogeneous areas depicted in the remote sensing images.

specifically true in the case where such images gather different radiometric values, that then carry various semantic information. On the other hand, using images at multiple spatial resolutions provides hierarchical links between their respective radiometric information. (In particular, the availability of a large range of spatial resolutions, from Medium Spatial Resolution (MR, 30–5 m) to VHR images, has already led to methods for the extraction of hierarchical patterns (Akcay and Aksoy, 2008; Wemmert et al., 2009; Gaetano et al., 2009). These considerations motivate our use of multiple images at multiple spatial resolutions (Chang et al., 2007). In particular, we propose to take advantage of this potential spatial (Sun et al., 2003) and radiometric enrichment to propose a multiresolution representation of the data, leading to a hierarchical unsupervised region-based approach.

1.3. Contributions

Based on these considerations, a top-down hierarchical region-based framework has been recently proposed by Kurtz et al. (2012) to segment and classify multiresolution images from the lowest to the highest resolution, and then finally extract complex patterns from VHR images. This top-down hierarchical approach (TDHA) constitutes a generic and versatile framework. In this work, we propose to adapt and improve it, in order to efficiently deal with the case of landslide mapping from remote sensing imagery.

From a methodological point of view, our contributions are twofold. Firstly, we propose to integrate topographic and morphometric *a priori* knowledge derived from non-spectral data, namely Digital Terrain Model (DTM), for

guiding the segmentation/classification process. This strategy has shown promising results in the context of mapping shallow landslides with pixel-based approaches (Mondini et al., 2013). In our approach, this knowledge is used for the construction of the hierarchical image representation data structure, namely a binary partition tree (BPT) (Salembier and Garrido, 2000) considered for the segmentation. It is also used to enrich the feature space in the context of a multiresolution classification procedure (Kurtz et al., 2 010). This enriched TDHA is described in Subsection 2.1. Secondly, we propose a new strategy that allows to reduce the experts' interaction when handling large image datasets. Indeed, the proposed TDHA uses experts' knowledge *via* the interactive definition of segmentation examples for each semantic class, and at each resolution. Based on the domain adaptation paradigm (Daumé III and Marcu, 2006), whose relevance has already been proved for classification of satellite images (Bruzzone et al., 1999; Bahirat et al., 2012), we propose a strategy that enables to reduce *a minima* the error-prone and time-consuming task that consists of providing segmentation examples. This domain adaptation strategy is described in Subsection 2.2.

From an applicative point of view, the proposed method has been used for analysing multiresolution image datasets and map different landslide sub-units. These experiments, described in Section 3, demonstrate the efficiency of the proposed method with different hierarchical levels of detail, addressing various operational needs.

To the best of our knowledge, this work constitutes the first application of a top-down hierarchical strategy, dealing with multiresolution images, for landslides monitoring purpose.

2. Methodology

This section presents the proposed workflow for the segmentation/classification of landslides from multiresolution satellite images. Subsection 2.1 provides a description of the TDHA, and how it is enhanced by integrating morphometric *a priori* knowledge derived from DTM. Subsection 2.2 details the proposed domain adaptation strategy enabling to use the knowledge learned on a reference source dataset to automatically process another dataset representing the same type of objects.

2.1. Top-down hierarchical approach for segmentation/classification of landslide multiresolution images

The proposed TDHA, summarized in Figure 2, takes as input a multiresolution set of n images $\mathcal{I}_1, \ldots, \mathcal{I}_n$ (in our case, $n = 3$ images are considered: one MR, one HR and one VHR image). It provides as output n segmentation/classification maps computed using a top-down strategy. It performs n successive steps (one per spatial resolution image). Each step is composed of:

- an example-based hierarchical segmentation (see Figure 2-①, and Subsection 2.1.1);
- a multiresolution clustering (see Figure 2-②, and Subsection 2.1.2).

This approach is mainly based on an iterative decomposition strategy: at each resolution/step, the output is embedded into the next resolution image to be processed as input of the next step (see Figure 2-③).

2.1.1. Example-based hierarchical segmentation

Overview. The segmentation of a satellite image is a complex task since the different objects of interest do not necessarily share the same spatial scale. To deal with this issue, the segmentation step is divided into k different substeps, applied independently to a group of similar thematic ground areas (k is set by the expert and corresponds to the

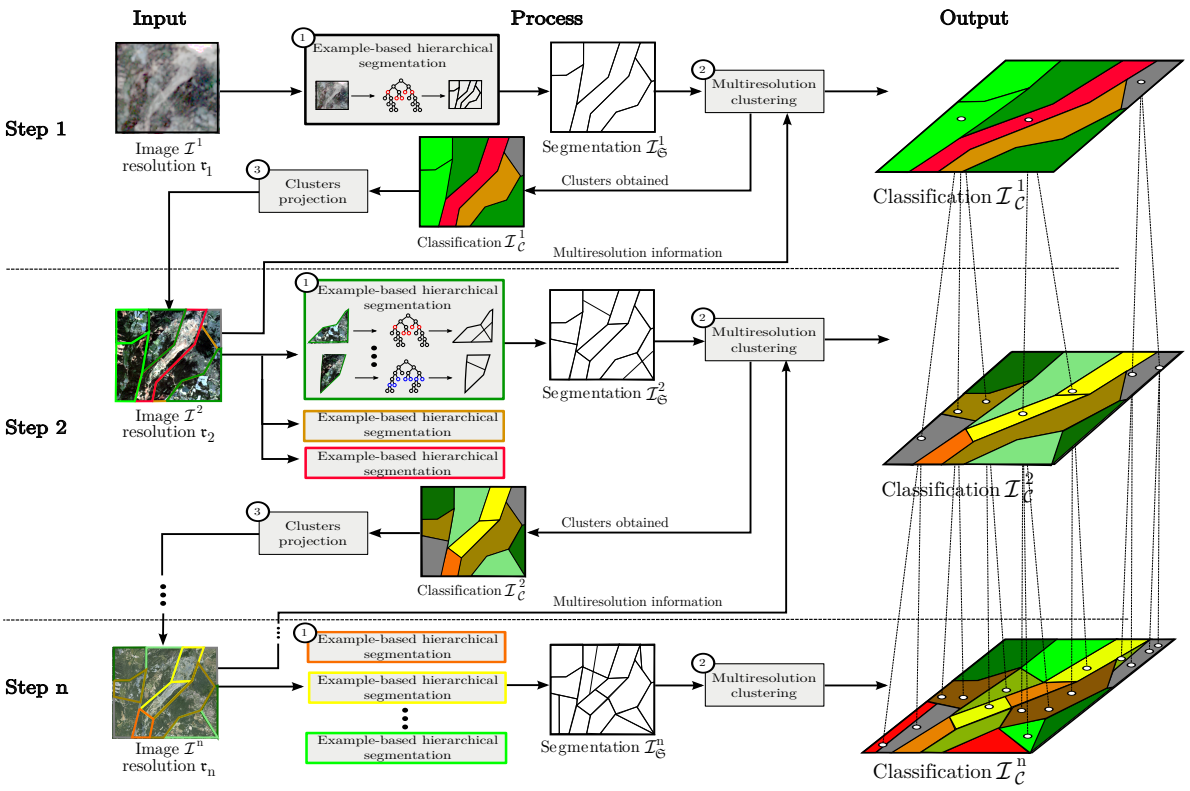


Figure 2: Workflow overview of the proposed TDHA (see text).

number of thematic ground areas extracted by the multiresolution clustering approach from the previous resolution image, see Subsection 2.1.2).

Each sub-step requires a segmentation example interactively defined by the expert (Figure 2-①). Segmentation is performed by using a hierarchical strategy based on Binary Partition Trees (BPTs) (Salembier and Garrido, 2000; Salembier and Wilkinson, 2009; Valero et al., 2010; Benediktsson et al., 2011; Alonso-González et al., 2012). Thus, defining a segmentation example consists of providing a cut of the BPT. To this end, a software has been designed to visualize the segmentation results for different cuts in real-time (see Subsection 3.2). The expert can interactively parse the tree to select a set of nodes producing the required tree-cut example (Figure 3). The experts' behavior is then learned. More precisely, a tree-cut example is represented by a set of centroids (modeled by color histograms and spatial features) using a clustering algorithm. This example is automatically reproduced in the similar thematic ground areas of the image by using a cutting strategy that processes all the associated BPTs.

Technical description.

Building a BPT: The BPT of a multivalued image $\mathcal{I} : E \rightarrow V$ with B spectral bands V_b, is built in a bottom-up approach (*i.e.*, from its leaves to its root). The leaves of the tree are composed by the regions belonging to the initial partition. Practically, the initial partition of E is generally composed by the flat zones of $\mathcal{I}$. The remaining nodes are obtained by successively merging couples of (already defined) nodes of the tree. Such couples of nodes are generally

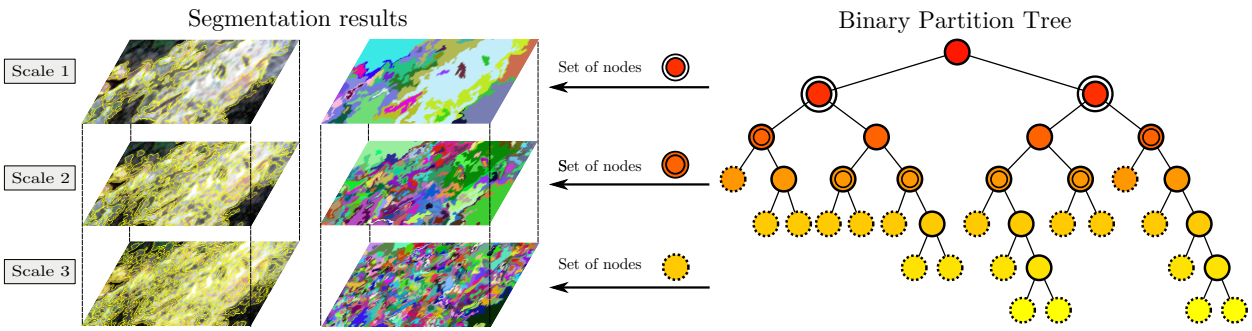


Figure 3: An example of BPT associated to an HR image (the number of nodes is significantly reduced, for sake of readability). The nodes of the tree are depicted by color disks (the root is the highest node). The colors of the nodes (from yellow to red) symbolize the decrease of the similarity measure O_r between two neighboring regions (and thus, also the decrease of the function α controlling the trade-off between O_r and O_g). For the sake of visualization, three partitions associated to three cuts of the BPT are depicted.

chosen as spatially adjacent, thus leading to a partition of E. The root node represents the entire image support E.

A huge number of distinct BPTs may be obtained for a same initial partition of E. In order to decide which BPT is the most relevant, it is then necessary to define the priority of the fusions between nodes. A BPT generation then relies on two main notions: a *region model* (which specifies how regions are characterized), and a *merging criterion* (which defines the similarity of neighboring regions and thus the merging order).

The basic models and criteria used in most image segmentation approaches generally rely on radiometric homogeneity hypotheses (Garrido et al., 1998). Such models are well adapted to process MR images, but when dealing with VHR images, other geometric and/or spatial features require consideration. Morphological features related to the sensed ground surface, can provide relevant information for the delineation of the landslides and their sub-parts. They can be obtained from DTM and are typically linked with distinct geomorphological processes (Anders et al., 2011). Morphological features such as the slope and the curvature of the regions (see Figure 1) can be considered to characterize the nodes during the construction of the BPT.

We propose to rely on both the increase of the ranges of the intensity values (for each spectral band) and on the morphological features of the regions in order to merge in priority objects composing the landslide sub-parts. In the sequel, the chosen region model and merging criterion are defined.

Region model: A node/region $R_i \subseteq E$ is modeled here by a couple of values

$$M_r(R_i) = \langle (v_b^-(R_i), v_b^+(R_i)) \rangle_{b=1}^{B} \quad (1)$$

$$M_g(R_i) = (slop(R_i), curv(R_i)) \quad (2)$$

where $v_b^\star$ provides the extremal values for the b-th spectral band in $\mathcal{I}$, while $slop$ and $curv$ are functions providing the slope and the curvature, respectively. Broadly speaking, M_r and M_g provide (low-level) spectral and geometrical information. During the merging process, the region model of two merged regions R_i and R_j is then provided by

$$M_r(R_i \cup R_j) = \langle (\min\{v_b^-(R_i), v_b^-(R_j)\}, \max\{v_b^+(R_i), v_b^+(R_j)\}) \rangle_{b=1}^{B} \quad (3)$$

$$M_g(R_i \cup R_j) = (slop(R_i \cup R_j), curv(R_i \cup R_j)) \quad (4)$$

While M_r is derived directly from the image, values for the slope $slop$ and the curvature $curv$ are obtained as the first and the second order derivatives of a DTM representing the topographic surface. (Further details on the computation of the derivatives from the DTM at different spatial scales are provided in Subsection 3.1.2.)

Merging criterion: At each step, the algorithm determines the pair of most similar connected regions minimizing the increase of the ranges of the intensity values (for each spectral band) and having similar morphological properties. This leads to the following merging criteria

$$O_r(R_i, R_j) = \frac{1}{B} \sum_{b=1}^{B} \frac{\max\{v_b^+(R_i), v_b^+(R_j)\} - \min\{v_b^-(R_i), v_b^-(R_j)\}}{v_b^+(E) - v_b^-(E)} \tag{5}$$

$$O_g(R_i, R_j) = \frac{1}{2}(|slop(R_i) - slop(R_j)| + |curv(R_i) - curv(R_j)|) \tag{6}$$

The similarity measure between two neighboring regions R_i and R_j are then computed as:

$$O(R_i, R_j) = \alpha.O_r(R_i, R_j) + (1 - \alpha).O_g(R_i, R_j) \tag{7}$$

with $\alpha \in [0, 1]$, where $O_r(R_i, R_j)$ and $O_g(R_i, R_j)$ have been normalized. In practice, the closer the nodes are to the root, the less relevant O_r is. Consequently, the weight α can be defined as a function depending directly on the value of O_r (and decreasing when O_r increases). In particular, it has been experimentally observed that a standard Gaussian formulation

$$\alpha(O_r) = \exp(-O_r^2) \tag{8}$$

provides a satisfactory behavior of the merging function O (Kurtz et al., 2011).

Based on these chosen model regions and merging criterion, the BPT is built, as exemplified in Figure 3. As a concern of the merging criterion, a part of the BPT nodes may be affected by outlier regions composed of very small sets of pixels with extremely low or high intensity values (Equation (5)). This behavior could seriously affect the resulting partitions extracted from the BPT, then leading to over-segmentation results. This issue is handled by the classification step that will aggregate smaller regions to larger ones based on their spectral and morphological similarities (*e.g.*, , a small region with extreme values surrounded by a larger region with different values but with the same altitude and slope).

2.1.2. Multiresolution clustering

Overview. The classification of the segments created by the previous step is performed by using a multiresolution clustering approach (Kurtz et al., 2010) to automate as much as possible the global process (which is already affected by a semi-automatical segmentation step). Instead of characterizing the segments extracted at the current resolution by using classical features (*e.g.*, color, texture, shape), we use their decompositions at the next resolution (Figure 2-②). To this end, for each segment, a class-based histogram is computed modeling its composition in terms of radiometric clusters into the next resolution image. A classical clustering approach – here, the K-MEANS (MacQueen, 1967) – is then performed to create c groups of segments sharing similar features where c is set by the expert (we also offer the possibility to instantiate this methodology with a supervised classification approach). Once these groups have been created, the expert selects and recognizes them to match with potential thematic ground classes (*e.g.*, Cluster 1 $\rightarrow$ Vegetation, . . . , Cluster $c \rightarrow$ Landslide).

These c clusters are then embedded into the next resolution image to be processed as input of the next step (Figure 2-③) and segmented into sub-clusters of interest. Practically, the c clusters are then considered in the next resolution image as the k groups of thematic ground areas to be segmented, and new BPTs are built inside these groups of segments (see Subsection 2.1.1). This original strategy enables to cluster into similar groups (at the current resolution), regions sharing similar image characteristics at the next considered resolution.

Technical description.

As stated above, the multiresolution clustering approach used in the proposed TDHA is mainly based on the notion of color class-based histogram to characterize the segments to be classified. These multiresolution features are well adapted to deal with MR and HR images to enable the extraction/classification of the whole structure of the landslides.

In the specific case of the classification of the landslide sub-units (*i.e.*, source area, transport area, toe) from VHR images, it is also relevant to consider the morphological features of the data during their classification. Consequently, we consider both the radiometric and the spatial properties of the ground during the classification of the regions extracted from the VHR images. In this way, we propose to characterize each segment extracted from the VHR images using the following features:

- the radiometric average value of each spectral band $A_1, \ldots, A_B$;
- the slope value $slop$;
- the curvature value $curv$;
- the normalized altitude value alt.

Indeed, preliminary experiments have shown that morphological features could enable separation of the different classes related to the considered object of interest. In particular (see Figure 4) slope and normalized altitude seem to be powerful features to discriminate the landslide sub-units.

2.2. Domain adaptation strategy

The classical domain adaptation approaches act by transferring the learned model (in general, a classifier) from a training to a target dataset for which *a priori* information is not available. We propose hereinafter a wise extension of such approaches enabling the transfer of a segmentation established interactively by an expert on a geographic subset of a remote sensing dataset to automatically process the data in a target region of interest.

2.2.1. Transfer of the learned elements

Once a multiresolution set of images has been processed (*i.e.*, segmented and classified) in a training area, it becomes possible to transfer the learned segmentation examples and clusters to process the set of images for target regions of interest. The underlying idea is that a specific landslide (or sub-units of a landslide) extracted from a particular ground unit will present similar spatial and radiometric characteristics as other landslides within the same geographic ground unit. Consequently, we assume that the landslides (and their sub-components) from the training and the target area can be characterized with similar geometrical and radiometric (*i.e.*, slope, curvature, normalized altitude) features.

For the processing, we transfer the following elements from the training area to the target one:

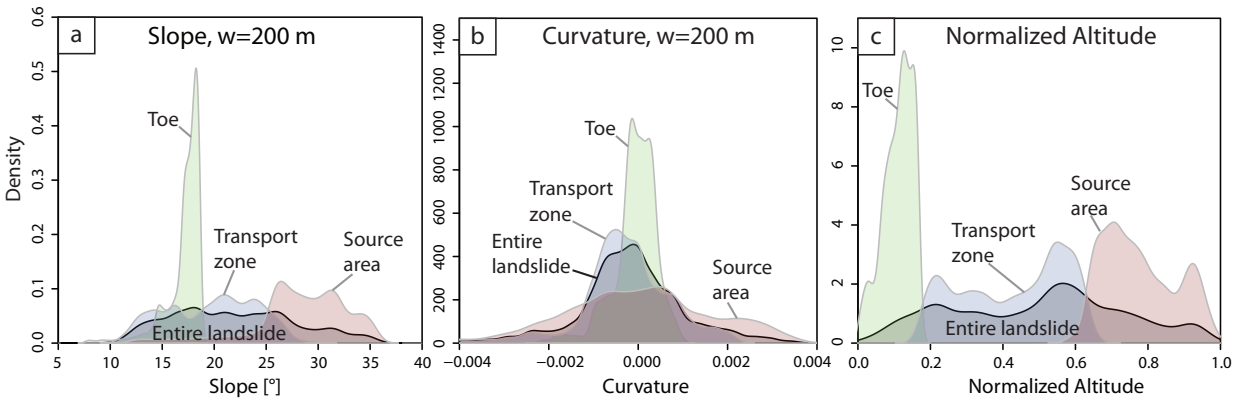


Figure 4: Probability density distributions of the morphological features for the landslide sub-units at the La Valette landslide. (a) Local slope at a window size of 200 m. (b) Local curvature at a window size of 200 m. (c) Altitude normalized by the range of altitude values over the entire landslide.

- the K tree-cut examples provided by the expert (for each thematic ground area of each resolution image in the training area) during the hierarchical example-based segmentation step

$$K = k_1 + \cdots + k_n \text{ sets of color histograms and geometrical features} \tag{9}$$

These K examples can be used as input of the example-based segmentation approach to run the cutting strategy on the BPTs built for the different thematic ground areas of the images. This approach enables to automatically segment the n images at the target area;

- the C centroids of the clusters (related to the different resolution images) learned during the multiresolution clustering step (and their potential associated semantic label, *e.g.*, Cluster 1 $\rightarrow$ Vegetation class)

$$C = c_1 + \cdots + c_n \text{ sets of centroids} \tag{10}$$

These C centroids can be used as input of the multiresolution clustering approach to classify the segments automatically created by the example-based segmentation approach.

2.2.2. *Robustness to statistical variability*

Due to potential differences in the environmental conditions (*e.g.*, differences in atmospheric conditions, sun incidence angle, variations of the lithology, possible changes in surface soil moisture), the statistical distributions of the data in the target area may differ from the training area in a similar manner than between different datasets (Bahirat et al., 2012). Consequently, the statistical properties of the transferred learned elements could not be relevant to efficiently segment and classify the images at the target area. For instance, if the distribution of the color values is shifted from the training to the target area, the tree-cut examples provided by the expert during the hierarchical example-based segmentation step (modeled by color histograms) could not be adapted to segment the images composing the target area. Indeed, the algorithm used to repeat the segmentation examples on the remaining parts of the images is

mainly based on the computation of a distance between pairs of histograms and most of the classical distance (*e.g.*, the Euclidean one) are sensitive to the translation of the color values.

To deal with this issue, a solution consists of using a more robust distance to compute the similarity between histograms. We propose to use the constrained Dynamic Time Warping (DTW) similarity measure to compare the considered histograms (Sakoe and Chiba, 1978). Indeed, this similarity measure enables small distortions on the radiometric axis (Petitjean et al., 2012a).

Let $H(A) = \langle H_1(A), H_2(A), \ldots, H_v(A) \rangle$ and $H(B) = \langle H_1(B), H_2(B), \ldots, H_v(B) \rangle$ be two histograms of v bins. The similarity measure is computed as

$$DTW_\lambda(H_{1\cdots i}(A), H_{1\cdots j}(B)) = \begin{cases} +\infty & \text{if } |i-j| \geq \lambda \\ |H_i(A) - H_j(B)| + \min \begin{cases} DTW_\lambda(H_{1\cdots i-1}(A), H_{1\cdots j-1}(B)) \\ DTW_\lambda(H_{1\cdots i}(A), H_{1\cdots j-1}(B)) \\ DTW_\lambda(H_{1\cdots i-1}(A), H_{1\cdots j}(B)) \end{cases} & \text{otherwise} \end{cases} \tag{11}$$

where $\lambda \in \mathbb{Z}$ represents the tolerance of distortions on the radiometric axis and $H_{x\cdots y}(\star)$ represents the sub-histogram $\langle H_x(\star), \ldots, H_y(\star) \rangle$. This similarity measure requires larger computation times than the Euclidean distance, but provides better results. Associated to this measure, different averaging methods have been recently proposed (Petitjean et al., 2011; Petitjean and Gançarski, 2012) in order to define efficiently the mean histograms.

Since the centroids of the clusters learned during the multiresolution clustering step are modeled by class-based histograms (and robust spatial features), they are more adapted to deal with the statistical variability among the considered areas. Consequently, these elements can be transferred without adapting them to the statistical distribution of the target area.

3. Experiments

The datasets used to test the proposed method are described in Subsection 3.1. In order to allow the expert to actually test the proposed multiresolution methodology, a software has been designed. This tool is described in Subsection 3.2. The evaluation protocol is presented in Subsection 3.3 while the experimental settings and the results obtained are presented in Subsection 3.4.

3.1. Data

3.1.1. Multiresolution images

The study area is the Barcelonnette basin (South East French Alps) which is characterized by specific slope, lithological, and climate conditions favoring the triggering of landslides of different sizes (Malet et al., 2005). Here ,three subsets comprising the La Valette, the Poche and the Super-Sauze slow-moving landslides are examined (Figure 5). These landslides have a similar spatial organization with the presence of a more or less circular (in the horizontal plane) scarp area and an elongated transit zone narrower than the source area. These geometrical features are characteristic of flow-like landslides. Moreover, the three landslides have developed in the same lithology (*e.g.*, black marls) and feauture bare soils at the surface; the radiometric values observed on

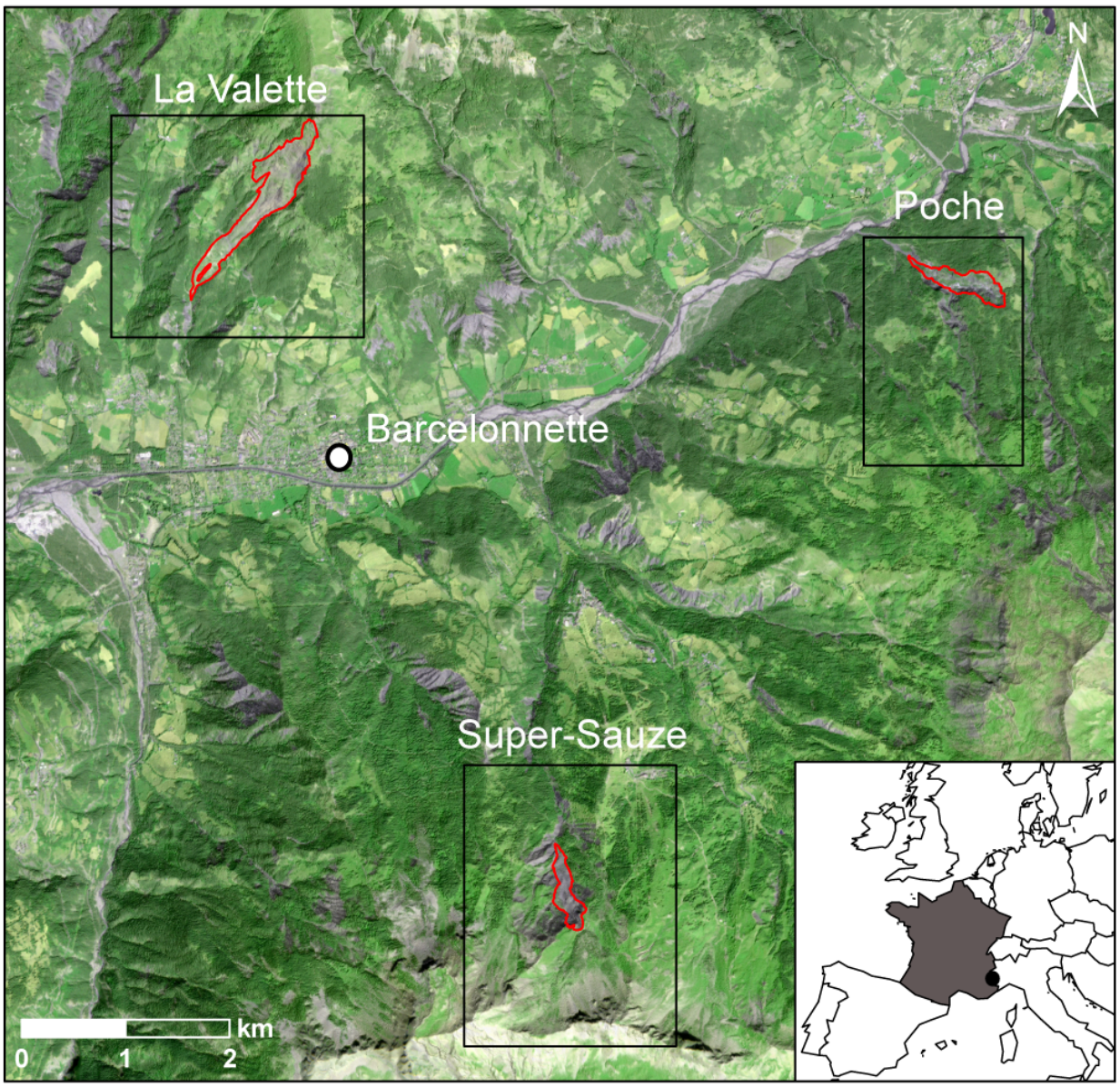


Figure 5: Study area: overview of the Barcelonnette basin (South East France) and of the three large active landslides considered in this experiment. This map has been built from an HR image (5 m) acquired in 2009 (© RAPIDEYE).

each landslide are therefore in the same range. We then consider three datasets (denoted **LA VALETTE**, **POCHE** and **SUPER-SAUZE**) composed each of:

- a LANDSATTM MR (30 m) multispectral image with seven spectral bands (denoted as $\mathcal{I}_1$);
- a RAPIDEYE HR (5 m) multispectral image with five spectral bands (denoted as $\mathcal{I}_2$);
- a VHR (50 cm) ortho-image, geometrically corrected, with three spectral bands (denoted as $\mathcal{I}_3$).

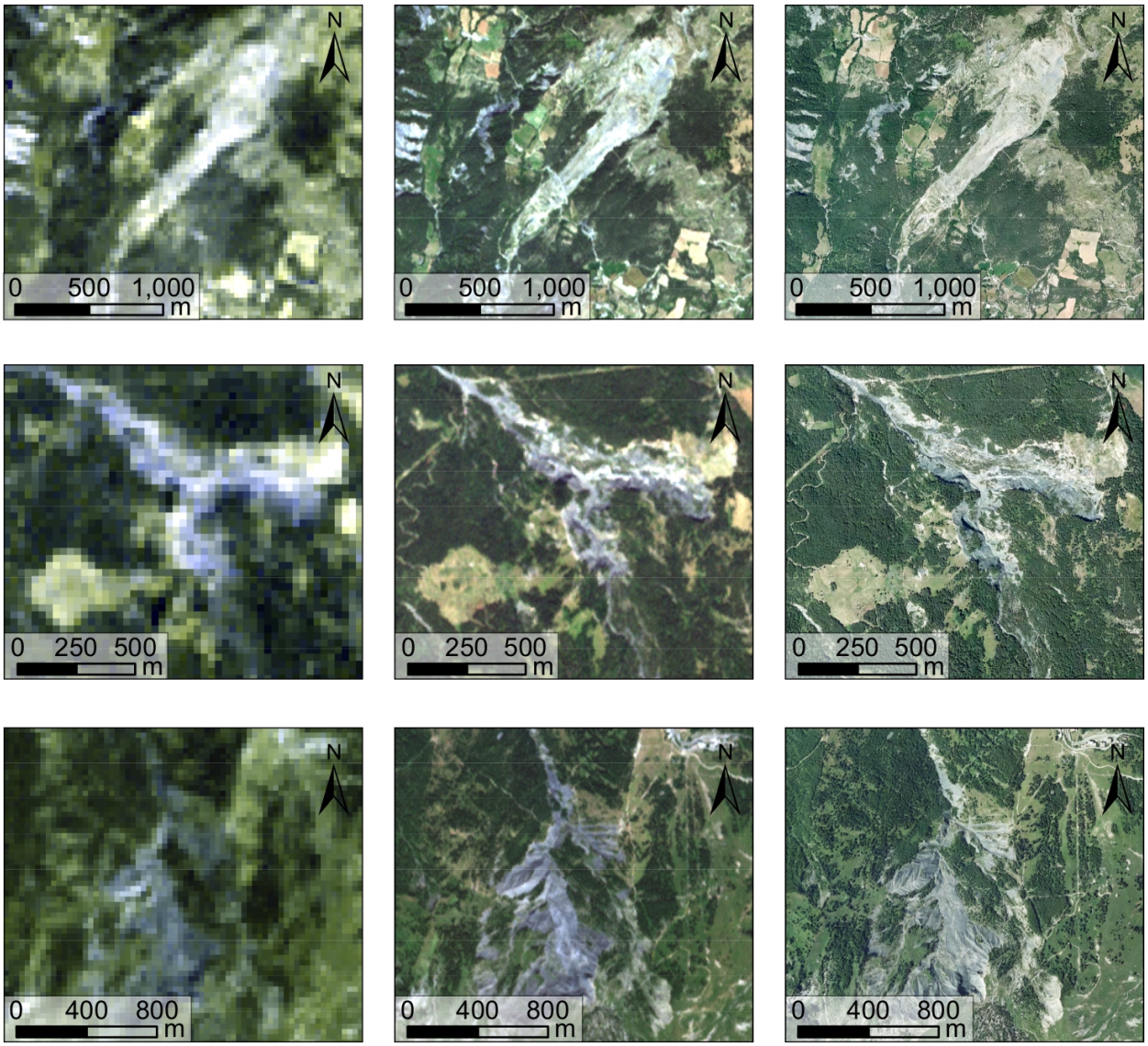


Figure 6: Multiresolution images of the three datasets. Top to bottom lines: **La Valette**, **Poche**, **Super-Sauze**. Left column: MR images (30 m) acquired in 2009 (© LandsatTM). Center column: HR images (5 m) acquired in 2009 (© RapidEye). Right column: VHR images (50 cm) acquired in 2009 (© IGN).

In order to avoid temporal issues, all these images have been acquired in Summer 2009. All the data are georeferenced in the same local cartographic projection (Lambert I). This guarantees the feasibility of the multiresolution approach. Figure 6 presents the images composing the three datasets considered in these experiments.

3.1.2. Terrain parameters

For each dataset, a DTM with a spatial resolution of 3.4 m is available. These models have been derived from three LiDAR surveys acquired in Summer 2009. From these DTMs, a slope gradient image and a curvature image are computed. In the classical case, the computation of the raster files corresponding to these gradient images is usually

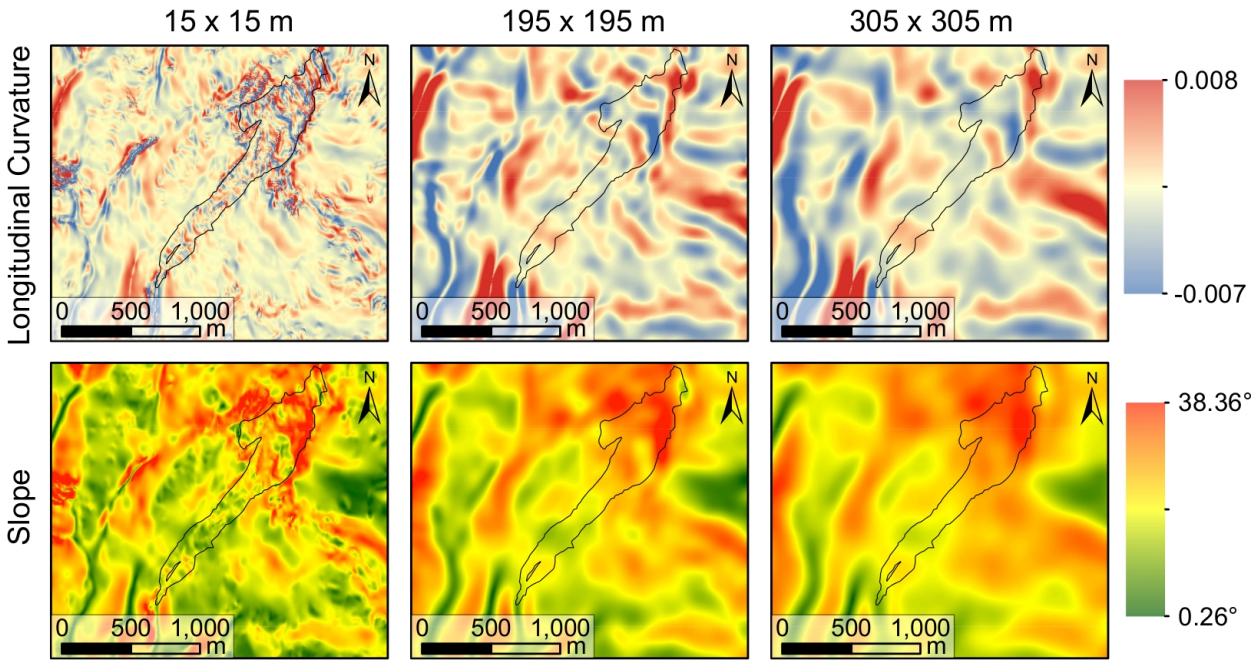


Figure 7: Illustration of the multiscale resampling approach of the DTM used on the **La Valette** dataset. This figure shows the influence of the window size parameter ϱ on the computation of the slope and curvature of the terrain.

done in a 3 m $\times$ 3 m neighborhood moving window. However, when dealing with a high spatial resolution DTM, this small size of neighborhood captures a lot of local details that are not relevant for the general partitioning of the landslide. To deal with this issue a solution consists of resampling the obtained raster files to enable the consideration of more general structures (*e.g.*, scarp, track, toe). Since simple resampling to coarser resolution would lead to a loss of information, we propose in these experiments to use a multiscale approach (Wood, 1996) that allows the computation of resampled images with variable window sizes *via* a least square fitting strategy. We have experimentally assessed the impact of the main parameter ϱ (*i.e.*, the longitudinal curvature window size – expressed in meters) by setting it to different successive values (3, ..., 115) and found that $\varrho \geq 39$ corresponds to a scale adapted to deal with the different structures of interest that need to be extracted (Figure 7).

These gradient images have been resampled successively to 5.0 m and 0.5 m spatial resolution images in order to process the HR and VHR images composing the datasets. In addition, these resulting images have been georeferenced in the same local cartographic projection as the one used for the considered multiresolution optical images.

3.1.3. Levels of analysis

To evaluate the efficiency of the proposed methodology, several tests have been performed to extract hierarchies of complex patterns in the three datasets. The TDHA was run with three images as input ($n = 3$) in order to extract three levels of details:

- **Level 1.** Large natural areas from the MR images $\mathcal{I}_1$: spatial separation between mineral and vegetated areas.
- **Level 2.** Landslides structures from the HR images $\mathcal{I}_2$: landslide bodies and surrounding crests and gullies.

- **Level 3.** Landslides sub-parts from the VHR images $\mathcal{I}_3$: source area, transport area and toe.

For Levels 1 and 2, the principal task was mainly to distinguish the landslide bodies from the surrounding crests and gullies, and since the affected area is one compact object this rather corresponds to an image segmentation task.

3.2. Software

In order to allow the expert to test the proposed multiresolution methodology, a software tool has been designed to interactively browse each BPT in a "threshold-like" approach in order to determine the most satisfactory segmentation examples (globally, and/or by refining one or several branches).

Due to the pre-processing of the data structures, the short computation time (less than 30 seconds of CPU) enables, in particular, to carry out several segmentations to finally select the best one. We have also developed (and integrated in this software) a TIFF library which allows to load only the subdivisions of the images that are necessary to the current segmentations. This library enables to reduce the memory resources required by the application.

This tool has been implemented using the Java MUSTIC library and the Orfeo Toolbox (OTB) framework. Both are open source libraries and are freely available[1]. It is planned to fully integrate the proposed methodology into these libraries and to distribute this software under a free license.

3.3. Evaluation protocol

Two evaluation protocols have been used to assess the efficiency of the TDHA:

1. visual analysis of the segmentation/classification results (for Level 1): a grade between 1 (bad accuracy) and 10 (high accuracy);

2. quantitative accuracy assessment with comparisons to ground-truth maps (for Levels 2 and 3).

To perform quantitative accuracy assessment, the classification results have been compared to certified ground-truth maps by using different quality indexes. The ground-truth maps are based on field work and visual interpretation of (V)HR satellite imagery. Figure 8 presents the ground-truth maps used in the experiments to evaluate the accuracy of the classification results.

We describe hereinafter the quality indexes considered in this study. The first index that has been used is the average F-measure $\overline{\mathcal{F}}$ corresponding to the mean, for each class, of the obtained F-measures. To this end, for each thematic class, the best corresponding clusters (in terms of partitions) were extracted. Then, we have computed: the percentage of false positives (denoted by $f^{(p)}$), the percentage of false negatives (denoted by $f^{(n)}$) and the percentage of true positives (denoted by $t^{(p)}$). These measures are used to estimate the precision $\mathcal{P}$ and the recall $\mathcal{R}$ of the results:

$$\mathcal{P} = \frac{t^{(p)}}{t^{(p)} + f^{(p)}} \tag{12}$$

$$\mathcal{R} = \frac{t^{(p)}}{t^{(p)} + f^{(n)}} \tag{13}$$

[1]The MUSTIC Java library, developed by some of the authors, can be downloaded at the following url: http://icube-bfo.unistra.fr/fr/index.php/Plateformes. The OTB framework is an open source set of tools for remote sensing data exploitation. It has been developed by the French Space Agency (CNES) to promote the use and the exploitation of the images derived from the PLEIADES systems (de Lussy et al., 2005). It can be downloaded at the following url: http://otb.cnes.fr.

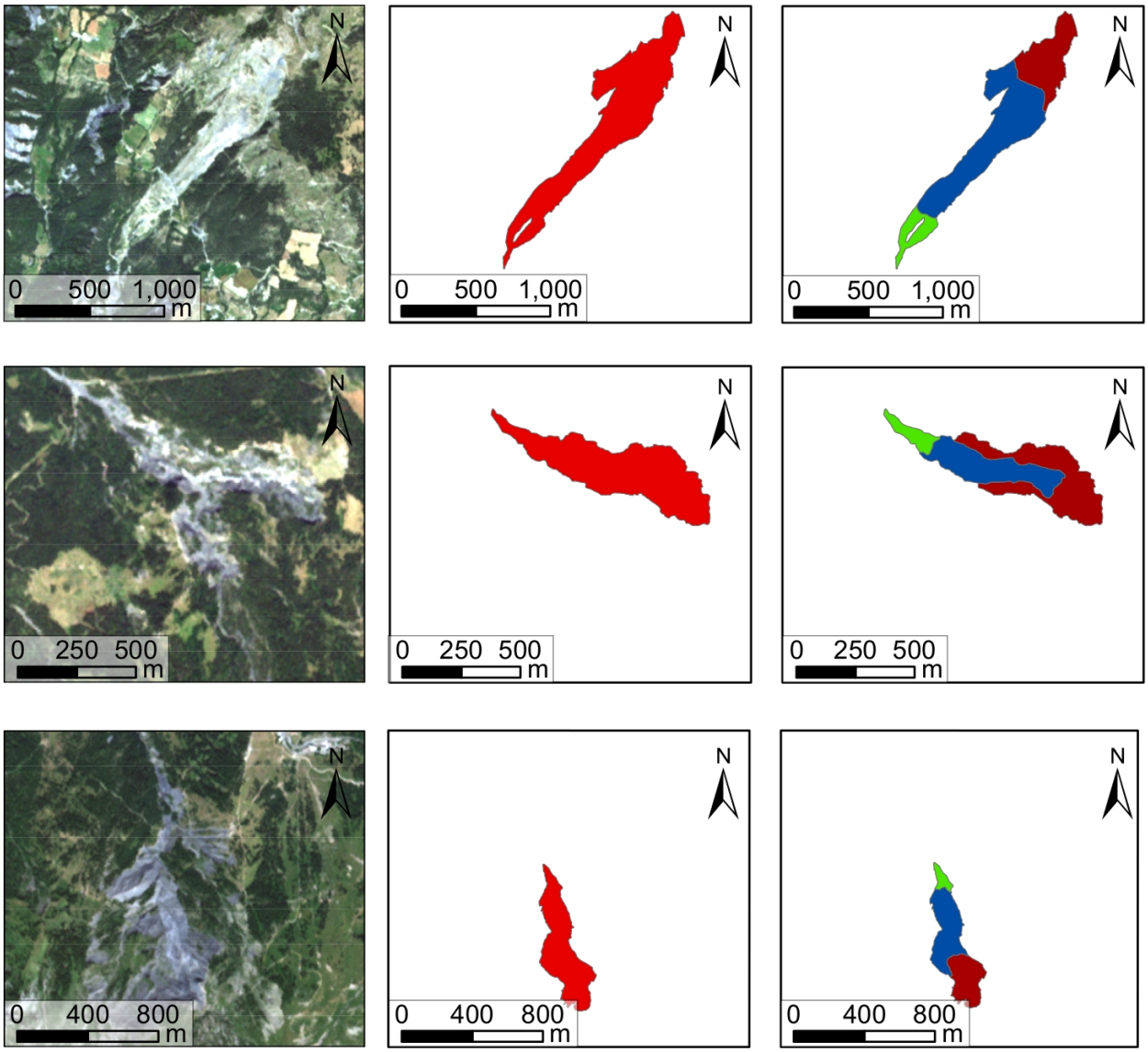
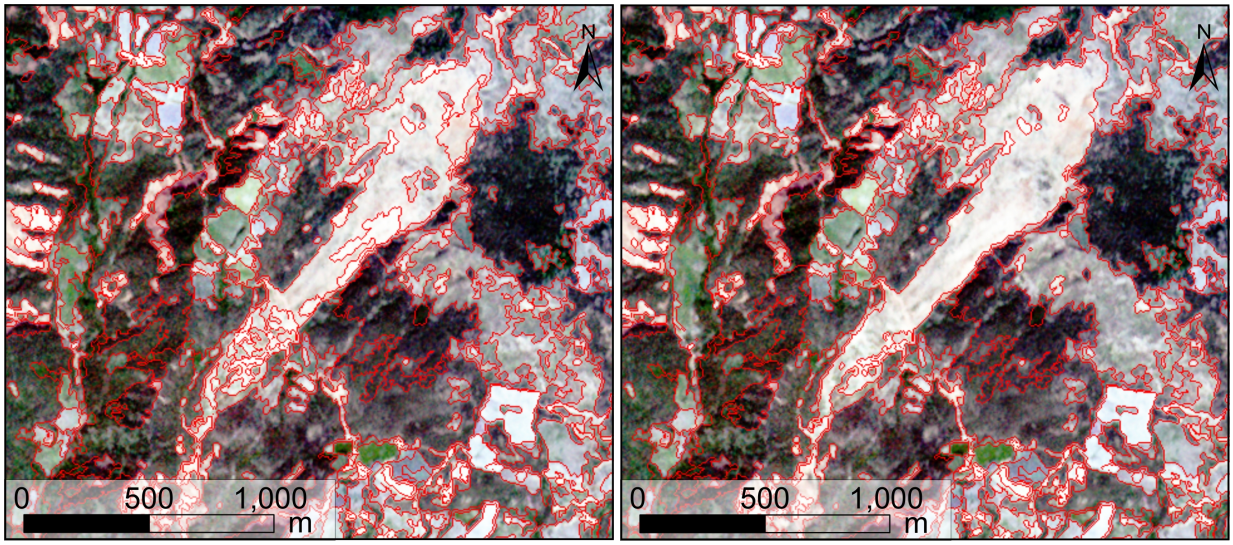


Figure 8: Ground-truth maps associated to the three datasets. Top to bottom lines: **La Valette**, **Poche**, **Super-Sauze**. Left column: HR images (5 m) acquired in 2009 (© RapidEye). Center column: Level 2. Landslide structures (in red). Right column: Level 3. Landslide sub-parts: source area (in brown), transport area (in blue) and toe (in light green).

For each experiment, we have then computed the geometrical mean $\overline{\mathcal{P}}$ of the obtained precisions and the geometrical mean $\overline{\mathcal{R}}$ of the obtained recalls. Finally, we have computed the mean F-measure $\overline{\mathcal{F}}$ which is the harmonic mean of the mean precision and the mean recall:

$$\overline{\mathcal{F}} = 2 \cdot \frac{\overline{\mathcal{P}}.\overline{\mathcal{R}}}{\overline{\mathcal{P}} + \overline{\mathcal{R}}} \tag{14}$$

The second index that has been computed is the Kappa index $\mathcal{K}$, which is a measure of global classification accuracy (Congalton, 1991). As the results provided by the K-means algorithm are sensitive to the initialization step of the algorithm, each classification run has been repeated ten times. We have then computed the variance value σ obtained



(a) (b)

Figure 9: Extract of the results (Level 2) obtained from the HR image of the **LA VALETTE** dataset. (a) Segmentation result obtained by considering only the radiometric information. (b) Segmentation result obtained by considering both the radiometric and the spatial information. The borders of the regions are depicted in red.

for each considered evaluation index and for each series of runs.

The computation of these class-specific indexes requires the matching of classes of interest with clusters extracted by the unsupervised classification approach. To this end, we have used an automatic strategy, which consists of selecting the clusters that maximize the overlapping with the corresponding class.

3.4. Experimental study

3.4.1. General settings

To process the three considered datasets, the TDHA has been parametrized as follows. The example-based hierarchical segmentation step and the multiresolution clustering step have been iteratively run three times ($n = 3$) to extract the objects of interest linked to the three considered scales. To perform, $K = 11$ tree-cut examples have been provided by the expert ($k_1 = 1$, $k_2 = 4$, $k_3 = 6$) and $C = 20$ clusters have been extracted ($c_1 = 4$, $c_2 = 6$, $c_3 = 10$). The values of these parameters have been chosen in agreement with the experts in landslide analysis to enable the extraction of the different land-cover classes. For instance for the MR images ($c_1 = 4$), four clusters have been extracted corresponding respectively to four land-cover classes: high reflecting mineral areas, low reflecting mineral areas, herbaceous vegetation areas and forest areas.

Preliminary experiments with the example-based hierarchical segmentation approach have shown that using tree-cut examples modeled by 10 color histograms, each one composed of 100 bins is sufficient to produce relevant results. One may notice that the multiresolution clustering approach did not directly find all the appropriate clusters with respect to the different classes of object of interest. To tackle this problem, the standard solution consists of extracting a higher number of clusters than the number of thematic classes contained in the expected results. Once these clusters have been extracted, the expert could recognize and manually associate them to natural landcover classes.

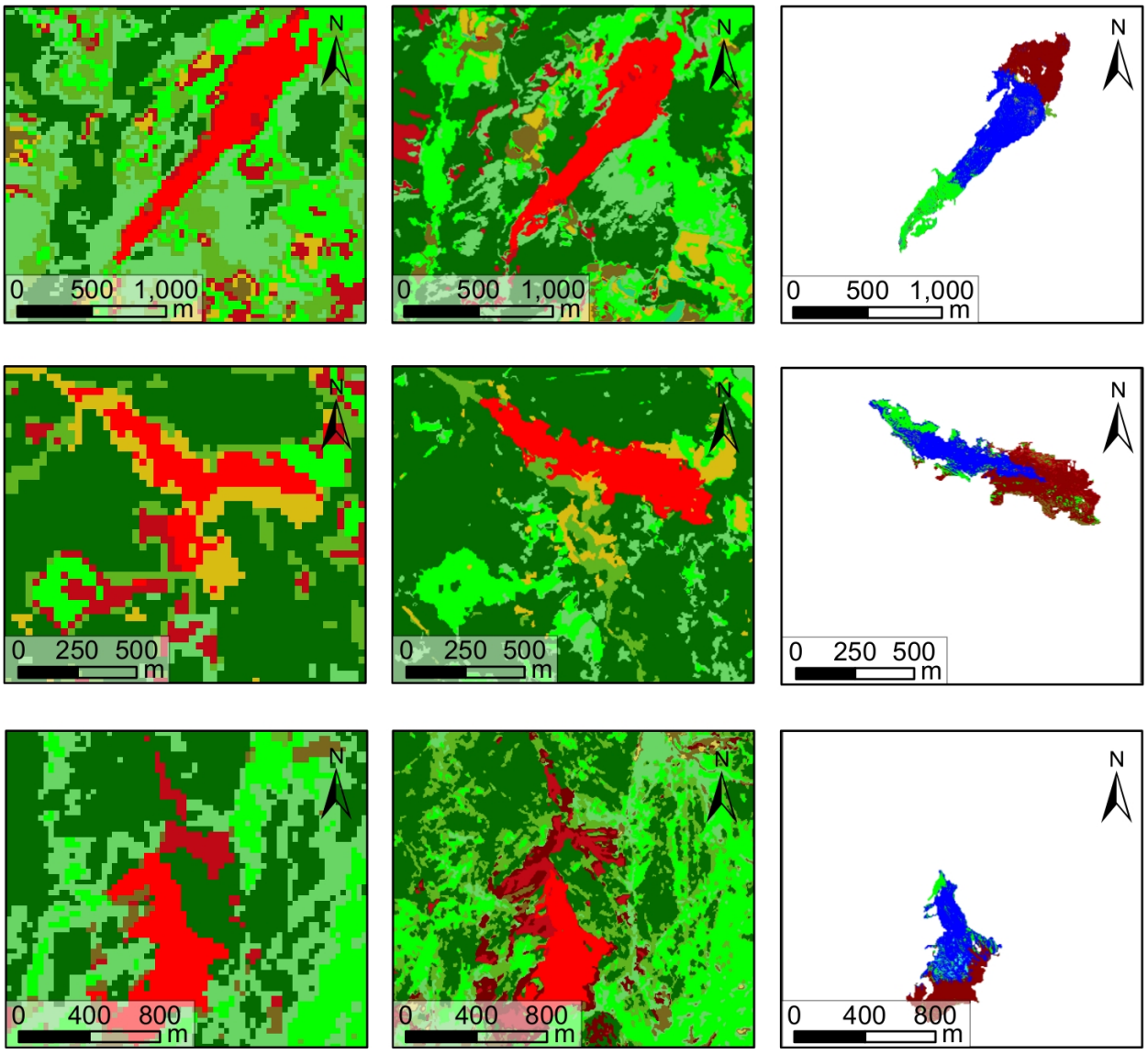


Figure 10: Classification results obtained from the three datasets (that have been processed independently). Top to bottom lines: **La Valette**, **Poche**, **Super-Sauze**. Left column: Level 1 obtained from MR images. Center column: Level 2 obtained from HR images. Right column: Level 3 obtained from VHR images.

3.4.2. Extension of TDHA for landslide detection

To validate the extension of TDHA to landslide detection, and in particular the interest of the use of spatial features, we have compared the results obtained by introducing or not these features both in the segmentation and in the classification steps.

Segmentation step. In order to evaluate the interest of constraining the way to build the BPTs by using or not the spatial features, we have run the segmentation step by varying the α parameter of the merging criterion (see Equation (7)) to different values.

Table 1: Evaluation results on the **LA VALETTE**, **POCHE**, **SUPER-SAUZE** datasets. Each dataset has been processed independently.

Levels	Features	Visual evaluation					
1. Landslide position	Multires. histograms	8/10		7/10		5/10	
		Evaluation with ground-truth maps					
		$\overline{\mathcal{F}} \pm \sigma$	$\mathcal{K} \pm \sigma$	$\overline{\mathcal{F}} \pm \sigma$	$\mathcal{K} \pm \sigma$	$\overline{\mathcal{F}} \pm \sigma$	$\mathcal{K} \pm \sigma$
2. Landslide body	Multires. histograms	0.63 ± 0.01	0.41 ± 0.02	0.57 ± 0.01	0.39 ± 0.02	0.61 ± 0.01	0.40 ± 0.02
3. Landslide sub-parts	Exp1 (A_1, A_2, A_3)	0.65 ± 0.01	0.38 ± 0.01	0.63 ± 0.02	0.37 ± 0.01	0.64 ± 0.02	0.38 ± 0.01
	Exp2 (A_1, A_2, A_3, $slop$)	0.67 ± 0.02	0.41 ± 0.02	0.65 ± 0.01	0.40 ± 0.02	$\mathbf{0.67} \pm 0.01$	$\mathbf{0.42} \pm 0.02$
	Exp3 (A_1, A_2, A_3, $curv$)	0.65 ± 0.02	0.40 ± 0.01	0.64 ± 0.02	0.38 ± 0.02	0.63 ± 0.02	0.38 ± 0.01
	Exp4 (A_1, A_2, A_3, $slop$, $curv$)	0.67 ± 0.01	0.41 ± 0.02	0.67 ± 0.01	$\mathbf{0.48} \pm 0.02$	0.66 ± 0.02	0.41 ± 0.01
	Exp5 (A_1, A_2, A_3, $slop$, $curv$, alt)	$\mathbf{0.69} \pm 0.02$	$\mathbf{0.42} \pm 0.01$	$\mathbf{0.68} \pm 0.02$	0.42 ± 0.01	$\mathbf{0.67} \pm 0.02$	$\mathbf{0.42} \pm 0.02$
		LA VALETTE		**POCHE**		**SUPER-SAUZE**	

The first segmentation experiment has been performed by only considering the radiometric region model of the nodes during their fusions. To this end, the weight α has been set to 1.0 in the equation of the merging criterion. The second segmentation experiment has been performed by considering both the radiometric and the geometrical region models of the nodes during their fusions. To this end, the weight α has been defined as stated in Equation (8). Once the images of the different datasets have been segmented by the hierarchical segmentation step, the resulting regions have been classified using the proposed multiresolution clustering approach. To evaluate the quality of these segmentation results, we focus on the results of the HR images (Level 2) for the three datasets.

After classification, the comparisons between the classified regions resulting from the first experiment (*i.e.*, radiometric model) and the ground-truth maps have shown Kappa values and F-Measures of ($\mathcal{K} = 0.38$ and $\overline{\mathcal{F}} = 0.61$) for the **LA VALETTE** dataset, ($\mathcal{K} = 0.36$ and $\overline{\mathcal{F}} = 0.54$) for the **POCHE** dataset and ($\mathcal{K} = 0.37$ and $\overline{\mathcal{F}} = 0.58$) for the **SUPER-SAUZE** dataset.

After classification, the comparisons between the classified regions resulting from the second experiment (*i.e.*, both radiometric and geometric models) and the ground-truth maps have shown Kappa values and F-Measures of ($\mathcal{K} = 0.41$ and $\overline{\mathcal{F}} = 0.64$) for the **LA VALETTE** dataset, ($\mathcal{K} = 0.40$ and $\overline{\mathcal{F}} = 0.58$) for the **POCHE** dataset and ($\mathcal{K} = 0.41$ and $\overline{\mathcal{F}} = 0.62$) for the **SUPER-SAUZE** dataset.

To assess visually the differences between the segmentation results obtained by considering or not the spatial information, Figure 9 focuses on the segmentation results of the HR image of the **LA VALETTE** dataset. One can see that the segmentation results obtained with the proposed landslides extension of the TDHA are spatially more consistent and regular than the results obtained without considering the spatial information. These quantitative and visual accuracy assessments show that considering spatial features during the segmentation step of the TDHA enables to enhance the segmentation of landslides.

Classification step. After the segmentation step, the resulting regions have been classified using the proposed multiresolution clustering approach. The regions extracted at Level 1 and Level 2 have been characterized using composition histograms and the classification maps obtained have been visually or quantitatively evaluated. For Level 3, different combinations of radiometric and spatial features have been used to characterize the regions to be classified:

- Experiment 1: (A_1, A_2, A_3) the radiometric mean value of each spectral band;
- Experiment 2: $(A_1, A_2, A_3, slop)$ the radiometric mean value of each spectral band combined with the slope value;
- Experiment 3: $(A_1, A_2, A_3, curv)$ the radiometric mean value of each spectral band combined with the curvature value;
- Experiment 4: $(A_1, A_2, A_3, slop, curv)$ the radiometric mean value of each spectral band combined with the slope and the curvature values;
- Experiment 5: $(A_1, A_2, A_3, slop, curv, alt)$ the radiometric mean value of each spectral band combined with the slope, the curvature and the altitude values.

As the altitude information of the regions represent absolute values extracted from the DTM, they can differ from one landslide to another and then cannot be sufficiently discriminative to separate the regions composing the different sub-units of the landslides. To deal with this issue, we have normalized the altitude values of the regions by using the global structure of the landslides extracted at Level 2. To this end, we assume that a landslide has a slope structure. By noting alt_{max} and alt_{min} its highest and lowest altitude, respectively, we then affect to each region $R_i \subseteq E$ composing the landslide a normalized altitude value computed as:

$$alt_{norm}(R_i) = \frac{alt(R_i) - alt_{min}}{alt_{max} - alt_{min}} \tag{15}$$

Results obtained on the three datasets are shown on Figure 10. For each level of analysis, the classification maps obtained are represented. The results of the visual and quantitative accuracy assessments are presented in Table 1.

Figure 10 (left column) presents the results for the extraction of large natural areas (Level 1) from the MR images. The mineral parts (orange, red clusters) and the vegetation parts (light green, dark green clusters) are correctly separated by the process. Furthermore, the landslides are correctly located in the image.

Figure 10 (center column) presents the results for the extraction of the landslides structures (Level 2) from the HR images. The landslide body is correctly delineated in the classification result and the different kinds of vegetation (light green, dark green clusters) are correctly separated by the process.

Finally, Figure 10 (right column) shows the results for the extraction of the landslide sub-parts (Level 3) from the VHR images.

To compare the impact of the different spatial features used during the classification of the regions extracted at Level 3, Figure 11 shows the classification results obtained on the **La Valette** using the different combinations of features. One can note that the more discriminative spatial features are the slope and the altitude values, that enable to accurately separate the landslide structure in the different sub-parts of interest. This visual study is also confirmed by the results of the comparisons with the ground-truth maps presented in Table 1.

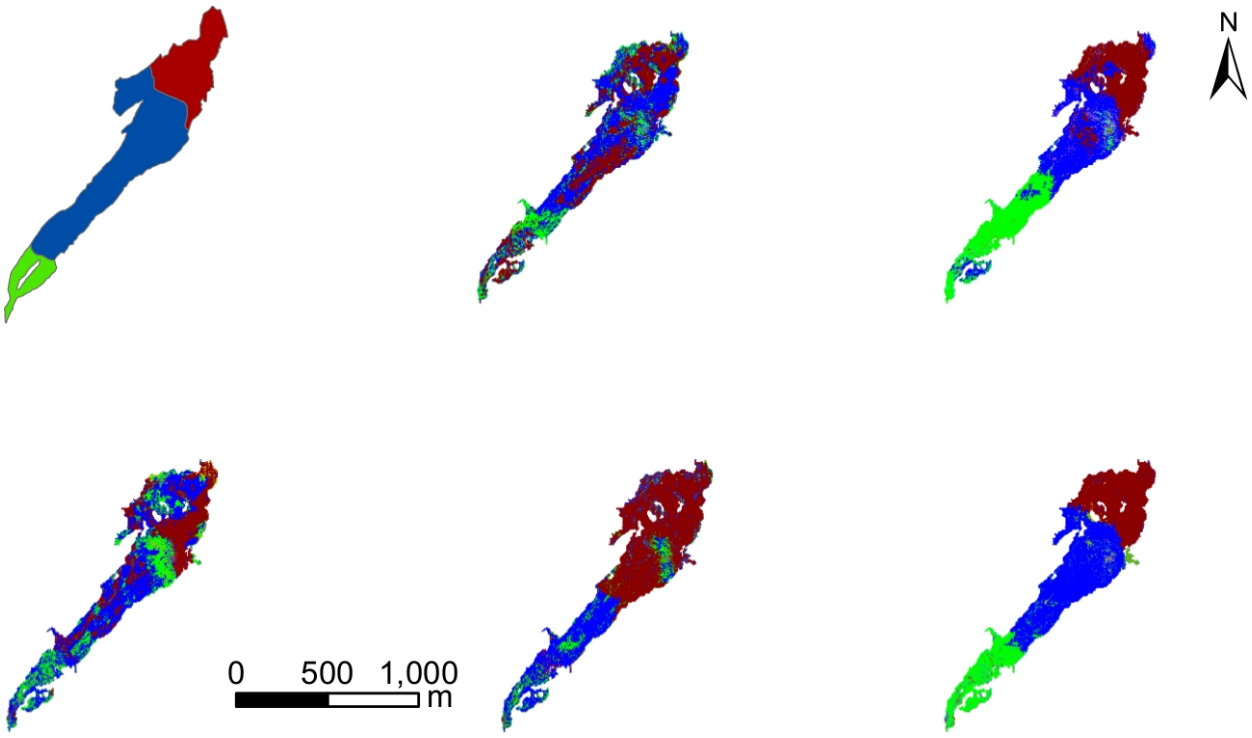


Figure 11: Classification results at Level 3 obtained from the VHR image of the **LA VALETTE** dataset. Top left: ground-truth map. Top center: experiment 1 (A_1, A_2, A_3). Top right: experiment 2 ($A_1, A_2, A_3, slop$). Bottom left: experiment 3 ($A_1, A_2, A_3, curv$). Bottom center: experiment 4 ($A_1, A_2, A_3, slop, curv$). Bottom right: experiment 5 ($A_1, A_2, A_3, slop, curv, alt$).

3.4.3. Domain adaptation experiment

Experimental protocol. To evaluate the interest of the presented domain adaptation strategy, learning transfer experiments have been carried out:

- **SUPER-SAUZE** $\rightarrow$ **LA VALETTE**: the **SUPER-SAUZE** area has been used as training area while the **LA VALETTE** area has been used as target area;
- **LA VALETTE** $\rightarrow$ **POCHE**: the **LA VALETTE** area has been used as training area while the **POCHE** area has been used as target area;
- **POCHE** $\rightarrow$ **SUPER-SAUZE**: the **POCHE** area has been used as training area while the **SUPER-SAUZE** area has been used as target area.

Practically, the $K = 11$ tree-cut examples (provided by the expert during the example-based segmentation step) and the $C = 20$ centroids of the clusters (learned during the multiresolution clustering step) which are resulting from the processing of the training area, have been transferred to automatically process the target dataset. However, as different specific thematic classes (*i.e.*, mineral and/or vegetated clusters) are not present in the considered target datasets, the segmentation examples and the learned clusters corresponding to these classes have not been considered during the transfer step.

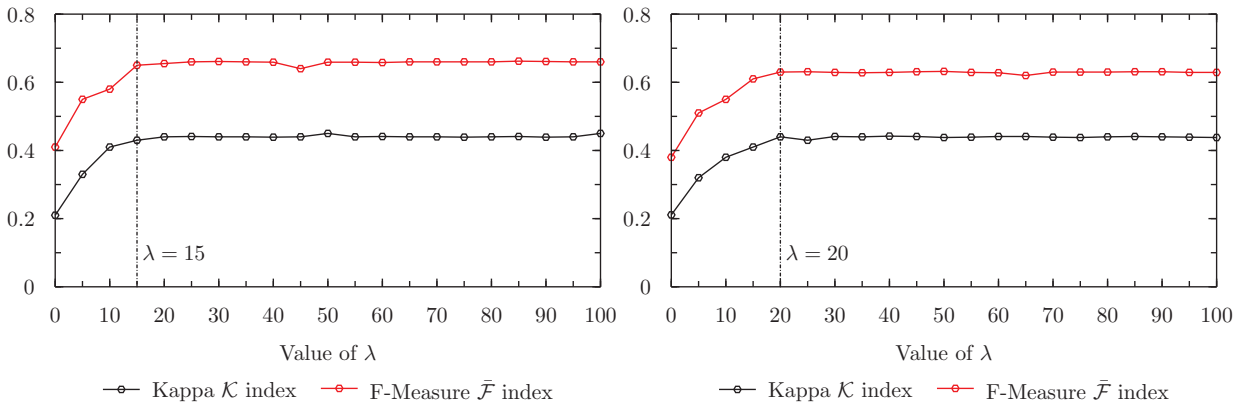
Table 2: Evaluation results of the domain adaptation experiments. For each experiment one dataset has been used as training dataset while another one has been used as target one.

Levels	Features	Visual evaluation					
1. Landslide position	Multires. histograms	7/10		6/10		5/10	
		Evaluation with ground-truth maps					
		$\overline{\mathcal{F}} \pm \sigma$	$\mathcal{K} \pm \sigma$	$\overline{\mathcal{F}} \pm \sigma$	$\mathcal{K} \pm \sigma$	$\overline{\mathcal{F}} \pm \sigma$	$\mathcal{K} \pm \sigma$
2. Landslide body	Multires. histograms	0.61 ± 0.02	0.38 ± 0.02	0.56 ± 0.01	0.38 ± 0.02	0.59 ± 0.01	0.38 ± 0.01
3. Landslide sub-parts	Exp1 (A_1, A_2, A_3)	0.63 ± 0.01	0.37 ± 0.01	0.59 ± 0.02	0.36 ± 0.02	0.62 ± 0.01	0.37 ± 0.01
	Exp2 (A_1, A_2, A_3, $slop$)	0.65 ± 0.01	0.40 ± 0.02	0.61 ± 0.01	0.39 ± 0.02	$\mathbf{0.65} \pm 0.02$	0.41 ± 0.02
	Exp3 (A_1, A_2, A_3, $curv$)	0.65 ± 0.02	0.40 ± 0.01	0.61 ± 0.01	0.37 ± 0.01	0.61 ± 0.01	0.39 ± 0.02
	Exp4 (A_1, A_2, A_3, $slop$, $curv$)	0.65 ± 0.01	0.39 ± 0.01	0.58 ± 0.02	$\mathbf{0.48} \pm 0.02$	$\mathbf{0.65} \pm 0.02$	0.40 ± 0.01
	Exp5 (A_1, A_2, A_3, $slop$, $curv$, alt)	$\mathbf{0.66} \pm 0.02$	$\mathbf{0.42} \pm 0.01$	$\mathbf{0.63} \pm 0.02$	0.41 ± 0.01	$\mathbf{0.65} \pm 0.02$	$\mathbf{0.42} \pm 0.02$
		Super-Sauze → La Valette		La Valette → Poche		Poche → Super-Sauze	

The impact of the λ parameter (required for the computation of the constrained DTW similarity measure) on the quality of the obtained segmentation/clustering results has been tested. We recall that this parameter represents the tolerance of distortions on the radiometric axis between the bins of pairs of color histograms during their comparisons (see Equation (11)). It enables to deal with the statistical variability of the datasets during the transfer of the learned knowledge. To study the impact of λ, we have run the domain adaptation experiments on images at different resolutions, by varying the parameter ($\lambda = 0, 5, 10, 15, 20, 25, \ldots, 100$). As the considered color histograms are composed each of 100 bins, the maximal value of λ is 100. Each experiment has then been assessed by using the evaluation measures presented previously.

Domain adaptation results. Evaluation results obtained on the domain adaptation experiments are presented in Table 2. The results obtained on the three datasets are slightly lower when the segmentation examples and the clusters learned are transferred from another dataset than when they are learned directly. However, such results are sufficient to detect landslides. Furthermore, these experiments show that it is possible to maximize the potential of the expert-defined segmentation examples (and thus to reduce the time spent by the expert) to process several large datasets sensed over similar hazardous territories.

Impact of the λ parameter. We discuss hereinafter the impact of the choice of the λ parameter value during the learning transfer step. The result of this impact-study is presented in Figure 12. From $\lambda = 0$ to $\lambda = 20$, the quality of the segmentation and clustering results increases. In particular, best results are obtained when $\lambda = 15$ for the (**La Valette** → **Poche**) experiment and when $\lambda = 20$ for the (**Poche** → **Super-Sauze**) experiment. After $\lambda = 20$ the quality of the results remains stable.



(a) Results for the **La Valette → Poche** experiment.

(b) Results for the **Poche → Super-Sauze** experiment.

Figure 12: Impact study of the λ parameter for the computation of the constrained Dynamic Time Warping similarity measure.

Such experiments confirm our original assumption: the automatic reproduction step of the tree-cut examples is directly affected by the differences between the statistical distributions of the data composing the training area and the target one. However, the results obtained show that the use of the constrained DTW similarity measure to compare the considered histograms is a relevant strategy to deal with this issue.

Nevertheless, a too high value of λ can increase the computation times. Experiments show that $\lambda = 15$ is a good balance between time-cost constraints and the desired accuracy of the segmentation results for the considered images.

Discussion. We show in this study that the hierarchical segmentation and classification models learned on a first dataset sensed over a particular landslide can be re-used in order to automatically process other datasets sensed over the same kind of landslide. An actual limit is that the proposed approach requires that the multiresolution images used for the mapping only cover the landslide surface (and not a larger heterogeneous region with different types of landslides). A potential solution to this issue would consist of integrating the TDHA to a landslide mapping framework initialized with a landslide detection approach (Stumpf et al., 2013). Such a detection step could be performed from a low resolution image covering a large heterogeneous area composed of various landslides. The TDHA approach could then be used for the fine mapping of a particular landslide selected by the detector. The domain adaptation strategy could then be used to automatically process other landslides of the large area.

3.4.4. Comparisons with related works

To highlight the relevance of the proposed top-down multiresolution methodology for landslide mapping, we compared the results obtained within this framework to those obtained with other related methodologies. In particular, we considered two specific segmentation approaches, based on a bottom-up strategy and that can be used to extract the sub-parts of the landslides directly from the VHR images (*i.e.*, in a monoresolution fashion).

More precisely, the TDHA has been compared to the standard Mean-Shift algorithm (Comaniciu and Meer, 2002) and a region-merging algorithm (based on both image spectral and textural characteristics) (Baatz and Schape, 2000) which have shown satisfactory results in the context of remote sensing image segmentation. We have used the OTB implementation of the Mean-Shift segmentation algorithm with manually selected parameter values and an implemen-

tation of the region-merging segmentation algorithm provided in the Definiens eCognition software. The segmentation maps provided by these two algorithms have been classified in the same way as we did in the TDHA.

The results obtained on the three datasets (**LA VALETTE**, **POCHE**, **SUPER-SAUZE**) have shown that the average percentages of sub-parts of the landslides well recognized are comparable to those obtained with the TDHA. Indeed, the proposed multiresolution methodology enables to extract a large part of the objects of interest ($\overline{\mathcal{K}} = 0.43$ and $\overline{\mathcal{F}} = 0.69$) while these scores reach ($\overline{\mathcal{K}} = 0.40$ and $\overline{\mathcal{F}} = 0.64$) with the Mean-Shift algorithm and ($\overline{\mathcal{K}} = 0.38$ and $\overline{\mathcal{F}} = 0.61$) with the region-merging algorithm. A qualitative analysis carried by two experts in landslide mapping has also confirmed these quantitative results.

This comparative study suggests that the proposed top-down multiresolution methodology provides better results than the ones obtained with a bottom-up strategy applied on VHR images. Furthermore, the TDHA enables to extract different levels of objects of interest while most of the other related approaches enable only the extraction at a single semantic level. This multiscale property leads to a better understanding of the observed scenes and their compositional structures which is a crucial need when dealing with the mapping of geohazards. In addition, our methodology is interactive and thus intuitive to configure for the expert who needs to process rapidly huge mass of data.

4. Conclusion

This article has introduced a novel approach for the detection of landslides from multiresolution sets of images. It is based on a top-down multiresolution hierarchical approach that takes advantage both of geomorphological information for guiding the segmentation/classification steps, and of a domain adaptation strategy to optimize and minimize the expert's interaction.

From a methodological point of view, it constitutes an original solution to address the issues linked to the analysis of VHR images. Based on a multiresolution paradigm, the proposed approach enables to deal with such images without being convoluted by the large size and the level of details of these data. Furthermore, this partitioning approach authorizes to adapt the scale of the segmentation process to restricted areas of the images instead of segmenting the whole dataset using only one segmentation parameter. To reduce the amount of time spent by the expert to supervise the global extraction process, it has been proposed to re-use the knowledge learned on a first dataset in order to automatically process other datasets. We have seen that considering the DTW criteria as similarity measure when comparing the knowledge extracted from a training area to the one extracted from a target area, enables to deal with the variability in the statistical distribution of the data, and then to enhance the classification process.

From an applicative point of view, this article has proposed a novel strategy to address the issues linked to the analysis of geohazards. Experiments performed on three sets of multiresolution satellite images were carried out to evaluate the robustness of this methodology in the context of the extraction of natural landslides. These experiments have provided satisfactory and accurate classification results which may be further used for operational needs, in particular for the assessment of natural disasters in hazardous areas.

This work opens up various research perspectives. As methodological perspectives, the choice of the considered spatial features in the segmentation and in the classification processes has to be deeply studied. For instance, textural and topological features could be used. We also plan to validate the proposed methodology by using other segmentation strategies. Finally, we plan to study more formally the robustness of the learning transfer step to deal with

different kinds of landslides and images. As technological perspectives, the coming pair of SENTINEL-2 satellites will provide at the same time images with different spatial and radiometric resolutions (four bands at 10 m, six bands at 20 m and three bands at 60 m at a high temporal frequency. Coupled with VHR images sensed by the PLEIADES satellites, these images will offer important properties of temporal and multiresolution complementarity adapted to deal with landslides extraction. In this context, we plan to study new trends to extend the proposed approach to deal with such properties by using dedicated multitemporal approaches (Petitjean et al., 2012b).

Acknowledgements

The research leading to these results has received funding from the French *Agence Nationale de la Recherche* (Grant Agreement ANR-10-COSI-012-03), and from the *Service National d'Observation* CNRS-INSU OMIV (*Observatoire Multidisciplinaire des Instabilités de Versants*).

References

Akcay, H.G., Aksoy, S., 2008. Automatic detection of geospatial objects using multiple hierarchical segmentations. IEEE Transactions on Geoscience and Remote Sensing 46, 2097–2111.

Alonso-González, A., Valero, S., Chanussot, J., López-Martinez, C., Salembier, P., 2012. Processing multidimensional SAR and hyperspectral images with binary partition tree. Proceedings of the IEEE 101, 723–747.

Anders, N.S., Seijmonsbergen, A.C., Bouten, W., 2011. Segmentation optimization and stratified object-based analysis for semi-automated geomorphological mapping. Remote Sensing of Environment 115, 2976–2985.

Baatz, M., Schape, A., 2000. Multiresolution segmentation–An optimization approach for high quality multi-scale image segmentation, in: Angewandte Geographische Informationsverarbeitung Symposium, pp. 12–23.

Bahirat, K., Bovolo, F., Bruzzone, L., Chaudhuri, S., 2012. A novel domain adaptation Bayesian classifier for updating land-cover maps with class differences in source and target domains. IEEE Transactions on Geoscience and Remote Sensing 50, 2810–2826.

Barlow, J., Franklin, S., Martin, Y., 2006. High spatial resolution satellite imagery, DEM derivatives, and image segmentation for the detection of mass wasting processes. Photogrammetric Engineering and Remote Sensing 72, 687–692.

Benediktsson, J.A., Bruzzone, L., Chanussot, J., Dalla Mura, M., Salembier, P., Valero, S., 2011. Hierarchical analysis of remote sensing data: Morphological attribute profiles and binary partition trees, in: Soille, P., Pesaresi, M., Ouzounis, G.K. (Eds.), Proceedings of the International Symposium on Mathematical Morphology – ISMM, Springer. pp. 306–319.

Benediktsson, J.A., Chanussot, J., Moon, W., 2012. Very High-Resolution remote sensing: Challenges and opportunities. Proceedings of the IEEE 100, 1907–1910.

Benz, U.C., Hofmann, P., Willhauck, G., Lingenfelder, I., Heynen, M., 2004. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. ISPRS Journal of Photogrammetry and Remote Sensing 58, 239–258.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 65, 2–16.

Bruzzone, L., Prieto, D., Serpico, S., 1999. A neural-statistical approach to multitemporal and multisource remote-sensing image classification. IEEE Transactions on Geoscience and Remote Sensing 37, 1350–1359.

Chang, Y.L., Liang, L.S., Han, C.C., Fang, J.P., Liang, W.Y., Chen, K.S., 2007. Multisource data fusion for landslide classification using generalized positive Boolean functions. IEEE Transactions on Geoscience and Remote Sensing 45, 1697–1708.

Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis.IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 603–619.

Congalton, R., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37, 35–46.

Daumé III, H., Marcu, D., 2006. Domain adaptation for statistical classifiers. Journal of Artificial Intelligence Research 26, 101–126.

Fiorucci, F., Cardinali, M., Carlà, R., Rossi, M., Mondini, A.C., Santurri, L., Ardizzone, F., Guzzetti, F., 2011. Seasonal landslide mapping and estimation of landslide mobilization rates using aerial and satellite images. Geomorphology 129, 59–70.

Gaetano, R., Scarpa, G., Poggi, G., 2009. Hierarchical texture-based segmentation of multiresolution remote-sensing images. IEEE Transactions on Geoscience and Remote Sensing 47, 2129–2141.

Galli, M., Ardizzone, F., Cardinali, M., Guzzetti, F., Reichenbach, P., 2008. Comparing landslide inventory maps. Geomorphology 94, 268–289.

Garrido, L., Salembier, P., Garcia, D., 1998. Extensive operators in partition lattices for image sequence analysis. Signal Processing 66, 157–180.

Guzzetti, F., Mondini, A.C., Cardinali, M., Fiorucci, F., Santangelo, M., Chang, K.T., 2012. Landslide inventory maps: New tools for an old problem. Earth-Science Reviews 112, 42–66.

Hölbling, D., Füreder, P., Antolini, F., Cigna, F., Casagli, N., Lang, S., 2012. A semi-automated object-based approach for landslide detection validated by persistent scatterer interferometry measures and landslide inventories. Remote Sensing 4, 1310–1336.

Kayastha, P., Dhital, M.R., De Smedt, F., 2013. Application of the analytical hierarchy process (AHP) for landslide susceptibility mapping: A case study from the Tinau watershed, west Nepal. Computers & Geosciences 52, 398–408.

Kurtz, C., Passat, N., Gançarski, P., Puissant, A., 2010. Multiresolution region-based clustering for urban analysis. International Journal of Remote Sensing 31, 5941–5973.

Kurtz, C., Passat, N., Gançarski, P., Puissant, A., 2012. Extraction of complex patterns from multiresolution remote sensing images: A hierarchical top-down methodology. Pattern Recognition 45, 685–706.

Kurtz, C., Passat, N., Puissant, A., Gançarski, P., 2011. Hierarchical segmentation of multiresolution remote sensing images, in: Soille, P., Pesaresi, M., Ouzounis, G.K. (Eds.), Proceedings of the International Symposium on Mathematical Morphology – ISMM, Springer. pp. 343–354.

Lahousse, T., Chang, K.T., Lin, Y.H., 2011. Landslide mapping with multi-scale object-based image analysis: A case study in the Baichi watershed, Taiwan. Natural Hazards and Earth System Science 11, 2715–2726.

Lu, P., Stumpf, A., Kerle, N., Casagli, N., 2011. Object-oriented change detection for landslide rapid mapping. IEEE Geoscience and Remote Sensing Letters 8, 701–705.

de Lussy, F., Kubik, P., Greslou, D., Pascal, V., Gigord, P., Cantou, J., 2005. PLEIADES-HR image system products and geometric accuracy, in: Proceedings of the ISPRS Hannover Workshop on High-Resolution Earth Imaging for Geospatial Information – WHREI, pp. 50–57.

MacQueen, J.B., 1967. Some methods of classification and analysis of multivariate observations, in: Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability – BSMSP, pp. 281–297.

Malet, J.P., Laigle, D., Remaître, A., Maquaire, O., 2005. Triggering conditions and mobility of debris flows associated to complex earthflows. Geomorphology 66, 215–235.

Mallinis, G., Koutsias, N., Tsakiri-Strati, M., Karteris, M., 2008. Object-based classification using Quickbird imagery for delineating forest vegetation polygons in a Mediterranean test site. ISPRS Journal of Photogrammetry and Remote Sensing 63, 237–250.

Martha, T.R., Kerle, N., Jetten, V., van Westen, C.J., Kumar, K.V., 2010. Characterising spectral, spatial and morphometric properties of landslides for semi-automatic detection using object-oriented methods. Geomorphology 116, 24–36.

Martha, T.R., Kerle, N., van Westen, C.J., Jetten, V., Kumar, K.V., 2012. Object-oriented analysis of multi-temporal panchromatic images for creation of historical landslide inventories. ISPRS Journal of Photogrammetry and Remote Sensing 67, 105–119.

Martha, T.R., Kumar, K.V., 2013. September, 2012 landslide events in Okhimath, India – An assessment of landslide consequences using very high resolution satellite data. Landslides 10, 469–479.

Mondini, A.C., Guzzetti, F., Reichenbach, P., Rossi, M., Cardinali, M., Ardizzone, F., 2011. Semi-automatic recognition and mapping of rainfall induced shallow landslides using optical satellite images. Remote Sensing of Environment 115, 1743–1757.

Mondini, A.C., Marchesini, I., Rossi, M., Chang, K.T., Pasquariello, G., Guzzetti, F., 2013. Bayesian framework for mapping and classifying shallow landslides exploiting remote sensing and topographic data. Geomorphology (In press). doi:10.1016/j.geomorph.2013.06.015.

Nichol, J., Wong, M.S., 2005. Satellite remote sensing for detailed landslide inventories using change detection and image fusion. International Journal of Remote Sensing 26, 1913–1926.

Petitjean, F., Gançarski, P., 2012. Summarizing a set of time series by averaging: From Steiner sequence to compact multiple alignment. Theoretical Computer Science 414, 76–91.

Petitjean, F., Inglada, J., Gançarski, P., 2012a. Satellite image time series analysis under time warping. IEEE Transactions on Geoscience and Remote Sensing 50, 3081–3095.

Petitjean, F., Ketterlin, A., Gançarski, P., 2011. A global averaging method for Dynamic Time Warping, with applications to clustering. Pattern Recognition 44, 678–693.

Petitjean, F., Kurtz, C., Passat, N., Gançarski, P., 2012b. Spatio-temporal reasoning for the classification of satellite image time series. Pattern

Recognition Letters 33, 1805–1815.

Raucoules, D., de Michele, M., Malet, J.P., Ulrich, P., 2013. Time-variable 3D ground displacements from High-Resolution Synthetic Aperture Radar (SAR). Application to La Valette landslide (South French Alps). Remote Sensing of Environment 139, 198–204.

Sakoe, H., Chiba, S., 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing 26, 43–49.

Salembier, P., Garrido, L., 2000. Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. IEEE Transactions on Image Processing 9, 561–576.

Salembier, P., Wilkinson, M.H.F., 2009. Connected operators: A review of region-based morphological image processing techniques. IEEE Signal Processing Magazine 26, 136–157.

Stumpf, A., Kerle, N., 2011. Object-oriented mapping of landslides using random forests. Remote Sensing of Environment 115, 2564–2577.

Stumpf, A., Lachiche, N., Malet, J.P., Puissant, A., 2013. Active learning in the spatial domain for remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing (In press). doi:10.1109/TGRS.2013.2262052.

Sun, W., Heidt, V., Gong, P., Xu, G., 2003. Information fusion for rural land-use classification with High-Resolution satellite imagery. IEEE Transactions on Geoscience and Remote Sensing 41, 883–890.

Thiery, Y., Malet, J.P., Sterlacchini, S., Puissant, A., Maquaire, O., 2007. Landslide susceptibility assessment by bivariate methods at large scales: Application to a complex mountainous environment. Geomorphology 92, 38–59.

Townshend, J.R.G., Huang, C., Kalluri, S.N.V., Defries, R.S., Liang, S., Yang, K., 2000. Beware of per-pixel characterization of land cover. International Journal of Remote Sensing 21, 839–843.

Valero, S., Salembier, P., Chanussot, J., 2010. New hyperspectral data representation using binary partition tree, in: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium – IGARSS, pp. 80–83.

Wemmert, C., Puissant, A., Forestier, G., Gançarski, P., 2009. Multiresolution remote sensing image clustering. IEEE Geoscience and Remote Sensing Letters 6, 533–537.

Wood, J.D., 1996. The geomorphological characterisation of digital elevation models. Ph.D. thesis. University of Leicester (UK).

Yilmaz, I., 2009. Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: A case study from Kat landslides (Tokat–Turkey). Computers & Geosciences 35, 1125–1138.

Appendix 3: Lampert et al. In Review. An Empirical Study into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation.

Thomas A. Lampert, Stumpf, A., Gançarski, P., In Review. An Empirical Study into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation. International Journal of Computer Vision, 23 pages.

Noname manuscript No.
(will be inserted by the editor)

An Empirical Study into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation

Thomas A. Lampert · André Stumpf · Pierre Gançarski

the date of receipt and acceptance should be inserted later

Abstract Although agreement between annotators has been studied in the past from a statistical viewpoint, little work has attempted to quantify the extent to which this phenomenon affects the evaluation of computer vision object detection algorithms. Many researchers utilise ground truth in experimentation and more often than not this ground truth is derived from one annotator's opinion. How does the difference in opinion affect an algorithm's evaluation?

Four examples of typical computer vision problems are chosen, and a methodology is applied to each to quantify the inter-annotator variance and to offer insight into the mechanisms behind agreement and the use of ground truth. It is found that when detecting linear objects annotator agreement, in terms of the number of pixels, is very low. The agreement in object position, linear or otherwise, can be partially explained through basic image properties. Automatic object detectors are compared to annotator agreement and it is found that there is a clear relationship between the two. Several methods for calculating ground truths from a number of annotations are applied and the resulting differences in the performance of the object detectors are quantified. It is found that the rank of a detector is highly dependent upon the method used to form the ground truth. It is also found that although the STAPLE and LSML ground truth estimation methods appear to represent the mean of the performance measured using the individual annotations, when there are few annotations, or there is a large variance in them, these estimates tend to degrade. Furthermore, one of the most commonly adopted annotation combination methods—consensus voting—accentuates more obvious features, which results in an overestimation of the algorithm's performance. Finally, it is concluded that in some datasets it may not be possible to state with any confidence that one algorithm outperforms another when evaluating upon one ground truth and a method for calculating confidence bounds is discussed.

Keywords Evaluation · Ranking · Performance · Feature Detection · Agreement · Annotation · Ground Truth · Gold Standard Ground Truth · Expert Agreement · ROC Analysis · Precision · Recall

T. Lampert
ICube, University of Strasbourg, France
Tel.: +33 (0)3 68 85 45 10
Fax: +33 (0)3 68 85 44 45
E-mail: tlampert@unisrta.fr

A. Stumpf
LIVE, University of Strasbourg, France

Pierre Gançarski
ICube, University of Strasbourg

1 Introduction

The evaluation of computer vision algorithms often requires ground truth data. The difficulty presented by this is that a gold standard ground truth can be costly to obtain (if possible at all). It is therefore commonly assumed that the opinion of one (or more) annotator(s) approximates this gold standard ground truth. Nevertheless, annotators rarely agree completely when giving their opinion and this disagreement can be characterised as bias, the tendency of an annotator to prefer one decision over another, and variance, the natural variation that one annotator will have to the next (or themselves at a later date) (Warfield et al, 2008). This poses a problem when evaluating computer vision algorithms: how does the difference in the annotator's opinion affect an algorithm's evaluation?

This work is intended to highlight the effects of variability in the ground truth (GT) on the design, training and evaluation of objects detectors. Thus it provides an empirical investigation into the effects of using different ground

truths when evaluating an object detector or segmentation algorithm.

This investigation focusses on four case studies, all of which embody a typical computer vision problem—object detection. These four case studies are: the segmentation of natural images (referred to as the Segmentation case study), the identification of fissures in aerial imagery (referred to as the Fissure case study), the identification of landslides in satellite imagery (referred to as the Landslide case study), and the identification of blood vessels in medical imagery (referred to as the Blood Vessel case study). The true ground truth of these data sets (which are referred to as the gold standard ground truth) can never be deduced from the imagery. For example in the fissure detection problem, this information cannot be known without full knowledge of the geophysical forces acting upon the terrain and similar limitations can be found in each of the datasets. This limitation is typical in many computer vision applications: medical imaging, remote sensing, and natural scene analysis, to name but a few. Furthermore, there exist many objects in these datasets that can cause false-positive and false-negative errors, making them ideal to study annotator and detector agreements.

Many statistical studies investigate the modelling of agreement and disagreement, particularly in ranking and classification problems, in which several objects are to be categorised. Nevertheless, the literature is lacking studies into annotator agreement in image segmentation and object detection. Those that do exist focus on developing methods for estimating the gold standard ground truth from a number of annotations (Biancardi and Reeves, 2009; Burl et al, 1994; Kauppi et al, 2009; Langerak et al, 2010; Li et al, 2011; Smyth et al, 1994; Warfield et al, 2004, 2008). These methods are rarely employed in real-world algorithm evaluation, where often experimentation is limited to one annotation, which is taken to be the gold standard. Some public datasets do, however, offer segmentations obtained from different annotators: The Berkeley Segmentation Dataset (Arbelaéz et al, 2011), the Digital Retinal Images for Vessel Extraction Database (DRIVE), the Lung Image Database Consortium image collection (Armato et al, 2011), and the STructured Analysis of the Retina (STARE) project database (Hoover et al, 2000), for example.

Through performance evaluation ground truth data often influences an algorithm's design, the choice of an algorithm's parameter values, and also influences the structure of the training data itself. It is therefore important to quantify the effect that different ground truths have on the algorithm's reported performance. Relying on the opinion of one annotator allows for the learning of that annotator's bias in the problem, but it does not necessarily result in a model that is effective at locating the true detections. This problem can, of course, be circumvented when the images are captured in tightly controlled conditions or are synthetically generated from a model (Lampert and O'Keefe, 2011), in these cases a gold standard ground truth is relatively trivial to calculate. In remote sensing and medical imaging problems, and those concerning natural images, however, this is not the case.

The following assumptions regarding the problem's characteristics are implicitly made within this study. In computer vision problems, true positive locations tend to be spatially correlated (objects tend not to be lone pixels but a number of pixels within close proximity to each other) and are also correlated with some image properties (those that tend to be used as features for classification algorithms). Furthermore, it is assumed that the annotators are not malicious in producing their annotation, are not producing annotations at random, and are not simply following low-level cues in the image but are instead able to draw upon some higher-level knowledge that allows them to distinguish between objects that belong to the negative class but share the same low-level image properties as those objects that constitute the positive class.

Therefore the objectives of this study are:

- to highlight the fact that the evaluation of a detector may be biased when only one annotation is used;
- to provide a general comparison between algorithms designed to infer the gold standard ground truth;
- and to investigate the effect that different ground truths have upon a detector's reported performance.

This paper is organised as follows. The following section presents a review of the most relevant work found in the literature. Section 3 describes the experimental methodology that will be followed in each of the case studies included in this paper. This methodology is followed and applied to several different datasets in Sections 4.2, 4.3, 4.1, and 4.4, and a discussion of these results is presented in Section 5. Finally the conclusions of the study are presented in Section 6.

2 Related Work

In a classic study Smyth et al (1994) analyse the uncertainty of an annotator's judgement in marking volcanoes in synthetic aperture radar images taken from the Magellan spacecraft as it orbited Venus. The authors assume a stochastic labelling process, to account for intra-annotator variability, and outline the probabilistic free-response ROC analysis that integrates the uncertainty of an annotator's judgement directly into the performance measure.

As previously mentioned, there exists a number of methods for estimating the gold-standard ground truth from two or more annotations. There also exists a body of work from the medical domain in which practitioners manually segment anatomical scans, which are subsequently warped to match novel scans in order to estimate their segmentations

(termed multi-atlas segmentation). Although this isn't strictly estimating gold-standard ground truth, the methods for combining multiple annotations are relevant and therefore they are included in the following review (those citations originating from this domain have squared brackets). Kauppi et al (2009) take GTs as the intersection (consensus), fixed size neighbourhoods of the points marked by each annotator, and a combination of the two. The authors conclude that the intersection method is preferential as the highest detector performance is achieved using it. Numerous weighted extensions to the voting framework have been proposed based upon global [Sabuncu et al, 2010], local [Artaechevarria et al, 2009; Isgum et al, 2009; Sabuncu et al, 2010], semi-local [Sabuncu et al, 2010; Wang et al, 2013], and non-local [Coupé et al, 2011] information.

Probably the most popular gold-standard ground truth estimation method originating form the medical domain is proposed by Warfield et al (2004), named simultaneous truth and performance level estimation (STAPLE) in which the annotator performances (sensitivity and specificity) and the gold-standard ground truth are simultaneously estimated within a maximum-likelihood setting; the optimisation being solved using expectation-maximisation (a variant for handling continuous labels has been proposed by Warfield et al (2008) and Xing et al (2011)). The same authors also propose an approach in which the bias and variance of each annotator is estimated instead of the performance measure (Warfield et al, 2008) and another variant that account for instabilities in the annotator performance measures (Commowick and Warfield, 2010). Much subsequent work has concentrated on the STAPLE algorithm: removing its assumption that annotator performances are constant throughout the data [Asman and Landman, 2011a] (Asman and Landman, 2012a; Commowick et al, 2012), and COLLATE (Asman and Landman, 2011b), which accounts for spatial variability in task difficulty. Landman et al (2010) point out that in research and clinical environments it is not often possible to obtain multiple annotations made over the whole dataset. Extensions to handle multiple partial but overlapping annotations have therefore been proposed (Commowick and Warfield, 2010; Landman et al, 2010, 2013).

Kamarainen et al (2012) propose a simpler alternative to STAPLE by maximising the mutual agreement of annotator ratings. This approach avoids the use of priors, and does not introduce areas that did not appear in the original annotations. Langerak et al (2010) argue, however, that STAPLE fails when annotator uncertainty varies considerably due to the fact that the STAPLE algorithm combines all of the annotators' labellings. Instead they propose the selective and iterative method for performance level estimation (SIMPLE) algorithm in which only labels that are deemed reliable are taken into account. Li et al (2011) propose a probabilistic approach that uses level sets in which the likelihood function is inspired by the STAPLE algorithm (LSML). To overcome the susceptibility of the STAPLE algorithm to strongly diverging annotations, however, they accept that the contribution of an annotator's judgement should be dependent upon their performance but differently to STAPLE they measure the amount of detail in an annotator's marking and add a constraint to the energy function that imposes a prior model on the shape of the outcome, thus forming the LSMLP algorithm. Biancardi and Reeves (2009) state that the STAPLE algorithm (even with the Markov random field extension) and simple voting strategies assume that the pixels are spatially independent. A novel voting procedure is introduced to overcome this. It is preceded by a distance transformation that attributes positive values to the inside boundary of a GT segmentation, which increase towards its centre, and decreases negatively outside the segment border; and thus the truth estimate from self distances (TESD) algorithm is introduced (Biancardi and Reeves, 2009).

A new direction that has recently gained interest is to combine the information derived from the manual annotations with that derived from the image to imply the location of objects-of-interest. Yang and Choe (2011) follow this path and propose a method that incorporates the warping error to preserve topological disagreements between the estimated gold-standard ground truth and the annotations. A number of extensions to the STAPLE algorithm have also been proposed [Asman and Landman, 2012b, 2013; Liu et al, 2013] which incorporate the image's intensity values, as well as the performance of multiple experts, to transfer the labelling of one image onto that of another. Moreover, Asman and Landman [2012c] propose to combine a locally weighted voting strategy with information derived form the image's intensity.

The Berkeley segmentation dataset contains five-hundred images, each having five GTs. The authors include the level of annotator agreement within their evaluations (Arbelaéz et al, 2011), which provides a valuable reference when interpreting the results. Using the earlier Berkeley 300 database, (Martin et al, 2001) present a statistical analysis of the variation observed within the annotations (Martin et al, 2001). They notice that independent annotators tend to be consistent, with low inter-annotator error (the same pixel tends to be included in the same region by different annotators). Although it was also observed that the number of segments in the same image identified by different annotators can vary by a factor of ten.

Finally, a novel branch of supervised machine learning that implicitly exploits multiple annotations has come into focus. Either using many annotators or by using increasingly popular crowd sourcing systems such as Amazon's Mechanical Turk, which presents its own problems (Raykar et al, 2009).

3 Methodology

The methodological evaluation will be centred around four aspects: Annotator Agreement; Annotator Analysis; Annotator Agreement and Detector Performance; and Ground Truths and Reported Detector Performance. Scripts to recreate the results presented henceforth are available on-line[1].

3.1 Data

The data used in each of the case studies can be modelled as an image, $I : \{0, 1, \ldots, X-1\} \times \{0, 1, \ldots, Y-1\} \mapsto \mathbb{R}$ where X is the image's width and Y its height.

For each study, N annotators have provided manual markings containing the locations of the objects that are of importance to the study. The case studies are binary detection problems therefore each annotator marks each image pixel to indicate the presence of the object under study (therefore each annotation has the value one where the annotator perceives the object to exist and zero otherwise). The result of this is N binary maps describing the location of the objects according to each annotator. As such, each annotator's output is modelled as a function $M_n : \{0, 1, \ldots, X-1\} \times \{0, 1, \ldots, Y-1\} \mapsto \{0, 1\}$, where 0 and 1 represent the absence and presence of the object respectively and $n = 1, \ldots, N$.

3.2 Annotator Agreement

The first stage of analysis is intended to test the level of agreement between the annotators in each case study, and to expose the image properties that promote this agreement.

Smyth (1996) presents a method for calculating the lower bound on error that can occur in a set of annotations relative to the (unknown) gold-standard ground-truth. This bound is defined to be

$$\bar{e} \geq \frac{1}{XYN} \sum_{y=0}^{Y-1} \sum_{x=0}^{X-1} N - \max\{A(x,y), N - A(x,y)\} \quad (1)$$

where $A(x, y)$ is the number of annotators that labelled pixel (x, y) as containing the object-of-interest, as defined in Equation (2). The minimum of Equation (1) is reached when all annotators agree and the maximum (which can only ever reach 0.5) when there is maximum disagreement—when the decision is evenly split—it is therefore a measure closely related to the entropy of the annotators' agreement. The minimum acceptable value quoted by the author is 10% and as such this measure provides a method to validate the quality of the experimental data used within each case study.

[1] https://sites.google.com/site/tomalampert

Also to this end, the per-pixel annotator agreement is calculated. The agreement is simply the number of annotators that have marked each pixel, such that

$$A(x, y) = \sum_{n=1}^{N} M_n(x, y), \quad (2)$$

and the agreement as a function of the number of annotators, $1 \leq n \leq N$, is calculated such that

$$\hat{A}(n) = \frac{1}{|C|} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} \chi_B(x, y) \quad (3)$$

where $B = \{(x, y) \mid A(x, y) \geq n\}$, χ_B is the indicator function, and $C = \{(x, y) \mid A(x, y) > 0\}$.

These functions allow for the testing of significant correlations between annotator agreement and different properties of the image—a means to uncover at least part of the reason behind the variance of agreement. Each of the datasets present different features but where applicable the following features will be tested: intensity, contrast, and each of the colour channels. The Pearson's r correlation coefficient will be used and since the sample size for the analysis is extremely large it will be tested for significance to 99% confidence.

In the case that the image is colour, intensity is calculated such that $I(x, y) = 0.2989 \cdot R(x, y) + 0.5870 \cdot G(x, y) + 0.1140 \cdot B(x, y)$. Image contrast in a colour image is calculated using the Michelson contrast measure within a 3×3 local neighbourhood such that

$$c(x, y) = \frac{\max_{(i,j) \in W_{xy}} L(i, j) - \min_{(i,j) \in W_{xy}} L(i, j)}{\max_{(i,j) \in W_{xy}} L(i, j) + \min_{(i,j) \in W_{xy}} L(i, j)} \quad (4)$$

where $L(i, j)$ is the image's tone component, obtained by converting the colour image into the CIELAB colour space, and W_{xy} is the set of co-ordinates that define the neighbourhood of $L(x, y)$. Image contrast in a grey scale image is calculated as above but $L(x, y) = I(x, y)$. As contrast is taken within a local neighbourhood, to make a fair comparison the maximum agreement is taken within the same neighbourhood when their correlation is calculated.

3.3 Annotator Analysis

A number of the gold-standard ground-truth estimation methods evaluated in this research weight annotations based upon the 'performance' of each annotator. This is based upon the assumption that some annotators may produce more accurate annotations when compared to others, and that the more reliable annotators can be identified through inter-annotator comparisons.

To examine the inter-annotator variability, cluster analysis using the pairwise F_1-score between the annotator markings is conducted. The F_1-score (He and Garcia, 2009), calculated between participants i and j, is defined as

$$F_{ij} = 2\frac{p_{ij}r_{ij}}{p_{ij} + r_{ij}}, \quad (5)$$

and this quantity is therefore the harmonic mean of precision (p_{ij}) and recall (r_{ij}). Note that the F_1-score is robust in the presence of class-imbalance (in most of the case studies, the number of non-object pixels greatly outnumbers those indicating the presence of an object) since it does not take into account true-negative classifications (He and Garcia, 2009). Hierarchical clustering is performed using Ward's minimum variance implemented with the Lance-Williams dissimilarity update formula by linking pairs of annotations with the highest pair-wise F_1-score and repeating this until all annotations are included.

As a principled way of identifying outliers within the group of annotations, the mean F_1-score difference $(1 - F_{ij})$ between each annotator and all other annotators is calculated. Those that have a mean difference greater than the average plus one standard deviation are labelled as outliers.

Following the example of Saur et al (2010), and to highlight any individual differences between the annotators, each is compared to the group's consensus, calculated such that

$$\kappa(x, y) = \begin{cases} 1 \text{ if } \frac{1}{N}A(x, y) \geq \tau, \\ 0 \text{ otherwise,} \end{cases} \quad (6)$$

where $\tau = 0.5$, by calculating the Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV) and Cohen's kappa coefficient. This will allow for specific tendencies of the outliers to be identified.

3.4 Annotator Agreement and Detector Performance

After analysing the properties of agreement and the annotators, it follows to investigate the relationship between annotator agreement and detector performance. To this end four detectors are selected from the case study domains and applied to the object detection problem at hand (every effort was made to select the best performing detectors within each domain). Each of these detectors is evaluated using ground truths calculated at increasing levels of agreement according to Eq. (6), $\tau = 1/N, 1/(N-1), \ldots, 1$.

It is common to measure detector performance through ROC curve analysis, however, recent literature points out that this may overestimate performance when applied to highly skewed datasets and therefore precision-recall (P-R) curves are preferable (Davis and Goadrich, 2006; He and Garcia, 2009). Nevertheless, precision is sensitive to the ratio of positive to negative instances in the dataset, $\phi = N_p/N_n$. To overcome this Flach (2003) proposes to analytically vary the class skew in the precision measure and Lampert and Gançarski (submitted) to integrate this added dimension, thus forming a $\bar{\text{P}}$-R, curve. This allows $\bar{\text{P}}$-R curves derived from GTs containing different class skews to be compared, i.e. GTs derived from different levels of agreement, and for a fair representation of detector performance in problems in which the class skew is a priori unknown. This measure is defined such that

$$\bar{P}(\theta) = \frac{1}{\pi'_2 - \pi'_1} \int_{\pi'_1}^{\pi'_2} \frac{\pi' \text{TP}(\theta)}{\pi' \text{TP}(\theta) + (1 - \pi')\phi \text{FP}(\theta)} \, d\pi' \quad (7)$$

where θ is the threshold on the detector's output, $\text{TP}(\theta)$ and $\text{FP}(\theta)$ are the number of true positive and false positive detections, and $\phi = N_p/N_n$ is the ratio of positive to negative instances in the dataset. Interpolation between $\bar{\text{P}}$-R points (Lampert and Gançarski, submitted) allows accurate area under curve (AUC) measures to be taken.

To assess the relationship between annotator agreement and detector output two correlation coefficients will be measured (to 99% confidence). The first being the correlation calculated within locations identified as objects by any annotator (CCO) and the second in the whole image (CCI). The first of these highlights the relationship between the detector output and annotator agreement in positive locations of the image. The second includes any false positive detections that the detector may make, and therefore the absolute value of these correlations in addition to the difference between them indicate how reliable the detector is.

3.5 Ground Truths and Reported Detector Performance

The final question that this research intends to investigate is: by how much is the reported performance of an algorithm affected by using different ground truths?

To this end several GTs are calculated according to Eq. (6): the combined annotations where $\tau = 1/N$, i.e. objects of interest that any annotator marked (Any-GT); the consensus of half of the annotators, or majority vote, in which $\tau = 0.5$ (0.5-GT); and the consensus of three-quarters of the annotators, where $\tau = 0.75$ (0.75-GT). Also included are gold standard GT estimations calculated using STAPLE (Warfield et al, 2004) (without assigning consensus votes (Commowick et al, 2012)), SIMPLE (Langerak et al, 2010), and LSML (Li et al, 2011) (using the 50% agreement as an initial estimate and 1000 iterations). Furthermore, an additional GT is determined by excluding those outliers identified in Section 3.3 and then combining the remaining according to Eq. 6 using $\tau = 0.5$ (Excl-0.5-GT).

Two forms of evaluation are investigated. The first being the relative detector ranking, ranked according to the area under the $\bar{\text{P}}$-R curve. And the second being the variability observed in the absolute value of the the $\bar{\text{P}}$-R curves.

Fig. 1 An example of the natural images used within this case study.

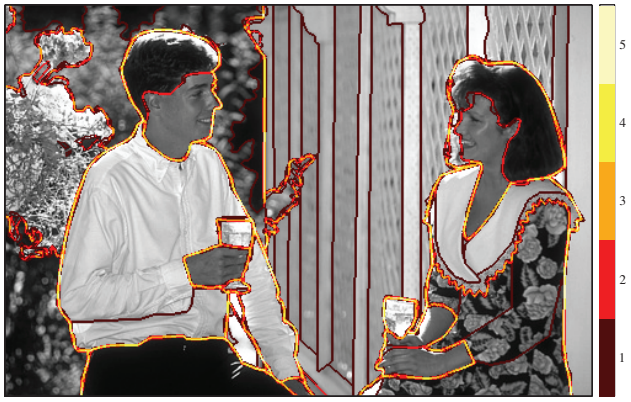


Fig. 2 Annotator agreement of segmentation boundary locations.

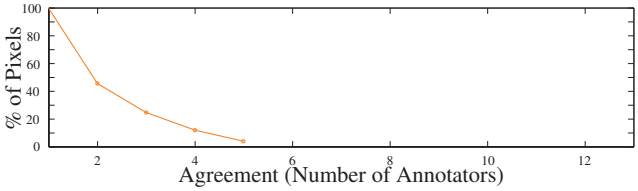


Fig. 3 Percentage of agreement (in pixels) as a function of the number of annotators.

Table 1 Pearson's r correlation coefficients between image features and agreement.

Feature	r	p
Intensity	0.0017	0.1403
Contrast	**0.3245**	0.0000
Red	−0.0017	0.1293
Green	0.0026	0.0249
Blue	0.0326	0.0000

4 Experimental Results and Analyses

In this section is presented the results of applying the methodology to each of the case studies included in this investigation.

4.1 Segmentation Case Study

We start the analysis using the standard dataset for evaluating segmentation algorithms.

4.1.1 Data

The Berkeley 300 (colour) dataset is used herein, however, as numerous annotators were used to annotate the images the whole dataset cannot be used. Instead, the largest subset of images for which the same annotators performed the segmentation was found (the user IDs are not available for the Berkeley 500 dataset, which is why the older Berkeley 300 dataset was used). This subset was determined to be the images: 65033.jpg, 157055.jpg (presented in Figure 1), 385039.jpg, 368016.jpg, and 105019.jpg. For which the same five annotators (user IDs 1123, 1105, 1109, 1115 and 1121) performed the manual segmentations. Each image was concatenated to form one large image, in which $X = 1595$ and $Y = 479$, and the same process was used to form one GT for each of the annotators.

4.1.2 Annotator Agreement

The lower-bound on error according to Smyth's calculation is low at $\bar{e} \geq 2.6611\%$. A pictorial example of the agreement upon segmentation boundaries is presented in Figure 2. In which it is obvious that there is a high level of agreement on segmentation boundaries that correspond to important aspects of the image (the people for example), however, there are other large segmentation boundaries that are only marked by one annotator (those on the trellis forming the right half of the image's background, for example). The level of agreement as a function of the number of experts is presented in Figure 3. It should be noted that annotator agreement attributed to the Berkeley dataset falls in an exponential manner and the level of agreement between all of the annotators is very low.

In the segmentation problem the objects to be detected (the segments) do not fit into a definable feature set, they are instead defined as the boundary between two objects, and the features of these objects are also not strictly defined. This is reflected in the statistical study presented in Table 1 in which it is found that agreement is correlated with the image's contrast and not intensity nor colour profile.

4.1.3 Annotator Analysis

The relatively low levels of agreement are also reflected in the pairwise differences in F_1-scores upon which the dendrogram in Figure 4 is based. The differences are relatively

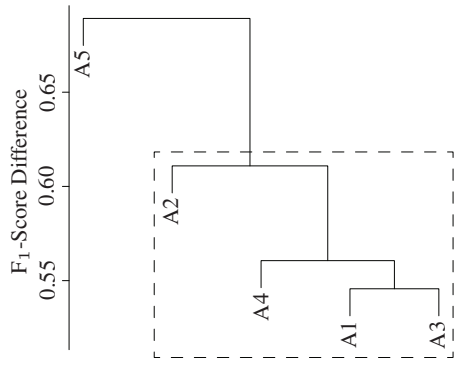


Fig. 4 Dendrogram describing the F_1-score difference relationships between each annotator's marking. The dashed box depicts the inliers (see Section 4.1.3).

Table 2 Sensitivity (Sens.), specificity (Spec.), positive predictive value (PPV), negative predictive value (NPV) and Cohen's kappa coefficient of the participants when compared to the consensus (rounded to four decimal places).

	Sens.	Spec.	PPV	NPV	kappa
A1	0.7694	0.9845	0.5634	0.9939	0.6399
A2	0.6373	0.9886	0.5921	0.9905	0.6034
A3	0.7853	0.9785	0.4882	0.9943	0.5892
A4	0.7309	0.9822	0.5166	0.9929	0.5933
A5	0.7275	**0.9649**	**0.3509**	0.9927	**0.4548**

high and range from 0.545 to 0.68. One outlier is identified, A5 (the mean F_1-score difference was found to be 0.6016, its standard deviation 0.0280 and A5 resulted in a difference of 0.6454), who also results in the lowest specificity, positive predictive value, and kappa coefficient as demonstrated in Table 2. The variance in the annotations are again underlined by the lowest specificities observed in all of the case studies.

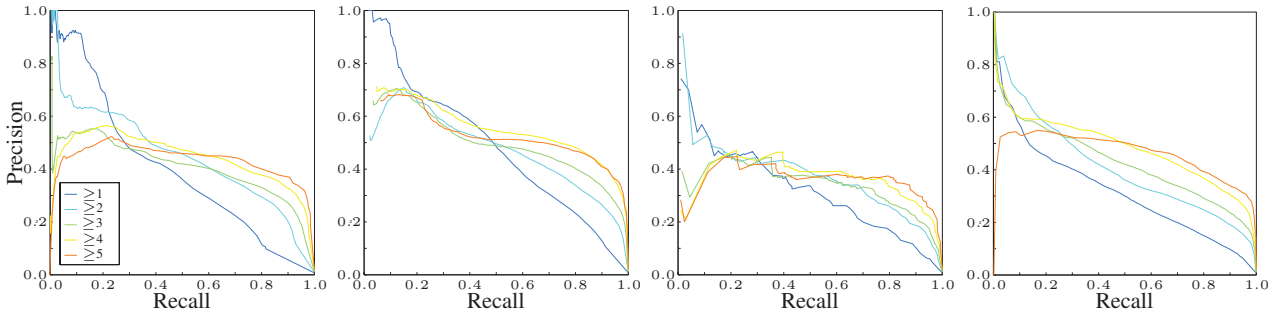
4.1.4 Agreement and Detector Performance

The top four performing segmentation algorithms that are listed on the Berkeley dataset web page[2] were selected to form part of this case study. These are: REN (Ren and Bo, 2012), gPb-ucm (UCM) (Arbelaéz et al, 2011), Global Probability of Boundary (GP) (Maire et al, 2008), and XREN (Ren, 2008). The integration limits of the $\bar{P}$-R curves were $\pi_1' = 0.0000$ and $\pi_2' = 0.0428$, which were found to be $\pi_1' = \mu - 3\sigma$ and $\pi_2' = \mu + 3\sigma$ where μ is the mean of the skew found within the Berkeley dataset and σ its standard deviation (Lampert and Gançarski, submitted). This case study deviates slightly from the prescribed methodology as it is common when evaluating segmentation algorithms to loosen the definition of true-positive detections to account for deviations in the location of detected boundaries, as discussed by Martin et al (2004). True-positive detections are accumulated if a detection is within a certain distance of a ground truth boundary (or multiple ground truths). In these experiments the allowed distance is taken to be the default found with the Berkeley benchmark code—0.0075 times the length of the image's diagonal (this matching is performed in the original individual images, not the composite image, to avoid boundaries being matched between separate images). The images that are part of the training set (images 105019.jpg and 368016.jpg) are removed from this point forward.

One further modification to the methodology was made to better suite the definition of segmentation. In the following case studies the act of taking low agreement GTs (≥ 1, for example) resulted in the delineation of objects that have been marked by the specified number of annotators, however, in this problem the result of this threshold on agreement is a GT with segmentation boundaries that have multiple pixel widths (as annotators may agree upon the boundary's existence but not on its exact location). This causes an unfair penalty on the algorithm because a segmentation algorithm is designed to detect single pixel segmentation boundaries. Therefore, each GT is thinned prior to its use to reduce any boundaries that are more than one pixel wide to widths of a single pixel, and in doing so any individual, low agreement, markings that the annotators may have made are preserved.

Although the evaluation methodology has been modified, Figure 5 demonstrates that the same trends as will be found in the subsequent case studies are visible. The GTs used in these experiments are calculated according to Eq. (6) by setting $\tau = 1/N, 1/(N-1), \ldots, 1$, and thus each curve represents the performance of the detector in identifying objects with a certain minimum level of annotator agreement. It is observed that in the higher recall ranges, the performance of all the detectors increases in line with agreement in a predictable manner. As will also be seen in the landslide case study, however, the lower recall range produces a different picture and this phenomenon will be explored further in the aforementioned case study. The effects of the large variability in annotator opinion start to be noticed, the correlations, although significant, are much lower than in the following case studies, and the $\bar{P}$-R curves are clearly less predictable. Another factor that has a detrimental impact upon the correlation coefficients presented in Table 3 is that the detectors in this study tend to result in responses that deviate from the true location of the segment (as defined by the annotations). This, coupled with the fact that the detectors output detection boundaries that have a width of one pixel and not a smooth decrease in response with distance from the boundary, lowers the overall correlation values (and this is the reason for which the segmentation community has proposed the modification to the evaluation framework that has been followed in this case study).

[2] `http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/bench/html/algorithms.html`



(a) GP (b) REN (c) UCM (d) XREN

Fig. 5 P̄recision-recall curves describing the detectors' performances using different levels of agreement as the ground truth.

Table 3 Pearson's r correlation coefficients between detector outputs and the annotator agreement; CCO is calculated within the pixels marked as a segment by the annotators, and CCI the whole image. The p-values are all 0.0000 (to four decimal places).

Detector	CCO	CCI	CCI−CCO
UCM	**0.2686**	**0.3663**	+0.0977
GP	0.1603	0.2746	**+0.1143**
XREN	0.2633	0.3206	+0.0573
REN	0.2089	0.3119	+0.1030

Table 4 Detector rankings evaluated using each ground truth (measured by the area under the P̄-R curve). The GTs that result in these are: Ranking #1 — Berkeley evaluation framework (A1–A5); Ranking #2 — A4, A5, Any-GT, LSML-GT, STAPLE-GT; Ranking #3 — A1, A2, A3, 0.5-GT, 0.75-GT, Excl-0.5-GT, SIMPLE-GT.

Rank 1	Rank 2	Rank 3
REN	REN	REN
GP	GP	XREN
UCM	XREN	GP
XREN	UCM	UCM

4.1.5 Ground Truths and Reported Detector Performance

It appears from Figure 5 that the REN detector is the best performing detector in these images (which is in accordance with the overall Berkeley ranking). This is corroborated using the rankings of the detectors when evaluated using different ground truths, see Table 4, the REN detector is consistently at the top. Overall, however, three rankings are observed depending upon which GT is used for evaluation (or which evaluation strategy is used). The first being produced by the Berkeley framework (BF) (Martin et al, 2004) alone, which includes all five annotations and relaxes the TP condition by matching a detection to any positive marking that is within the allowed distance in any of the annotations. The second ranking is produced by six of the GTs and the third by seven, although the best performing detector remains the same in all three. To illustrate these ranks, the P̄-R curves for all four detectors evaluated using the BF, STAPLE-GT, and 0.75-GT are plotted in Figure 6, there are therefore representative ground truths from each ranking presented in Table 4.

Interestingly, the worst performing method, UCM, produces the highest correlations in Table 3 and the best performing produces the third highest correlations, however, as has been discussed in the previous subsection,this could be due to a peculiarity of the data (low annotator agreement on segment locations). Due to the discrepancy between the segmentation evaluation framework (with tolerance on the exact location of the boundary) and the computation of the correlation coefficients (at the exact pixel location) a concurrent interpretation must be regarded with care. As it will be shown in the following case studies, the absolute values and differences of CCO and CCI can be related to the performance of a detector in the P̄-R framework but the relationship is not always linear.

To finalise the current case-study the evaluation of the REN algorithm against a number of differently estimated GTs is presented. Figure 7 demonstrates that there is a large variation in the level of performance, which is to be expected as there is a large variation between each annotation (see Section 4.1.3). At the extreme of this variance is the evaluation methodology commonly used to evaluate segmentation algorithms using the Berkeley dataset. Because of the generous leniency that is given when calculating true positive detections over multiple annotations, the performance is vastly greater than when testing against any of the individual annotations alone.

The 0.75-GT, 0.5-GT and SIMPLE-GT overestimate performance (particularly in the higher Recall ranges) and Any-GT underestimates performance when compared to the remaining GTs. The STAPLE-GT and LSML-GT are also within the lower performance estimates in the upper recall range, but they seem to model the mean of the individual annotations in the lower recall ranges. It is assumed that this is a consequence of the large variance observed in the annotations. It was found that the Excl-0.5-GT and SIMPLE-GT ground truths were very similar as they are both derived the same principle (removing the outliers and then voting), and therefore the detector's performance relative to both are also similar.

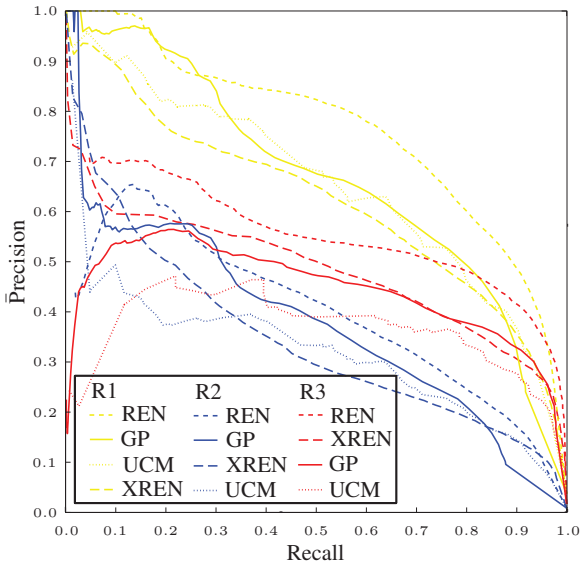


Fig. 6 $\bar{\text{P}}$recision-recall curves of all four detectors evaluated using the Berkeley Framework (A1–A5) giving Ranking #1 (R1), STAPLE-GT giving Ranking #2 (R2), and 0.75-GT giving Ranking #3 (R3).

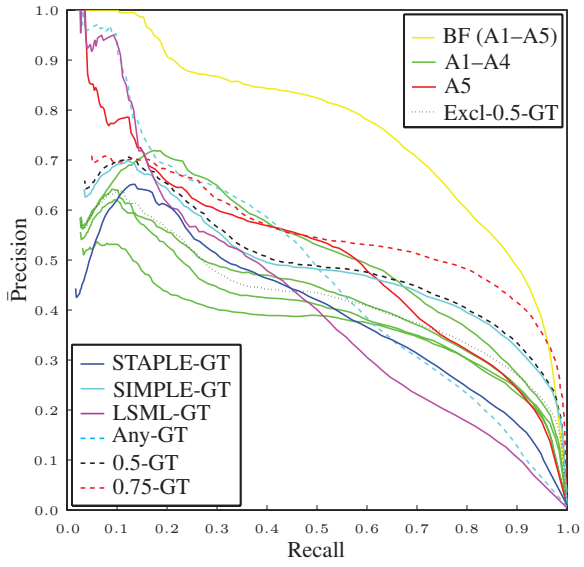


Fig. 7 $\bar{\text{P}}$recision-recall curves of the REN detector assuming differing ground truths.

4.2 Fissure Case Study

The following case study is concerned with the detection of fissures in remotely sensed images.

4.2.1 Data

The data is obtained from the Super-Sauze landslide in the Barcelonnette basin, southern French Alps, using an unmanned aerial vehicle to obtain high resolution images. Further information regarding this dataset is present in the literature (Niethammer et al, 2011). An area of interest, where $X = 1425$ and $Y = 906$, was extracted from the data and is presented in Figure 8. Very little colour information is present in this type of image and it was therefore converted to grey scale using the standard formula: $I(x, y) = 0.2989 \cdot R(x, y) + 0.5870 \cdot G(x, y) + 0.1140 \cdot B(x, y)$.

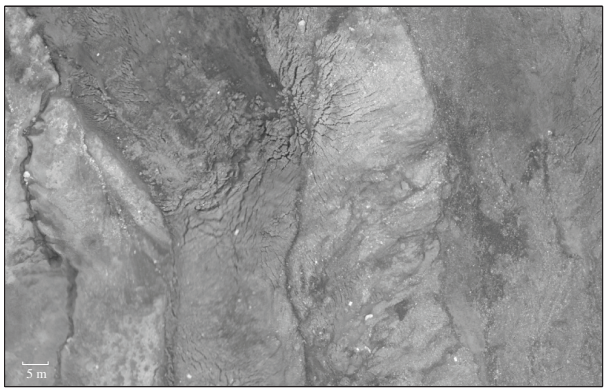


Fig. 8 The area of interest used within this case study.

Thirteen annotators ($N = 13$) were enlisted to manually mark the pixels in the (RGB) image that formed part of a fissure. Within this section, each of these participants will be referred to as A1–A13. The level of expertise ranged from expert geomorphologists familiar with the study site (2), non-experts familiar with fissure formation and/or detection (5), and contributors without any *a priori* knowledge (6). Prior to beginning the marking experiment, all the annotators were given a twenty minute presentation to introduce the basic physics of crack formation and the characteristics of the targeted fissures. The annotators then independently marked the pixels which they believed to form part of a fissure, taking as much time as they required to complete the task (this ranged from 2–3 h). The annotators were instructed to perform the marking on a level in which they could see individual pixels clearly, but were encouraged to zoom out to gain information related to the context of the area being marked.

4.2.2 Annotator Agreement

Smyth's lower error bound estimate is found to be $\bar{e} \geq 1.26\%$, i.e. the average error rate amongst the thirteen annotators. This value is well within the 10% limit that is recommended (Smyth, 1996). This value is also considerably lower than the error bound of 20% stated by Smyth in the study of agreement within labelling volcanoes in satellite images of Venus, in which the signal-to-noise ratio of the object is much lower than in this study.

The annotator agreement attributed to this case study is presented in Figure 9 and Figure 10. The first thing to notice is how little agreement exists between all annotators (a similar finding to that of the Segmentation case study)—out of all of the pixels that were marked as fissures only 0.6979%

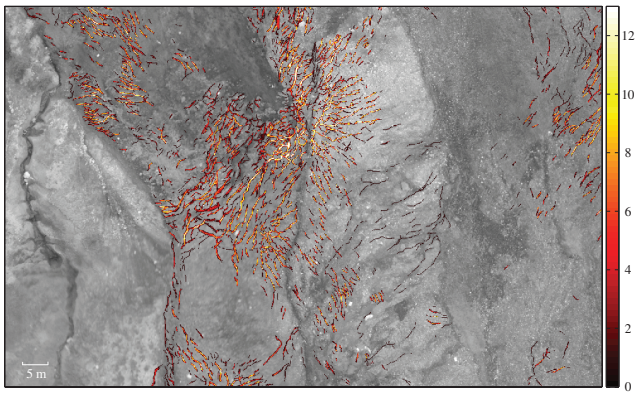


Fig. 9 Annotator agreement of fissure locations, the agreement of each pixel is calculated according to Eq. (2).

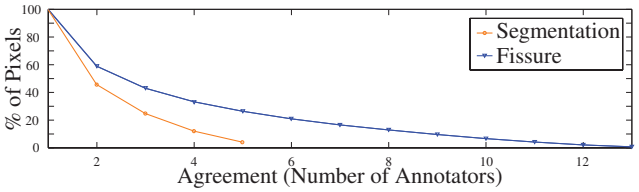


Fig. 10 Percentage of agreement (in pixels) as a function of the number of annotators.

Table 5 Pearson's r correlation coefficients between image features and agreement. The p-values are both 0.0000 (to four decimal places).

Feature	r
Intensity	-0.2245
Contrast	$\mathbf{0.4027}$

are agreed upon by all thirteen annotators. It is also worth noting that the level of agreement decreases exponentially as a function of the number of annotators. This is also common to the Segmentation dataset in which the objects of interest also form linear structures.

The correlation coefficients between agreement and a number of the image's features are presented in Table 5. There is a large and significant correlation between contrast and agreement, indicating that fissures on a lighter background are easier to see and are therefore marked by a greater number of annotators. A negative correlation is found between annotator agreement and image intensity as fissures are dark features within the image. These correlations both indicate that relative and absolute intensity values are important features in this problem, perhaps because darker fissures are easier to see and so attract more agreement. This is in contrast to the previous case study, in which the object features are not strictly defined and therefore only contrast resulted in a considerable correlation.

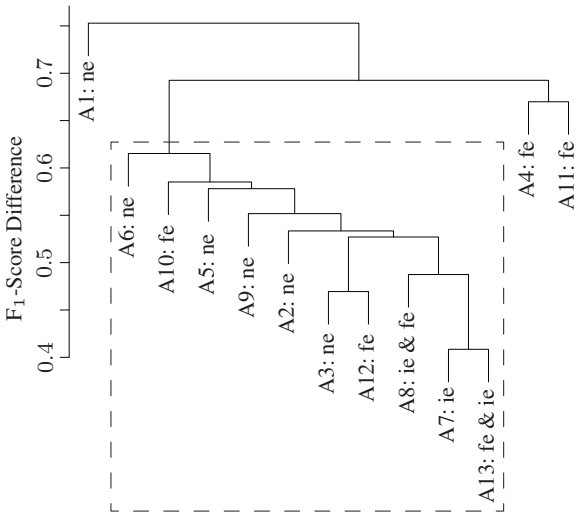


Fig. 11 Dendrogram describing the F_1-score difference relationships between each annotation. Key: ne — non-expert; ie — expert with previous experience of fissure mapping in imagery; and fe — expert with experience in the recognition of such fissures in the field. The dashed box depicts the inliers (see Section 4.2.3).

4.2.3 Annotator Analysis

A dendrogram describing the relationship between the annotators' pairwise F_1-scores is presented in Figure 11. It would be expected that more than one cluster emerges from the data, splitting the different experience levels; however, this isn't the case and annotators of varying levels of expertise are quite homogeneously mixed. This indicates that no group is overly biased in favour of one particular decision.

Annotators A1, A4, and A11 are identified as falling outside of one standard deviation of the mean F_1-score difference to all other annotators. These same annotators achieve considerably lower sensitivity when compared to the consensus. They also achieve lower kappa coefficients, and PPVs—indicating that, when compared to the consensus, these annotators fail to identify a majority of the fissures and/or produce more 'false negative' and 'false positive' detections. The mean F_1-score difference $(1-F_{ij})$ is found to be 0.5765 and the standard deviation 0.0459, these annotators fall outside this threshold having a mean difference of 0.6716, 0.6321, and 0.6287 (corresponding to A1, A4, and A11 respectively).

It is illustrated by these results that all of the annotators are reliable in detecting negative instances of fissures, indicated by high specificity and negative predictive values, due to the highly skewed nature of the problem in which negative instances constitute a high proportion of the data. Highlighting the difficulty and uncertainty in detecting positive instances however are low sensitivity and PPVs.

Table 6 Sensitivity (Sens.), specificity (Spec.), positive predictive value (PPV), negative predictive value (NPV) and Cohen's kappa coefficient of the participants when compared to the consensus (rounded to four decimal places).

	Sens.	Spec.	PPV	NPV	kappa
A1	**0.5595**	0.9847	**0.2893**	0.9950	**0.3722**
A2	0.7518	0.9911	0.4860	0.9972	0.5848
A3	0.7526	0.9945	0.6018	0.9972	0.6647
A4	**0.5705**	0.9906	**0.4032**	0.9952	**0.4656**
A5	0.6429	0.9938	0.5362	0.9960	0.5797
A6	0.6244	0.9926	0.4834	0.9958	0.5392
A7	0.9380	0.9866	0.4377	0.9993	0.5907
A8	0.7897	0.9906	0.4828	0.9976	0.5937
A9	0.6894	0.9926	0.5106	0.9965	0.5814
A10	0.6659	0.9925	0.4969	0.9963	0.5636
A11	**0.5799**	0.9899	**0.3905**	0.9953	**0.4596**
A12	0.7461	0.9937	0.5672	0.9972	0.6399
A13	0.8738	0.9836	0.3719	0.9986	0.5143

4.2.4 Annotator Agreement and Detector Performance

During this case study a number of detectors were selected and their ability to detect fissures in the area of interest was evaluated by calculating $\bar{\text{P}}$-R curves. The current state-of-the-art linear feature detectors were selected from the literature, namely:

- CrackTree (Zou et al, 2012);
- EDLines (Akinlar and Topal, 2011);
- a linear classifier trained using 2D Gabor wavelet ($\epsilon = 4$, $a = 2, 3, 4, 5$, and $k_0 = 3$) and inverted grey-scale features (2D GWLC) (Soares et al, 2006);
- LSD (von Gioi et al, 2010);
- Percolation (Yamaguchi and Hashimoto, 2010);
- grey scale thresholding;
- Gaussian filter matching (Stumpf et al, 2012), $\sigma = 1$ (Gauss);
- Top-Hat transform (4 pixel radius circular structuring element);
- and the Centre-Surround (C-S) transform (using a 3×3 pixel neighbourhood) (Vonikakis et al, 2008).

Where public source code was not available the respective authors kindly agreed to run the algorithm on the data and provide a number of outputs, calculated using a range of parameter values. Using publicly available code ensured that the implementations were true to the author's intentions and also allows for reproducibility of these results. As the 2D GWLC method is a supervised learning algorithm a random subset of the image, 569×362 pixels in size, was used as a training set (16% of the image), the GT was defined according to Eq. (6) using $\tau = 1/N$, and the training area was excluded from the test set. It should therefore be noted that the 2D GWLC results are derived using less data than the comparisons, however, they have low standard deviation and should therefore be comparable. Within this case study the $\bar{\text{P}}$-R integration limits are set to be $\pi'_1 = 0.1$ and $\pi'_2 = 0.5$ (from ten times as many negative as positive instances to a balanced dataset) to reflect the large range of skews that can be observed in a remote sensing application.

Table 7 Pearson's r correlation coefficients between detector outputs and annotator agreement; CCO is calculated within the pixels marked as a fissure by the annotators, and CCI the whole image. The p-values are all 0.0000 (to four decimal places).

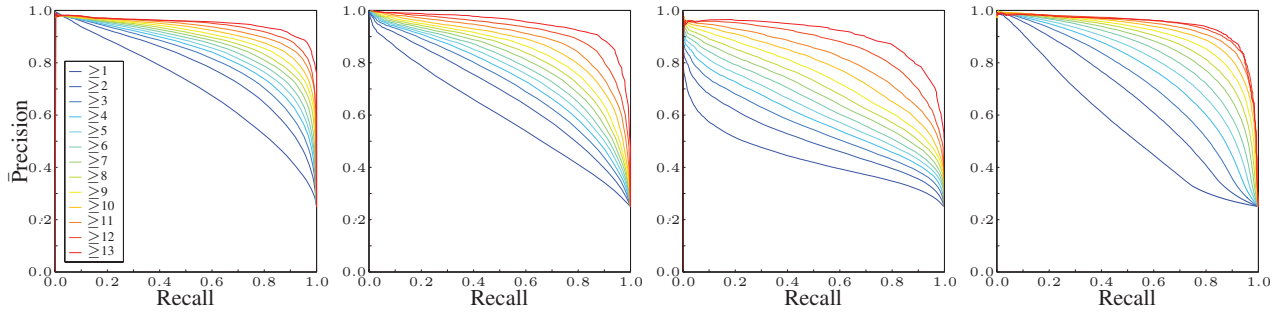
Detector	CCO	CCI	CCI−CCO
2D GWLC	0.5563	0.5166	**−0.0397**
Gauss	0.5293	0.4711	−0.0582
C-S	**0.6387**	**0.5259**	−0.1128
Top-Hat	0.5187	0.2780	−0.2407

Out of the evaluated detectors four clearly demonstrated superior performance over the others (2D GWLC, Top-Hat, C-S, and Gauss). The $\bar{\text{P}}$-R curves derived from these detectors are presented in Figure 12. A striking observation is that the performance of all the detectors increases in line with agreement in a predictable manner. Assuming that the more agreed upon fissures are the most obvious, this result indicates that the detectors extract similar features to those that aid an annotator's decision. There is, however, a large difference between the detection rate of high and low agreement fissures—detection of the lower is not a trivial matter and the decision most likely needs to be augmented with high-level information that is not exploited by these detectors.

Regarding the correlation between detector output and annotator agreement, the C-S detector produces the strongest CCO and CCI correlations. It does have, however, one of the largest drops between the two, indicating that the detector has low sensitivity. The $\bar{\text{P}}$-R curves give greater depth to this finding: the low sensitivity dominates at the low agreement decision boundary but the detector results in the highest performance at the high agreement decision boundaries. The 2D GWLC detector's output results in the second highest correlation with agreement over the whole image, and also exhibits the lowest drop in correlation between the two tests, indicating (relatively) low false positive rates. A large drop in correlation, along with a low absolute correlation, is observed with the top-hat detector, and indeed in Figure 12 the curves are skewed towards lower precision values. Overall, these correlation coefficients indicate that a detector's performance increases as the agreement upon the object increases and those detectors resulting in the lowest drop in correlation result in a tighter spread of $\bar{\text{P}}$-R curves.

4.2.5 Ground Truths and Reported Detector Performance

The rankings of the detectors' performance (measured as AUC) when evaluated using different GTs was determined, and three rankings emerged, as described in Table 8. These rankings the results of the correlation analyses, the 2D GWLC



(a) 2D GWLC (b) Single Scale Gaussian (c) Top-Hat (d) Centre-Surround

Fig. 12 P̄recision-recall curves describing the detectors' performances using different levels of agreement as the ground truth.

Table 8 Rankings of detectors evaluated using each ground truth (measured by the area under the P̄-R curve). The GTs that result in these ranks are: Ranking #1 — A2, A4, A6, A11, Any-GT, LSML-GT, STAPLE-GT, Excl-0.5-GT; Ranking #2 — A1, A3, A5, A7–A10, A12, A13, 0.5-GT, SIMPLE-GT; Ranking #3 — 0.75-GT.

Rank 1	Rank 2	Rank 3
2D GWLC	2D GWLC	C-S
Gauss	C-S	2D GWLC
C-S	Gauss	Gauss
Top-Hat	Top-Hat	Top-Hat

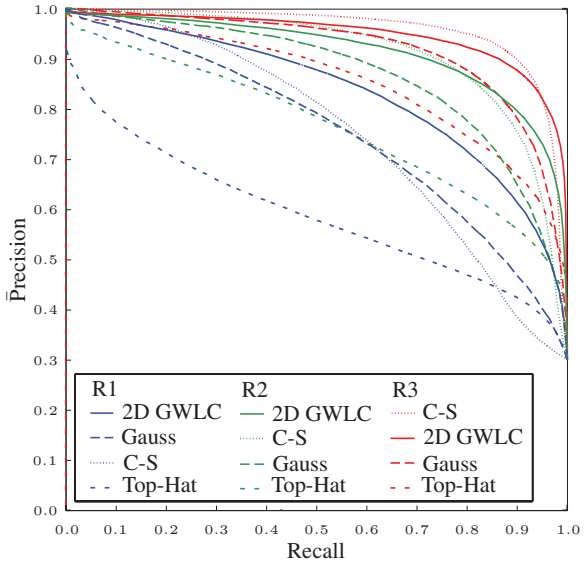


Fig. 13 P̄recision-recall curves of all four detectors evaluated using the STAPLE-GT giving Ranking #1 (R1), 0.5-GT giving Ranking #2 (R2), and 0.75-GT giving Ranking #3 (R3).

detector is consistently ranked first (by 19 of the GTs) and the top-hat detector last. These correspond to the highest and lowest drops in correlation observed in the previous section, see Table 7. Furthermore, a majority of the individual annotations give the same ranking as the SIMPLE-GT, and 0.5-GT ground truths, however, when the 0.75-GT, Any-GT, STAPLE-GT, and LSML-GT ground truths are under consideration, the ranking is perturbed. Therefore, the method of calculating the GT influences the detectors' ranking. More importantly, the ranking derived using a 75% voting strategy is in disagreement with that obtained using all of the thirteen annotator judgements individually, and this appears to be in contradiction to what should be expected. To illustrate these ranks, the P̄-R curves for all four detectors evaluated using the STAPLE-GT, 0.5-GT, and 0.75-GT are plotted in Figure 13, each colour represents one of the rankings presented in Table 8 (to be concise, subsequent case studies only present the table of rankings).

The P̄-R curves of the 2D GWLC, obtained using each of the GTs, are presented in Figure 14. The effects of the voted GTs (0.5-GT, 0.75-GT and SIMPLE-GT) become evident; these P̄-R curves overestimate the performance of the detector (in comparison to the other GTs), and seem to act as generous estimations of the upper bound on the detector's performance derived from the individual annotations. The curves calculated using the GTs of the outlying annotators (Section 4.2.3) give relatively lower estimates of performance. Apart from these, the curves of all the annotators are tightly clustered, in which the Any-GT appears to act as a lower bound on the performance and those obtained using the STAPLE-GT and LSML-GT ground truths appear to approximately model the mean performance obtained using the remaining individual annotations. It should be noted, however, that the LSML technique is highly dependent upon the initial estimation.

It was found that when applied to this dataset the Excl-0.5-GT and SIMPLE-GT ground truths were essentially the same (99.84% of the pixels were identical) as they are both derived the same principle (removing the outliers and then voting), the detector's performances relative to both were therefore approximately equal and therefore only the SIMPLE-GT ground truth is included in this analysis.

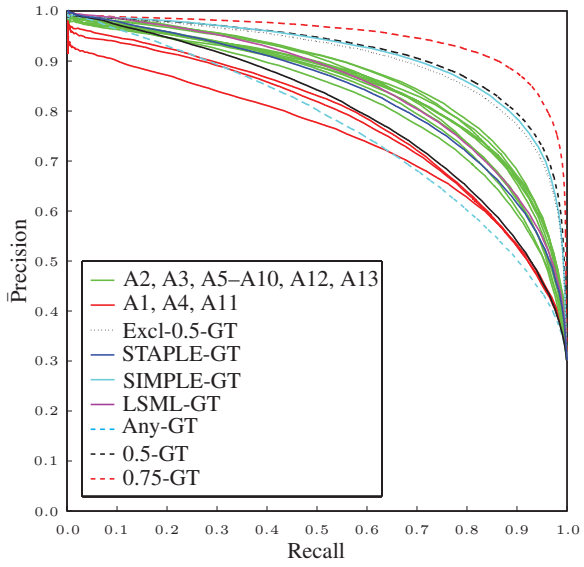


Fig. 14 $\bar{\text{P}}$recision-recall curves of the 2D GWLC detector assuming differing ground truths.

Fig. 15 The Geoeye-1 satellite image used within this case study (RGB channels).

4.3 Landslide Case Study

In this section we analyse another geographic remote sensing dataset, the goal of which is to identify landslides in satellite imagery.

4.3.1 Data

The dataset is derived from Geoeye-1 satellite images with four spectral bands (blue, green, red near infra-red) and a nominal ground resolution of 50 cm. The image presented in Figure 15 was captured at Nova Friburgo, Brazil shortly after a major landslide event in January 2011 and covers approximately 10 km^2 ($X = 5960$ and $Y = 5960$ pixels). A second image was recorded by the same satellite in May 2010 and depicts the ground conditions before the event.

Five annotators ($N = 5$), who were all familiar with landslide mapping in remote sensing images, were asked independently to mark the outlines of the landslide affected areas. For the image interpretation pre-event and the post-event satellite images were visualized using a natural color scheme on the RGB bands. Additionally a digital elevation model (ASTER-GDEM-VALIDATION-TEAM, 2011) with a nominal resolution of 30 m was available to the annotators to visualize the terrain characteristics. The annotators were free to zoom in and out as needed and no time limit was given.

4.3.2 Annotator Agreement

The overall error bound according to Smyth is $\bar{e} \geq 1.1012\%$, a similar level to that found in the Fissure dataset. The agreement of the annotators in the location of the landslides is presented in Figure 16. It can be noticed that the characteristics of the objects of interest differ to those of the previous problem. Here a landslide forms an two-dimensional area and previously the fissures typically form linear structures. The effect of this becomes clear in Figure 17, in which the level of agreement falls linearly with respect to the number of annotators. Indicating that disagreement typically occurs along the borders of the objects (indeed, if the outlines of the GT annotations are used agreement drops approximately exponentially).

Table 9 Pearson's r correlation coefficients between image features and agreement. The p-values are all 0.0000 (to four decimal places).

Feature	r
Intensity	0.0609
Contrast	0.0310
Near-IR	$\mathbf{-0.2766}$
Red	0.1841
Green	-0.0115
Blue	0.0200

The image features that produce the highest correlation with annotator agreement are the near infra-red and the red colour channels. This follows what would be expected as the near infra-red channel is typically used for vegetation identification, which is removed during a landslide. Furthermore, the soil in this area is typically reddish brown in appearance and therefore the red channel gives good distinction between landslide and non-landslide affected areas.

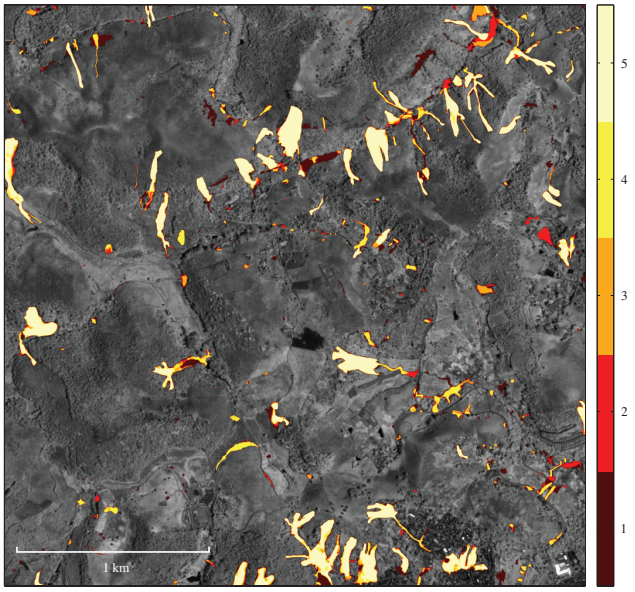


Fig. 16 Annotator agreement of landslide locations (overlaid on top of the near infra-red image band).

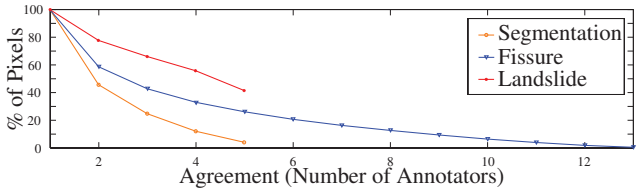


Fig. 17 Percentage of agreement (in pixels) as a function of the number of annotators.

4.3.3 Annotator Analysis

Each annotator was compared to the others by calculating pairwise F_1-scores and the resulting dendrogram is presented in Figure 18. In this case-study, each of the annotators were geographers familiar with the detection of landslides in remotely sensed imagery. This is reflected in the low inter F_1-score difference ($1 - F_{ij}$), which ranges from 0.14 to 0.28 (by comparison this range was approximately 0.4 to 0.75 in the previous case study). Nevertheless, by taking the mean difference between all other annotators and thresholding to within one standard deviation of this value, one outlier is identified and this is A2 (the mean difference was found to be 0.2044 and its standard deviation 0.0275, A2 resulted in a mean difference of 0.2438). This annotator also results in the lowest of the sensitivity and negative predictive values (when compared to the consensus opinion) presented in Table 10. On average, sensitivity, PPV and kappa are higher than in the previous case study, indicating that the features used for the identification of landslides are more clearly defined and understood by the annotators.

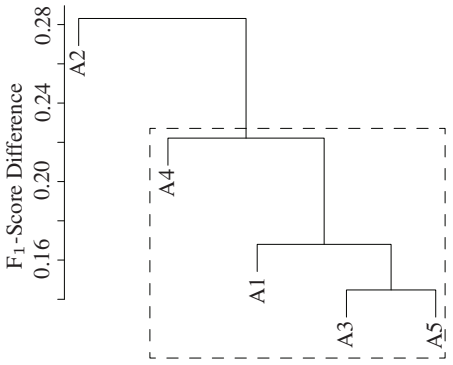


Fig. 18 Dendrogram describing the F_1-score difference relationships between each annotation. The dashed box depicts the inliers (see Section 4.3.3).

Table 10 Sensitivity (Sens.), specificity (Spec.), positive predictive value (PPV), negative predictive value (NPV) and Cohen's kappa coefficient of the participants when compared to the consensus (rounded to four decimal places).

	Sens.	Spec.	PPV	NPV	kappa
A1	0.9280	0.9942	0.8837	0.9966	0.9007
A2	**0.7499**	0.9972	0.9276	0.9883	**0.8222**
A3	0.8797	0.9978	0.9502	0.9943	0.9097
A4	0.9713	0.9837	0.7380	0.9986	0.8300
A5	0.9419	0.9945	0.8893	0.9972	0.9107

4.3.4 Agreement and Detector Performance

Implementations of four of the most popular classification algorithms were selected (due to their proven strength in real-world applications) and were applied to this problem. Namely the random forest (RF) (Liaw and Wiener, 2002), support-vector machine (SVM) (Meyer, 2009), k-nearest neighbours (KNN) (Li, 2012), and a neural network (ANN) (Venables and Ripley, 2002) algorithms. After fine scale image segmentation, 101 object features describing the spectral characteristics, texture, shape, topographic variables and neighbourhood contrast were extracted. The resulting dataset is available on-line[3] and a detailed description of the feature extraction methods are given in the literature (Stumpf et al, accepted).

Each classifier was trained upon samples from a randomly selected square subset covering 10% of the area of interest (each classifier was trained using the same subset). The number of trees in the RF were fixed at 500 and 10 variables were tested for the splits at each node. The SVM was employed with a radial basis kernel and parameters $C = 10$ and $\sigma = 0.004$ determined through an exhaustive grid search. The ANN was single layer network with a logistic activation function. An exhaustive grid search to optimize the weight decay function and the number of nodes resulted

[3] `http://eost.unistra.fr/recherche/ipgs/dgda/dgda-perso/andre-stumpf/data-and-code/`

Table 11 Pearson's r correlation coefficients between detector outputs and annotator agreement; CCO is calculated within the pixels marked as a landslide by the annotators, and CCI the whole image. The p-values are all 0.0000 (to four decimal places).

Detector	CCO	CCI	CCI−CCO
RF	0.6497	0.7829	+0.1332
KNN	0.6072	0.7551	+0.1479
SVM	**0.6503**	**0.7992**	**+0.1489**
ANN	0.6417	0.7565	+0.1148

in values of 0.1 and 7, respectively. Likewise, a grid search for the number of nearest neighbours resulted in $k = 23$ for the KNN algorithm. The parameter tuning was performed through bootstrap resampling of the training data and the area under the ROC curve as a performance measure. The $\bar{\text{P}}$-R integration limits were set to $\pi'_1 = 0.01$ and $\pi'_2 = 0.10$ to reflect typical ratios of affected and unaffected areas after large scale landslide triggering events (Malamud et al, 2004; Parker et al, 2011).

The $\bar{\text{P}}$-R curves resulting from each of these classifiers are presented in Figure 19. They largely follow the trend that was found in the previous case study—as agreement increases the performance of the classifier also increases. Except that, a similar trend to that observed in the Segmentation case study is observed in the lower recall range, in which the tendency for precision to increase with agreement is reversed. This phenomenon can be explained by analysing the correlations between annotator agreement and the detector outputs presented in Table 9. It should be noticed that in all of the cases CCI is higher than CCO. Indicating that the detector output strengths agree with annotator agreement within landslide locations and even more over the whole image. This implies that there is a relatively low FP detection rate, which at the lower recall ranges result in high precision values. As the agreement threshold is increased, however, the landslide areas that have increasingly stronger features form the GT, and these also have the highest detection strengths according to each detector. The high overall CCI correlations imply that as the lower agreement objects are removed from the GT they are instead being detected as false positive detections, thus reducing precision in the lower recall ranges as annotator agreement increases.

4.3.5 Ground Truths and Reported Detector Performance

In this case study only one ranking emerges: SVM, RF, ANN, and KNN. The SVM is ranked as the best detector and the KNN the worst, and these correspond to the highest and lowest CCO and CCI correlations observed in the previous section (see Table 7). There is a united consensus in the detector ranking—which was not observed in the previous case studies. This could be due to the lower inter-annotator variability observed in this problem, which allows the gold-standard

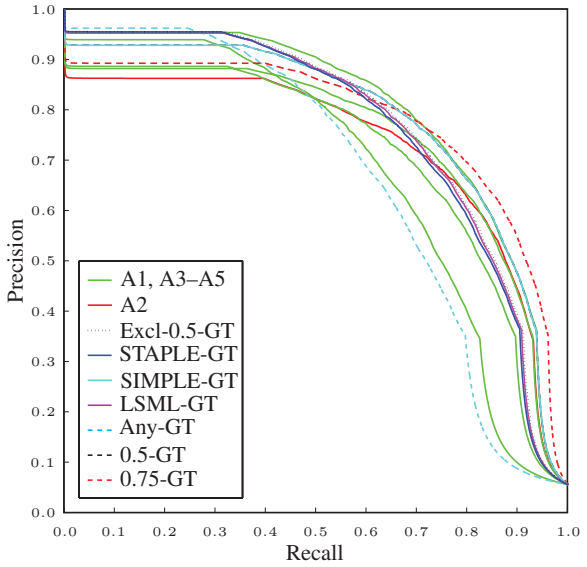


Fig. 20 $\bar{\text{P}}$recision-recall curves of the RF detector assuming differing ground truths (the 0.5-GT and SIMPLE-GT ground truths are identical).

GT estimation methods to better model all of the opinions. The performance of the classifiers in this study is largely similar and with greater annotator variance this could result in more perturbations to the rankings, however due to the low annotation variance this is not the case.

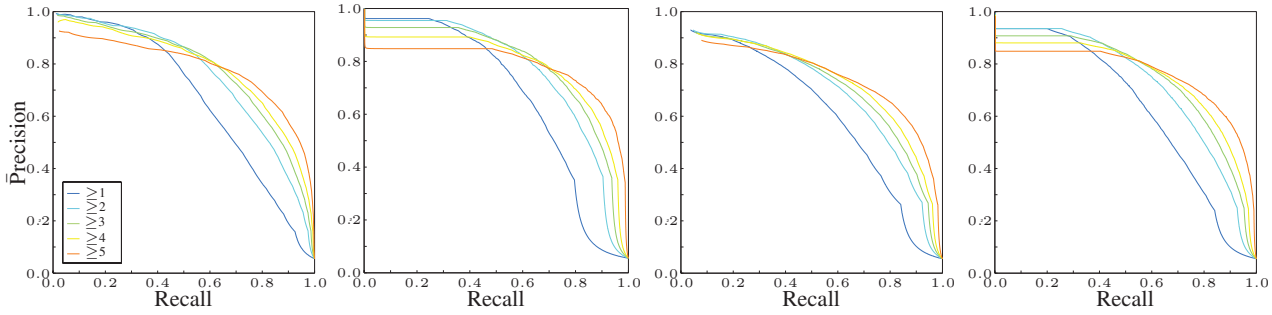
The lower degree of variance is also observed in the $\bar{\text{P}}$-R curves resulting from each GT—as presented in Figure 20—the variance in performance, although existing, is much lower than in the previous case studies. Similar trends can still be implied however. The GT derived from 75% agreement gives a higher estimate of the detector's performance compared to the other GT estimation methods. The 50% agreement GT and that calculated using SIMPLE produce identical performance curves. STAPLE and LSML tend to produce ground truths that model the performance within the bounds of that estimated by the individual annotations. And Any-GT gives a (relatively speaking) pessimistic outlook of the detector's performance.

4.4 Blood Vessel Case Study

In this section we move onto the third domain of study, the medical domain. The STructured Analysis of the Retina (STARE) dataset was created for the evaluation of retinal blood vessel detection algorithms.

4.4.1 Data

The dataset consists of twenty colour retinal images, which for the purposes of this study are treated as a single image in which $X = 2800$ and $Y = 3025$. An example image is presented in Figure 21. A mask was formed which delineates



(a) RF (b) SVM (c) KNN (d) ANN

Fig. 19 $\bar{\text{P}}$recision-recall curves describing the detectors' performances using different levels of agreement as the ground truth.

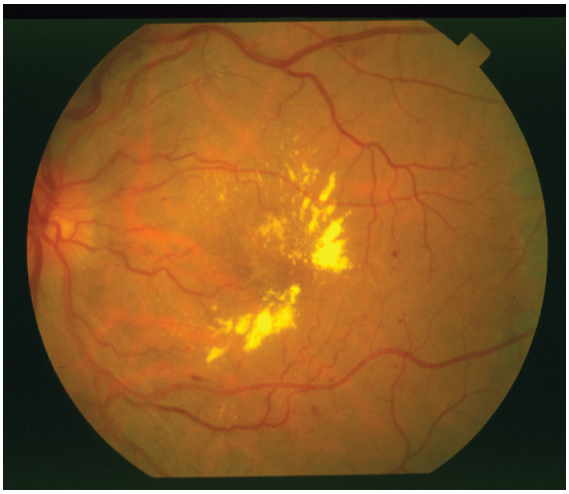


Fig. 21 An example of the retinal images used within this case study.

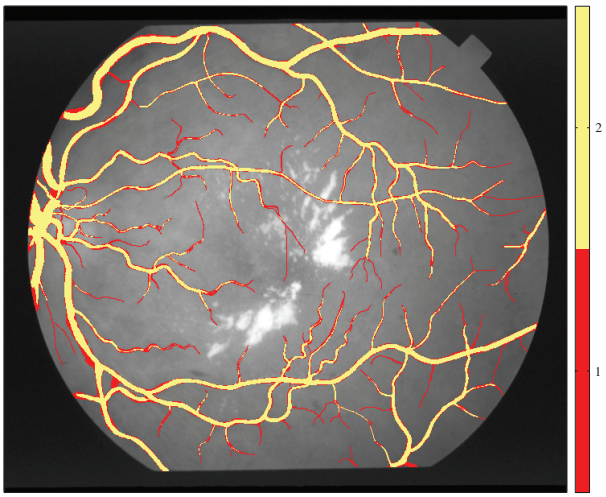


Fig. 22 Annotator agreement of blood vessel locations.

the pixels that fall outside the retina by thesholding the intensity of the red channel at a value of 40 (the black area in Figure 21) and these pixels were excluded from the experiments. The dataset contains two annotations which delineate the pixels that are part of the blood vessels.

4.4.2 Annotator Agreement

This dataset produces the highest lower error-bound according to Smyth, at $\bar{e} \geq 3.1123\%$. The reason for this will be discussed in Section 5, however, it is well below the 10% limit recommended by Smyth. An image depicting an example of the level of annotator agreement observed in this dataset is presented in Figure 22. The percentage of agreement is presented in Figure 23, with only two data points it is hard to make any general observations, however, the decrease in agreement seems to follow that observed in the fissure dataset, in which the objects of interest have similar characteristics (both being networks of linear structures).

There is a low correlation between all of the image features and annotator agreement, except in the green colour channel, in which a negative correlation is found indicating that blood vessels are suitably identified by the absence of the green component.

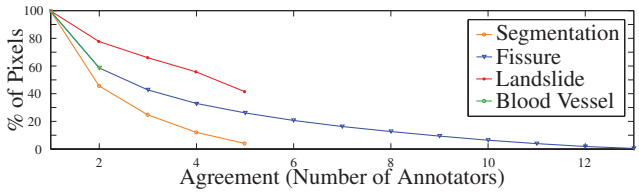


Fig. 23 Percentage of agreement (in pixels) as a function of the number of annotators.

Table 12 Pearson's r correlation coefficients between image features and agreement.

Feature	r	p
Intensity	−0.0861	0.0000
Contrast	−0.0026	0.0000
Red	0.0050	0.0000
Green	**−0.1495**	0.0000
Blue	−0.0007	0.0087

Table 13 Sensitivity (Sens.), specificity (Spec.), positive predictive value (PPV), negative predictive value (NPV) and Cohen's kappa coefficient of the participants when compared to the consensus (rounded to four decimal places).

	Sens.	Spec.	PPV	NPV	kappa
A1	0.6536	1.0000	1.0000	0.9417	0.4956
A2	0.9358	1.0000	1.0000	0.9887	0.5702

Table 14 Pearson's r correlation coefficients between detector outputs and annotator agreement; CCO is calculated within the pixels marked as a blood vessel by the annotators, and CCI the whole image. The p-values are all 0.0000 (to four decimal places).

Detector	CCO	CCI	CCI−CCO
MSF	0.3923	0.3573	−0.0350
GMM	**0.5833**	**0.8133**	**+0.2300**
LMSE	0.4168	0.5950	+0.1782
KNN	0.4361	0.6952	+0.2591

4.4.3 Annotator Analysis

The dendrogram is not included in this case study as no outliers can be identified with only two annotations. The F_1-score difference $(1 - F_{ij})$ calculated between the two annotations was found to be 0.2583 meaning that they give fairly consistent markings. The statistics in Table 13 are not as informative as in the previous case studies due to the low number of annotators and this highlights one of the issue of estimating ground truths using few annotations and such statistical comparisons. Nonetheless, we can infer from them that A2 marked a much larger number of blood vessels compared to A1 due to A2 having a high sensitivity and A1 not (in this case the 50% agreement GT that these statistics are calculated according to contains locations that any of the annotators marked, hence the specificity and PPV being one).

4.4.4 Agreement and Detector Performance

The four detectors selected for this case-study were the Matched-Filter Response (MSF) (Hoover et al, 2000), Linear Classifier (LMSE), k-nearest neighbours (KNN), and Gaussian Mixture Model (GMM). The LMSE, KNN and GMM classifiers were implemented using the MLVessel software package (Soares et al, 2006), which extracted features based upon the inverted green channel, and the response of Gabor wavelets at scales 2–5 applied to the inverted green channel. The first five images (im0001–5) of the dataset were used exclusively for training. The integration limits of the $\bar{\text{P}}$-R curves were $\pi'_1 = 0.023$ and $\pi'_2 = 0.235$, which were found to be $\pi'_1 = \mu - 3\sigma$ and $\pi'_2 = \mu + 3\sigma$ where μ is the mean of the skew found within a number of retinal image datasets and σ its standard deviation (Lampert and Gançarski, submitted).

Although only two annotations exist in this dataset, the same trend is visible as with all the other datasets—as the threshold on agreement increases, so does performance. Furthermore, it can be observed that three of the $\bar{\text{P}}$-R curves overlap within the lower recall ranges (LMSE, KNN, and GMM), and one does not (MSF). These three correspond to the correlations observed in Table 14 that increase from CCO to CCI and the fourth that does not overlap results in a decrease between CCO and CCI.

Table 15 Detector rankings evaluated using each ground truth (measured by the area under the $\bar{\text{P}}$-R curve). The GTs that result in these are: Ranking #1 — A1, 0.75-GT, SIMPLE-GT; Ranking #2 — A2, 0.5-GT/Any-GT, STAPLE-GT; Ranking #3 — LSML-GT.

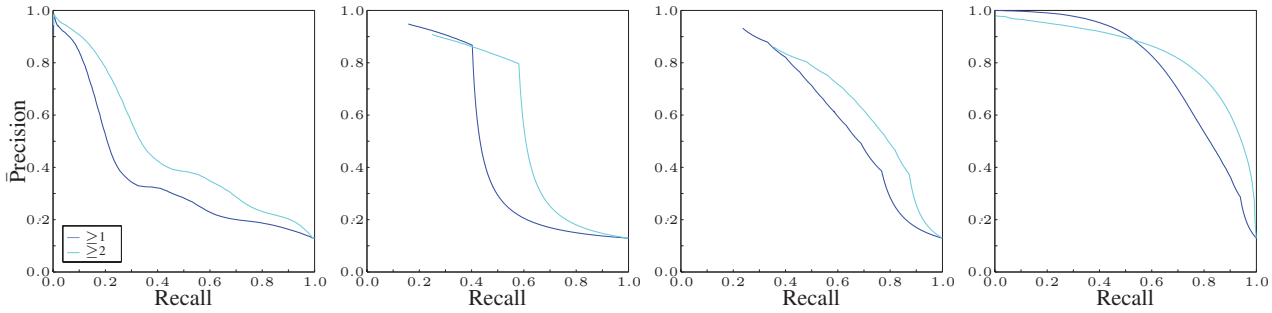
Rank 1	Rank 2	Rank 3
GMM	GMM	GMM
MSF	KNN	KNN
LMSE	MSF	LMSE
KNN	LMSE	MSF

4.4.5 Ground Truths and Reported Detector Performance

The better performing detector is GMM and Figure 25 presents its performance measured using the ground truth estimation methods selected for this study. Due to the limited number of annotations the Any-GT, 0.5-GT, and Excl-0.5-GT ground truths are identical (as no outliers can be identified).

Once again, three rankings emerge from the evaluation of detector performance using different ground truths, and these are presented in Table 15. Having a large difference in performance puts the GMM detector consistently on the top. Nevertheless, there is still a large difference in the performance of the remaining three detectors (see Figure 24) and the ranking of these is not consistent. The MSF detector, for example, achieves the lowest performance in Figure 24 and the lowest correlation with annotator agreement (Table 14), however, depending upon the GT that is taken, this detector is placed second, third, or last!

The results presented in Figure 25 reveal different findings to those in the other case studies. The LSML-GT forms the lower bound on the reported performance and the STAPLE-GT is equal to the 0.5-GT (and also the Any-GT), and therefore forms a lower bound estimate. Whereas previously (but to a lesser extent in the Segmentation case study) the STAPLE-GT and LSML-GT ground truths represented a mean estimate of the performance measured using the individual annotations. Once more the 0.75-GT ground truth results in a higher estimate of performance than that obtained using each of the individual annotations.



(a) MSF (b) LMSE (c) KNN (d) GMM

Fig. 24 P̄recision-recall curves describing the detectors' performances using different levels of agreement as the ground truth.

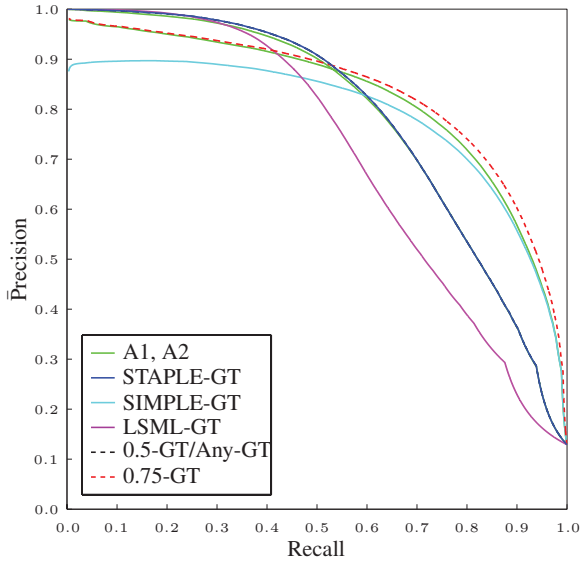


Fig. 25 P̄recision-recall curves of the GMM detector assuming differing ground truths (the curve obtained using STAPLE-GT overlaps that obtained using 0.5-GT).

5 Discussion

The following discussion is divided into two parts, the first provides a summary of the results presented in the previous section along with their implications, and the second part presents general recommendations that can be derived from these implications .

5.1 Summary of Results

It has been shown that the performance of a classifier increases as GTs are formed using increasingly higher agreement levels. Forming a GT using an agreement of 50% generally increases the reported performance of a detector to a range that is far greater than that obtained using all of the individual annotations. Kauppi et al (2009) conclude that the intersection method (consensus) is preferential as it results in the highest detector performance. Nevertheless, this study gives indication that the method in fact over estimates performance and focusses on evaluating a detector against the most obvious objects in the image. Raising the level of agreement at which the GT is calculated simply exaggerates this tendency.

One factor that has a stabilising effect on reported performance is a lower variance of the annotations. The Landslide dataset contains the lowest variance between annotations (because landslides are areal objects and not linear structures) and this is reflected in the spread of the performance curves and in the stability of the detector ranking. In this study the curves are relatively tightly clustered, and choosing any of the GTs for evaluating an algorithm would have resulted in similar reported performance. On the other end of the scale the Segmentation dataset contained the largest variance of annotations, and the reported performances also contain the largest variance. This is in contrast to the findings of Martin et al (2001) who found a large amount of agreement by comparing the regions that the segmentations contain and not the segments themselves. This also affected the gold-standard ground truth estimation methods, where in the other case studies the STAPLE and LSML methods typically modelled the 'mean' performance of the individual annotators, whereas in this dataset they actually resulted in the lowest performance curves. These methods both combine annotations based upon the annotator's statistical profile and given that there is a large variance in this dataset this may not be appropriate. In this situation removing the outlier annotations and performing consensus voting appears to be more stable (see the Segmentation case study). In other case studies this method also reported similar performances to that obtained using the STAPLE and LSML algorithms (except in the Fissure case study).

By and large, when the variance between annotations is relatively low (for example in the Landslide case study in which the F_1-score differences range from 0.14 to 0.28) the STAPLE and LSML methods provide GTs that report a performance within the middle of that reported by each of the individual annotations. Nevertheless, as noted above, this is not the case when the variance increases or few annotations are available (as in the Blood Vessel case study) and

this seems to be in line with other studies (Langerak et al, 2010). The SIMPLE algorithm was proposed to overcome these limitations in situations in which annotator uncertainty varies considerably (Langerak et al, 2010), and indeed, in these situations it does seem to offer an improvement (see, for example, the Segmentation and Blood Vessel case studies). Nevertheless, when the variance in annotator agreement is not so extreme SIMPLE seems to result in an overestimation of performance (see the Fissure dataset for example).

All of the detectors produced medium to high correlations between their output and the agreement of the annotators. It can be stated that a detector's performance increases as the agreement upon the object increases and those detectors resulting in the lowest drop in correlation result in a tighter spread of $\bar{\text{P}}$-R curves. This seems intuitive as agreement should be higher for more obvious objects and, assuming that the detector is effective, these should also elicit the highest detector responses. This translates to increasingly higher $\bar{\text{P}}$-R curves as GTs with higher levels of agreement are used. Unexpectedly however, when the correlation of the detector output and agreement increases from within object locations (CCO) to the whole image (CCI), precision decreases in lower recall ranges. Surprisingly, this reduction in precision indicates an accurate detector because as agreement increases lower-agreement objects are removed from the GT but the detector still detects them as false positive detections. This could be an indication that some of the annotators have missed important objects in the image, which the detector considers to be true positives. If this is correct, it indicates that the algorithm is performing better than the annotators at detecting these objects and by feeding back these locations to the annotators for confirmation, this could be a way of improving the ground truth reliability.

The image features included in this study account for a high proportion of the observed agreement (it should be kept in mind these features are not independent of each other), but these only capture local low-level information, ignoring any higher level and global queues and knowledge that the annotators exploit. This is compounded by the agreement level GT curves, which generally show that there is a large difference between the detection rate of high and low agreement objects—detection of the lower is not a trivial matter and the decision most likely needs to be augmented with high-level information that is not exploited by these detectors.

In all but one of the case studies it has been shown that the rank of a detector is dependent upon the GT used in the evaluation. It can therefore be stated that the variance in performance observed when evaluating two detectors using different ground truths is not equal and furthermore, the position of the performance measured using the same GTs within this range is not constant between detectors. Three different rankings were observed in three of the four case studies. In one occasion the top ranked detector changed depending upon the GT, however, in most cases the top ranked detector remained constant. This is partly due to the fact that these top ranked detectors are considerably superior to the remaining three and, had their performance been closer, this would not have been the case. The effects of ranking become more obvious in the Blood Vessel case study, in which the detector that produces the worst correlation with annotator agreement (MSF: CCO = 0.3923 and CCI = 0.3573) was placed second, third and fourth in each of the three emergent rankings, even though it is clearly the worst performing of the evaluated detectors. Moreover, taking the 50% or 75% consensus GTs does not necessarily result in a detector ranking that is the consensus of the ranks obtained using the individual annotations (see, for example, Tables 8 and 15). In fact, it can produce a ranking that has nothing in common with these individual rankings (Table 8).

The most consistent ranking was observed using the Landslide dataset—the case study that presented the most stable set of annotations (those with the least variance between them). The lower inter-annotator variability observed in this problem allowed the gold-standard GT estimation methods to better model all of the annotations.

The largest minimum bound on error, $\bar{e}$, was found in the Blood Vessel case study although the Segmentation and Fissure case studies produced the lowest pairwise F_1 scores (in fact the agreement between the two annotators in the Blood Vessel case study is relatively high). This uncovers two peculiarities with Smyth's calculation (see Equation (1)) when used with only two, and an odd number of, annotators: the maximum of $\bar{e}$ is reached when the maximum disagreement amongst the annotators takes place. On either side of this maximum $\bar{e}$ decreases symmetrically. First, when only two annotators are present, $N = 2$, any disagreement results in the maximum of the function since $[N - \max\{A(x,y), N - A(x,y)\}]/N \in \{0, 0.5\}$. Secondly, when an odd number of annotators are present this term can not reach the theoretical maximum of 0.5, and therefore all disagreements contribute less than in the case of two annotators. Thus although the F_1 score attests to greater agreement in the Blood Vessel case study, it receives a higher minimum bound on the error.

As has been shown in the Segmentation case study. The evaluation framework adopted in the segmentation domain, through accounting for variances observed in the annotations, yields an overly optimistic algorithm performance when compared to the traditional precision-recall evaluation framework. Moreover, the Berkeley framework produces an unique ranking that is not observed using any of the individual annotations or combinations therefore.

5.2 Recommendations

Comparing annotators and deciding upon outliers based solely upon inter-annotator performance is not a reliable method even though it offers reasonable modelling of—what could be described as—the average performance when correctly implemented (the SIMPLE, and to some extent the LSML, algorithms for example). Several counter examples can be easily proposed, such as a situation in which all but one annotator is inaccurate, a case in which the accurate annotator would be deemed an outlier and removed. Furthermore, an inaccurate annotation could in fact contain all of the true positive positions but have low specificity, other annotations may have low sensitivity and therefore removing the 'outlier' implies discarding valuable information that may not be possible to infer using other means. As Smyth (1996) states "Without knowing ground truth one can not make any statements about the errors of an individual labeller".

Over simplistic methods to utilise all of the available annotations (voting) have been shown to fail. More sensitive algorithms, such as STAPLE, take a step in the right direction. Nevertheless, these algorithms still rely on the assumption that the gold-standard ground-truth can be inferred through measuring the performance of the annotators in relation to each other. The most promising advances have started to integrate information derived from image properties into the process, and it has been shown herein that these properties indeed correlate with annotator agreement. Care should be taken, however, as this produces a somewhat circulatory solution in which the image features used by the detection algorithms are also used to decide upon which objects the algorithms are evaluated. Furthermore, in some domains correlation strengths between annotator agreement and image features decrease when moving from within object locations to the whole image. Demonstrating that these properties are not uniquely tied to the objects of interest and employing this source of information risks introducing false positive locations to the inferred ground truth.

In other fields of science, progress has been made on improving the rating of annotator performance by gathering meta-data along with the annotations. The Cooke method (Cooke, 1991) prescribes that the annotators are asked to estimate a credible interval of probable values along with their concrete answer, and furthermore they are also asked to answer multiple questions on topics from their field that have known answers. This information is used to weight the annotator's contribution in relation to their accuracy in this estimation and thus, has been shown to be more accurate than consensus voting (Aspinall, 2010).

It is clear that evaluating upon different ground truths, whether these are annotations or some merging thereof, reveals different trends in the performance of classification algorithms. Synonymously different images reveal different algorithm strengths during evaluation and, as such, large datasets are used to smooth the differences and reveal the best overall performing algorithm. However laborious it may be, the presented work implies that an algorithm should also be evaluated using different ground truths, the spread of measured performance quantified and used as a test as to whether the observed differences in performance are significant or not.

The variance of the annotations, and thus the variance of the algorithm's measured performance, is indicative of the number of annotations that should be collected to give an accurate measure of performance. The Landslide dataset, for example, exhibits low annotator variance and this is reflected in the spread of $\bar{\text{P}}$-R curves, which are relatively tightly clustered. Performance bounds can therefore be reliably estimated with few annotations. The Segmentation annotations, in contrast, exhibit large variance and so do the resulting $\bar{\text{P}}$-R curves. Under these conditions (and those in which few annotations are available, such as in the Blood Vessel case study) it may not be possible to state whether one algorithm outperforms another with any certainty. It is clear therefore that additional study into the nature of the problem should be conducted, and more annotations collected, before an algorithm is deemed to outperform another.

A viable approach to achieve this does appear to be possible. In all of the evaluated datasets the Any-GT and high agreement level ground truths (0.5-GT or 0.75-GT) appear to model the lower and upper bounds (respectively) on the spread of measured performance. This may offer a means of measuring the performance overlap between two algorithms, which would be characteristic of the confidence that can be attributed to any measured differences in performance.

This approach accepts that there exists imperfections in the individual annotations, which are included in the Any-GT but assuming that a perfect detector is created these imperfections cause the performance to degrade and simply decreases the lower bound on performance (and therefore represents the uncertainty inherent in the problem). Furthermore, there is a high likelihood that these imperfections are removed at high agreement levels (since they are variations of individual annotators). The upper bound, therefore is stable with respect to these and the true, unknown, detector performance is contained somewhere within these bounds.

6 Conclusions

This paper set out to quantify the effects of obtaining ground truth data from multiple annotators in a computer vision setting. It has also taken some steps towards identifying which properties of the image are related to agreement amongst the annotators. Statistical analyses of the GTs in each case study lead to the quantification of the differences between the annotations. A number of gold-standard GT estimation

methods were evaluated, including removing the outlier annotations, and it was found that the STAPLE and LSML algorithms find a balance between all annotations when their variance is low. The other GTs that were evaluated, formed by taking objects that any of the annotators marked, and thresholding at 50% and 75% agreement, tend to form lower and upper bounds on detector performance. The performance measured when using the GT derived by removing outlier annotations and then taking the consensus vote approaches that of STAPLE and LSML in all but one of the case study. It does, however, appear to be more stable when the annotations have high variability.

It can be concluded that the rank of a detector is highly dependent upon which GT estimation algorithm is used. In some cases the GTs calculated by voting result in detector ranks that are in discordance with each of the individual annotations. The $\bar{\text{P}}$-R curves obtained using the voted GTs also appear to be outliers when compared to those of the remaining GTs, suggesting that these commonly employed GT estimation methods overemphasise detector performance in comparison to individual annotator opinions. Furthermore, under some conditions, a detector with a low correlation between its output and annotator agreement can be placed above those that have vastly better correlated outputs.

Similarly to evaluating an algorithm over a data set that contains multiple images, it is concluded that an algorithm should be evaluated using multiple ground truths. The variance of performance that is observed using these different ground truths can then be used to quantify the confidence in the observed performance differences. In situations in which there are few annotations available, or when the inter-annotator variance is high, further study into the nature of the problem should be conducted as these conditions imply that it is not possible to state that one algorithm outperforms another with any confidence. Therefore, whenever possible the intrinsic uncertainties of annotator judgements should be assessed before the evaluation of object detectors, since the absolute performance measure and the relative ranking of detectors may vary considerably according to the employed GT.

The possibility to estimate the true detector performance through the variability of annotator opinion would be an interesting avenue to follow. Assuming that the performances derived using different GTs are observations of a hidden variable, it may be possible to estimate its true value—the gold standard performance. Much research is dedicated to inferring the gold-standard GT, however, this is a complex problem in which many assumptions need to be made, and the proposed approach may bypass some of these.

An additional question that arises from this study is: which metric should rate an estimated gold standard? Generally speaking the gold standard is unknown and therefore comparison is impossible. Restricting the evaluation to the individual annotations assumes high specificity and sensitivity. Removing annotators, however, assumes inability compared to the consensus, but do those removed have true insight into the problem? It is clear however that detector performance should not be used to evaluate a gold standard estimation.

Acknowledgements This work is part of the FOSTER project, which is funded by the French Research Agency (Contract ANR Cosinus, ANR-10-COSI-012-03-FOSTER, 2011–2014). The participating annotators from LIVE, IPGS, and ICube (University of Strasbourg), and ITC (University of Twente) are gratefully acknowledged.

References

Akinlar C, Topal C (2011) EDLines: A real-time line segment detector with a false detection control. Pattern Recogn Lett 32(13):1633–1642

Arbelaéz P, Maire M, Fowlkes C, Malik J (2011) Contour detection and hierarchical image segmentation. IEEE Trans PAMI 33(5):898–916

Armato S, McLennan G, Bidaut L, McNitt-Gray M, Meyer C, Reeves A, Zhao B, Aberle D, Henschke C, Hoffman E, Kazerooni E, MacMahon H, Van Beeke E, Yankelevitz D, Biancardi A, Bland P, Brown M, Engelmann R, Laderach G, Max D, Pais R, Qing D, Roberts R, Smith A, Starkey A, Batrah P, Caligiuri P, Farooqi A, Gladish G, Jude C, Munden R, Petkovska I, Quint L, Schwartz L, Sundaram B, Dodd L, Fenimore C, Gur D, Petrick N, Freymann J, Kirby J, Hughes B, Casteele A, Gupte S, Sallamm M, Heath M, Kuhn M, Dharaiya E, Burns R, Fryd D, Salganicoff M, Anand V, Shreter U, Vastagh S, Croft B (2011) The lung image database consortium (LIDC) and image database resource initiative (IDRI) : A completed reference database of lung nodules on CT scans. Medical Physics 38:915–931

Artaechevarria X, Munoz-Barrutia A, de Solorzano CO (2009) Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans Med Imag 28(8):1266–1277

Asman A, Landman B (2011a) Characterizing spatially varying performance to improve multi-atlas multi-label segmentation. In: Proc. of the 22nd int. conf. on Information processing in medical imaging, pp 85–96

Asman A, Landman B (2011b) Robust statistical label fusion through COnsensus level, Labeler Accuracy, and Truth Estimation (COLLATE). IEEE Trans Med Imag 30:1179–1794

Asman A, Landman B (2012a) Formulating spatially varying performance in the statistical fusion framework. IEEE Trans Med Imag 31:1326–1336

Asman A, Landman B (2012b) Non-local STAPLE: An intensity-driven multi-atlas rater model. In: Proc. of the 15th Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, vol 3, pp 426–434

Asman A, Landman B (2012c) Simultaneous segmentation and statistical label fusion. In: Proc. SPIE Medical Imaging 2012: Image Processing, vol 8314

Asman A, Landman B (2013) Non-local statistical label fusion for multi-atlas segmentation. Medical Image Analysis 17(2):194–208

Aspinall W (2010) A route to more tractable expert advice. Nature 463:294–295

Biancardi A, Reeves A (2009) TESD: A novel ground truth estimation method. In: Medical Imaging 2009: Computer-Aided Diagnosis, vol 7260, pp 72,603V–72,603V–8

Burl MC, Fayyad UM, Perona P, Smyth P (1994) Automated analysis of radar images of Venus: Handling lack of ground truth. In: ICIP, vol 3, pp 236–240

Commowick O, Warfield S (2010) Incorporating priors on expert performance parameters for segmentation validation and label fusion: a maximum a posteriori STAPLE. In: Proc. of the 13th Int. Conf. on Medical Image Computing and Computer Assisted Intervention, pp 25–32

Commowick O, Akhondi-Asl A, Warfield S (2012) Estimating a reference standard segmentation with spatially varying performance parameters: Local MAP STAPLE. IEEE Trans Med Imag 31(8):1593–1606

Cooke R (1991) Experts in Uncertainty: Opinion and Subjective Probability in Science. Oxford University Press

Coupé P, Manjón J, Fonov V, Pruessner J, Robles M, Collins D (2011) Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage 54(2):940–954

Davis J, Goadrich M (2006) The relationship between precision-recall and ROC curves. In: ICML, pp 233–240

Flach P (2003) The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In: ICML, pp 194–201

von Gioi RG, Jakubowicz J, Morel JM, Randall G (2010) LSD: A fast line segment detector with a false detection control. IEEE Trans PAMI 32(4):722–732

He H, Garcia E (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284

Hoover A, Kouznetsova V, Goldbaum M (2000) Locating blood vessels in retinal images by piece-wise threhsold probing of a matched filter response. IEEE Trans Med Imag 19(3):203–210

Isgum I, Staring M, Rutten A, Prokop M, Viergever M, van Ginneken B (2009) Multi-atlas-based segmentation with local decision fusion–application to cardiac and aortic segmentation in CT scans. IEEE Trans Med Imag 28(7):1000–1010

Kamarainen JK, Lensu L, Kauppi T (2012) Combining multiple image segmentations by maximizing expert agreement. In: Proc. of the 3rd Int. Workshop on Machine Learning in Medical Imaging, pp 193–200

Kauppi T, Kamarainen JK, Lensu L, Kalesnykiene V, Sorri I, Kälviäinen H, Uusitalo H, Pietilä J (2009) Fusion of multiple expert annotations and overall score selection for medical image diagnosis. In: Image Analysis, LNCS, vol 5575, Springer, pp 760–769

Lampert T, Gançarski P (submitted) The bane of skew: Uncertain ranks and unrepresentative precision. Machine Learning

Lampert T, O'Keefe S (2011) A detailed investigation into low-level feature detection in spectrogram images. Pattern Recognition 44(9):2076–2092

Landman B, Bogovic J, Prince J (2010) Simultaneous truth and performance level estimation with incomplete, overcomplete, and ancillary data. In: Proc. SPIE Medical Imaging 2010: Image Processing, vol 7623

Landman B, Asman A, Scoggins A, Bogovic J, Xing F, Prince J (2013) Robust statistical fusion of image labels. IEEE Trans Med Imag 31(2):512–522

Langerak T, van der Heidean U, Kotte A, Viergever M, van Vulpen M, Pluim J (2010) Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE Trans Med Imag 29(12):2000–2008

Li S (2012) FNN: Fast nearest neighbor search algorithms and applications. R package version 0.6-3

Li X, Aldridge B, Fisher R, Rees J (2011) Estimating the ground truth from multiple individual segmentations incorporating prior pattern analysis with application to skin lesion segmentation. In: ISIB, pp 1438–1441

Liaw A, Wiener M (2002) Classification and regression by randomForest. Rnews 2:18–22

Liu X, Montillo A, Tan E, Schenck J (2013) iSTAPLE: improved label fusion for segmentation by combining STAPLE with image intensity. In: Proc. SPIE Medical Imaging 2013: Image Processing, vol 8669

Maire M, Arbelaez P, Fowlkes C, Malik J (2008) Using contours to detect and localize junctions in natural images. In: IEEE Conf. CVPR, pp 1–8

Malamud B, Turcotte D, Guzzetti F, Reichenbach P (2004) Landslide inventories and their statistical properties. Earth Surface Processes and Landforms 29:687–711

Martin D, Fowlkes C, Tal D, Malik J (2001) A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV, vol 2, pp 416–423

Martin D, Fowlkes C, Malik J (2004) Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Trans PAMI 26(5):530–549

Meyer D (2009) Support Vector Machines - The Interface to libsvm in package e1071. R package

Niethammer U, James M, Rothmund S, Travelletti J, Joswig M (2011) UAV-based remote sensing of the Super-Sauze landslide: Evaluation and results. Eng Geol 128(1):2–11

Parker R, Densmore A, Rosser N, Michele Md, Li Y, Huang R, Whadcoat S, Petley D (2011) Mass wasting triggered by the 2008 wenchuan earthquake is greater than orogenic growth. Nature Geoscience 4:449–452

Raykar V, Yu S, Zhao L, Jerebko A, Florin C, Hermosillo-Valadez G, Bogoni L, Moy L (2009) Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In: ICML, pp 889–896

Ren X (2008) Multi-scale improves boundary detection in natural images. In: ECCV, pp 533–545

Ren X, Bo L (2012) Discriminatively trained sparse code gradients for contour detection. In: NIPS, pp 593–601

Sabuncu M, Yeo B, Leemput KV, Fischl B, Golland P (2010) A generative model for image segmentation based on label fusion. IEEE Trans Med Imag 29(10):1714–1729

Saur S, Alkadhi H, Stolzmann P, Baumüller S, Leschka S, Scheffel H, Desbiolles L, Fuchs T, Székely G, Cattin P (2010) Effect of reader experience on variability, evaluation time and accuracy of coronary plaque detection with computed tomography coronary angiography. Eur Radiol 20(7):1599–1606

Smyth P (1996) Bounds on the mean classification error rate of multiple experts. Pattern Recogn Lett 17(12):1253–1257

Smyth P, Fayyad U, Burl M, Perona P, Baldi P (1994) Inferring ground truth from subjective labelling of Venus images. In: NIPS, pp 1085–1092

Soares J, Leandro J, Cesar-Jr R, Jelinek H, Cree M (2006) Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification. IEEE Trans Med Imag 25(9):1214–1222

Stumpf A, Lampert T, Malet JP, Kerle N (2012) Multi-scale line detection for landslide fissure mapping. In: IGARSS, IEEE, pp 5450–5453

Stumpf A, Lachiche N, Malet N, Malet JP, Kerle N, Puissant A (accepted) Active learning in the spatial domain: Application to landslide mapping with VHR satellite images. IEEE Trans Geosci Remote Sens

Venables W, Ripley B (2002) Modern applied statistics with S-Plus. Springer, New York

Vonikakis V, Andreadis I, Gasteratos A (2008) Fast centre-surround contrast modification. IET Image Process 2(1):19–34

Wang H, Suh J, Das S, Pluta J, Craige C, Yushkevich P (2013) Multi-atlas segmentation with joint label fusion. IEEE Trans PAMI 35(3):611–623

Warfield S, Zou K, Wells W (2004) Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans Med Imag 23(7):903–921

Warfield S, Zou K, Wells W (2008) Validation of image segmentation by estimating rater bias and variance. Phil Trans R Soc A 366(1874):2361–2375

Xing F, Soleimanifard S, Prince J, Landman B (2011) Statistical fusion of continuous labels: identification of cardiac landmarks. In: Proc. SPIE Medical Imaging 2011: Image Processing, vol 7962

Yamaguchi T, Hashimoto S (2010) Fast crack detection method for large-size concrete surface images using percolation-based image processing. Mach Vision and Appl 21(5):797–809

Yang HF, Choe Y (2011) Ground truth estimation by maximizing topological agreements in electron microscopy data. In: Proc. of the 7th Int. Conf. on Advances in visual computing, pp 371–380

Zou Q, Cao Y, Li Q, Mao Q, Wang S (2012) Cracktree: Automatic crack detection from pavement images. Pattern Recogn Lett 33(3):227–238