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.

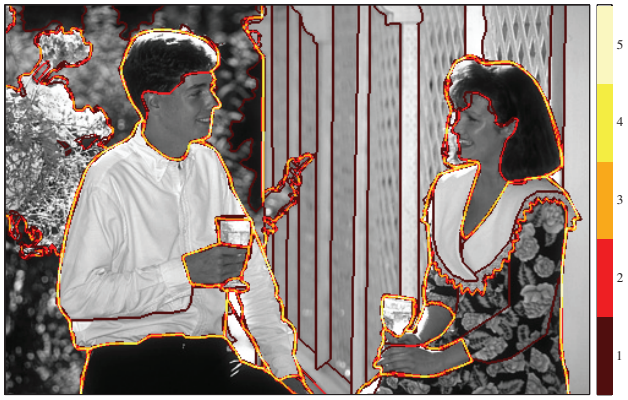


Fig. 2 Annotator agreement of segmentation boundary locations.

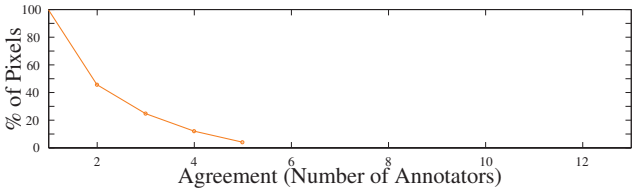


Fig. 3 Percentage of agreement (in pixels) as a function of the number of annotators.

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

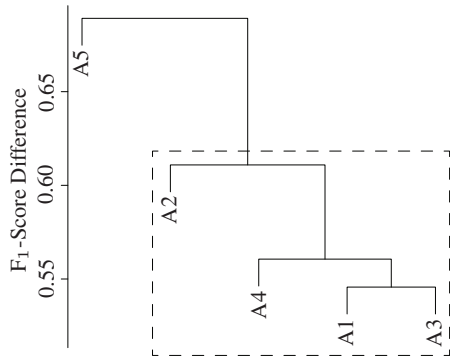


Fig. 4 Dendrogram describing the F_1-score difference relationships between each annotator's marking. The dashed box depicts the inliers (see Section 4.1.3).

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.

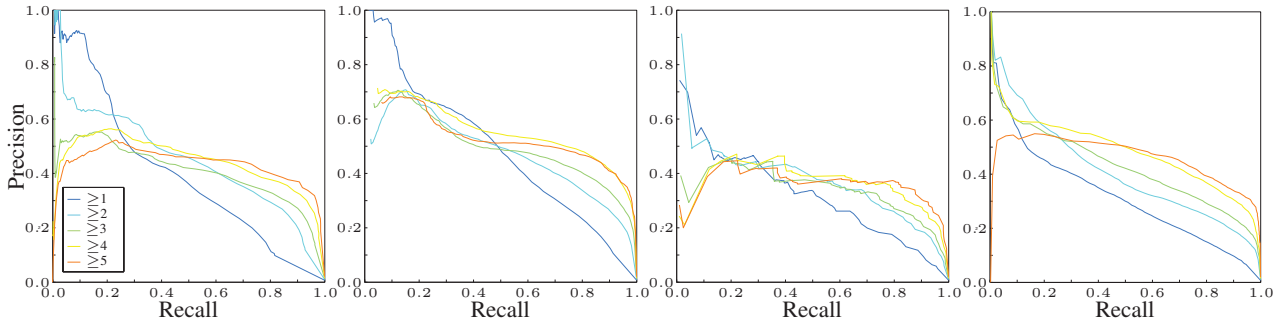
4.1.4 Agreement and Detector Performance

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

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.

4.2 Fissure Case Study

The following case study is concerned with the detection of fissures in remotely sensed images.

Table 4 Detector rankings evaluated using each ground truth (measured by the area under the $\bar{P}$-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

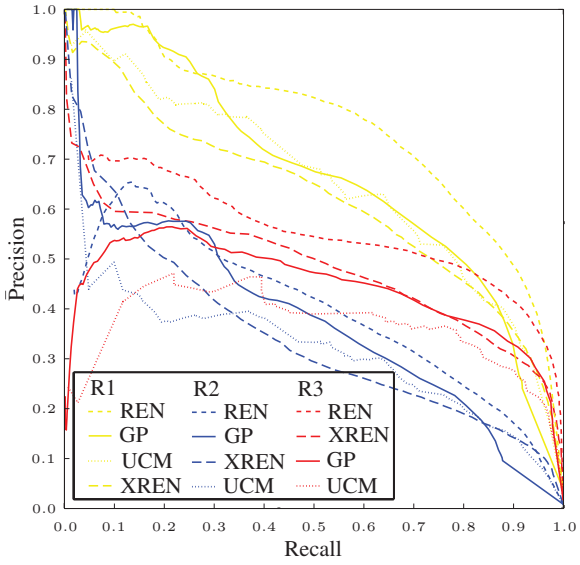


Fig. 6 $\bar{P}$recision-recall curves of all four detectors evaluated using the Berkeley Framework (A1–A5) giving Ranking #1 (R1), STAPLE-GT giving Ranking #2 (R2), and 0.75-GT giving Ranking #3 (R3).

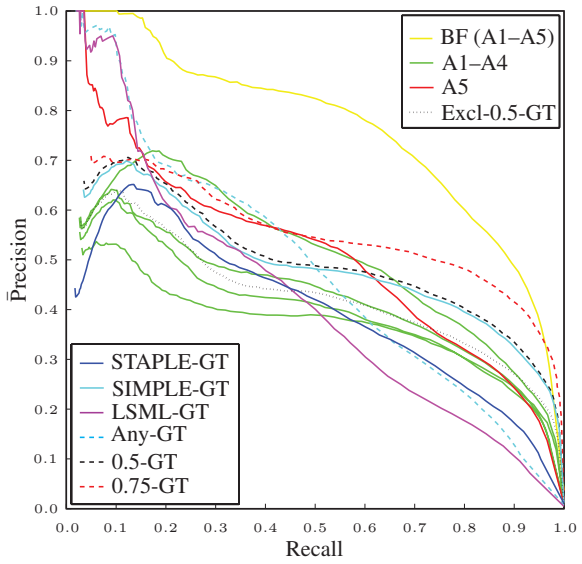


Fig. 7 $\bar{P}$recision-recall curves of the REN detector assuming differing ground truths.

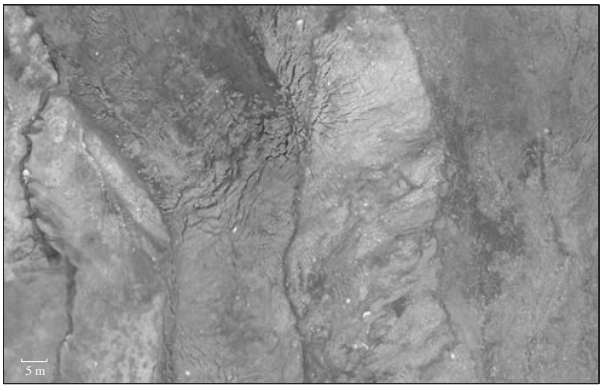


Fig. 8 The area of interest used within this case study.

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: $I(x, y) = 0.2989 \cdot R(x, y) + 0.5870 \cdot G(x, y) + 0.1140 \cdot B(x, y)$.

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.

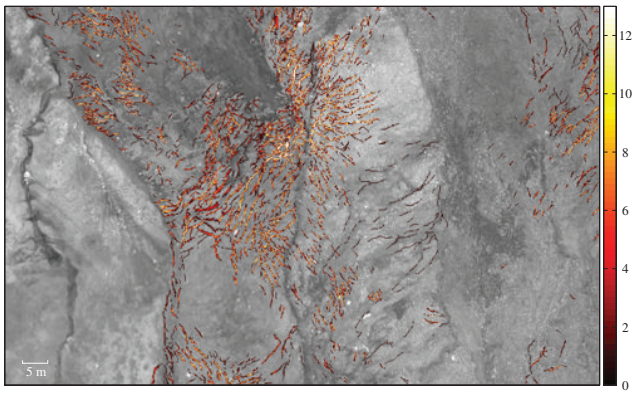


Fig. 9 Annotator agreement of fissure locations, the agreement of each pixel is calculated according to Eq. (2).

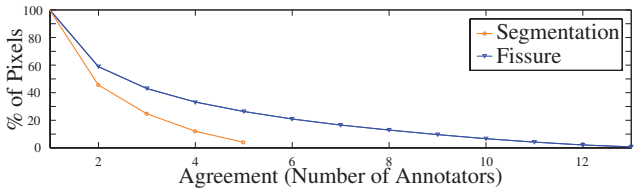


Fig. 10 Percentage of agreement (in pixels) as a function of the number of annotators.

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% 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.

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.

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	$\mathbf{0.4027}$

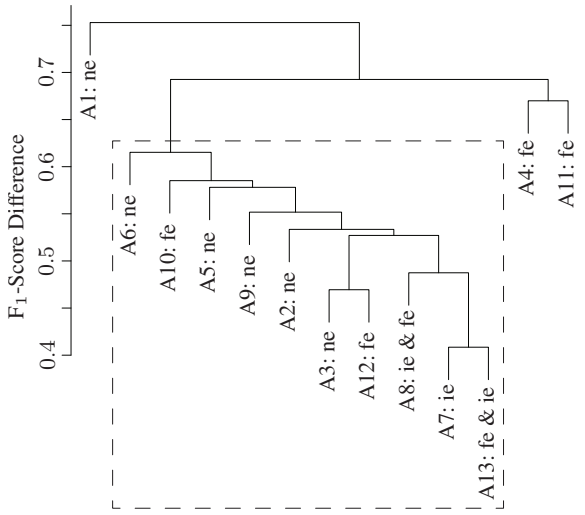


Fig. 11 Dendrogram describing the F_1-score difference relationships between each annotation. Key: ne — non-expert; ie — expert with previous experience of fissure mapping in imagery; and fe — expert with experience in the recognition of such fissures in the field. The dashed box depicts the inliers (see Section 4.2.3).

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).

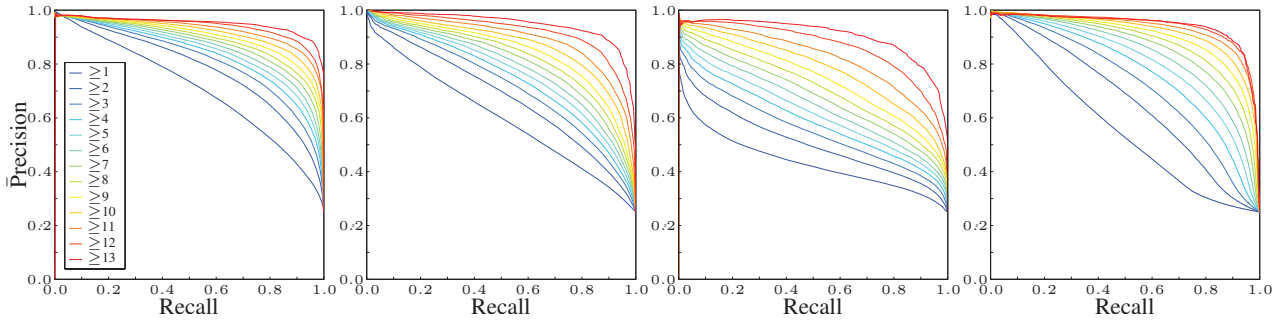
Detector	CCO	CCI	CCI−CCO
2D GWLC	0.5563	0.5166	**−0.0397**
Gauss	0.5293	0.4711	−0.0582
C-S	**0.6387**	**0.5259**	−0.1128
Top-Hat	0.5187	0.2780	−0.2407

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

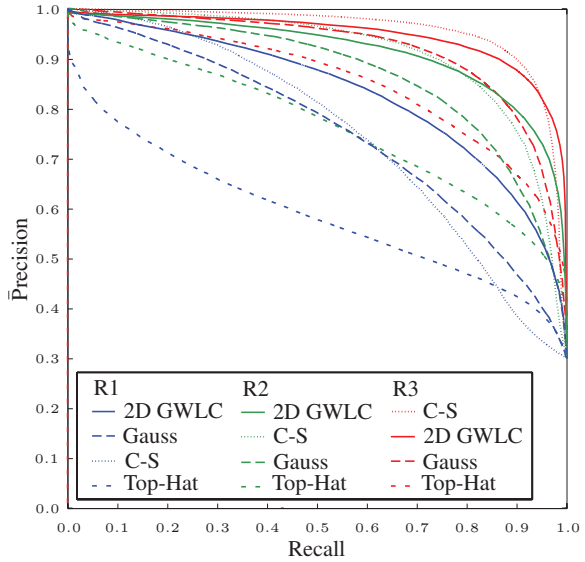


Fig. 13 P̄recision-recall curves of all four detectors evaluated using the STAPLE-GT giving Ranking #1 (R1), 0.5-GT giving Ranking #2 (R2), and 0.75-GT giving Ranking #3 (R3).

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.

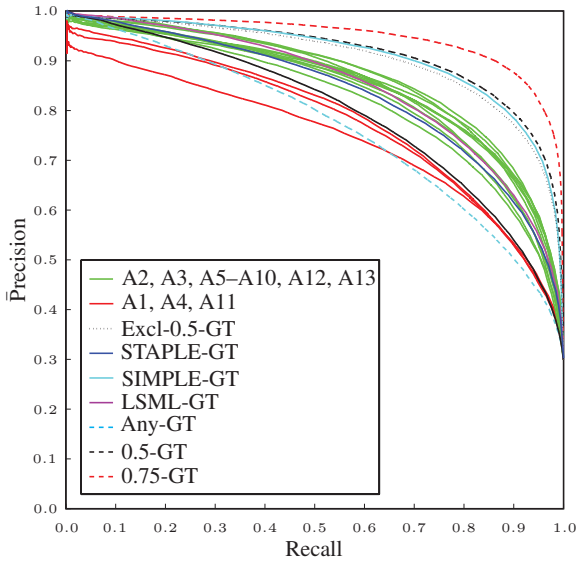


Fig. 14 $\bar{P}$recision-recall curves of the 2D GWLC detector assuming differing ground truths.

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.

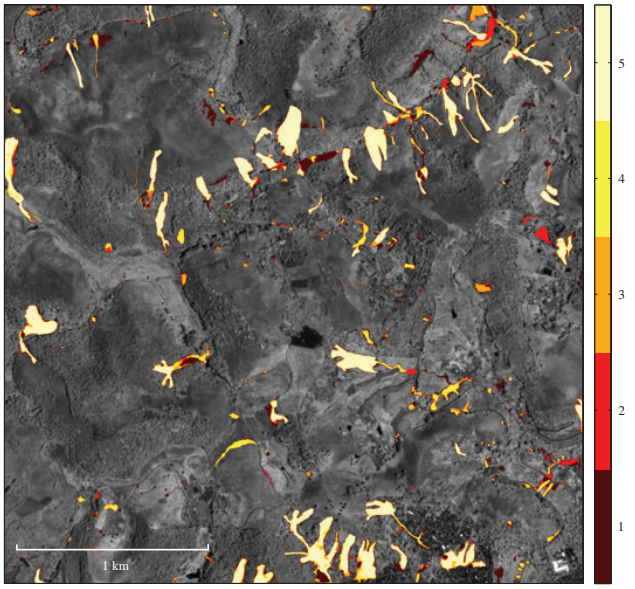


Fig. 16 Annotator agreement of landslide locations (overlaid on top of the near infra-red image band).

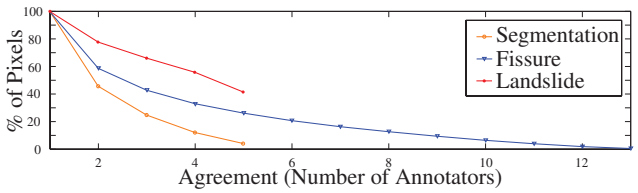


Fig. 17 Percentage of agreement (in pixels) as a function of the number of annotators.

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.

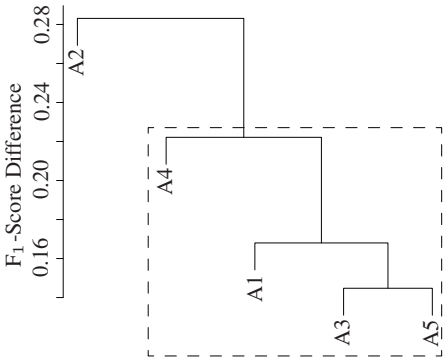


Fig. 18 Dendrogram describing the F_1-score difference relationships between each annotation. The dashed box depicts the inliers (see Section 4.3.3).

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, 2013).

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

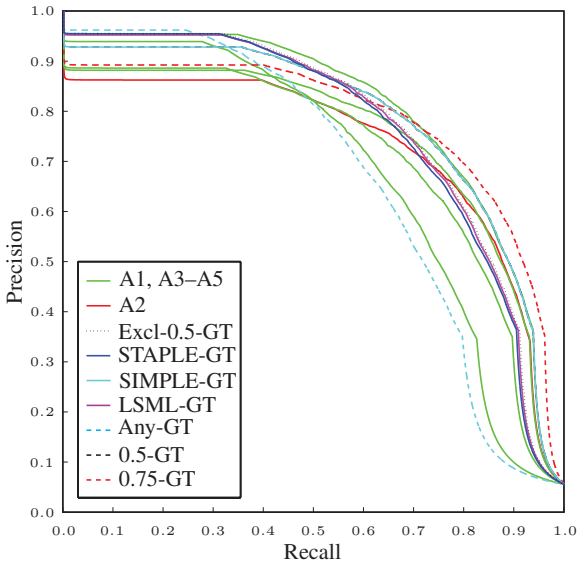


Fig. 20 $\bar{\text{P}}$recision-recall curves of the RF detector assuming differing ground truths (the 0.5-GT and SIMPLE-GT ground truths are identical).

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.

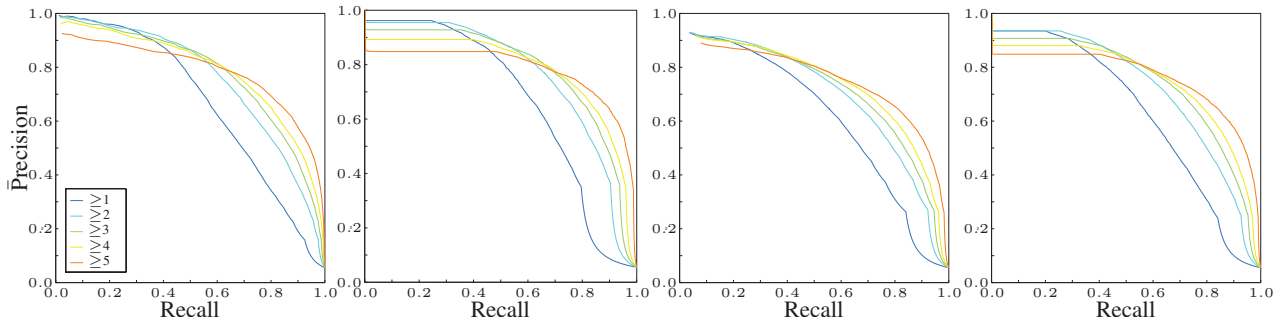
The lower degree of variance is also observed in the $\bar{\text{P}}$-R curves resulting from each GT—as presented in Figure 20—the variance in performance, although existing, is much lower than in the previous case studies. Similar trends can still be implied however. The GT derived from 75% agreement 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.

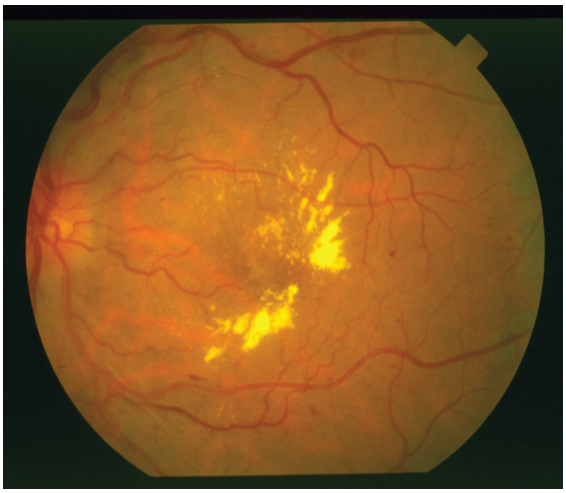


Fig. 21 An example of the retinal images used within this case study.

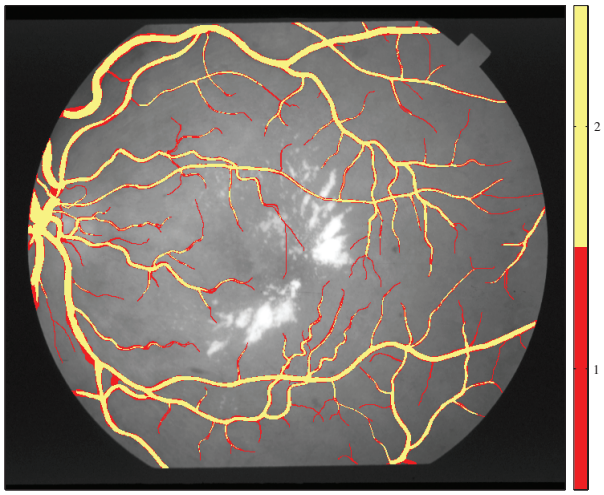


Fig. 22 Annotator agreement of blood vessel locations.

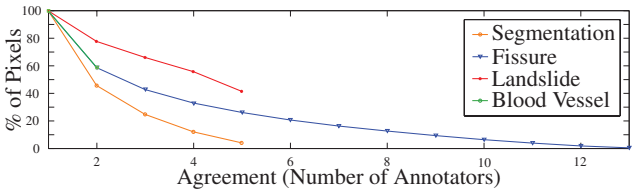


Fig. 23 Percentage of agreement (in pixels) as a function of the number of annotators.

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.

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

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	$\mathbf{-0.1495}$	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

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 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	$\mathbf{0.5833}$	$\mathbf{0.8133}$	$\mathbf{+0.2300}$
LMSE	0.4168	0.5950	$+0.1782$
KNN	0.4361	0.6952	$+0.2591$

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.

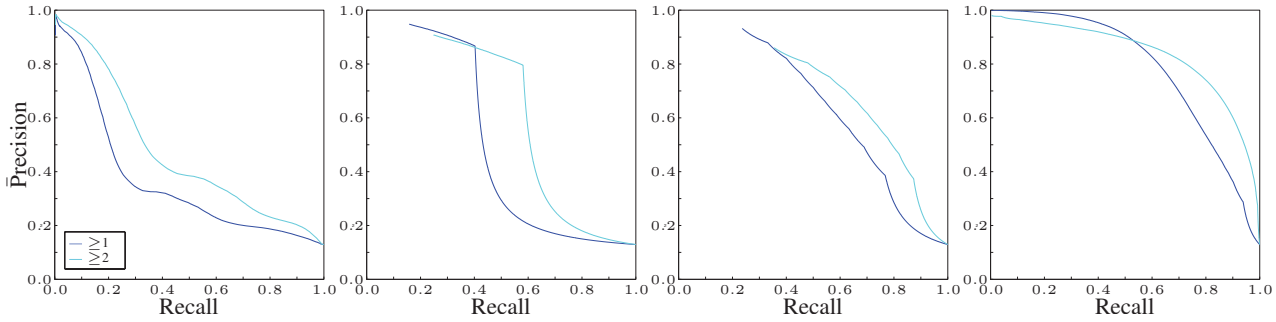
Rank 1	Rank 2	Rank 3
GMM	GMM	GMM
MSF	KNN	KNN
LMSE	MSF	LMSE
KNN	LMSE	MSF

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.

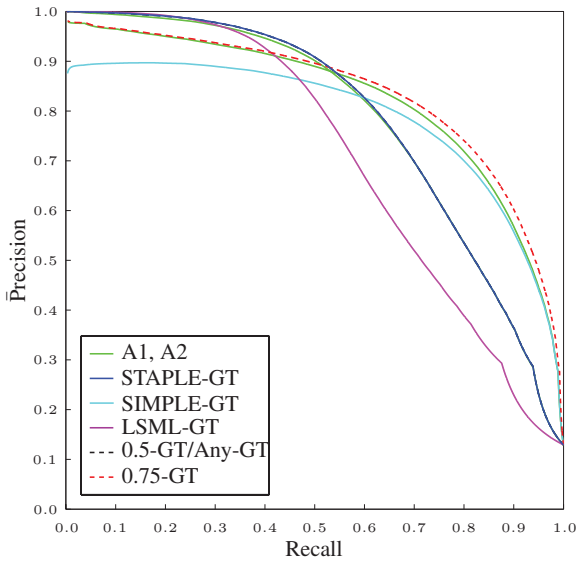


Fig. 25 P̄recision-recall curves of the GMM detector assuming differing ground truths (the curve obtained using STAPLE-GT overlaps that obtained using 0.5-GT).

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 (2013) Active Learning in the Spatial Domain for Remote Sensing Image Classification. IEEE Trans Geosci Remote Sens PP(99): 1–16

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