An Empirical Study into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation

Thomas A. Lampert*, André Stumpf, and Pierre Gançarski

***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 foreground-background segmentation 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? A methodology is applied to four image processing problems 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 structures annotator agreement is very low. The agreement in a structure's position can be partially explained through basic image properties. Automatic segmentation algorithms are compared to annotator agreement and it is found that there is a clear relationship between the two. Several ground truths estimation methods are used to infer a number of algorithm performances. It is found that: the rank of a detector is highly dependent upon the method used to form the ground truth; and that although STAPLE and LSML 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, resulting in an overestimation of performance. It is concluded that in some datasets it is not possible to confidently infer an algorithm ranking state when evaluating upon one ground truth.**

***Index Terms*—Evaluation, ranking, performance, feature detection, agreement, annotation, ground truth, gold standard ground truth, expert agreement, receiver operating characteristic analysis, precision, recall.**

I. Introduction

The evaluation of computer vision algorithms often requires ground truth (GT) data. The difficulty presented by this is that a gold standard GT 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 GT. 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) [1]. This poses a problem when evaluating computer vision algorithms: how does the difference in the annotator's opinion affect an algorithm's evaluation?

T. Lampert and P. Gancarski are with ICube, University of Strasbourg, France e-mail: tomalampert@outlook.com.

A. Stumpf is with LIVE, University of Strasbourg, France

Manuscript received September 15, 2014; revised September 15, 2014.

This work is intended to quantify the effects of variability in the GT on the design, training and evaluation of segmentation algorithms. To this end, selected supervised and unsupervised algorithms are evaluated on four different GT datasets, all of which embody typical computer vision problems. These four investigated problems 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 GT of these data sets (gold standard GT) cannot be deduced from the imagery alone and image annotations by human experts are used as the best approximation available. This limitation is typical in many computer vision applications such as medical imaging, remote sensing, and natural scene analysis. 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.

Several previous studies have developed statistical methods for estimating the gold standard GT from a number of annotations [2, 3, 4, 5, 6, 7, 8, 1]. Although some public datasets offer segmentations obtained from different annotators [9, 10, 11] these methods are rarely employed in real-world algorithm evaluation, where experimentation is typically limited to one annotation. Consequently, little is known about the effect of different GTs and estimated gold standards on the performance comparison among different algorithms.

Through performance evaluation GT 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 GTs 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 target. This problem can be circumvented when the images are captured in tightly controlled conditions or are synthetically generated from a model [12] and a gold standard GT is 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 (segments 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. 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 segments that belong to the negative class but share the same low-level image properties as those segments that constitute the positive class.

Therefore the objectives of this study are to:

- empirically show the possible bias in the evaluation of an algorithm when only one annotation is used;
- quantify the effect that different GTs may produce when comparing algorithms' performances;
- and provide a general comparison between algorithms designed to infer the gold standard GT.

The following section reviews the most relevant work from the literature. Section III prescribes the experimental methodology to be followed. The analysed datasets and the results are described in Section IV and a discussion of these results is presented in Section V. Finally the conclusions of the study are presented in Section VI.

II. Related Work

In a classic study Smyth et al. [7] analyse the uncertainty of an annotator's judgement in marking volcanoes in synthetic aperture radar images of 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.

Also more recently a number of methods for combining image annotations from two or more annotations are proposed. This includes works from the medical domain in which practitioners manually segment anatomical scans. The annotations are subsequently warped to match novel scans in order to estimate their segmentations. Kauppi et al. [4] 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 [13], local [14, 15, 13], semi-local [13, 16], and non-local [17] information.

Probably the most popular gold-standard GT estimation method originating from the medical domain has been proposed by Warfield et al. [8], named simultaneous truth and performance level estimation (STAPLE) in which the annotator performances (sensitivity and specificity) and the gold-standard GT 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. [1] and Xing et al. [18]). The same authors also propose an approach in which the bias and variance of each annotator is estimated instead of the performance measure [1] and another variant that account for instabilities in the annotator performance measures [19]. Much subsequent work has concentrated on the STAPLE algorithm: removing its assumption that annotator performances are constant throughout the data [20, 21, 22], and COLLATE [23], which accounts for spatial variability in task difficulty. Landman et al. [24] 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 [19, 24, 25].

Kamarainen et al. [26] 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. [5] 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. [6] 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 they accept that the contribution of an annotator's judgement should be dependent upon their performance but differently to STAPLE the energy function is constrained by a shape prior that is dependent upon the amount of detail in the annotator's marking, forming the LSMLP algorithm. Biancardi and Reeves [2] 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 GT segmentation's inside boundary, which increase towards its centre, and decreases negatively outside the segment border; thus the truth estimate from self distances (TESD) algorithm is introduced [2].

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 segments-of-interest. Yang and Choe [27] follow this path and propose a method that incorporates the warping error to preserve topological disagreements between the estimated gold-standard GT and the annotations. A number of extensions to the STAPLE algorithm have also been proposed [28, 29, 30] 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 [31] propose to combine a locally weighted voting strategy with information derived form the image's intensity.

The widely used Berkeley segmentation dataset contains five-hundred images, each having five GTs. The authors include the level of annotator agreement within their evaluations [9], which provides a valuable reference when interpreting the results. Using the earlier Berkeley 300 database, Martin et al. [32] present a statistical analysis of the variation observed within the annotations [32]. They notice that independent annotators tend to include the same pixel in the same region by different annotators but also that the number of segments

in the same image can vary by a factor of ten. The impact of GTs from different annotators on the ranking of segmentation algorithms has not been investigated yet.

III. Methodology

The methodological evaluation will be centred around four aspects: Annotator Agreement; Relation between Annotator Agreement and Detector Performance; and Ground Truths and Reported Detector Performance. Scripts to recreate the results presented henceforth are available on-line[1].

A. 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 foreground target in each study. All case studies are binary detection problems and each annotation has the value one where the annotator perceived the object to exist and zero otherwise. The result of this are 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 $n = 1, \ldots, N$.

B. Annotator Agreement

The first stage of analysis tests the level of agreement between the annotators in each case study, and exposes the image properties that promote this agreement.

Smyth [33] 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

$$\bar{e} \geq \frac{1}{XYN} \sum_{y=0}^{Y-1} \sum_{x=0}^{X-1} \min\{N - A(x,y), A(x,y)\} \quad (1)$$

where $A(x, y)$ is the number of annotators that labelled pixel (x, y) as containing the segment-of-interest, as defined in Equation (2). The minimum of Equation (1) is reached when all annotators agree and the maximum (0.5) when the decisions are evenly split. It is therefore a measure closely related to the entropy of the annotators' decisions. As a minimum value for an acceptable quality of experimental data the author suggests 10%.

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 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 $L(x, y) = I(x, y)$. For the comparison of contrast and agreement the maximum agreement within the local neighbourhood is used.

A number of the gold-standard ground-truth estimation methods evaluated in this research weight annotations based upon the assumption that the more reliable annotators can be identified through inter-annotator comparisons.

To examine the inter-annotator variability, a cluster analysis using the pairwise F_1-score between the annotator markings is conducted in this study. The F_1-score [34], 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 since it does not take into account true-negative classifications [34]. 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. [35], 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

[1] https://sites.google.com/site/tomalampert

(NPV) and Cohen's kappa coefficient. This will allow for specific tendencies of the outliers to be identified.

C. Relation between 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 segmentation problem at hand (every effort was made to select the best performing detectors within each domain). Each of these detectors is evaluated using GTs 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 [36, 34]. Nevertheless, precision is sensitive to the ratio of positive to negative instances in the dataset, $\phi = N_p/N_n$. To overcome this Flach [37] proposes to analytically vary the class skew in the precision measure and Lampert and Gançarski [38] 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 [38] 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 significant segments 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.

D. 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. segments 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 [8] (without assigning consensus votes [22]), SIMPLE [5], and LSML [6] (using the 50% agreement as an initial estimate and 1000 iterations). Furthermore, an additional GT is determined by excluding those outliers identified in Section III-B 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.

IV. Experimental Results and Analyses

This section presents the results of applying the methodology to each of the case studies included in this investigation.

A. Data

The case studies presented in this section are concerned with[2]:

Image segmentation Most of the images within the Berkeley 300 (colour) dataset have been annotated by numerous different annotators. Only for a small subset of five images did the same annotators perform the segmentation (annotator IDs for the Berkeley 500 dataset are not available). These images are: 65033.jpg, 157055.jpg (Figure 1a), 385039.jpg, 368016.jpg, and 105019.jpg. 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.

Fissures in remotely sensed images 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 [39, 40]. An area of interest, where $X = 1425$ and $Y = 906$, was extracted from the data and is presented in Figure 1b. 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 the marking experiment, all the annotators were given a basic introduction on the characteristics of the targeted fissures. The annotators then independently marked all pixels which they believed to form part of a fissure, taking as much time as they required (this ranged from 2–3 h). The annotators were encouraged to perform the marking on a level in which they could see individual pixels clearly and zoom in and out as needed for assessing the context of the area being marked.

Landslides in satellite imagery The dataset is derived from Geoeye-1 satellite images with four spectral bands (blue,

[2]Colour copies of the images can be found in the multimedia material that accompanies this paper.

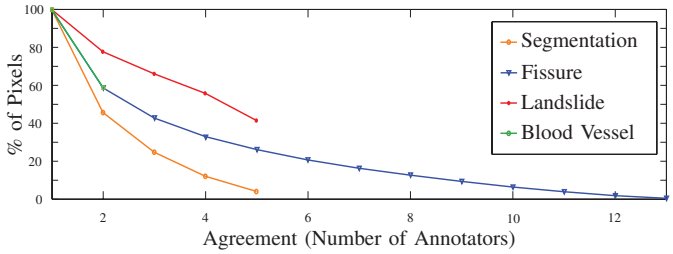


Fig. 2. Agreement (in pixels) as a function of the number of annotators.

green, red near infra-red) and a nominal ground resolution of 50 cm. The image presented in Figure 1c 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. Further detailed information on the dataset can be found in the literature [41].

Retinal blood vessels The STructured Analysis of the Retina (STARE) dataset was used in this case study. 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 1d. A mask was formed which delineates the pixels that fall outside the retina by thesholding the intensity of the red channel at a value of 40 (the black area) and these pixels were excluded from the experiments. The dataset contains two annotations which delineate the pixels that are part of the blood vessels.

B. Annotator Agreement

Smyth's lower error bound estimate, i.e. the average error rate amongst the annotators, for each dataset was found to be $\bar{e} \geq 2.6611\%$ (Segmentation), $\bar{e} \geq 1.26\%$ (Fissure), $\bar{e} \geq 1.1012\%$ (Landslide), and $\bar{e} \geq 3.1123\%$ (STARE). These values are well within the 10% limit that is recommended [33] and considerably lower than an error bound of 20% for labelling volcanoes in satellite images of Venus [33], in which the signal-to-noise ratio of the segments is much lower than in these studies.

The annotator agreements for each case study are presented in Figures 1 and 2. For the Segmentation, Fissure and Blood Vessel case studies the level of agreement decreases approximately exponentially as a function of the number of annotators. For the Fissure dataset the thirteen annotators agree only about 0.6979% of all of the pixels that were marked as fissures by any of the annotators. The decrease of degreement with the number of annotators is strongest for the Segmentation case, whereas the Landslide dataset exposes a rather linear trend. These different trends result from the combination of the geometric structure of the targeted objects with the fact that disagreement occurs mainly along the object borders.

TABLE I
PEARSON'S r CORRELATION COEFFICIENTS BETWEEN IMAGE FEATURES AND AGREEMENT. CORRELATIONS IN ITALIC FONT ARE NOT SIGNIFICANT AT P=0.0001.

Feature	r
Intensity	0.002
Contrast	**0.325**
Red	−0.002
Green	0.003
Blue	0.033

(a) Segmentation

Feature	r
Intensity	−0.2245
Contrast	**0.4027**

(b) Fissure

Feature	r
Intensity	0.0609
Contrast	0.0310
Near-IR	**−0.2766**
Red	0.1841
Green	−0.0115
Blue	0.0200

(c) Landslide

Feature	r
Intensity	−0.0861
Contrast	−0.0026
Red	0.0050
Green	**−0.1495**
Blue	−0.0007
	(p = 0.0087)

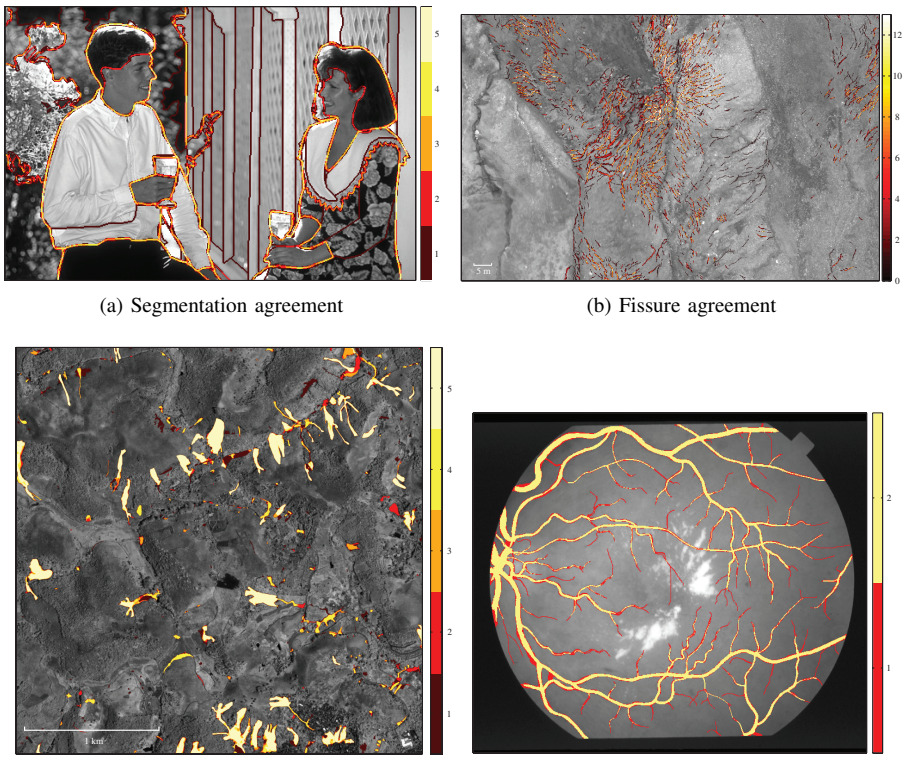
(d) STARE

Equvialent uncertainties about the outline of a feature lead to a stronger disagreement (in % of pixels) if the targeted features are only one pixel wide (Segmentation) or several pixels wide (Fissure, Blood Vessel) than when exhibiting a rather blob-like shape (Landslide). Indeed, if the outlines of the Landslide GT annotations are used agreement also drops approximately exponentially.

The correlation coefficients between agreement and image properties that are relevant for each case study are presented in Table I. In each case study there exists at least one significant correlations between image properties and agreement: contrast in the Segmentation dataset indicating that the annotators agree more on stronger edges; contrast and intensity in the Fissure dataset indicating that dark fissures on a lighter background attract greater agreement; near-IR and red which exhibit a strong response if vegetation is removed during a landslide and the reddish soil is exposed; and Green in the Blood Vessel dataset which is in line with the use of the green channel in many blood vessel detection studies.

Dendrograms describing the relationship between the annotators' pairwise F_1-scores for each case study are presented in Figure 3 and the full statistics of each annotator compared to the average annotation are presented in Table II.

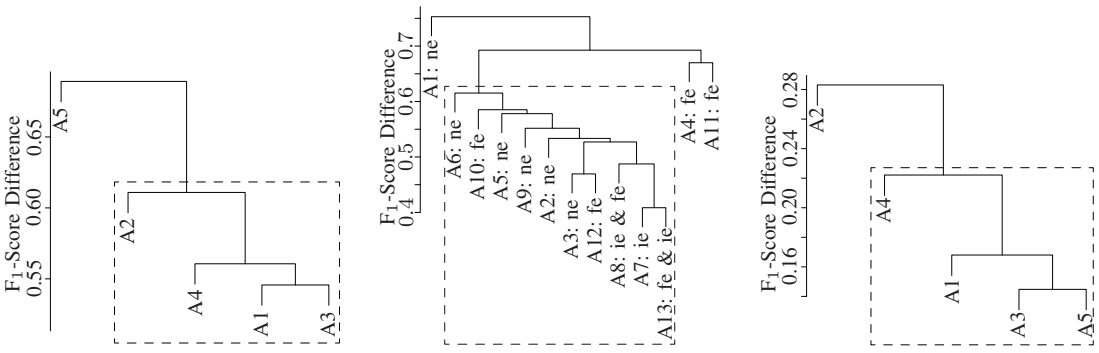
The relatively low levels of agreement in the segmentation problem are reflected in the pairwise differences in F_1-scores upon which the dendrogram in Figure 3a is based. The differences are relatively high and range from 0.545 to 0.68. One outlier is identified, A5 (the mean F_1-score difference was 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 IIIa. The variance in the annotations are underlined by the lowest specificities observed in all of the case studies. A dendrogram describing the STARE dataset is not included as no outliers can be identified with only two annotations. Nevertheless, the F_1-score difference $(1 - F_{ij})$ calculated between the two annotations was found to



(a) Segmentation agreement (b) Fissure agreement

(c) Landslide agreement (d) STARE agreement

Fig. 1. Annotator agreement according to Eq. (2) from the four analysed datasets.



(a) Segmentation (b) Fissure (c) Landslide

Fig. 3. Dendrogram describing the F_1-score difference relationships between each annotation. Fissure 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.

be 0.2583 meaning that they give fairly consistent markings. The statistics in Table IIId are not as informative as in the other case studies due to the low number of annotators and this highlights one of the issues of estimating GTs 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).

It would be expected that more than one cluster emerges within the Fissure dataset, 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 none of the groups is overly biased in favour of one particular decision. Annotators A1, A4, and A11 are identified as falling outside of one standard deviation of the mean F_1-score difference to all other annotators. These same annotators achieve considerably lower sensitivity when compared to the consensus (see Table IIIb). 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 in this dataset however are low sensitivity and PPVs.

In the Landslide 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 Fissure case study). Nevertheless, 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 IIIc. On average, sensitivity, PPV and kappa are higher than in the Fissure case study, indicating that the features used for the identification of landslides are more clearly defined and understood by the annotators.

C. Agreement and Detector Performance

During these case studies 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:

Image segmentation The top four performing segmentation algorithms listed on the Berkeley dataset web page[3] were selected to form part of this case study. These are: REN [42], gPb-ucm (UCM) [9], Global Probability of Boundary (GP) [43], and XREN [44]. The integration limits of the $\bar{\text{P}}$-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 [38]. As discussed by Martin et al. [45] 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. True-positive detections are accumulated if a detection is within a certain distance of one or multiple GT boundaries. 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. The images 105019.jpg and 368016.jpg are used as the training set and removed from this point forward. One further modification to the methodology was made to better suite the definition of segmentation. The low agreement GTs (≥ 1, for example) resulted in the delineation of segment boundaries that are multiple pixel wide (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 the width of the boundaries to one pixel, while any individual, low agreement, markings that the annotators are preserved.

[3]http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/bench/html/algorithms.html

TABLE II
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**

(a) Segmentation

	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

(b) Fissure

	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

(c) Landslide

	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

(d) STARE

Fissures in remotely sensed images Current state-of-the-art linear feature detectors were selected from the literature: 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) [46]; Gaussian filter matching [47], $\sigma = 1$ (Gauss); Top-Hat transform (4 pixel radius circular structuring element); and the Centre-Surround (C-S) transform (using a 3×3 pixel neighbourhood) [48]. 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 (to ensure that the implementations were true to the author's intentions and to allow reproducibility of the 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. 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.

Landslides in satellite imagery For popular classification algorithms were applied (due to their proven strength in real-world applications): random forest (RF) [49], support-vector machine (SVM), k-nearest neighbours (KNN), and a neural network (ANN) algorithms. After fine scale image segmentation, 101 features describing the spectral characteristics, texture, shape, topographic variables and neighbourhood contrast were extracted. The resulting dataset is available on-line[4] and a detailed description of the feature extraction methods are given in the literature [41]. Each classifier was trained upon samples from the same randomly selected square subset covering 10% of the area of interest. 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 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 [50].

Retinal blood vessels The four detectors selected for this case-study were the Matched-Filter Response (MSF) [11], 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 [46], 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 of the dataset (im0001–5) 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 skew found within a number of retinal image datasets and σ its standard deviation [38].

The $\bar{\text{P}}$-R curves derived from these detectors are presented in Figure 4. A striking observation is that the performance of all the detectors increases in the higher recall ranges with agreement in a predictable manner. Assuming that the more agreed upon features are the most obvious, this result indicates that the detectors extract similar features to those that aid an annotator's decision. Regarding the Fissure dataset, there is 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. For the lower recall range in the Segmentation, Landslide and STARE datasets 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 III and is discussed below.

Several general tendencies can be drawn from the correlations results presented in Table III. The detectors that exhibit a large drop between the CCO and CCI correlations exhibit low sensitivity. This is reflected in the $\bar{\text{P}}$-R curves of the C-S detector (Figure 4h) for example: 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 detectors that exhibit a high correlation with agreement over the whole image, and also exhibit the lowest drop in correlation between the two tests (2D GWLC, Gauss, SVM, and GMM detectors for example), have (relatively) low false positive rates and result in high $\bar{\text{P}}$-R curves. A large drop in correlation, along with a low absolute correlation, is observed with the top-hat detector, and indeed in Figure 4 the curves are skewed towards lower precision values. The detectors that result in the lowest drop in correlation (or an increase in correlation) result in a tighter spread of $\bar{\text{P}}$-R curves.

The $\bar{\text{P}}$-R curves resulting from the Segmentation, Landslide and STARE datasets largely follow the trends that as agreement increases the performance of the algorithm also increases. However, there is a tendency for precision to be inversely proportional to agreement in the lower recall range. This phenomenon can be explained by analysing the correlations between annotator agreement and the detector outputs presented in Table III. It should be noticed that in all of the cases in which this trend is observed CCI is higher than CCO. Indicating that the detector output strengths agree with annotator agreement within feature locations and more so 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 image locations having 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 segments 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.

D. Ground Truths and Reported Detector Performance

The rankings of the detectors' performance (measured as AUC) when evaluated using different GTs were determined, and in three of the case studies (Segmentation, Fissure and STARE) three rankings emerged, which are described in Table IV. In the Landslide dataset only one emerged due to the lower inter-annotator variance: SVM, RF, ANN, and KNN. For the Fissure, Landslide and STARE datasets these rankings reflect the results of the correlation analyses, the top ranked detectors (2D GWLC, SVM, and GMM) and the bottom ranked detectors (Top-Hat and KNN) correspond to either the highest correlations or the lowest drops in correlation observed in the previous section, see Table III.

[4]http://eost.unistra.fr/recherche/ipgs/dgda/dgda-perso/andre-stumpf/data-and-code/

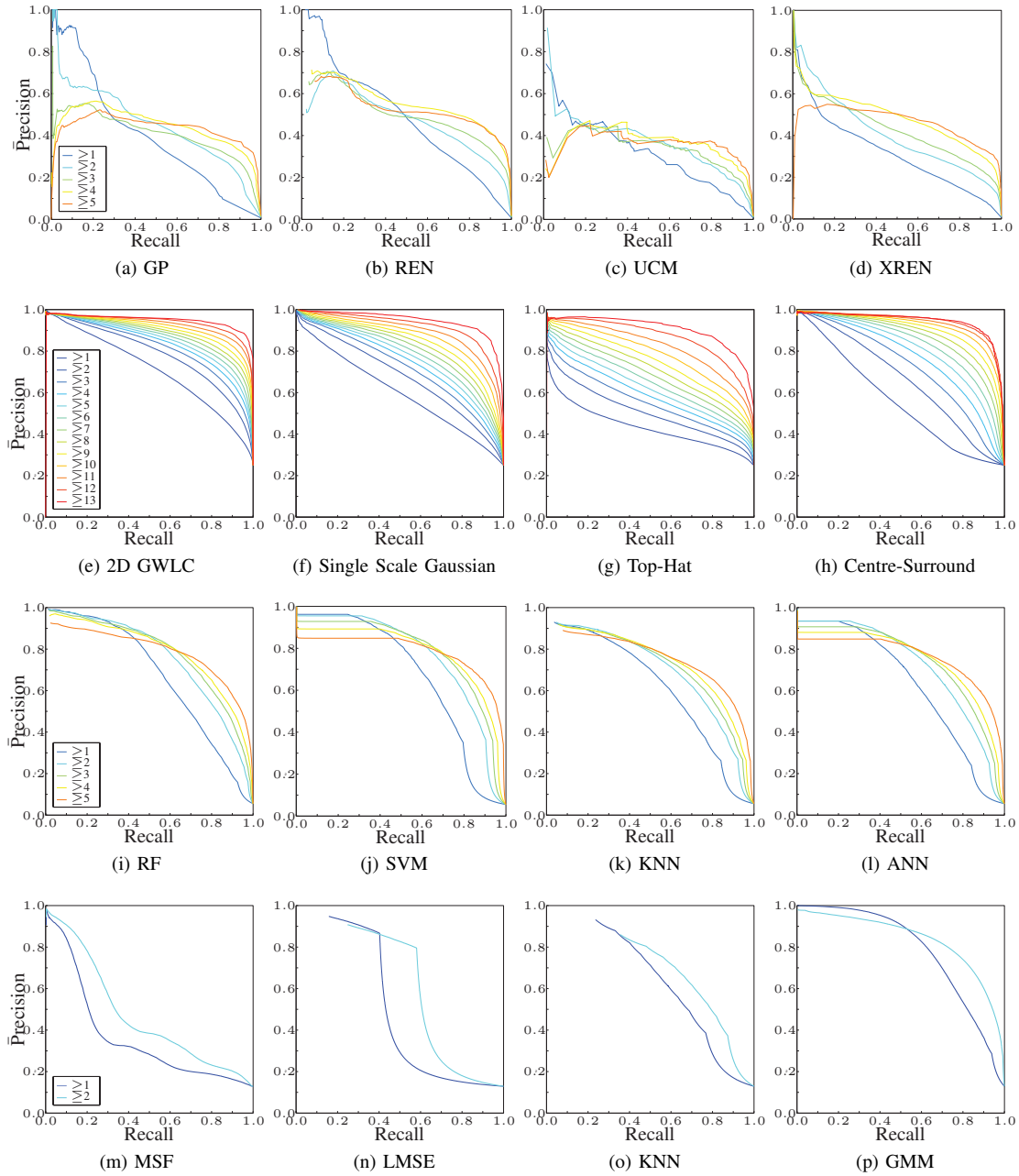


Fig. 4. $\bar{\text{P}}$recision-recall curves describing the detectors' performances using different levels of agreement as the ground truth.

For the Segmentation dataset, however, the algorithms with the highest correlation (UCM) and the highest increase in correlation from CCO to CCI (GP) only rank with medium to low performance. On the one hand this can be attributed to the relative high variance of the annotations and the overall low correlation between detector output and the annotator agreement (Figure III). On the other hand it must also be considered that the correlations are derived using all annotated pixels while the $\bar{\text{P}}$-R curves are based on GTs that were thinned to one pixel and TP rates calculated taken into account a distance tolerance to the object boundaries. This highlights not only the sensitivity of algorithm evaluation when using different GTs with high variance but also illustrates how different evaluation strategies can provoke different outcomes.

In the Fissure dataset, a majority of the individual annotations give the same ranking as SIMPLE-GT, and 0.5-GT however, when the 0.75-GT, Any-GT, STAPLE-GT, and LSML-GT are under consideration, the ranking changes. The method of calculating the GT influences the detectors' ranking. More importantly, the ranking derived using a 75% voting strategy (Fissure) and LSML (STARE) are in disagreement with that obtained using the annotator judgements individually, and this appears to be in contradiction to what should be expected. To illustrate these ranks for the Fissure dataset, the $\bar{\text{P}}$-R curves for all four detectors evaluated using the STAPLE-GT, 0.5-GT, and 0.75-GT are plotted in Figure 5, each colour represents one of the rankings presented in Table 4b.

For the STARE dataset the ranking of the lower three

TABLE III
PEARSON'S r CORRELATION COEFFICIENTS BETWEEN DETECTOR OUTPUTS AND ANNOTATOR AGREEMENT; CCO IS CALCULATED WITHIN THE PIXELS MARKED AS A POSITIVE INSTANCE 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
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
MSF	0.3923	0.3573	−0.0350
GMM	**0.5833**	**0.8133**	**+0.2300**
LMSE	0.4168	0.5950	+0.1782
KNN	0.4361	0.6952	+0.2591

TABLE IV
RANKINGS OF DETECTORS EVALUATED USING EACH GROUND TRUTH (MEASURED BY THE AREA UNDER THE $\bar{P}$-R CURVE). (A) FOR THE SEGMENTATION CASE STUDY 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. (B) FOR THE FISSURE CASE STUDY 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. (C) FOR THE BLOOD VESSEL CASE STUDY 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
REN	REN	REN
GP	GP	XREN
UCM	XREN	GP
XREN	UCM	UCM

(a) Segmentation

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

(b) Fissure

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

(c) Landslide

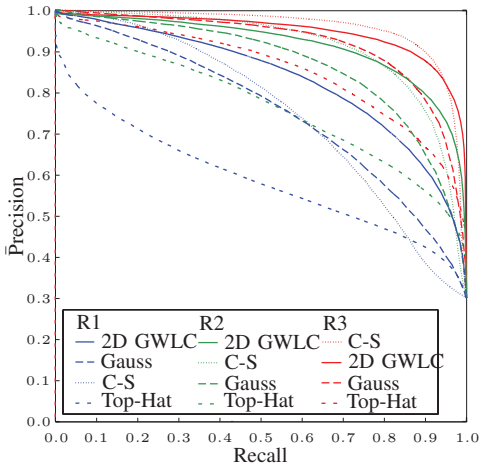


Fig. 5. $\bar{P}$recision-recall curves of all four detectors evaluated using the STAPLE-GT giving Ranking #1 (R1), 0.5-GT giving Ranking #2 (R2), and 0.75-GT giving Ranking #3 (R3).

detectors is not consistent. The MSF detector, for example, achieves the lowest performance in Figure 4 and the lowest correlation with annotator agreement (Table III), however, depending upon the GT that is taken, this detector is placed second, third, or last.

An overview of the performance variance with different GT estimation methods and evaluation frameworks can be can be obtained from Figure 6 presenting the $\bar{P}$-R curves of four selected detectors for each of the respective case studies. The $\bar{P}$-R curves for the REN segmentation algorithm Figure 6a shows the largest level of performance variation caused by the large variation between each annotation (see Section 4.3). At the positive extreme of this variance is the evaluation methodology commonly used to evaluate segmentation algorithms on the Berkeley dataset using a distance tolerance when evaluating TP detections. The 0.75-GT, 0.5-GT and SIMPLE-GT yield higher performance (particularly in the higher Recall ranges) and Any-GT relatively low 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 model a mean of the individual annotations in the lower recall ranges. This is a consequence of the large variance observed in the annotations. Excl-0.5-GT and SIMPLE-GT were very similar as they are both derived using the same principle (removing the outliers and then voting).

The $\bar{P}$-R curves of the 2D GWLC, obtained using each of the GTs in each case-study, are presented in Figure 6b (due to the limited number of annotations in the STARE dataset Any-GT, 0.5-GT, and Excl-0.5-GT are identical). The effects of the voted GTs (0.5-GT, 0.75-GT and SIMPLE-GT) become evident: these $\bar{P}$-R curves estimate a relatively higher performance of the detector and seem to act as generous estimations of the upper bound on the detector's performance derived from the individual annotations. The Any-GT appears to act as a lower bound on the performance of the individual annotators and when sufficient annotations are available (Fissure and Landslide) the results obtained using STAPLE-GT and LSML-GT appear to approximately model the mean performance obtained using the inlying individual annotations. It should be noted, however, that the LSML technique is highly dependent upon the initial estimation.

Similarly, for the Landslide case study (Figure 6c) 0.5-GT, 0.75-GT and SIMPLE-GT yield $\bar{P}$-R curves at the upper bound of the performance range, STAPLE and LSML tend to produce GTs that model the performance within the bounds the individual annotations, and Any-GT marks the lower bound of the detectors performance. Overall it can be observed that the lower annotator variance leads to a significantly lower spread of the $\bar{P}$-R curves.

On the contrary, for the STARE dataset (Figure 6d) the LSML-GT forms a lower bound on the reported performance. The STAPLE-GT (equal to the 0.5-GT and the Any-GT) marks

the mean of the obtained curves, whereas previously (but to a lesser extent in the Segmentation case study) the STAPLE-GT and LSML-GT represented a mean estimate of the performance measured using the individual annotations. Once more 0.75-GT results in a higher estimate of performance than that obtained using each of the individual annotations.

V. 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.

A. Summary of Results

It has been shown that the performance of classifiers and detectors increases as GTs are formed using increasingly higher agreement levels. Forming a GT using an agreement of 50% generally increases a detector's reported performance to a range far greater than that obtained using all of the individual annotations. Kauppi et al. [4] conclude that the intersection method (consensus) is preferential as it results in the highest performance. Nevertheless, this study gives indication that the method focusses on evaluating a detector against the most obvious segments in the image and provides overly optimistic performance estimates. Raising the level of agreement at which the GT is calculated increases 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 which is reflected in the tight spread of the performance curves and in the stability of the detector ranking. Hence 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. [32] who found a large amount of agreement by comparing the regions that the segmentations contain and not the outlines themselves. This also affected the gold-standard GT 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. In all but the Fissure case study this method also reported similar performances to that obtained using the STAPLE and LSML algorithms.

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 [5]. The SIMPLE algorithm was proposed to overcome these limitations in situations in which annotator uncertainty varies considerably [5], 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 segment increases and those detectors resulting in the lowest drop in correlation (from CCO to CCI) result in a tighter spread of $\bar{P}$-R curves. This seems intuitive as agreement should be higher for more obvious segments and, assuming that the detector is effective, these should also elicit the highest detector responses. This translates to increasingly higher $\bar{P}$-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 segment 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 segments 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 segments in the image, which the detector considers to be true positives and feeding back these locations to the annotators for confirmation, could be a way of improving the GT 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 capture only 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 segments—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 the Landslide 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 GTs 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 are the most obvious in the Blood Vessel case study, in which the detector that produces the worst correlation

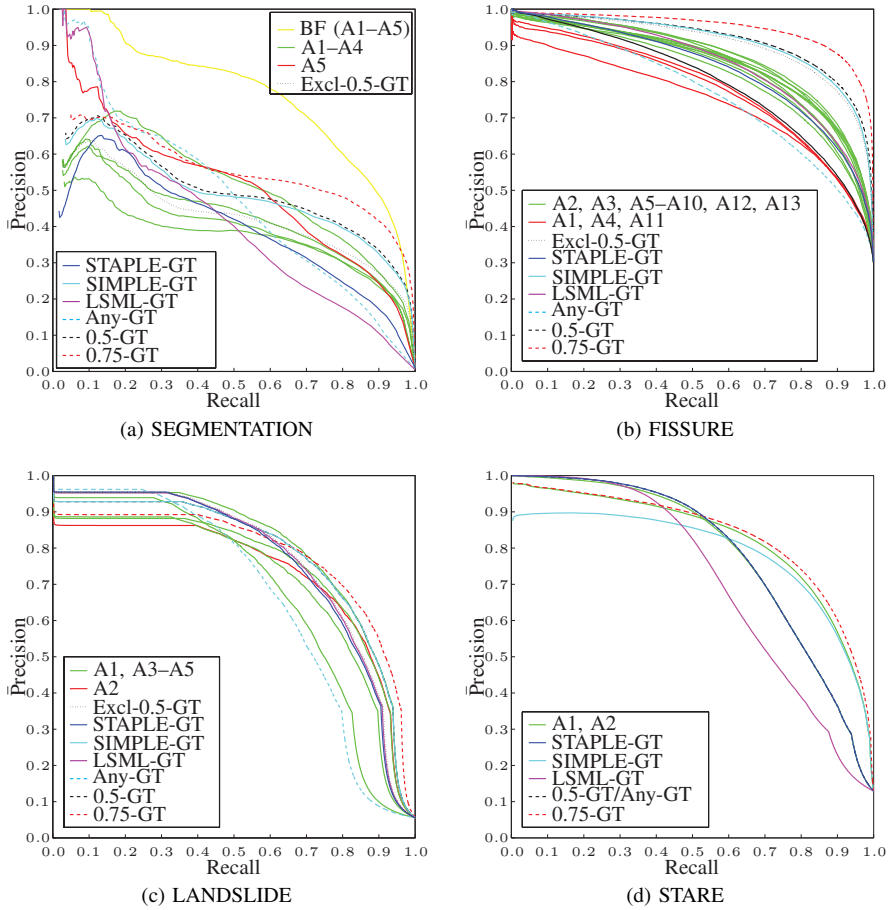


Fig. 6. $\bar{\text{P}}$recision-recall curves with differing estimated ground truths for (a) the REN detector (b) the 2D GWLC detector (c), RF (0.5-GT and SIMPLE-GT are identical) (d), GMM (the curve obtained using STAPLE-GT overlaps that obtained using 0.5-GT and Any-GT, 0.5-GT, and Excl-0.5-GT are identical).

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 Vb and Vc). In fact, it can produce a ranking that has nothing in common with these individual rankings (Table Vb).

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.

Finally, 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 a very optimistic estimate of the algorithm performance when compared to the traditional precision-recall evaluation framework.

B. 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 [33] states "Without

knowing GT 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 segments the algorithms are evaluated. Furthermore, in some domains correlation strengths between annotator agreement and image features decrease when moving from within segment locations to the whole image. Demonstrating that these properties are not uniquely tied to the segments of interest and employing this source of information risks introducing false positive locations to the inferred GT.

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 [51] 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 [52].

It is clear that evaluating upon different GTs, 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 GTs. While the presented study does not offer an ultimate solution for how those GTs should be combined the described analysis framework provides means to quantify the spread of measured performance and test 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 with certainty whether one algorithm outperforms another and further studies with more annotations should be conducted.

Considering that in all of the evaluated datasets the Any-GT and high agreement level GTs (0.5-GT or 0.75-GT) appear to model the lower and upper bounds (respectively) on the spread of measured performance 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.

VI. 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 segments 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 segmentation algorithms, 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.

ACKNOWLEDGEMENT

This work is part of the FOSTER project, 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

[1] S. Warfield, K. Zou, and W. Wells, "Validation of image segmentation by estimating rater bias and variance," *Phil. Trans. R. Soc. A*, vol. 366, no. 1874, pp. 2361–2375, 2008.

[2] A. Biancardi and A. Reeves, "TESD: A novel ground truth estimation method," in *Medical Imaging 2009: Computer-Aided Diagnosis*, vol. 7260, February 2009, pp. 72 603V–72 603V–8.

[3] M. C. Burl, U. M. Fayyad, P. Perona, and P. Smyth, "Automated analysis of radar images of Venus: Handling lack of ground truth," in *ICIP*, vol. 3, 1994, pp. 236–240.

[4] T. Kauppi, J.-K. Kamarainen, L. Lensu, V. Kalesnykiene, I. Sorri, H. Kälviäinen, H. Uusitalo, and J. Pietilä, "Fusion of multiple expert annotations and overall score selection for medical image diagnosis," in *Image Analysis*, ser. LNCS. Springer, 2009, vol. 5575, pp. 760–769.

[5] T. Langerak, U. van der Heidean, A. Kotte, M. Viergever, M. van Vulpen, and J. Pluim, "Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE)," *IEEE Trans. Med. Imag.*, vol. 29, no. 12, pp. 2000–2008, 2010.

[6] X. Li, B. Aldridge, R. Fisher, and J. Rees, "Estimating the ground truth from multiple individual segmentations incorporating prior pattern analysis with application to skin lesion segmentation," in *ISIB*, 2011, pp. 1438–1441.

[7] P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labelling of Venus images," in *NIPS*, 1994, pp. 1085–1092.

[8] S. Warfield, K. Zou, and W. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, 2004.

[9] P. Arbelaéz, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. PAMI*, vol. 33, no. 5, pp. 898–916, 2011.

[10] S. A. et al., "The lung image database consortium (LIDC) and image database resource initiative (IDRI) : A completed reference database of lung nodules on CT scans," *Medical Physics*, vol. 38, pp. 915–931, 2011.

[11] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piece-wise threhsold probing of a matched filter response," *IEEE Trans. Med. Imag.*, vol. 19, no. 3, pp. 203–210, 2000.

[12] T. Lampert and S. O'Keefe, "A detailed investigation into low-level feature detection in spectrogram images," *Pattern Recognition*, vol. 44, no. 9, pp. 2076–2092, 2011.

[13] M. Sabuncu, B. Yeo, K. V. Leemput, B. Fischl, and P. Golland, "A generative model for image segmentation based on label fusion," *IEEE Trans. Med. Imag.*, vol. 29, no. 10, pp. 1714–1729, 2010.

[14] X. Artaechevarria, A. Munoz-Barrutia, and C. O. de Solorzano, "Combination strategies in multi-atlas image segmentation: application to brain MR data," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1266–1277, 2009.

[15] I. Isgum, M. Staring, A. Rutten, M. Prokop, M. Viergever, and B. van Ginneken, "Multi-atlas-based segmentation with local decision fusion–application to cardiac and aortic segmentation in CT scans," *IEEE Trans. Med. Imag.*, vol. 28, no. 7, pp. 1000–1010, 2009.

[16] H. Wang, J. Suh, S. Das, J. Pluta, C. Craige, and P. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE Trans. PAMI*, vol. 35, no. 3, pp. 611–623, 2013.

[17] P. Coupé, J. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. Collins, "Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation," *NeuroImage*, vol. 54, no. 2, pp. 940–954, 2011.

[18] F. Xing, S. Soleimanifard, J. Prince, and B. Landman, "Statistical fusion of continuous labels: identification of cardiac landmarks," in *Proc. SPIE Medical Imaging 2011: Image Processing*, vol. 7962, 2011.

[19] O. Commowick and S. Warfield, "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*, 2010, pp. 25–32.

[20] A. Asman and B. Landman, "Characterizing spatially

varying performance to improve multi-atlas multi-label segmentation," in *Proc. of the 22nd int. conf. on Information processing in medical imaging*, 2011, pp. 85–96.

[21] ——, "Formulating spatially varying performance in the statistical fusion framework," *IEEE Trans. Med. Imag.*, vol. 31, pp. 1326–1336, 2012.

[22] O. Commowick, A. Akhondi-Asl, and S. Warfield, "Estimating a reference standard segmentation with spatially varying performance parameters: Local MAP STAPLE," *IEEE Trans. MI*, vol. 31, no. 8, pp. 1593–1606, 2012.

[23] A. Asman and B. Landman, "Robust statistical label fusion through COnsensus level, Labeler Accuracy, and Truth Estimation (COLLATE)," *IEEE Trans. Med. Imag.*, vol. 30, pp. 1179–1794, 2011.

[24] B. Landman, J. Bogovic, and J. Prince, "Simultaneous truth and performance level estimation with incomplete, over-complete, and ancillary data," in *Proc. SPIE Medical Imaging 2010: Image Processing*, vol. 7623, 2010.

[25] B. Landman, A. Asman, A. Scoggins, J. Bogovic, F. Xing, and J. Prince, "Robust statistical fusion of image labels," *IEEE Trans. MI*, vol. 31, no. 2, pp. 512–522, 2013.

[26] J.-K. Kamarainen, L. Lensu, and T. Kauppi, "Combining multiple image segmentations by maximizing expert agreement," in *Proc. of the 3rd Int. Workshop on Machine Learning in Medical Imaging*, 2012, pp. 193–200.

[27] H.-F. Yang and Y. Choe, "Ground truth estimation by maximizing topological agreements in electron microscopy data," in *Proc. of the 7th Int. Conf. on Advances in visual computing*, 2011, pp. 371–380.

[28] A. Asman and B. Landman, "Non-local STAPLE: An intensity-driven multi-atlas rater model," in *Proc. of the 15th Int. Conf. on Med. Image Computing and Computer-Assisted Intervention*, vol. 3, 2012, pp. 426–434.

[29] ——, "Non-local statistical label fusion for multi-atlas segmentation," *Medical Image Analysis*, vol. 17, no. 2, pp. 194–208, 2013.

[30] X. Liu, A. Montillo, E. Tan, and J. Schenck, "iSTAPLE: improved label fusion for segmentation by combining STAPLE with image intensity," in *Proc. SPIE Medical Imaging 2013: Image Processing*, vol. 8669, 2013.

[31] A. Asman and B. Landman, "Simultaneous segmentation and statistical label fusion," in *Proc. SPIE Medical Imaging 2012: Image Processing*, vol. 8314, 2012.

[32] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV*, 2001, pp. 416–423.

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

[34] H. He and E. Garcia, "Learning from imbalanced data," *IEEE Trans. KDE*, vol. 21, no. 9, pp. 1263–1284, 2009.

[35] S. Saur, H. Alkadhi, and et al., "Effect of reader experience on variability, evaluation time and accuracy of coronary plaque detection with computed tomography coronary angiography," *Eur. Radiol.*, vol. 20, no. 7, pp. 1599–1606, 2010.

[36] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *ICML*, 2006, pp. 233–240.

[37] P. Flach, "The geometry of ROC space: understanding machine learning metrics through ROC isometrics," in *ICML*, 2003, pp. 194–201.

[38] T. Lampert and P. Gançarski, "The bane of skew: Uncertain ranks and unrepresentative precision," *Machine Learning*, vol. 97, no. 1–2, pp. 5–32, 2014.

[39] U. Niethammer, M. James, S. Rothmund, J. Travelletti, and M. Joswig, "UAV-based remote sensing of the Super-Sauze landslide: Evaluation and results," *Eng. Geol.*, vol. 128, no. 1, pp. 2–11, 2011.

[40] A. Stumpf, J.-P. Malet, N. Kerle, U. Niethammer, and S. Rothmund, "Image-based mapping of surface fissures for the investigation of landslide dynamics," *Geomorphology*, vol. 186, pp. 12–27, 2013.

[41] A. Stumpf, N. Lachiche, N. Malet, J.-P. Malet, N. Kerle, and A. Puissant, "Active learning in the spatial domain for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. PP, no. 99, pp. 1–16, 2013.

[42] X. Ren and L. Bo, "Discriminatively trained sparse code gradients for contour detection," in *NIPS*, 2012, pp. 593–601.

[43] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik, "Using contours to detect and localize junctions in natural images," in *IEEE Conf. CVPR*, 2008, pp. 1–8.

[44] X. Ren, "Multi-scale improves boundary detection in natural images," in *ECCV*, 2008, pp. 533–545.

[45] D. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Trans. PAMI*, vol. 26, no. 5, pp. 530–549, 2004.

[46] J. Soares, J. Leandro, R. Cesar-Jr., H. Jelinek, and M. Cree, "Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification," *IEEE Trans. Med. Imag.*, vol. 25, no. 9, pp. 1214–1222, 2006.

[47] A. Stumpf, T. Lampert, J.-P. Malet, and N. Kerle, "Multi-scale line detection for landslide fissure mapping," in *IGARSS*. IEEE, 2012, pp. 5450–5453.

[48] V. Vonikakis, I. Andreadis, and A. Gasteratos, "Fast centre-surround contrast modification," *IET Image Process.*, vol. 2, no. 1, pp. 19–34, 2008.

[49] A. Liaw and M. Wiener, "Classification and regression by randomForest," *Rnews*, vol. 2, pp. 18–22, 2002.

[50] B. Malamud, D. Turcotte, F. Guzzetti, and P. Reichenbach, "Landslide inventories and their statistical properties," *Earth Surface Processes and Landforms*, vol. 29, pp. 687–711, 2004.

[51] R. Cooke, *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford Uni. Press, 1991.

[52] W. Aspinall, "A route to more tractable expert advice," *Nature*, vol. 463, pp. 294–295, 2010.