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 who mark feature locations within images 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 relation between the two. Several ground truth 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 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 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 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 at all possible). For example, determining gold-standard GT in remote sensing experiments would typically require field surveys over large and sometimes remote areas and for medical scans difficulties arise since it would require invasive surgery. 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 annotator opinion affect an algorithm's evaluation?

This work intends to quantify the effects of GT variability on the design, training, and evaluation of segmentation algorithms. To this end, supervised and unsupervised algorithms are evaluated in four case studies, all of which embody typical computer vision problems: 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 (the gold-standard GT) cannot be deduced from the imagery alone and annotations by human experts are used as the best available approximation. 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 [1, 2, 3, 4, 5, 6, 7, 8]. 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 how different GTs and estimated gold standards affect the performance comparison of 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 reporting an algorithm's performance. Relying on one annotator's opinion allows an algorithm to learn the annotator's bias, and 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

Manuscript received September 28, 2014; revised October 26, 2015 and March 15, 2016; accepted March 17, 2016. This work was supported by the French Research Agency through the COCLICO Project: ANR Modèles Numéeriques Program under Grant ANR-12-MN-001-COCLICO 2012–2016.

T. A. Lampert and P. Gançarski are with the ICube Laboratory, University of Strasbourg, France (e-mail: tlampert@unistra.fr).

A. Stumpf is with the Laboratoire Image, Ville, et Environment (LIVE), University of Strasbourg, France.

tightly controlled conditions or are synthetically generated [12] because the gold-standard GT is trivial to calculate. In remote sensing and medical imaging problems, and those concerning natural images, 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 connected pixels) and 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. This 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 demonstrate any bias that results from evaluating an algorithm with a single annotation;
- quantify the effect that different GTs may have on the evaluation of multiple algorithms;
- and provide a general comparison between algorithms designed to infer the gold-standard GT.

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

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.

More recently a number of methods for combining multiple image annotations are proposed. These include work 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 it results in the highest detector performance. 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 is proposed by Warfield et al. [8], named simultaneous truth and performance level estimation (STAPLE) in which annotator performance (measured as 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 their sensitivity and specificity [1] and another variant that accounts 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 for 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 segments 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) 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 inside of the GT segmentation's 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 features-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 in-

clude 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, 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 yet been investigated.

III. Methodology

To recapitulate, this work aims to demonstrate the effects of GT variability on the design, training, and evaluation of segmentation algorithms by studying their performance measured using single annotations, comparing multiple algorithms using different GTs, and comparing gold-standard GT inference algorithms. To achieve these aims, the methodological evaluation will be centred around three aspects: annotator agreement; the 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 specific to each study. All case studies are binary detection problems and each annotation has the value one where the annotator perceived the feature-of-interest to exist and zero otherwise. The result of this process is are N binary maps describing the location of the features-of-interest 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 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 feature-of-interest, 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 closely related to the entropy of the annotators' decisions. The maximum value for an acceptable level of experimental data quality suggested by the author is $10\,\%$.

Also to this end, the per-pixel annotator agreement is calculated, which is simply the number of annotators that have marked each pixel, such that

$$A(x,y) = \sum_{n=1}^{N} M_n(x,y). \quad (2)$$

The ratio of the number of pixels that are have a minimum level of agreement to the number of pixels that belong to annotated regions can therefore be calculated as follows

$$\hat{A}(n) = \frac{1}{|C|} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} \chi_B(x,y) \quad (3)$$

where $B = \{(x,y) \mid A(x,y) \geq n\}$, χ_B is the indicator function, $C = \{(x,y) \mid A(x,y) > 0\}$, and $1 \leq n \leq N$ is the range of values for the minimum level of agreement.

These functions allow for the testing of correlations between annotator agreement and different image properties—a means to uncover at least part of the reason behind the variance of agreement. Each dataset presents different features, but where applicable the following 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 by substituting $I(x, y)$ for $L(x, y)$. For the comparison of contrast and agreement the maximum agreement within the local neighbourhood is used.

Ground truths at different levels of agreement are calculated such that

$$\gamma_\tau(x,y) = \begin{cases} 1 & \text{if } \frac{1}{N} A(x,y) \geq \tau, \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

where τ represents the level of annotator agreement. Additionally, a number of the gold-standard ground-truth estimation methods are evaluated. These weight annotations based upon the assumption that 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 annotator markings is performed. The F_1-score [34], calculated between participants i and j, is defined as

$$F_{ij} = 2\frac{p_{ij} r_{ij}}{p_{ij} + r_{ij}}, \quad (6)$$

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

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 (image pixels that $50\,\%$ or more of the annotators marked as containing a relevant feature), calculated using Eq. (5) where $\tau = 0.5$. This is achieved by calculating:

Sensitivity, which measures the proportion of positives that are correctly identified as such;

Specificity, which measures the proportion of negatives that are correctly identified as such;

Positive Predictive Value (PPV), which measures the proportions of positives that are true positives;

Negative Predictive Value (NPV), which measures the proportions of negatives that are true negatives;

Cohen's kappa coefficient, which measures the inter-rater agreement correcting for agreement that occurs by chance.

C. Relation between Agreement and Detector Performance

After analysing the properties of annotator agreement, it follows to investigate its relation to detector performance. Therefore four detectors are selected from each of the case study domains and applied to the detection problem at hand (every effort was made to select the best performing detectors within each domain). Each of the detectors is evaluated using GTs calculated at increasing levels of agreement according to Eq. (5), where $\tau = 1/N, 2/N, \ldots, N/N$.

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 (those in which the number of positive, N_p, and negative, N_n, examples are not balanced) and therefore precision-recall (P-R) curves are preferable [36, 34]. Nevertheless, precision is sensitive to the skew ratio, $\phi = N_p/N_n$. To overcome this Flach [37] proposes to analytically vary the skew ratio 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 skew ratios, i.e. GTs derived from different levels of agreement, to be compared and for a fair representation of detector performance in problems in which the skew ratio is a priori unknown. The 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 a threshold on the detector's output, π'_1 and π'_2 are the lower and upper bounds of the problem's estimated range of skew ratios, and $\text{TP}(\theta)$ and $\text{FP}(\theta)$ are the number of true positive and false positive detections. Interpolation between $\bar{\text{P}}$-R points [38] enables accurate area under the curve (AUC$\bar{\text{P}}$R) measurements to be taken.

To assess the relation 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 features by any of the annotator (CCO) and the second the whole image (CCI). The first of these highlights the relation between the detector output and annotator agreement in positive feature locations. 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 are indicative of a detector's reliability.

D. Ground Truths and Reported Detector Performance

The final question that this research intends to investigate is: how great is the influence of different ground truths on an algorithm's reported performance?

To this end several GTs are calculated according to Eq. (5): 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 0.5-GT as an initial estimate and 1000 iterations). Furthermore, an additional GT is determined by excluding outlying annotations (these will be identified in Section III-B) and combining those remaining according to Eq. (5) with $\tau = 0.5$ (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 values of the $\bar{\text{P}}$-R curves.

IV. Experimental Results and Analyses

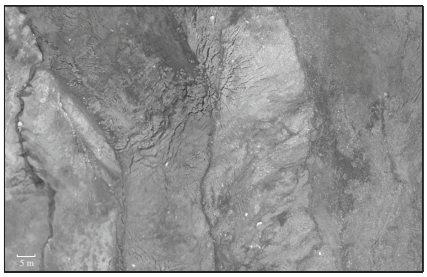
This section presents the results of applying the described methodology to each of the case studies included in this investigation.

A. Data

The case studies presented in this section are concerned with:

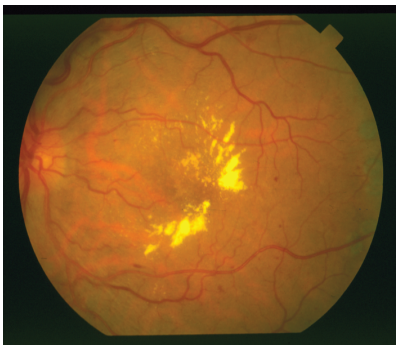
Image segmentation Most of the images within the Berkeley 300 (colour) dataset have been annotated by numerous different annotators. Only for a small subset of five images did the same annotators perform the segmentation (annotator IDs for the Berkeley 500 dataset are not available). These images are: 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.

(a) Segmentation



(b) Fissure

(c) Landslide



(d) Blood Vessel

Fig. 1. Images used in the four case studies, (a) and (d) are random examples taken from the dataset.

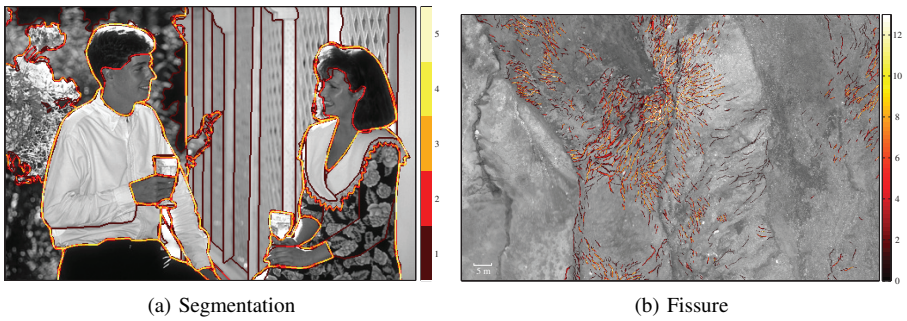
Fissures in remotely sensed images The data is obtained from the Super-Sauze landslide in the Barcelonnette basin, southern French Alps, using an unmanned aerial vehicle to obtain high resolution images. Further information regarding this dataset is present in the literature [39, 40]. An area of interest, 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 to fissure characteristics. The annotators then independently marked all the pixels that they believed to form part of a fissure, taking as much time as they required (this ranged from 2 to 3 hours). 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 to assess 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, green, red, and near infra-red) and a nominal ground resolution of $50\,\mathrm{cm}$. The image presented in Figure 1c was captured at Nova Friburgo, Brazil shortly after a major landslide event in January 2011 and covers approximately $10\,\mathrm{km}^2$ ($X = 5960$ 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 to independently mark the outlines of the regions affected by landslide activity. To achieve this, the RGB components of the pre-event and the post-event satellite images were visualised using a natural color scheme. Detailed information regarding this dataset exists in the literature [41].

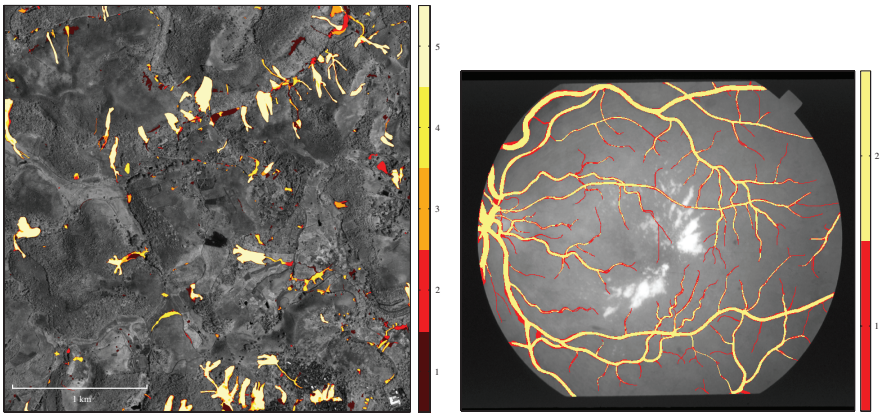
Retinal blood vessels The STructured Analysis of the Retina (STARE) dataset was used in this case study. The dataset consists of twenty colour retinal images, which for the purposes of this study are treated as a single image ($X = 2800$ and $Y = 3025$). An example image is presented in Figure 1d. A 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 blood vessels in the image.

B. Annotator Agreement

The pixel-level annotator agreements for each case study are presented in Figure 2. To verify that these are acceptable for experimental use Smyth's lower error bound estimate, i.e. the average error rate amongst the annotators, was calculated and found to be $\bar{e} \geq 2.6611\,\%$ (Segmentation), $\bar{e} \geq 1.26\,\%$ (Fissure), $\bar{e} \geq 1.1012\,\%$ (Landslide), and $\bar{e} \geq 3.1123\,\%$ (Blood Vessel). These values are well within the $10\,\%$ limit



(a) Segmentation (b) Fissure



(c) Landslide (d) Blood Vessel

Fig. 2. Pixel-level annotator agreement in each case study, calculated according to Eq. (2). Colour describes the level of agreement on the location of the case study's targeted feature in the image. The images have been converted into grey-scale to better represent agreement.

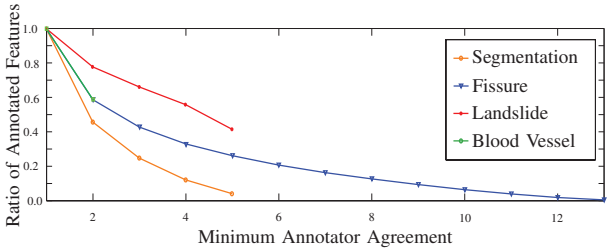


Fig. 3. Ratio of the number of pixels having a minimum level of annotator agreement to the number of pixels that belong to annotated regions as the level of agreement increases, calculated according to Eq. (3).

that is recommended [33] and considerably lower than the error bound of 20 % found in the volcano labelling experiment presented by the author [33], in which the signal-to-noise ratio of the features is much lower than in the presented case studies.

The ratio of pixels having a minimum level of annotator agreement to the number of pixels that belong to annotated regions is presented in Figure 3. For the Segmentation, Fissure and Blood Vessel case studies the ratio decreases approximately exponentially as a function of minimum annotator agreement. For the Fissure dataset the thirteen annotators agree on only approximately 0.6979 % of all of the pixels that were marked as fissures by any of the annotators. The ratio decreases most rapidly in the Segmentation case study, whereas the Landslide case study exhibits a rather linear trend. These differences are due to a combination of the geometric structure of the targeted objects, and the fact that disagreement

TABLE I
PEARSON'S r CORRELATION COEFFICIENTS BETWEEN IMAGE FEATURES AND AGREEMENT. CORRELATIONS IN ITALICS 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) Blood Vessel

tends to occur along object borders. As such, uncertainties in the outline of a feature lead to a stronger disagreement if the targeted features are only one pixel wide (Segmentation) or several pixels wide (Fissure, Blood Vessel) when compared to the rather blob-like regions exhibited in the Landslide case study. Indeed, if the outlines of the Landslide annotations are analysed, agreement also drops approximately exponentially.

The correlation coefficients between annotator agreement

and image properties are presented in Table I. These offer an explanation for the relation between detector performance and annotator agreement that will be explored in the remainder of this paper, i.e. stronger image features tend to be more confidently detected by a detection algorithm, and also attract higher levels of annotator agreement. In each case study there exists at least one significant correlation between image properties and agreement: contrast in the Segmentation case study indicating that the annotators tend to agree on stronger edges; contrast and intensity in the Fissure case study 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 because the reddish soil is exposed; and green in the Blood Vessel case study, which is the principal channel used for discrimination in many blood vessel detection studies.

Dendrograms describing the annotator pairwise F_1-scores in each case study are presented in Figure 4 and the full statistics of each annotator compared to the average annotation (Eq. (5), $\tau = 0.5$) 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 used to form the dendrogram in Figure 4a. The differences are relatively high, ranging from 0.545 to 0.680, and one outlier is identified: A5 (the mean F_1-score difference was found to be 0.6016, with a standard deviation of 0.0280, and A5 resulted in a difference of 0.6454). This annotator also results in the lowest specificity, positive predictive value, and kappa coefficient as shown in Table IIIa. The variance in the annotations are emphasised by the lowest specificities observed in all of the case studies. A dendrogram describing the Blood Vessel 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 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. Nevertheless, we can infer 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 with which these statistics are calculated 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 case study, Figure 4b, partitioning 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. These same annotators achieve considerably lower sensitivity when compared to the consensus (see Table IIIb). They also result in 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.

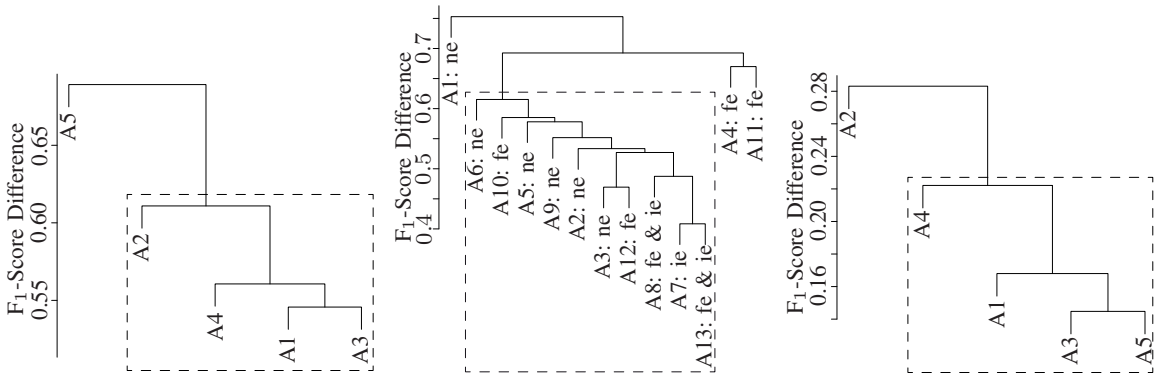
In the Landslide case study, each of the annotators were geographers familiar with the detection of landslides in remotely sensed imagery. This is reflected in the low inter F_1-score difference $(1 - F_{ij})$, which ranges from 0.14 to 0.28 (by comparison this range was approximately 0.40 to 0.75 in 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 features in the area of interest was evaluated by calculating $\bar{\mathrm{P}}$-R curves:

Segmentation The top four performing segmentation algorithms listed on the Berkeley dataset web page[2] were selected to form part of this case study. These were: REN [42], gPb-ucm (UCM) [9], Global Probability of Boundary (GP) [43], and XREN [44]. The integration limits of the $\bar{\mathrm{P}}$-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 skew found within the Berkeley dataset and σ its standard deviation [38]. As discussed by Martin et al. [45], when evaluating segmentation algorithms it is common to loosen the definition of true-positive detections to account for deviations in detected boundary location. True-positive detections are accumulated if a detection is within a defined distance of one or more 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 randomly selected for use 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 ($\tau = 1/N$, for example) result in multiple pixel wide segmentations (as annotators may agree upon the boundary's existence, but not on its exact location), which causes an unfair

[2]http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/bench/html/algorithms.html, accessed 23rd October 2015



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

Fig. 4. Dendrograms describing the F_1-score difference between each annotation for the Segmentation, Fissure, and Landslide case-studies (only two annotations are present in the Blood Vessel case study and therefore this analysis cannot be completed). Fissure case study dendrogram 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.

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) Blood Vessel

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 boundary widths to one pixel whilst preserving any individual, low agreement markings.

Fissure Current state-of-the-art linear feature detectors were selected from the literature: a linear classifier trained using 2D Gabor wavelet (elongation $\epsilon = 4$, scales $a = 2, 3, 4, 5$, and frequency $k_0 = 3$) and inverted grey-scale features (2D GWLC) [46]; Gaussian filter matching, where $\sigma = 1$ [47] (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. (5) 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 were set to $\pi_1' = 0.1$ 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.

Landslide Four 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). 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[3] 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

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

selected square subset covering $10\,\%$ of the area of interest. The number of trees in the RF was fixed to 500 and 10 variables were tested for the splits at each node. The SVM used a radial basis kernel having parameters $C = 10$ and $\sigma = 0.004$, determined through an exhaustive grid search. The ANN was a 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. Parameter tuning was performed through bootstrap resampling of the training data using 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 to unaffected areas after large-scale landslide triggering events [50].

Blood Vessel 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], the features were taken to be the inverted green channel, and the responses of Gabor wavelets (elongation $\epsilon = 4$, scales $a = 2, 3, 4, 5$, and frequency $k_0 = 3$) applied to the inverted green channel. The first five images of the dataset (im0001–5) were used exclusively for training. The $\bar{\text{P}}$-R curve integration limits 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 was 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 5. A striking observation is that the performance of all detectors increases with annotator agreement in a predictable manner in the higher recall ranges. It was shown in Table I that there is a tenancy for annotators to agree upon more obvious image features, and these results indicate that the detectors extract similar features. 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 most likely needs to be augmented with high-level information that is not exploited by the evaluated detectors. In the lower recall ranges of the Segmentation, Landslide, and Blood Vessel case studies the tendency for precision to increase with agreement is reversed. This phenomenon can be explained by analysing the correlations between annotator agreement and detector output presented in Table III.

Several general tendencies can be drawn from these correlations. The detectors that exhibit a large drop between CCO and CCI also exhibit low sensitivity (i.e. produce a high false-positive rate). For example, this is reflected in the $\bar{\text{P}}$-R curves of the C-S detector (Figure 5h): low sensitivity dominates the low agreement ground truths (for example, $\tau = 1/13$), but the detector results in the highest performance when using the high agreement ground truths (for example, $\tau = 1$). 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, & GMM detectors for example), have (relatively) low false positive rates and result in high $\bar{\text{P}}$-R curves (Figures 5e, 5f, 5j, & 5p). A large drop in correlation, along with a low absolute correlation, is observed with the Top-Hat detector, and indeed in Fig. 5g the curves are skewed towards lower precision values. The detectors that result in the lowest drop or an increase in correlation (2D GWLC, Gauss, RF, KNN, SVM, & ANN) result in a tighter spread of $\bar{\text{P}}$-R curves (Figures 5e, 5f, 5i, 5k, 5j, & 5l).

TABLE III
PEARSON'S r CORRELATION COEFFICIENTS BETWEEN DETECTOR OUTPUTS AND ANNOTATOR AGREEMENT $A(x, y)$ AS DEFINED BY EQ. (2); CCO IS CALCULATED WITHIN THE PIXELS MARKED AS A POSITIVE INSTANCE BY ANY OF THE ANNOTATORS, AND CCI THE WHOLE IMAGE. THE P-VALUES ARE ALL 0.0000 (TO FOUR DECIMAL PLACES).

Case Study	Detector	CCO	CCI	CCI−CCO
Segmentation	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
Fissure	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
Landslide	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
Blood Vessel	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

The $\bar{\text{P}}$-R curves from the Segmentation, Landslide and Blood Vessel case studies largely follow the trend: as agreement increases, algorithm performance also increases. There is, however, a tendency for precision to be inversely proportional to agreement in lower recall ranges. This phenomenon can be explained by analysing the correlations between annotator agreement and detector outputs presented in Table III and noting that in all the cases in which this trend is observed CCI is higher than CCO. This indicates that detector outputs agree with annotator agreement within feature locations and more so over the whole image, implying that there is a relatively low FP detection rate, which at the lower recall ranges results in high precision. As the threshold on agreement increases, image locations having increasingly stronger features form the GT and these locations also result in the highest detector responses. Furthermore, high CCI values imply that as lower agreement segments are removed from the GT they are instead classified as false positive detections, thus reducing precision in the lower recall ranges as the threshold on annotator agreement is increased.

D. Ground Truths and Reported Detector Performance

The detector ranks (measured as AUC$\bar{\text{P}}$R [38]) when evaluated using different GTs were determined, and in three of the case studies (Segmentation, Fissure, and Blood Vessel) three rankings emerged, which are described in Table IV. In the Landslide case study only one emerged due to the low inter-annotator variance. In the Fissure, Landslide, and Blood Vessel case studies these ranks reflect the results of the correlation

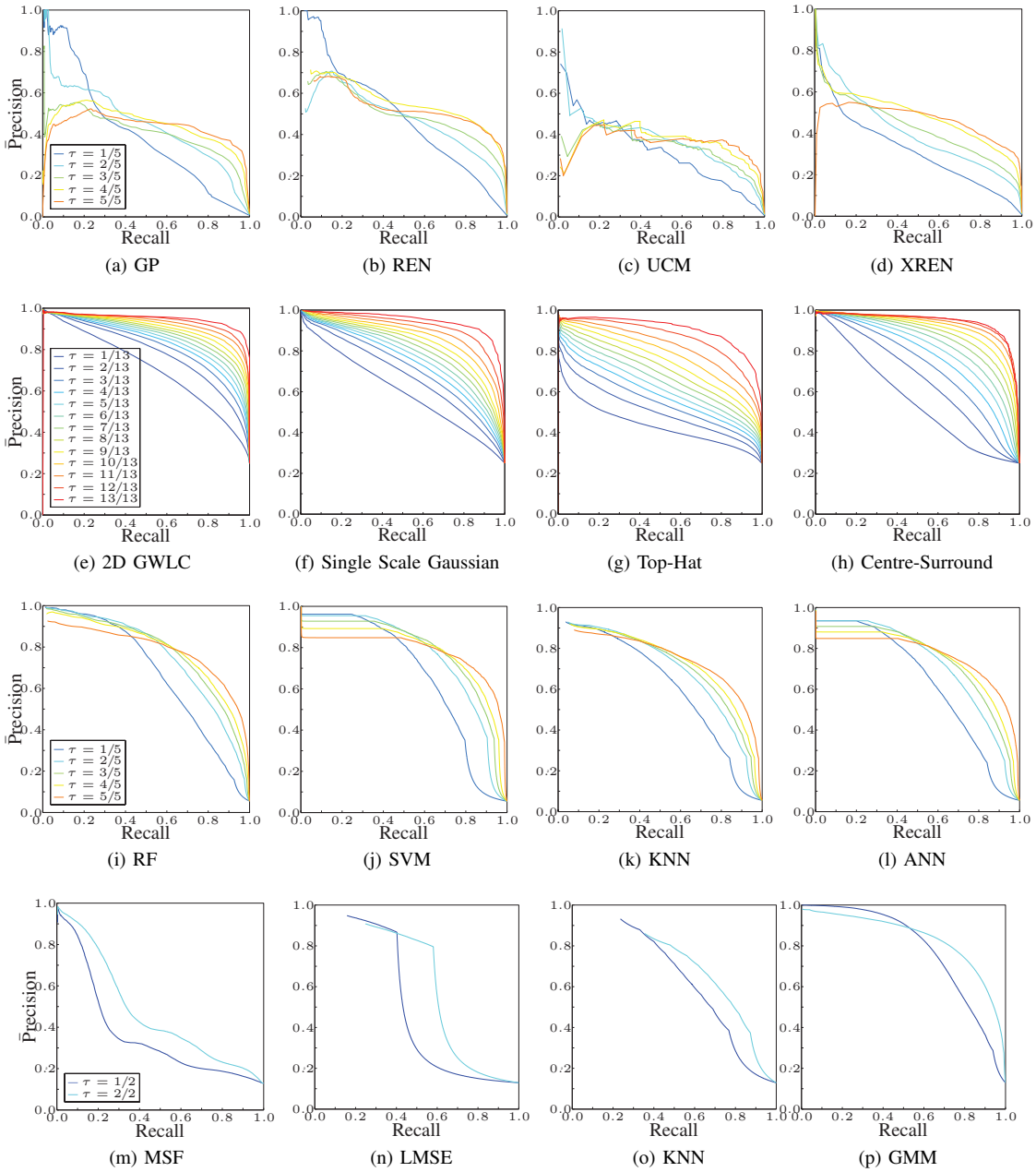


Fig. 5. P̄recision-recall curves describing detector performance. The curves in each subfigure are determined using ground truths calculated with increasing levels of agreement, according to Eq. (5). Figures 5a–5d are from the Segmentation case study, Figures 5e–5h from the Fissure case study, Figures 5i–5l from the Landslide case study, and Figures 5m–5p from the Blood Vessel case study.

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). Furthermore, in the Fissure case study the ranking observed is not determined by the level of annotator expertise. This corroborates the lack of distinction between different expertise levels in the dendrogram presented in Figure 4b.

In the Segmentation case study, however, the algorithms with the highest correlation (UCM) and the highest CCO to CCI increase (GP) are ranked at the middle or bottom. On one hand this can be attributed to the relatively high annotation variance and the overall low correlation between detector output and the annotator agreement (Table III). On the other hand it should also be considered that the correlations are derived using all of the annotated pixels, while the P̄-R curves are calculated using GTs that were thinned to a width of one pixel and the TP rates calculated with a tolerance to small mismatches of the segmentation boundary. This not only highlights the sensitivity of evaluating algorithms using different GTs that exhibit high variance, but also illustrates how different evaluation strategies can provoke different outcomes.

In the Fissure case study, a majority of the individual annotations give the same ranking as obtained using the

TABLE IV
RANKINGS OF DETECTORS EVALUATED USING EACH GROUND TRUTH (MEASURED BY THE AREA UNDER THE $\bar{\text{P}}$-R CURVE). (A) SEGMENTATION CASE STUDY, THE GTS THAT RESULT IN EACH RANKING 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) FISSURE CASE STUDY, THE GTS THAT RESULT IN EACH RANKING ARE: RANKING #1 — A2(NE), A4(FE), A6(NE), A11(FE), ANY-GT, LSML-GT, STAPLE-GT, EXCL-0.5-GT; RANKING #2 — A1(NE), A3(NE), A5(NE), A7(IE), A8(FE & IE), A9 (NE), A10(FE), A12(FE), A13(FE & IE), 0.5-GT, SIMPLE-GT; RANKING #3 — 0.75-GT. (C) LANDSLIDE CASE STUDY, ALL GTS RESULT IN THE SAME RANKING. (D) BLOOD VESSEL CASE STUDY, THE GTS THAT RESULT IN EACH RANKING ARE: RANKING #1 — A1, 0.75-GT, SIMPLE-GT; RANKING #2 — A2, 0.5-GT/ANY-GT, STAPLE-GT; RANKING #3 — LSML-GT.

Position	Ranking #1	Ranking #2	Ranking #3
1	REN	REN	REN
2	GP	GP	XREN
3	UCM	XREN	GP
4	XREN	UCM	UCM

(a) Segmentation

Position	Ranking #1	Ranking #2	Ranking #3
1	2D GWLC	2D GWLC	C-S
2	Gauss	C-S	2D GWLC
3	C-S	Gauss	Gauss
4	Top-Hat	Top-Hat	Top-Hat

(b) Fissure

Position	Ranking #1
1	SVM
2	RF
3	ANN
4	KNN

(c) Landslide

Position	Ranking #1	Ranking #2	Ranking #3
1	GMM	GMM	GMM
2	MSF	KNN	KNN
3	LMSE	MSF	LMSE
4	KNN	LMSE	MSF

(d) Blood Vessel

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 detector ranking. More importantly, the ranking derived using a 75 % voting strategy (Fissure) and LSML (Blood Vessel) are in disagreement with that obtained using the individual annotations, which contradicts what should be expected. To illustrate ranks in the Fissure case study, 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 6, each colour represents one of the rankings presented in Table IVb.

In the Blood Vessel case study the ranks of the lower three detectors are not consistent. The MSF detector, for example, achieves the lowest performance in Figure 5 along with the lowest correlation with annotator agreement (Table III), however, depending upon which GT is used, this detector is placed second, third, or last.

An overview of the performance variations that result from

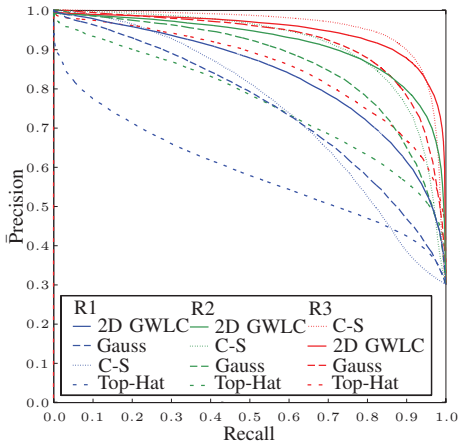


Fig. 6. $\bar{\text{P}}$recision-recall curves of all four detectors in the Fissure detection case study evaluated using: STAPLE-GT which results in Ranking #1 (R1); 0.5-GT which results in Ranking #2 (R2); and 0.75-GT which results in Ranking #3 (R3).

using different GT estimation methods and evaluation frameworks can be obtained from Figure 7, in which the $\bar{\text{P}}$-R curves obtained using the best performing detector in each of the case studies are presented. The $\bar{\text{P}}$-R curves for the REN segmentation algorithm (Figure 7a) exhibit the largest level of variance, a result of the high annotator variance observed in Section III-B. At the upper extreme of this variance is the methodology prescribed for evaluating segmentation algorithms upon the Berkeley datasets, which includes a tolerance for misalignments of TP detections. The 0.75-GT, 0.5-GT, and SIMPLE-GT yield higher performance curves (particularly in higher recall ranges) and Any-GT relatively low performance curves when compared to the remaining GTs. The STAPLE-GT and LSML-GT curves show low $\bar{\text{p}}$recision (when compared to the remaining curves) in the upper recall ranges, but model the mean of the individual annotation curves in the lower recall ranges. This is a consequence of the large variance observed in the annotations. The curves resulting from Excl-0.5-GT and SIMPLE-GT are very similar as they are both derived using the same principle (removing outliers and then voting).

The $\bar{\text{P}}$-R curves resulting from 2D GWLC are presented in Figure 7b. The effects of the voted GTs (0.5-GT, 0.75-GT, and SIMPLE-GT) become evident: these $\bar{\text{P}}$-R curves estimate a relatively high detector performance and seem to act as generous estimates of the upper bound of the performance derived from the individual annotations. Moreover, the Any-GT appears to act as an estimate of the lower bound of the performance derived from the individual annotations, and when sufficient annotations are available (Fissure and Landslide) the curves obtained using STAPLE-GT and LSML-GT appear to approximately model the mean of the performance obtained using the inlying individual annotations. It should be noted, however, that the LSML technique is highly dependent upon the estimate used for initialisation.

Similarly, in the Landslide case study (Figure 7c) 0.5-GT, 0.75-GT, and SIMPLE-GT yield $\bar{\text{P}}$-R curves that seem to model the upper bound of the performance obtained using the individual annotations, STAPLE and LSML tend to produce

GTs that result in $\bar{P}$-R curves that are within the range of those obtained using the individual annotations, and Any-GT marks the lower bound of the detector's performance. Overall it can be observed that the lower annotator variance observed in this case study leads to a significantly lower $\bar{P}$-R curve spread.

On the contrary, in the Blood Vessel case study (Figure 7d, due to the limited number of annotations the Any-GT, 0.5-GT, and Excl-0.5-GT are identical) the LSML-GT forms a lower bound on the reported performance. The STAPLE-GT (equal to the 0.5-GT and the Any-GT) delineates the mean of all the curves, whereas previously (but to a lesser extent in the Segmentation case study) the STAPLE-GT and LSML-GT represented an estimate of the mean of the curves obtained 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 summarises the results presented in the previous section and their implications, and the second 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 indicates 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.

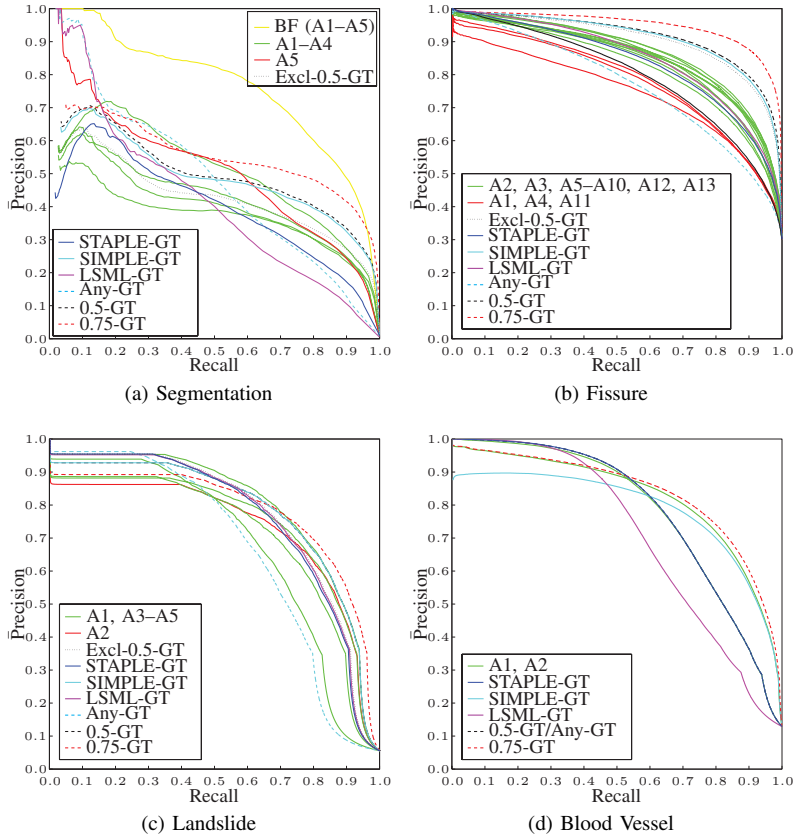
One factor that has a stabilising effect on reported performance is low annotation variance. The Landslide dataset contains the lowest variance between annotations and this is reflected in the tight spread of the performance curves and in the stability of the detector ranking. Hence choosing any of the 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 annotation variance, and the reported performances also exhibit the largest variance. This is in contrast to the findings of Martin et al. [32] who found a large amount of agreement between the segmented regions, but not the boundaries 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 derived using the individual annotations, in this case study they actually resulted in the lowest performance curves. Both of these methods combine annotations using 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 annotation 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 when 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).

The output of all of the detectors produced medium to high correlations with annotator agreement. 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 lower $\bar{P}$-R curve spread. 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 between 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—as agreement increases, lower-agreement segments are removed from the GT causing the detector to classify them as false positives. This could be an indication that some of the annotators have missed important segments, which the detector considers to be true positives, and providing these locations as feedback to the annotators for confirmation could be a way of improving 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. Further evidence for this is provided 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 study it has been shown that the rank of a detector is dependent upon the GT used for evaluation. It can therefore be stated that the variance in performance observed when evaluating two detectors using different GTs is not equal and therefore, the relative difference in performance between detectors is dependent upon the GT used for evaluation. Three different rankings were observed in



(a) Segmentation (b) Fissure

(c) Landslide (d) Blood Vessel

Fig. 7. $\bar{\text{P}}$recision-recall curves for one detector in each case study using different ground truth estimation methods: (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).

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 and had their performance been closer this would not have been the case. The effects are most 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 IVb and IVd). In fact, it can produce a ranking that has nothing in common with these individual rankings (Table IVb).

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 this domain, through accounting for variances observed in the annotations, yields a very optimistic estimate of 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".

Overly 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 assume that the gold-standard ground truth can be inferred by measuring the performance of annotators in relation to each other. The most promising advances have started to integrate information derived from the image into the process, and it has been shown herein that these properties do 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 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 a 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 for accurate evaluation. The Landslide case study, 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, as 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, 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.

Finally, to be able to use such an approach, and to understand annotator variance within standard evaluation datasets, it should be made possible to determine which annotations each annotator produced, and to ensure a sufficient coverage of the dataset by the same annotators.

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 outlier annotations, and it was found that the STAPLE and LSML algorithms find a balance between all annotations when their variance is low. Ground truths 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. 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 studies. It does, however, appear to be more stable when the annotations have high variance.

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 resulted in a detector rank that is 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 when compared to individual annotator opinion. Furthermore, under some conditions a detector whose output is poorly correlated with annotator agreement can be placed above those that have vastly better correlated outputs.

Therefore in addition to evaluating an algorithm over a data set that contains multiple images, it is concluded that an algorithm should also 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 differences between detectors. 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 detection algorithms, since measures of absolute performance and relative ranking of detectors may vary considerably according to the GT employed.

The possibility of estimating a detector's true performance through the variability of annotator opinion would be an interesting avenue to follow. Assuming that 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 avoid some of these.

An additional question that is raised by this study is: which metric should be used to evaluate an estimated gold standard? Generally speaking the gold standard is unknown and therefore comparison is impossible. Restricting evaluation to individual annotations assumes high specificity and sensitivity. Removing annotations, however, assumes inability compared to the consensus, but do those removed represent true insight into the problem? One thing is clear, detector performance should not be used to evaluate an estimated gold-standard ground truth.

ACKNOWLEDGEMENT

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 valida-

tion and label fusion: a maximum a posteriori STAPLE," in *Proc. of the 13th Int. Conf. on Med. Image Comput. Comput. Assist. Interv.*, 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," *Med. Image Anal.*, 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 Surf. Process. Landf.*, 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.