Combining Multiple Image Segmentations by Maximizing Expert Agreement

Joni-Kristian Kamarainen, Lasse Lensu, and Tomi Kauppi

Machine Vision and Pattern Recognition Laboratory
Department of Information Technology
Lappeenranta University of Technology
P.O. Box 20, FI-53851 Lappeenranta, Finland
lasse.lensu@lut.fi
http://www2.it.lut.fi/mvpr/

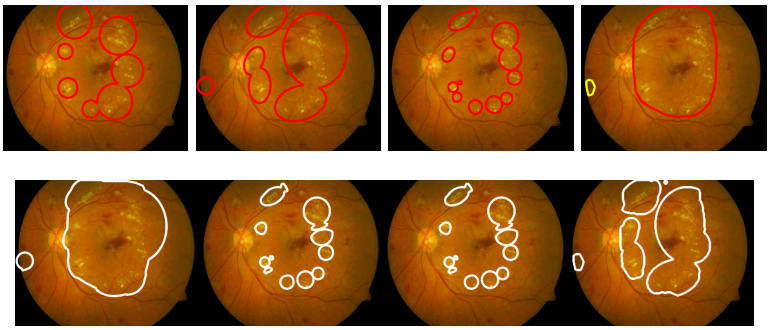
Abstract. A common characteristic of collecting the ground truth for medical images is that multiple experts provide only partially coherent manual segmentations, and in some cases, with varying confidence. As the result, there is considerable spatial variation between the expert segmentations, and for training and testing, the "true" ground truth is estimated by disambiguating (combining) the provided segments. STAPLE and its derivatives are the state-of-the-art approach for disambiguating multiple spatial segments provided by clinicians. In this work, we propose a simple yet effective procedure based on maximizing the joint agreement of experts. Our algorithm produces the optimal disambiguation by maximizing the agreement and no priors are used. In the experimental part, we generate a new ground truth for the popular diabetic retinopathy benchmark, DiaRetDB1, for which the original expert markings are publicly available. We demonstrate performance superior to the original and also STAPLE generated ground truth. In addition, the DiaRetDB1 baseline method performs better with the new ground truth.

1 Introduction

Image databases and expert ground truth are in common use in medical imaging research. The motivation for using expert ground truth arises from the fact that a variety of imaging modalities exist for medical examinations and generally the ground truth is not provided by a device. The common solutions to get spatial ground truth estimating the true segmentation of physiological details or lesions make use of synthetic images, physical or digital phantoms related to the imaging, or involve a group of experts to perform manual or semi-automatic segmentation. For studies striving for the true expert knowledge based on the visual evaluation of real *in vivo* images, the last approach is the best option since real images are used in clinical diagnosis and no special effort must be made to validate the synthetic data. In this case, the two relevant questions are the number of experts to use and how the ambiguous information represents the "truth". However, as the number of experts for the laborious manual work increases and the experts are qualified for medical diagnosis work, their joint

agreement should converge to the ground truth – to the gold standard in medical diagnosis which is based on the second opinion from a peer or several peers.

In many applications, method development is driven by publicly available and commonly accepted benchmark datasets. This practice enables the evaluation and comparison of image processing methods, and assessment of the state of the art. In a wider context, the practice is in accordance with other research fields, such as biometrics, where continuously developed benchmarks set pressure for developing better methods. However, to the authors' best knowledge, only the DiaRetDB1 benchmark dataset [1] provides the original expert segmentations used to generate the combined ground truth. This is surprising as a characteristic and preferred property of medical data for research purposes is that the provided ground truth originates from multiple experts and it should be possible to validate also that. Examples are shown in Figures 1. The expert segmentations vary significantly, and therefore, the expert information must be considered more as delineations than accurate manual segmentations, and there clearly exists the need for an appropriate combining (disambiguation) procedure.



(a) (b) (c) (d)

Fig. 1: 1st row: DiaRetDB1 expert markings for the lesion *Hard exudate* (red: high confidence, yellow: moderate, green: low). The disambiguated ground truth (white) produced by (a) minimal and (b) maximal confidence (Sec. 2.1). The disambiguated ground truth by (c) the proposed method and (d) STAPLE with default parameters.

For machine learning methods, the combined ground truth is required to represent the "truth" to be learned. However, it is the adopted disambiguation procedure which ultimately defines the truth. The authors of DiaRetDB1 proposed a heuristic method called "expert fusion", which has the good property that it was based on processing the whole image ensemble. The state-of-the-art approach is the STAPLE method by Warfield et al. [2] and its derivatives [3,4]. STAPLE is based on a probabilistic formulation of the segment combination problem with certain constancy priors and the Expectation-Maximization (EM) algorithm. The method performs well with toy and many true examples, but for multiple experts and varying confidences its performance is unclear. Our approach has the following good properties:

- The proposed method needs no prior models for the spatial distribution of structures or spatial homogeneity constraints.
- The proposed method does not estimate the "most likely segmentation" of a single image but the most likely segmentation of the whole image ensemble.
- The proposed method does not assume similar "behavior" of the experts, i.e., trained raters.

We propose a simple yet effective method to automatically find the optimal solution for a given data. We avoid complicated formulation requiring priors by using a simple guideline familiar to all clinicians: make the decision on a patient for which the peers (experts) jointly agree. The disambiguated ground truth corresponds to the maximal joint-agreement of experts, which is empirically demonstrated with DiaRetDB1 by presenting results superior to the original DiaRetDB1 method and STAPLE. In addition, the performance of the DiaRetDB1 baseline algorithm significantly improves when using our method.

2 Disambiguation method

Our disambiguation method stems from the evaluation practices in medical imaging, and the main application is automatic lesion detection/segmentation. Supervised methods require example images for training, validation, and testing and to extract meaningful features multi-valued ground truth must be correctly disambiguated. We adopt the following general guidelines:

- Ground truth is combined from the markings of N experts (raters) who have independently marked the images and whose opinions are equally important.
- The markings are represented as spatial segments.
- Each segment is associated with a single confidence value (e.g., low, moderate, or high) given by the raters.
- The main evaluation procedure is *image-based*, which supports the clinical practice (cf. patient-based diagnosis).
- Pixel-wise ground truth (spatial segments) is computed from the expert markings and is represented as a binary mask (*disambiguation*). [1]
- Image-wise ground truth is constructed from the binary masks and is binary (*true*/*false*) whether a specific lesion or finding is present or not.
- A detection algorithm must provide a "score value" for the all lesion types in each test image.
- The scores and image-wise ground truth are used to compute a ROC curve or other characteristic values for the comparison (e.g., the equal error rate).

The receiver operating characteristic (ROC) curve is a standard evaluation procedure used in computer vision, such as in the comparison of face recognition algorithms. For medical applications, the ROC axes directly correspond to the meaningful clinical measures specificity and sensitivity, and the equal error rate (EER) is a point on the ROC curve where the specificity and sensitivity are equal. If there is no prior knowledge of a desired sensitivity or specificity level, then the EER represents a neutral evaluation measure.

[1] Probability maps can be considered as the ideal result, but for straightforward training and testing, such masks should be finally converted into binary masks.

2.1 Average expert opinion

The input to this process is the segments marked by several medical experts, such as the four markings in Figure 1. Other marking types have also been tested, such as lesion-specific "representative points" [5], but the polygon segments seem to be the most intuitive for medical experts and result to the best performance.

The most straightforward combination procedure is averaging:

$$I_{conf_{i,j}}(x,y) = \frac{1}{N}\sum_{n=1}^{N} I_{exp_{i,j,n}}(x,y) \tag{1}$$

where $I_{conf_{i,j}}$ is the average segmentation of the input image i for the lesion type j. N is the total number of experts and $I_{exp_{i,j,n}}(x,y)$ is the segmentation mask of the nth expert. The expert segmentation masks $I_{exp_{i,j,n}}(x,y)$ are constructed from the expert images, such as the four examples in Figure 1, by filling the polygons with values of the expert-chosen confidence levels. The symbolic confidence levels converted into an ordinal scale are, for example, 0 for no marking, 1/3 for low, 2/3 for moderate, and 1 for high. The only requirement for the scale is that it is monotonically increasing. The average confidence image in (1) corresponds to the mean expert opinion in the same scale as the original annotations.

The average confidence image has two disadvantages: 1) it does not take into account the possible differences of the experts in their use of the scale and 2) for machine learning methods, the average expert segmentation does not produce binary values for foreground and background. A binary mask can be generated by thresholding the average expert segmentation image with the threshold $\tau \in [0,1]$:

$$I_{mask_{i,j}}(x,y) = \begin{cases} 1, \text{if } I_{conf_{i,j}}(x,y) \geq \tau \\ 0, \text{otherwise} \end{cases} . \tag{2}$$

The threshold parameter τ adjusts experts' joint-agreement: for $\tau \to 0$ the binary mask approaches to *set union*, and for $\tau \to 1$ to *set intersection* (see Fig. 1). If the resulting mask in (2) contains only zeros, then the image-wise ground truth is *false* for the lesion j in the image i. If any pixel is one, then the value is *true*.

In [5], the average segmentations we formally defined in (1), were implicitly used and the authors tested "plausible" confidence thresholds with their baseline method. Based on the tests, the confidence threshold was set to 0.75, which corresponds to the semantic "moderate confidence" in their scale. However, the baseline method undesirably tied the training and evaluation together. As a consequence, their selection of the moderate average confidence is questionable.

2.2 Optimal expert opinion by maximizing mutual agreement

The approach here is based on the following intuitive principle: "The ground truth should optimally represent the mutual agreement of all experts". A performance measure for the mutual agreement is needed. The performance depends only on the two factors: the *experts' markings* $I_{exp_{i,j,n}}$ and *ground truth* (g_t), and without loss of generality it is expected to output a real number

$$\text{perf} : \left\{ I_{exp_{i,j,n}}, g\_t_{i,j} \right\} \to \mathbb{R} \ . \tag{3}$$

$\{\cdot\}$ is used to denote that the performance is computed for a set of rated images. The definition reminds of the chicken-and-egg dilemma since the ground truth is inferred from the expert markings $I_{exp_{i,j,n}}$ using (1) and (2). This is justifiable since the true ground truth is a latent variable which should be inferred from the markings of the $n = 1, \ldots, N$ experts which are constant and the only inference variable is the threshold τ in (2). Generation of the image-wise ground truth is straightforward: if any of the pixels in $I_{mask_{i,j}}(x, y)$ for the lesion j is non-zero, the image is labeled to contain that lesion. A detection ROC curve can be automatically computed from the image-wise ground truth and image scores computed from the expert images. ("confidence score" in [6]). For the image-wise expert scores, we adopt the *summax rule*: pixel confidences of $I_{exp_{i,j,n}}$ are sorted, and 1% of the highest values are summed. We choose the average equal error rate (EER point on the ROC curve) as our performance measure in (3), which can now be given in the more explicit form:

$$\mathrm{perf}(\{I_{exp_{i,j,n}}\}, \{g\_t_{i,j}\}) = \frac{1}{N} \sum_n \mathrm{EER}\left(\{summax_{1\%}(I_{exp_{i,j,n}})\}, \{I_{mask_{i,j}}(x, y; \tau)\}\right) \quad . \tag{4}$$

A single EER value is computed for each expert n and over all images (i), and then the expert-specific EER values are summed for the lesion j.

The summax rule can be justified as a robust maximum rule by the multiple classifier theory [7], and the EER measure can be replaced with any other measure, for example, with a given sensitivity or specificity level. The only factor which affects the performance in (4) is the threshold τ which is used to form the ground truth. To maximize the mutual agreement, we should seek for the most appropriate threshold $\hat{\tau}$ which provides the highest average performance (EER) over all experts - the optimal disambiguation threshold. In addition, instead of a single threshold $\hat{\tau}$, separate thresholds $\hat{\tau}_j$ are selected for each lesion since different lesion types may significantly differ by their visual detectability. The optimal ground truth is ultimately equivalent to searching the optimal threshold:

$$\hat{\tau}_j \leftarrow \underset{\tau_j}{\mathrm{argmin}} \ \frac{1}{N} \sum_n EER(\cdot, \cdot) \quad . \tag{5}$$

The most straightforward approach to realize the *argmin* step is to iteratively test all possible values of τ from 0 to 1. The values can be enumerated from the expert confidence and converted to integers for numerical stability. Equation (5) maximizes the performance for each lesion type over all experts ($\sum_n$). The optimal thresholds $\hat{\tau}_j$ are lesion specific and they are guaranteed to produce the maximal mutual expert agreement according to the performance measure *perf*.

3 Results

We empirically test our method with the DiaRetDB1 dataset downloaded from `http://www2.it.lut.fi/project/imageret/` which to the authors' best knowledge is the only database which provides also the original expert markings.

STAPLE here represents the state-of-the-art and was downloaded from `http://crl.med.harvard.edu/software/STAPLE/`.

DiaRetDB1 [1] has recently become an important database for evaluating diabetic retinopathy detection algorithms. The database contains dedicatedly selected retinal images divided into fixed sets of training and test images both containing a representative set of *normal findings*, and from mild to severe non-proliferative findings including *microaneurysms* (MA), *haemorrhages* (HA), *hard exudates* (HE) and *soft exudates* (SE). A detailed description is available in [1].

3.1 Disambiguation

We report the performance using the EER which is justified since no auxiliary/prior information on the risks or penalties for false negatives or false positives is available. EER represents a "balanced error point" in the ROC curve and allows comparison to the previous works.

The results for DiaRetDB1 are presented in Table 1. It is noteworthy that the original confidence threshold (0.75) in [5] is not optimal for any of the lesion types, and is clearly incorrect for haemorrhages (HA, 0.60) and microaneurysms (MA, 0.10). The underlined values in the table are the best achieved performances. The average performance for all lesion types significantly varies depending on the threshold.

To compare our approach to the state-of-the-art, we combined the expert masks using the STAPLE algorithm described in [2,3,4]. The STAPLE algorithm does not utilize pixel-based confidence information, and therefore, the algorithm was run separately for all different confidence levels of the original data (S1: 0.25, S2: 0.75 and S3: 1.00). STAPLE outperforms the DiaRetDB1 method, but is inferior to the proposed method for all lesion types.

Table 1: Expert-specific equal error rate (EER) and the average EER performances for DiaRetDB1. EER for the different confidence thresholds τ_j reported and the algorithm found optimal values underlined. The last three rows are the results of the STAPLE algorithm and with the masks generated by using different confidence levels (S1: 0.25, S2: 0.75, S3: 1.00).

	Haemorrhage (HA)					Hard exudate (HE)					Microaneurysm (MA)					Soft exudate (SE)				
τ	e1	e2	e3	e4	Joint	e1	e2	e3	e4	Joint	e1	e2	e3	e4	Joint	e1	e2	e3	e4	Joint
.00	0.17	0.17	0.14	0.09	0.14	0.29	0.12	0.05	0.07	0.13	0.04	0.22	0.22	0.12	0.15	0.45	0.17	0.23	0.11	0.24
.10	0.12	0.12	0.12	0.06	0.10	0.26	0.09	0.05	0.02	0.11	0.03	0.20	0.22	0.11	**0.14**	0.42	0.10	0.19	0.07	0.20
.20	0.08	0.08	0.07	0.02	0.07	0.25	0.07	0.05	0.05	0.10	0.17	0.20	0.25	0.10	0.18	0.42	0.07	0.17	0.07	0.18
.30	0.07	0.07	0.02	0.02	0.05	0.24	0.05	0.05	0.05	0.09	0.17	0.17	0.18	0.07	0.15	0.32	0.05	0.08	0.08	0.13
.40	0.07	0.07	0.02	0.02	0.05	0.18	0.05	0.06	0.06	0.09	0.20	0.15	0.18	0.09	0.16	0.29	0.06	0.08	0.08	0.13
.50	0.05	0.05	0.05	0.05	0.05	0.16	0.03	0.06	0.06	0.08	0.22	0.11	0.16	0.12	0.15	0.26	0.09	0.09	0.09	0.13
.60	0.02	0.05	0.02	0.02	**0.03**	0.16	0.03	0.06	0.06	0.08	0.20	0.16	0.14	0.18	0.17	0.16	0.11	0.11	0.11	0.12
.70	0.03	0.06	0.05	0.05	0.05	0.16	0.03	0.06	0.06	0.08	0.20	0.20	0.10	0.18	0.17	0.07	0.14	0.13	0.14	0.12
.75	0.03	0.06	0.05	0.05	0.05	0.16	0.03	0.06	0.06	0.08	0.20	0.20	0.10	0.18	0.17	0.07	0.14	0.13	0.14	0.12
.80	0.09	0.06	0.08	0.11	0.09	0.03	0.07	0.03	0.07	**0.05**	0.15	0.18	0.18	0.18	0.17	0.09	0.09	0.12	0.13	**0.11**
.90	0.09	0.06	0.08	0.11	0.09	0.03	0.07	0.03	0.07	**0.05**	0.16	0.17	0.16	0.21	0.17	0.09	0.09	0.12	0.13	**0.11**
1.0	0.12	0.12	0.12	0.16	0.13	0.03	0.07	0.03	0.08	0.06	0.21	0.21	0.21	0.21	0.21	0.10	0.10	0.11	0.14	**0.11**
S1	0.07	0.07	0.02	0.02	0.05	0.27	0.09	0.02	0.02	0.10	0.17	0.17	0.18	0.07	0.15	0.35	0.05	0.12	0.08	0.15
S2	0.07	0.07	0.02	0.02	0.05	0.20	0.02	0.07	0.05	0.09	0.19	0.16	0.19	0.07	0.15	0.29	0.06	0.08	0.08	0.13
S3	0.05	0.05	0.05	0.05	0.05	0.16	0.03	0.06	0.06	0.08	0.21	0.10	0.18	0.10	0.15	0.20	0.10	0.10	0.10	0.13

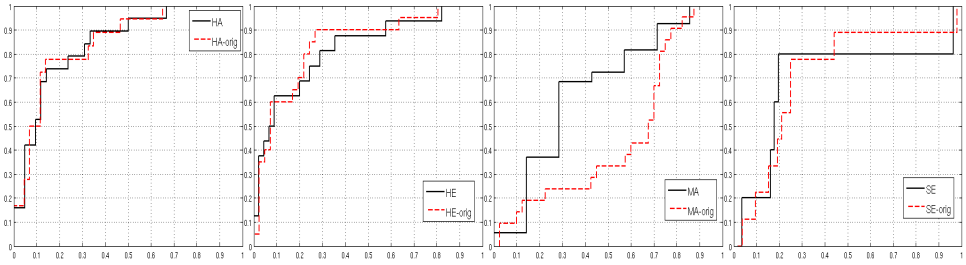
3.2 Revised DiaRetDB1 baseline results

The purpose of this experiment was to study whether the new combination procedure has any effect on the performance of the DiaRetDB1 baseline method. For the two lesion types, Hard exudate (HE) and Soft exudate (SE), the proposed method produced multiple optimal thresholds: HE $\in$ [0.8,0.9] and SE $\in$ [0.8,1.0], and thus the DiaRetDB1 baseline method was run multiple times with both the minimum and maximum values. The results are in Table 2.

Table 2: The minimum, maximum, and average EER (5 rand. iters) for the DiaRetDB1 baseline method and evaluation protocol. The results include the original and the proposed optimal ground truth.

	Haemorrhage (HA)			Hard exud. (HE)			Microaneurysm (MA)			Soft exud. (SE)			
	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Overall
In [5]	0.233	0.333	0.273	0.200	0.220	0.216	0.476	0.625	0.593	0.250	0.333	0.317	0.349
Our (min)	0.263	0.476	0.322	0.250	0.250	0.250	0.286	0.574	0.338	0.333	0.333	0.333	0.311
Our (max)	0.263	0.476	0.322	0.250	0.250	0.250	0.386	0.574	0.338	0.200	0.268	0.241	**0.288**

The minimum and maximum thresholds for the proposed disambiguation rule produce equal results except in the case of soft exudates, for which the maximum in the equally performing interval (1.0) is clearly better. The main difference to the original DiaRetDB1 method occurs with microaneurysms, as expected, since the optimal threshold (0.1) significantly differs from the original (0.75). For haemorrhages, the original result was too optimistic since the optimal confidence yields to worse minimum and average EER. On average, the optimal confidence provided 11–17% better performance. This can be explained by the fact that the training and testing segments are now consistent with experts' opinion. The findings can be verified by the ROC curves in Fig. 2.



(a) Haemorrhage (b) Hard exudate (c) Microaneurysm (d) Soft exudate

Fig. 2: ROC curves for the DiaRetDB1 baseline method using the original and proposed (max) method to generate training and testing data.

4 Discussion and conclusion

In this work, the multiple ground truth combination and disambiguation problem was studied in the context of medical images where lesions are delineated by multiple experts (raters). A simple yet effective combining procedure maximizing the mutual agreement of experts' rating was proposed. The procedure corresponds to the clinical practice of "expert/colleague consultation" as it provides the ground truth segments most consistent over all expert opinions. Therefore, the principle of the proposed procedure is well justified and it was exemplified with the DiaRetDB1 database for which our method produced clinically correct and more reliable ground truth. Our work deviates from the state-of-the-art (STAPLE) by being notably simpler, does not require spatial consistency priors and does not produce areas that do not appear in the original expert segments (the upper bound is the complete union). In our experiments, the proposed disambiguation yielded to 11–17% performance improvements compared to the DiaRetDB1 baseline method. For the experts themselves, the mutual agreement, depending on the lesion type, improved 9–40% compared to the original DiaRetDB1 ground truth, and 7–40% compared to the one produced by STAPLE.

Acknowledgements The authors wish to thank the research collaborators from the University of Eastern Finland and the University of Tampere, and the sponsors of the earlier ImageRet[2] project for their support.

References

1. Kauppi, T., Kalesnykiene, V., Kamarainen, J.K., Lensu, L., Sorri, I., A. Raninen, A., Voutilainen, R., Uusitalo, H., Kälviäinen, H., Pietilä, J.: The DIARETDB1 diabetic retinopathy database and evaluation protocol. In: BMVC. (2007)
2. Warfield, S., Zou, K., Wells, W.: Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans. on Medical Imaging **23**(7) (2004)
3. Commowick, O., Warfield, S.: A continuous staple for scalar, vector, and tensor images: An application to dti analysis. IEEE Trans. on Medical Imaging **28**(6) (2009)
4. Commowick, O., Warfield, S.: Estimation of inferential uncertainty in assessing expert segmentation performance from staple. IEEE Trans. on Medical Imaging **29**(3) (2010)
5. Kauppi, T., Kamarainen, J.K., Lensu, L., Kalesnykiene, V., Sorri, I., Kälviäinen, H., Uusitalo, H., Pietilä, J.: Fusion of multiple expert annotations and overall score selection for medical image diagnosis. In: Scandinavian Conf. on Image Analysis. (2009) 760–769
6. Everingham, M., Zisserman, A., Williams, C.K.I., Van Gool, L.: The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. online
7. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classfiers. IEEE Trans. on PAMI **20**(3) (1998)

[2] `http://www2.it.lut.fi/project/imageret/`