Fusion of Multiple Expert Annotations and Overall Score Selection for Medical Image Diagnosis

Tomi Kauppi[1], Joni-Kristian Kamarainen[2], Lasse Lensu[1],
Valentina Kalesnykiene[3], Iiris Sorri[3], Heikki Kälviäinen[1], Hannu Uusitalo[4],
and Juhani Pietilä[5]

[1] Machine Vision and Pattern Recognition Research Group (MVPR)
[2] MVPR/Computational Vision Group, Kouvola
Department of Information Technology,
Lappeenranta University of Technology (LUT), Finland
[3] Department of Ophthalmology, University of Kuopio, Finland
[4] Department of Ophthalmology, University of Tampere, Finland
[5] Perimetria Ltd., Finland

Abstract. Two problems especially important for supervised learning and classification in medical image processing are addressed in this study: i) how to fuse medical annotations collected from several medical experts and ii) how to form an image-wise overall score for accurate and reliable automatic diagnosis. Both of the problems are addressed by applying the same receiver operating characteristic (ROC) framework which is made to correspond to the medical practise. The first problem arises from the typical need to collect the medical ground truth from several experts to understand the underlying phenomenon and to increase robustness. However, it is currently unclear how these expert opinions (annotations) should be combined for classification methods. The second problem is due to the ultimate goal of any automatic diagnosis, a patient-based (image-wise) diagnosis, which consequently must be the ultimate evaluation criterion before transferring any methods into practise. Various image processing methods provide several, e.g., spatially distinct, results, which should be combined into a single image-wise score value. We discuss and investigate these two problems in detail, propose good strategies and report experimental results on a diabetic retinopathy database verifying our findings.

1 Introduction

Despite the fact that medical image processing has been an active application area of image processing and computer vision for decades, it is surprising that strict evaluation practises in other applications, e.g., in face recognition, have not been used that systematically in medical image processing. The consequence is that it is difficult to evaluate the state-of-the-art or estimate the overall maturity of methods even for a specific medical image processing problem. A step

A.-B. Salberg, J.Y. Hardeberg, and R. Jenssen (Eds.): SCIA 2009, LNCS 5575, pp. 760–769, 2009.

towards more proper operating procedures was recently introduced by the authors in the form of a public database, protocol and tools for benchmarking diabetic retinopathy detection methods [1]. During the course of work in establishing the DiaRetDB1 database and protocol, it became evident that there are certain important research problems which need to be studied further. One important problem is the optimal fusion strategy of annotations from several experts. In computer vision, ground truth information can be collected by using expert made annotations. However, in related studies such as in visual object categorisation, this problem has not been addressed at all (e.g., the recent LabelMe database [2] or the popular CalTech101 [3]). At least for medical images, this is of particular importance since the opinions of medical doctors may significantly deviate from each other or the experts may graphically describe the same finding in very different ways. This can be partly avoided by instructing the doctors while annotating, but often this is not desired since the data can be biased and grounds for understanding the phenomenon may weaken. Therefore, it is necessary to study appropriate fusion or "voting" methods.

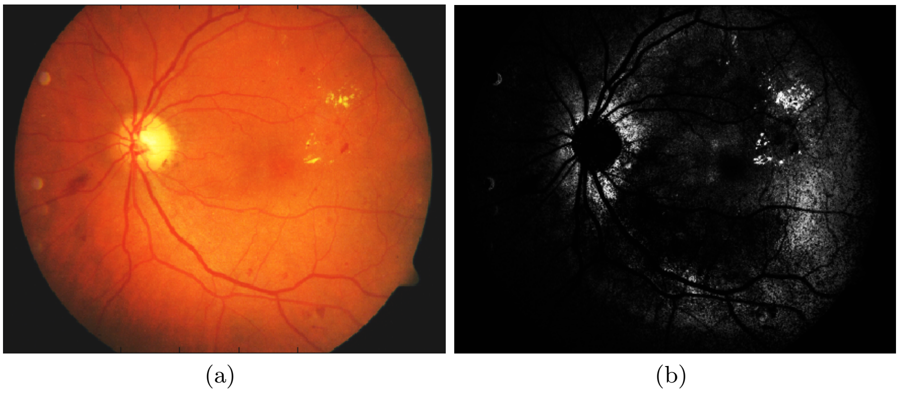
Another very important problem arises from the fact how medical doctors actually use medical image information. They do not see it as a spatial map which is evaluated pixel by pixel or block by block, but as a whole depicting supporting information for a positive or negative diagnosis result of a specific disease. In image processing method development, on the other hand, pixel- or block-based analysis is more natural and useful, but the ultimate goal should be kept in mind, i.e., supporting the medical decision making. This issue was discussed in [1] and used in the development of the DiaRetDB1 protocol. The evaluation protocol, which simulates patient diagnosis using medical terms (specificity and sensitivity), requires a single overall diagnosis score for each test image, but it was not explicitly defined how the multiple cues should be combined into a single overall score. We address this problem throughly in this study and search for the optimal strategy to combine the cues. Also this problem is less known in medical image processing, but a well studied problem within the context of multiple classifiers or classifier ensembles (e.g., [4,5,6]).

The two problems are discussed in detail in Sections 2 and 3, and in the experimental part in Section 4 we utilise the evaluation framework (ROC graphs and equal error rate (EER) / weighted error rate (WER) error measures) to experimentally evaluate different fusion and scoring methods. Based on the discussions and the presented empirical results, we draw conclusions, define best practises and discuss the restrictions implied by our assumptions in Section 5.

2 Overall Image Score Selection for Medical Image Diagnosis

Medical diagnosis aims to diagnose the correct disease of a patient, and it is typically based on background knowledge (prior information) and laboratory tests which today include also medical imaging (e.g., ultrasound, eye fundus imaging, CT, PET, MRI, fMRI). The outcome of the tests and image or video data (observations) is typically either positive or negative evidence and the final diagnosis is based on a combination of background knowledge and test outcomes under strong Bayesian decision making for which all clinicians have been trained in the medical school [7]. Consequently, medical doctors are interested in medical image processing similar to a patient-based tool which provides a positive or negative outcome with a certain confidence. The tool confidence is typically fixed by setting the system to operate at certain *sensitivity* and *specificity* levels ($[0\%, 100\%]$), and therefore, these two terms are of special importance in medical image processing literature. The sensitivity value depends on the diseased population and specificity on the healthy population. Since these values are defined by the true positive rate (sensitivity is true positives divided by the sum of true positives and false negatives) and false positive rates (specificity is true negatives divided by the sum of true negatives and false positives), receiver operating characteristic (ROC) analysis is a natural tool to compare any methods [1]. Fixing the sensitivity and specificity values corresponds to selecting a certain operating point from the ROC. In [1], the authors introduced automatic evaluation methodology and published a tool to automatically produce the ROC graph for data where a single score value representing the test outcome (a higher score value increases the certainty of the positive outcome) is assigned to every image. The derivation of a proper image scoring method was not discussed, but is a topic in this study.

We restrict our development work to pixel- and block-based image processing schemes which are the most popular. The implication is that, for example, every pixel in an input image is classified to as a positive or negative finding, or positive finding likelihoods are directly given (see Fig. 1). To establish the final overall image score, these pixel or block values must be combined.



(a)

(b)

Fig. 1. Example of pixel-wise likelihoods for hard exudates in eye fundus images (diabetic findings): (a) the original image (hard exudates are the small yellow spots in the upper-right part of the image); (b) probability density (likelihood) "map" for the hard exudates (estimated with a Gaussian mixture model from RGB image data)

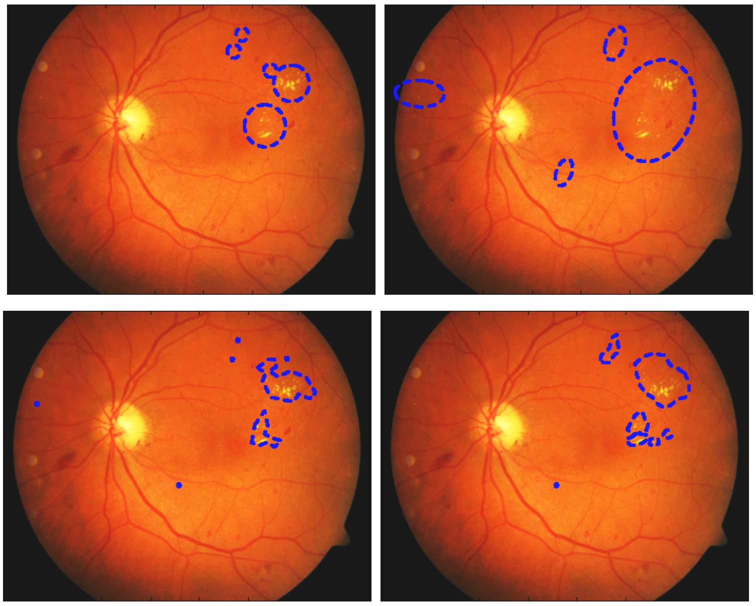


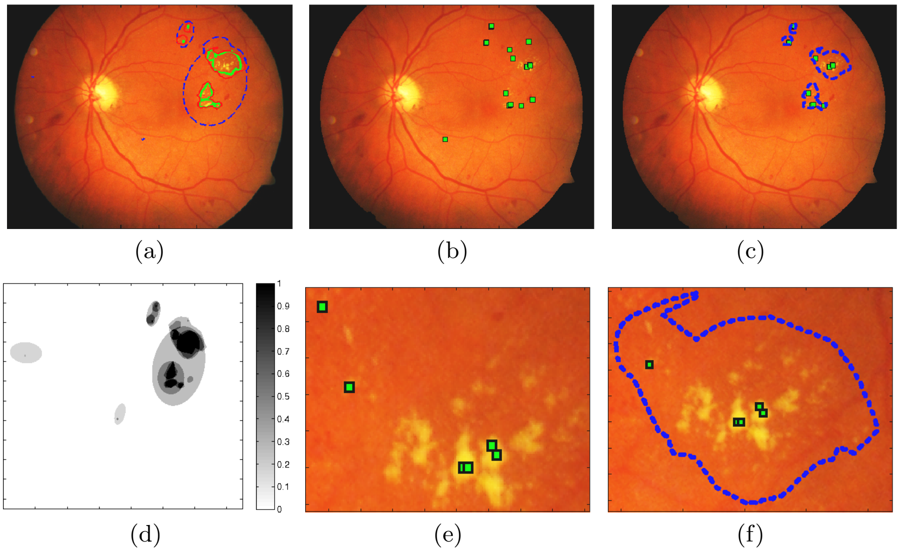
Fig. 2. Four independent expert annotations of hard exudates in one image

In the pixel- and block-based analyses, the final decision (score fusion) must be based on the fact that we have a (two-class) classification problem where the classifiers vote for positive or negative outcomes with a certain confidence. It follows that the optimal fusion strategy can easily be devised by exploring the results from a related field, combining classifiers (classifier ensembles), e.g., from the milestone study by Kittler et al. [4]. In our case, the "classifiers" act on different inputs (pixels) and therefore obey the distinct observations assumption in [4]. In addition, the classifiers have equal weights between the negative and positive outcomes. In [4], the theoretically most plausible fusion rules applicable also here were the product, sum (mean), maximum and median rules. We replaced the median rule with a more intuitive rank-order based rule for our case: "summax", i.e., the sum of some proportion of the largest values ($summax_{X\%}$). In our formulation, the maximum and sum rules can be seen as two extrema whereas summax operates between them so that X defines the operation point. Since any other straightforward strategies would be derivatives of these four, we restrict our analysis to them.

After the following discussion on fusion strategies, we experimentally evaluate all combinations of fusion and scoring strategies. Our evaluation framework and the DiaRetDB1 data is used for the purpose.

3 Fusing Multiple Medical Expert Annotations

It is recommended to collect medical ground truth (e.g., image annotations) from several experts within that specific field (e.g., ophthalmologists for eye

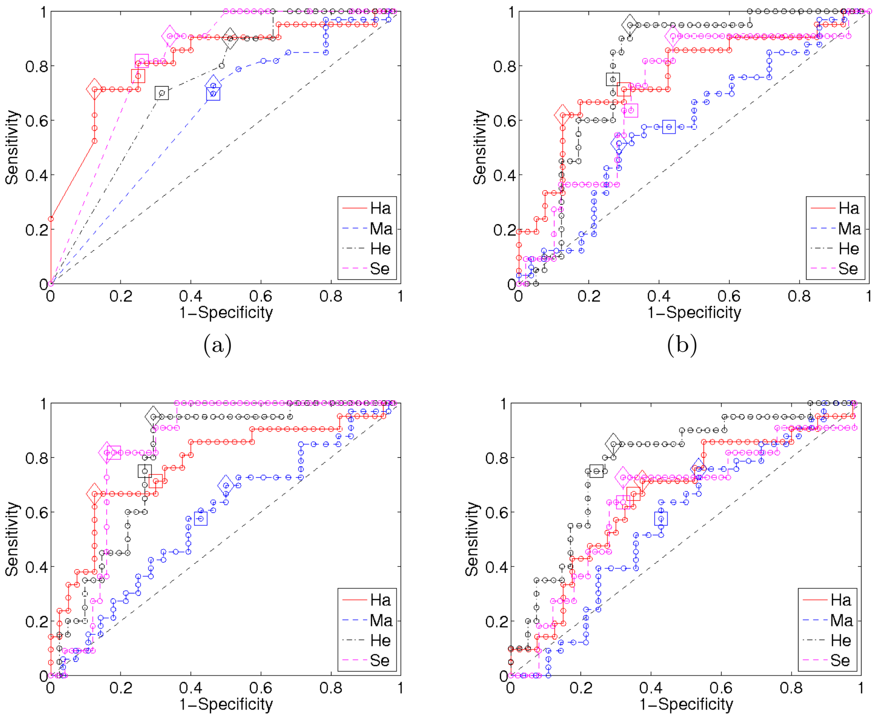

(a) (b) (c)

(d) (e) (f)

Fig. 3. Different annotation fusion approaches for the case shown in Fig. 2: (a) areas (applied confidence threshold for blue 0.25, red 0.75 and green 1.00); (b) representative points and their neighbourhoods (5 × 5); (c) representative point neighbourhoods masked with the areas (confidence threshold 0.75, blue colour); d) confidence map of areas in Fig. 3(a) e) close up image of representative point neighbourhoods in Fig. 3(b); f) close up image of masked representative point neighbourhoods in Fig. 3(c)

diseases). Note that this is not the practise in computer vision applications, e.g., only the eyes or bounding boxes are annotated by a single user in the face recognition databases (The FERET [8]) and rough segmentations in object category recognition (CalTech101 [3], LabelMe [2]). Multiple annotations is a necessity in medical applications where colleague consultation is the predominant working practise. Multiple annotations generate a new problem of how the annotations should be combined to a single ground truth (consultation outcome) for training a classifier. The solution certainly depends on the annotation tools provided for the experts, but it is not recommended to limit their expression power by instructions from laypersons which can harm the quality of ground truth.

For the DiaRetDB1 database, the authors introduced a set of graphical directives which are understandable for people not familiar of computer vision and graphics [1]. In the introduced directives, polygon and ellipse (circle) areas are used to annotate the spatial coverage of findings and at least one required (representative) point inside each area defining a particular spatial location that attracted expert's attention (colour, structure, etc.) With these simple but powerful directives, the independent experts produced significantly varying annotations for the same images, or even for the same finding in an image (see Fig. 2 for examples). The obvious problem is how to fuse equally trustworthy information from multiple sources to provide representative ground truth which retains



(a)

(b)

(c)

(d)

Fig. 4. Example ROC curves of "weighted expert area intersection" fusion with confidence 0.75 for two scoring rules, where EER and WER are marked with rectangle and diamond (best viewed in colours): (a) *max*; (b) *mean*; (c) $summax_{0.01}$; (d) *product*

the necessary within-class and between-class variation for supervised machine learning methods.

The available information to be fused is as follows: spatial coverage data by the polygon and ellipse areas, pixel locations (and possibly their neighbourhoods) of the representative points and the confidence levels for each marking given by each expert ("high", "moderate" or "low"). The available directives establish the available fusion strategies: intersections (sums) of the areas thresholded by a fixed average confidence (Fig. 3(a)), fixed size neighbourhoods of the representative points (Fig. 3(b)) and fixed size neighbourhoods of the representative points masked by the areas (combination of two) (Fig.3(c)).

All possible fusion strategies combined with all possible overall scoring strategies were experimentally evaluated as reported next.

4 Experiments

The experiments were conducted using the publicly available DiaRetDB1 diabetic retinopathy database [1]. The database comprises 89 colour fundus images

Table 1. Equal error rate (EER) for different fusion and overall scoring strategies

WEIGHTED EXPERT AREA INTERSECTION								
	0.75				1.00			
	max	mean	summax0.01	prod	max	mean	summax0.01	prod
Ha	0.2500	0.3000	0.3000	0.3500	0.5250	0.3810	0.4000	0.4762
Ma	0.4643	0.4286	0.4286	0.4286	0.3939	0.3636	0.3636	0.4286
He	0.3171	0.2683	0.2683	0.2500	0.2195	0.2500	0.2500	0.2500
Se	0.2600	0.3636	0.1818	0.3636	0.6600	0.2800	0.3000	0.2800
TOTAL	1.2914	1.3605	1.1787	1.3922	1.7985	1.2746	1.3136	1.4348
REP. POINT NEIGHBOURHOOD								
	1x1				3x3			
	max	mean	summax0.01	prod	max	mean	summax0.01	prod
Ha	0.6500	0.4762	0.4762	0.7250	0.7000	0.4286	0.4286	0.6750
Ma	0.7143	0.4643	0.4643	0.4643	0.6429	0.4643	0.4643	0.4643
He	0.3000	0.2000	0.2500	0.2500	0.1500	0.2000	0.2500	0.3000
Se	0.3636	0.3636	0.3636	0.3636	0.4545	0.3636	0.3636	0.3636
TOTAL	2.0279	1.5041	1.5541	1.8029	1.9474	1.4565	1.5065	1.8029
	5x5				7x7			
	max	mean	summax0.01	prod	max	mean	summax0.01	prod
Ha	0.6000	0.4762	0.4762	0.6750	0.7000	0.3810	0.5250	0.6750
Ma	0.6786	0.4286	0.4286	0.4643	0.4643	0.4286	0.4643	0.4286
He	0.2500	0.2000	0.2000	0.2195	0.2500	0.2500	0.2683	0.2000
Se	0.3800	0.3636	0.3636	0.5455	0.4545	0.3636	0.2800	0.3636
TOTAL	1.9086	1.4684	1.4684	1.9043	1.8688	1.4232	1.5376	1.6672
REP. POINT NEIGHBOURHOOD MASKED (AREA 0.75)								
	1x1				3x3			
	max	mean	summax0.01	prod	max	mean	summax0.01	prod
Ha	0.6500	0.4762	0.5714	0.7250	0.6500	0.4000	0.4762	0.6750
Ma	0.6429	0.4643	0.5000	0.4286	0.6071	0.5000	0.4643	0.4286
He	0.4000	0.2500	0.2000	0.2000	0.2683	0.2000	0.2000	0.2500
Se	0.5400	0.2800	0.3000	0.3636	0.2200	0.2800	0.2800	0.3636
TOTAL	2.2329	1.4705	1.5714	1.7172	1.7454	1.3800	1.4205	1.7172
	5x5				7x7			
	max	mean	summax0.01	prod	max	mean	summax0.01	prod
Ha	0.6500	0.5000	0.4286	0.6750	0.7250	0.4762	0.4762	0.6750
Ma	0.5152	0.4286	0.4286	0.4286	0.5455	0.4286	0.5000	0.4286
He	0.2500	0.2683	0.2500	0.2195	0.2500	0.3000	0.2195	0.2500
Se	0.2200	0.3000	0.2800	0.2800	0.4545	0.2800	0.3000	0.2727
TOTAL	1.6352	1.4969	1.3871	1.6031	1.9750	1.4848	1.4957	1.6263

of which 84 contain at least mild non-proliferative signs of diabetic retinopathy (haemorrhages (Ha), microaneurysms (Ma), hard exudates (He) and soft exudates (Se)). The images were captured with the same 50 degree field-of-view digital fundus camera[1], and therefore, the data should not contain colour distortions other than those related to the findings. The fusion and overall scoring strategies were tested using the predefined training set of 28 images and test set of 61 images.

Since this study is restricted to pixel- and block-based image processing approaches, photometric information (colour) was a natural feature for the experimental analysis. For the visual diagnosis of diabetic retinopathy, colour is also the most important single visual cue. Since the whole medical diagnosis is naturally Bayesian, we were motivated to address the classification problem with a standard statistical tool, estimating probability density functions (pdfs) of each finding given a colour observation (RGB), $p(r, g, b|finding)$. For the un-

[1] ZEISS FF 450^{plus} fundus camera with Nikon F5 digital camera.

Table 2. Weighted error rate [WER(1)] for different fusion and overall scoring strategies

WEIGHTED EXPERT AREA INTERSECTION								
	0.75				1			
	max	mean	summax0.01	prod	max	mean	summax0.01	prod
Ha	0.2054	0.2530	0.2292	0.3304	0.3577	0.3494	0.3440	0.4119
Ma	0.3685	0.3853	0.4015	0.3891	0.3685	0.3452	0.2998	0.3561
He	0.3061	0.1835	0.1713	0.2213	0.1829	0.1841	0.1963	0.1841
Se	0.2155	0.2655	0.1709	0.2964	0.4209	0.2309	0.2609	0.2718
TOTAL	1.0954	1.0872	0.9729	1.2371	1.3301	1.1097	1.1011	1.2239

REP. POINT NEIGHBOURHOOD								
	1x1				3x3			
	max	mean	summax0.01	prod	max	mean	summax0.01	prod
Ha	0.3964	0.3845	0.4417	0.5000	0.4238	0.3631	0.4018	0.5000
Ma	0.3902	0.4107	0.4015	0.3837	0.4080	0.4031	0.4042	0.3561
He	0.2476	0.1713	0.2220	0.1970	0.1482	0.1598	0.1591	0.2451
Se	0.3118	0.2755	0.3073	0.3264	0.3509	0.2809	0.2709	0.3264
TOTAL	1.3460	1.2420	1.3724	1.4070	1.3309	1.2069	1.2361	1.4275
	5x5				7x7			
	max	mean	summax0.01	prod	max	mean	summax0.01	prod
Ha	0.4190	0.3482	0.4179	0.5000	0.4113	0.3631	0.4554	0.5000
Ma	0.4302	0.4031	0.3988	0.3864	0.3231	0.3880	0.4318	0.3750
He	0.1988	0.1598	0.1854	0.1841	0.2091	0.1829	0.2207	0.1957
Se	0.2100	0.2655	0.2509	0.4127	0.3927	0.2355	0.2200	0.2709
TOTAL	1.2580	1.1766	1.2529	1.4832	1.3362	1.1695	1.3279	1.3416

REP. POINT NEIGHBOURHOOD MASKED (AREA 0.75)								
	1x1				3x3			
	max	mean	summax0.01	prod	max	mean	summax0.01	prod
Ha	0.4351	0.4369	0.4702	0.5000	0.4238	0.3631	0.4315	0.5000
Ma	0.4280	0.4291	0.4069	0.3723	0.4702	0.4383	0.4329	0.4085
He	0.2439	0.1963	0.1726	0.1976	0.1988	0.1976	0.1726	0.1963
Se	0.3609	0.2409	0.2609	0.3173	0.1555	0.2555	0.1600	0.3118
TOTAL	1.4680	1.3033	1.3106	1.3871	1.2483	1.2544	1.1970	1.4167
	5x5				7x7			
	max	mean	summax0.01	prod	max	mean	summax0.01	prod
Ha	0.4113	0.4333	0.3482	0.5000	0.3988	0.3756	0.4190	0.4524
Ma	0.4129	0.4004	0.4015	0.3544	0.4334	0.3907	0.4383	0.3701
He	0.2073	0.1963	0.2232	0.2098	0.1957	0.2713	0.1713	0.2323
Se	0.1555	0.2700	0.1855	0.2718	0.3927	0.1900	0.2609	0.2109
TOTAL	1.1870	1.3001	1.1584	1.3360	1.4207	1.2276	1.2896	1.2657

known distributions, Gaussian mixture models (GMMs) were natural models and the unsupervised Figueiredo-Jain algorithm a good estimation method [9]. We also tried the standard expectation maximisation (EM) algorithm, but since the Figueiredo-Jain always outperformed it without the need to explicitly define the number of components, it was left out from this study. For training, different fusion approaches for the expert annotations discussed in Section 3 were used to form a training set for the GMM estimates. For every test set image, our method provided a full likelihood map (see Fig. 1(b)) from which the different overall scores in Section 2 were computed.

Our interpretations of the results are based qualitatively on the produced ROC graphs and quantitatively on EER (equal error rate) and WER (weighted error rate) measures, both introduced in the evaluation framework proposed in [1]. The EER is a single point in a ROC graph and the WER takes a weighted average of the false positive and false negative rates. Here we used WER(1) which gives no preference to either failure type, i.e., a ROC point which provides the smallest average error was selected. All results are shown in Tables 1 and 2. The results indicate that better results were always achieved using the “weighted expert area intersection” fusion instead of using the “representative point neighbourhood” methods. This was at first surprising, but understandable because the areas cover the finding areas more thoroughly than the representative points which are concentrated only near the most salient points. Moreover, it is evident from the results that the product rule was generally bad for the obvious reasons discussed already in [4]. The summax rule always produced either the best results or results comparable to the best results as evident in Tables 1 and 2, and in example ROC curves in Fig. 4. Since the best performance was achieved using the “weighted expert area intersection” fusion, for which the pure sum (mean), max and product rules were clearly inferior to the summax, the summax rule should be preferred.

5 Conclusions

In this paper, the problem of fusing a united ground truth (consultation outcome) from multiple medical expert annotations (opinions) for classifier learning and the problem of forming an image-wise overall score for automatic image-based evaluation were studied. All the proposed fusion strategies and the overall scoring strategies were first discussed in the contexts of related works of different fields and then experimentally verified against a public fundus image database. As results from our more theoretical discussion and the experimental results, we conclude that the best ground truth fusion strategy is the “weighted expert area intersection” and the best overall scoring method the “summax” rule ($X = 0.01$, example proportion), both described in this study.

Acknowledgements

The authors would like to thank the Finnish Funding Agency for Technology and Innovation (TEKES) and partners of the ImageRet project[2] (No. 40039/07) for support.

References

1. Kauppi, T., Kalesnykiene, V., Kamarainen, J.K., Lensu, L., Sorri, I., Raninen, A., Voutilainen, R., Uusitalo, H., Kälviäinen, H., Pietilä, J.: The diaretdb1 diabetic retinopathy database and evaluation protocol. In: Proc. of the British Machine Vision Conference (BMVC 2007), Warwick, UK, vol. 1, pp. 252–261 (2007)
2. Russel, B., Torralba, A., Murphy, K., Freeman, W.: LabelMe: a database and web-based tool for image annotation. Int. J. of Computer Vision 77(1-3), 157–173 (2008)

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

3. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Trans. on PAMI 28(4) (2006)
4. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classfiers. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 20(3), 226–239 (1998)
5. Tax, D.M.J., van Breukelen, M., Duin, R.P.W., Kittler, J.: Combining multiple classifiers by averaging or by multipying. The Journal of the Pattern Recognition Society 33, 1475–1485 (2000)
6. Fumera, G., Roli, F.: A theoretical and experimental analysis of linear combiners for multiple classifier systems. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 27(6), 942–956 (2005)
7. Gill, C., Sabin, L., Schmid, C.: Why clinicians are natural bayesians. British Medical Journal 330(7) (2005)
8. Phillips, P., Moon, H., Rauss, P., Rizvi, S.: The feret evaluation methodology for face recognition algorithms. IEEE Trans. on PAMI 22(10) (2000)
9. Figueiredo, M., Jain, A.: Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3), 381–396 (2002)