A Framework for Constructing Benchmark Databases and Protocols for Retinopathy in Medical Image Analysis

Tomi Kauppi[a], Joni-Kristian Kamarainen[ab], Lasse Lensu[a], Valentina Kalesnykiene[c], Iiris Sorri[c], Hannu Uusitalo[d], and Heikki Kälviäinen[a]

[a]Machine Vision and Pattern Recognition Laboratory (MVPR), Lappeenranta University of Technology (LUT), P.O. Box 20, FI-53851, Lappeenranta, Finland
[b]MVPR Computational Vision Group, Kouvola Unit, LUT, Finland
[c]Department of Ophthalmology, University of Eastern Finland, Finland
[d]Department of Ophthalmology, University of Tampere, Finland

Abstract. We address performance evaluation practises for developing medical image analysis methods, and contribute to the practise to establish and to share databases of medical images with verified ground truth and solid evaluation protocols. This helps to develop better algorithms, to perform fair method comparisons, including the state-of-the-art methods, and consequently, supports technology transfer from research laboratories to clinical practise. For this purpose, we propose a framework consisting of reusable methods and tools for the laborious task of constructing a benchmark database. We provide a medical image annotation software tool which helps to collect and store ground truth for retinopathy lesions from experts, including the fusion of multiple annotations from several experts. The tool and all necessary functionality for method evaluation are provided as a public software package. For demonstration purposes, we utilise the framework and tools to establish the DiaRetDB01 V2.1 database for benchmarking diabetic retinopathy detection algorithms. The database contains all necessary images, collected ground truth from several experts, and a strawman algorithm for the detection of lesions.

Keywords: Diabetic retinopathy detection, eye fundus imaging, benchmarking image database, eye fundus image processing, eye fundus image analysis, medical image processing, medical image analysis.

1 Introduction

Image databases and expert ground truth are in a regular use in medical image processing. However, it is common that the data is not public, and therefore, reliable comparisons and state-of-the-art surveys are difficult to conduct. In contrast to, for example, biometrics including face, iris, and fingerprint recognition, the research has been driven by public databases and solid evaluation protocols. These databases have been extended and revised resulting to continuous

pressure for the development of better methods. This could be adopted more also in medical image processing and analysis. During our research on diabetic retinopathy [1], we have experienced that developing databases from the scratch is demanding, laborious and time consuming. However, certain tasks occur repeatedly and are reusable as such. Here, we discuss related practical issues, point out and solve repeatably occurring sub-tasks, and provide the solutions as open-source tools on our web site. In the experimental part, we utilise the proposed framework and devise a revised version of the diabetic retinopathy database DiaRetDB1 published originally in [2, 3].

2 Benchmarking and Related Work

Recently Thacker et al. [4] studied the performance characterisation of computer vision methods, also transfable to medical image processing. The eight general considerations are adopted from [4], referred as the key questions:
C1: "How is testing currently performed?": If a commonly used database and protocol are available, their validity for development and evaluation needs to be examined. In the worst case, a new database needs to be constructed.
C2: "Is there a data set for which the correct answers are known?": Such a data set can be used to report the results enabling comparisons.
C3: "Are there data sets in common use?": See C1 and C2. Common data sets facilitate fair comparisons.
C4: "Are there experiments which show that algorithms are stable and work as expected?": This can be realised, if the expert ground truth is available.
C5: "Are there any strawman algorithms?": A strawman algorithm sets the baseline for method performance.
C6: "What code and data are available?": By publishing the code of a method, other research groups can avoid repeating the same work.
C7: "Is there a quantitative methodology for the design of algorithms?": This depends on the medical problem, but the methodology can be typically devised by following corresponding clinical work and practises.
C8: "What should we be measuring to quantify performance? Which metrics are used?": At least in image-wise (patient-wise) experiments, the receiver operating characteristic (ROC) curve together with several measurement points on the curve provide the means for the design and comparisons as in medical practise where the sensitivity and specificity values (e.g., correctly classified normal images vs. correctly classified abnormal images) are in common use [5], [6].

There are three essential components for benchmarking medical image analysis algorithms: 1) true patient images, 2) ground truth from experts, and 3) an evaluation protocol. The key questions $C1 - C8$ are utilised here to acknowledge the current benchmarking practises in medical image analysis. In this paper, we focus on eye fundus images, containing a long-term research tradition. The most important public eye fundus databases are as follows: STARE (Structured analysis of the retina) [7], DRIVE (Digital retinal images for vessel extraction) [8], MESSIDOR (Methods to evaluate segmentation and indexing techniques in the

field of retinal ophthalmology) [9], CMIF (Collection of multispectral images of the fundus) [10], ROC (Retinopathy online challenge) [11], and REVIEW (Retinal vessel image set for estimation of width) [12]. A summary of the main properties (highlighted by the key questions) is given in Table 1. To compare the reference databases with the proposed framework and the DiaRetDB1 database in terms of the key questions, a corresponding summary is given in Table 2.

Table 1. Summary of the surveyed fundus image databases.

Key questions	STARE (vessel)	STARE (disc)	DRIVE	MESSIDOR	CMIF	ROC	REVIEW
$C2$: *"Is there a data set for which the correct answers are known?"*	x		x	x		x	x
$C3$: *"Are there data sets in common use?"*	x	x	x	x	x	x	x
$C4$: *"Are there experiments which show algorithms are stable and work as expected?"*	x		x			x	
$C5$: *"Are there any strawman algorithms?"*	x	x	x				
$C6.1$: *"What code is available?"*						x	
$C6.2$: *"What data is available?"*	x	x	x	x	x	x	x
$C7$: *"Is there a quantitative methodology for the design of algorithms?"*							
$C8.1$: *"What should we be measuring to quantify performance?"*	x	x	x			x	x
$C8.2$: *"What metrics are used?"*		x	x			x	x
$\sum$	6	5	7	3	2	7	5

Table 2. Summary of the DiaRetDB1 V2.1 database.

Key questions	DIARETDB1 V2.1
$C2$: *"Is there a data set for which the correct answers are known?"*	Yes
$C3$: *"Are there data sets in common use?"*	Yes (publicly available at [22]).
$C4$: *"Are there experiments which show algorithms are stable and work as expected?"*	Experimental results reported in Section 4.4 strawman algorithm.
$C5$: *"Are there any strawman algorithms?"*	Yes (description in Section 4)).
$C6.1$: *"What code is available?"*	Functionality for reading/writing images and ground truth, strawman algorithm, and annotation software (publicly available at [22, 23])
$C6.2$: *"What data is available?"*	Images and ground truth (xml) (publicly available at [22]).
$C7$: *"Is there a quantitative methodology for the design of algorithms?"*	Medical practise used as a guideline at each development step.
$C8.1$: *"What should we be measuring to quantify performance?"*	Image- and pixel-based ROC-analysis (description in Section 4).
$C8.2$: *"What metrics are used?"*	Equal error rate (EER) defined in Section 4.

3 Patient Images and Ground Truth

3.1 Collecting Patient Images

Patient images are captured and selected by medical doctors or other trained persons. For a selected image set, two issues should be justified: 1) distribution

correspondence with the desired population and 2) privacy protection of patient data. In DiaRetDB1, the ophthalmologists wanted to investigate the accuracy of automatic methods analysing fundus images of patients who are in a serious risk of having diabetic retinopathy, providing clear findings. This studied sub-population is routinely screened by the Finnish primary health-care. Privacy protection of patient data considers the ethics of clinical practise, medical research, including permissions from a national ethics committee and patients, and also to data security, meaning that all patient information, including hidden metadata, must be explicitly removed from images in a public database. In DiaRetDB1, the fundus images were acquired using a standard fundus camera and were converted to raw bitmaps and then saved to portable network graphics (PNG) format using lossless compression and hidden metadata removed.

3.2 Image Annotations as the Ground Truth

There is a need for computer assisted annotation as originally discussed in [3] and [2]. In general, image ground truth markings are essential for training supervised algorithms as well as for their evaluation and comparison. The information is typically constructed by manually annotating a set of images, and commonly, simple tailored tools are used to collect the data. Annotating medical images, the two essential considerations apply: 1) annotations must be performed by clinically qualified persons (specialised or specialising medical doctors, or other trained professionals for specific tasks), denoted as "experts", and 2) the ground truth should include annotations from several experts.

To avoid biasing the results, the experts should be given minimal guidance for their annotation work. Moreover, basic image manipulation for viewing the images is needed, and a set of geometric primitives is provided for making the spatial markings. Ophthalmologists found the following polygon derived primitives useful: *small circle*, which can be quickly put on a small lesion, and *circle area* and *ellipse area* which are described by their centroid, radius/radii, and orientation (ellipse). Our system also requires at least one *representative point* for each lesion (the most salient cue, such as its colour or texture, describing the specific lesion). Furthermore, *confidence* from a set of three discrete values, low, moderate or high, is required for every marking. It is wise to define beforehand the types of markings, i.e., the class labels for the lesions (e.g., in DiaRetDB1: Hard exudates, Soft exudates, Microaneurysms, Haemorrhages).

Our tool is available at `http://www.it.lut.fi/project/imgannotool/` as Matlab M-files and as a Windows executable. Matlab is not the optimal environment for developing GUI based applications, but it is widely used in scientific computing. The default GUI is shown in Fig. 1.

3.3 Medical Markings Data Format

To store the annotated markings and to be able to restore their graphical layout, we need to define a data format. The data is naturally structured, and therefore,

structural data description languages are preferred. Several protocols for describing medical data exist, such as HL7 based on the extensible markup language (XML) [13], but these are complex protocols designed for patient information exchange between organisations and information systems. Since our requirements are considerably less comprehensive, we adopt our own light-weight data format based on the XML data description language, the Document Type Definition (DTD) description. The proposed data format was not used in [3, 2], but the original data was converted to the XML format without loss of information.

3.4 Fusion of Annotations from Multiple Experts

Fusing multiple expert annotations was originally studied in [15], and is revised here. An important question for training, evaluating, and benchmarking is how the annotations from multiple experts should be combined: 1) How to resolve inconsistencies in the annotations from a single expert? 2) How to fuse equally trustworthy information from multiple sources (multiple expert co-fusion)?

In our data format, the available expert information is the following (Fig. 1): 1) spatial coverage (polygon area), 2) representative point(s) (small circle areas), and 3) the subjective confidence level. The representative points are distinctive "cue locations" that attracted the expert's attention to the lesion. The confidence level describes the expert's subjective confidence for the lesion to represent a specific class as shown in Fig. 2. Three intuitive solutions exist for the fusion problem: i) fixed size neighbourhoods around the representative points (Fig. 3(b)), ii) union of spatial coverage and thresholded by a fixed confidence (Fig. 3(a)), and iii) interest points masked with confidence regions (Fig. 3(c)).

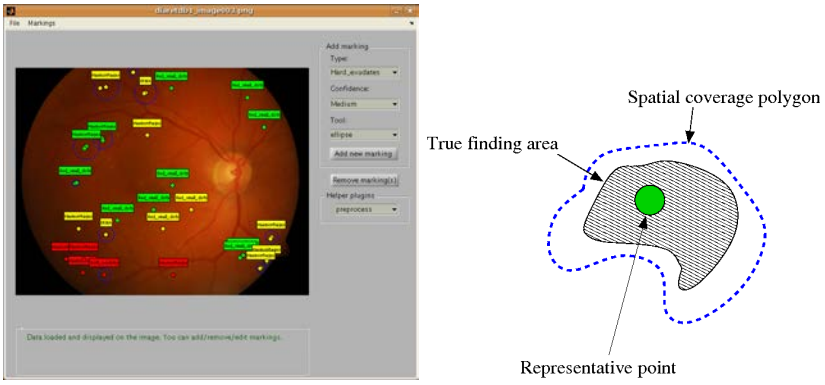


Fig. 1. GUI of the image annotation tool and parts of a single expert annotation.

The area intersection provided the best fusion results in all experimental setups [15], and is computed in a straightforward manner as a sum of expert

Fig. 2. Four independent markings (contours and representative points) for the same lesion (Hard exudates). The representative point markers denote the confidence level (*square* = 100%, *triangle* > 50% and *circle* < 50%).

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

Fig. 3. Illustration of the fusion approaches for the annotations in Fig. 2: (a) Area intersection (blue denotes areas for the confidence level 0.25, red for 0.75, and green for 1.00); (b) Representative point neighbourhoods (5×5); (c) Representative point neighbourhoods masked with the area corresponding 0.75 confidence; (d) Summed area confidences; (f) Close-up of masked representative points.

annotated confidence images divided by the number of experts (see Fig. 3(d)). For DiaRetDB1, the fused confidence of 0.75 yielded to the best results [15], resolving the inconsistencies of annotations either from a single expert or multiple expert co-fusion problems.

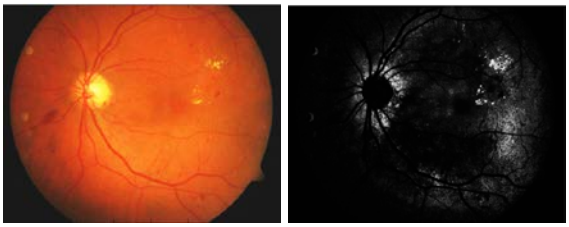
4 Algorithm Evaluation

4.1 Evaluation Methodology

The ROC based analysis perfectly suits to medical decision making, being the acknowledged methodology in medical research [14]. An evaluation protocol based on the ROC analysis was proposed in [3] for image-based (patient-wise) evaluation and benchmarking, and the protocol was further studied in [15]. In clinical medicine, the terms *sensitivity* and *specificity* defined in the range $[0\%, 100\%]$ or $[0, 1]$, are used to compare methods and laboratory assessments. The $sensitivity = \frac{TP}{TP+FN}$ depends on the diseased population whereas the $specificity = \frac{TN}{TN+FN}$ on the healthy population, defined by true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The x-axis of a ROC curve is 1-specificity, whereas the y-axis represents directly the sensitivity [6].

It is useful to form a ROC-based quality measure. The equal error rate (EER) [16] or weighted error rate (WER) [17] are prefered. The main difference between the two measures is that EER assumes equal penalties for both false positives and negatives, whereas in the WER, the penalties are adjustable.

In the image-based evaluation, a single likelihood value for each lesion should be produced for all test images. Using the likelihood values, a ROC curve can be automatically computed [15]. If a method provides multiple values for a single image, such as the full image likelihood map in Fig. 4(b), then the values must be fused to produce a single score.



(a) (b)

Fig. 4. Pixel-wise likelihoods for Hard exudates produced by the strawman algorithm (explained later): (a) Original image (hard exudates are the small yellow spots in the right part of the image); (b) "Likelihood map" for hard exudates.

4.2 Image-based Evaluation

We follow the medical practise where the decisions are "patient-wise" [18]. The image analysis system is treated as a black-box which takes an image as the input. The system produces a score that corresponds to the probability of the image being abnormal, and a high score corresponds to high probability. The objective of the image-based evaluation protocol is to generate a ROC curve by manipulating the score values of the test images.

Using the Bayesian theory, if we denote the score as likelihoods $p(I|abnorml$ and $p(I|normal)$, and let $\omega_k = [normal, abnormal]$ be the decision, the test image I is assigned to the class ω_j by the maximum a posterior rule

$$P(\omega_j|I) = \max_k P(\omega_k|I), \tag{1}$$

where

$$P(\omega_k|I) = \frac{p(I|\omega_k)P(\omega_k)}{\sum_{i=1}^{2} p(I|\omega_i)P(\omega_i)} \quad . \tag{2}$$

The prior values can be based on the training set or population characteristics, or they can be set as equal (default). The image score based evaluation method is presented in Algorithm 1.

Algorithm 1 Image-wise evaluation based on image scores

```
for each test images do
    curr_score ← image score
    for each test image do
        if curr_score ≥ test image score then
            assign "normal"
        else
            assign "abnormal"
        end if
    end for
    Compare test image assignments to the ground truth assignments and compute
    (sensitivity, specificity)-pair
    Add new ROC point (x, y) = (1-specificity, sensitivity)
end for
Return the final ROC curve (all points)
```

The image-based evaluation method is general since it requires only the scores for each test image. If we need to evaluate the performance in case of a method producing multiple values, e.g., a spatial likelihood map illustrated in Fig. 4, an additional procedure is needed to fuse the multiple values into a single score.

The score fusion is performed as follows: If we consider M medical evidences (features) extracted from the image, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$, where each evidence is a vector, then we can denote the score value of the image as $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M|abnormal)$. The joint probability is approximated from the classification results (likelihoods) in terms of decision rules using the combined classifier theory (classifier ensembles) [19]. The decision rules for deriving the score were compared in the recent study [15] where the rules were devised based on Kittler et al. [19] and a new intuitive rank-order based rule "summax" which defines the image score $p(\boldsymbol{x}_1 \ldots \boldsymbol{x}_M|abnormal)$ using the compared decision rules when the prior values of the population characteristics are equal ($P(normal) = P(abnormal)$) as follows:

$$\text{SCORE}_{summax} = \sum_{m \in N_{Y\%}} p(\boldsymbol{x}_m|abnormal) \tag{3}$$

where $N_{Y\%}$ are the indices of $Y\%$ top scoring pixel scores. Experimeting also with max, mean, and product rules, strong empirical evidence supports the rank-order based sum of maxims (summax; portion fixed to 0.01). [15]

4.3 Pixel-based Evaluation

To validate a design choice in method development, we also measure the spatial accuracy, i.e., whether the detected lesions are found in correct locations. Thefore, we propose also a pixel-based evaluation protocol (Algorithm 2) which is analogous to the image-based evaluation.

With a global pixel-wise score (curr_pix_score), the pixels in all test images are classified to either normal or abnormal. Now, the sensitivity and specificity

can be computed for each image by comparing the classified pixels to the disambiguated ground truth. Note that the sensitivity values are computed only for the abnormal test images whereas the specificity values are computed for all test images. This does not allow to determine the ROC curves for each test image, but if the procedure is repeated with a varying score, then the mean ROC curve can be computed.

Algorithm 2 Pixel-wise evaluation based on pixel scores

```
1: Form a list of tested pixel scores
2: for each tested pixel score (curr_pix_score) do
3:    for each test image do
4:       for each test image pixel score do
5:          if curr_pix_score ≥ pixel score then
6:             assign pixel is "normal"
7:          else
8:             assign pixel is "abnormal"
9:          end if
10:      end for
11:   end for
12:   Compare test image assignments to the spatial ground truth assignments and
      compute (sensitivity, specificity)-pair (over all pixels in all images)
13:   Add new ROC point (x, y) = (1-specificity, sensitivity)
14: end for
15: Return the final ROC curve (all points)
```

4.4 Strawman Algorithm

We provide a baseline method in the form of a strawman algorithm (Algorithm 3) [15] which users of the database may find it easier to start to use the data and to self-evaluate the maturity and applicability of their methods. The strawman algorithm is based on the use of photometric cue. The strawman results for DiaRetDB1 are show in Fig. 5 (ROC curves) and in Table 3 (EER values).

Algorithm 3 Strawman algorithm

1: Extract colour information (r, g, b) of the lesion from the train set images (Sec. 3.4).
2: Estimate $p(r, g, b|lesion)$ from the extracted colour information using GMM-FJ. [20, 21]
3: Compute $p(r, g, b|lesion)$ for every pixel in the test image (repeat step for every test image in the test set).
4: Evaluate the performance (Sec. 4).

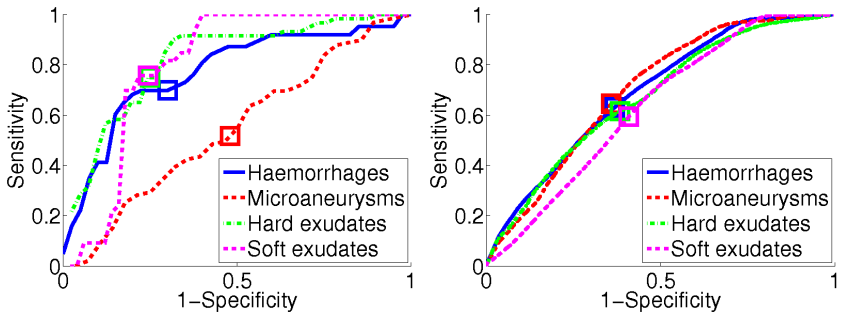


Fig. 5. ROC curves for the DiaRetDB1 strawman algorithm (*square* = *EER*): (a) image-based; (b) pixel-based.

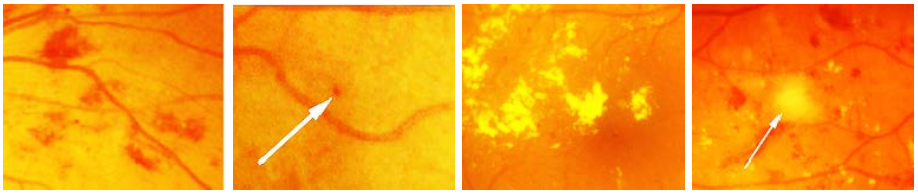
Table 3. EER results for the DiaRetDB1 strawman algorithm.

	Ha			*Ma*			*He*			*Se*			*All lesions avg.*		
	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean
Image-based	0.25	0.33	0.30	0.43	0.50	0.48	0.22	0.27	0.24	0.18	0.28	0.24	0.27	0.35	0.31
Pixel-based	0.37	0.37	0.37	0.35	0.36	0.36	0.37	0.44	0.39	0.40	0.41	0.40	0.37	0.40	0.38

5 Example – DiaRetDB01 diabetic retinopathy database and protocol V2.1

The authors have published two medical image databases with the accompanied ground truth: DiaRetDB0 and DiaRetDB1. The work on DiaRetDB0 provided us essential information how diabetic retinopathy data should be collected, stored, annotated and distributed. DiaRetDB1 was a continuation to establish a better database for algorithm evaluation. DiaRetDB1 contains eye fundus images selected by experienced ophthalmologists. The lesion types of interest were selected by medical doctors (see Fig. 6): microaneurysms (distensions in the capillary), haemorrhages (caused by ruptured or permeable capillaries), hard exudates (leaking lipid formations), soft exudates (microinfarcts), and neovascularisation (new fragile blood vessels). These lesions are symptoms of mild, moderate, and severe non-proliferative diabetic retinopathy, and they provide evidence for the early diagnosis. The images were annotated by four independent and experienced medical doctors inspecting similar images in their regular work.

The images and ground truth can be downloaded from [22]. The images are in PNG format, and the ground truth annotations follow the XML format. Moreover, we provide a DiaRetDB1 kit containing full Matlab functionality (M-files) for reading and writing images and ground truth, fusing expert annotations, and generating image based evaluation scores. The whole pipeline from images to evaluation results (including the strawman algorithm) can be tested using the



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

Fig. 6. Abnormal eye fundus findings caused by the diabetes (best viewed in colour): (a) haemorrhages; (b) microaneurysms (marked with an arrow); (c) hard exudates; (d) soft exudate (marked with an arrow).

provided functionality on the weg page. The annotation software (Matlab files and executables) is available at [23].

6 Conclusions

We have discussed the problem of establishing benchmark databases for the development of medical image analysis. We have pointed out the importance of commonly accepted and used databases. We have proposed reusable tools needed to solve the important sub-tasks, put our implementations publicly available, and established the diabetic retinopathy database DiaRetDB1 to promote and help other researchers to collect and publish their data. We believe that public databases and common evaluation procedures significantly support the development and enable the best methods to be adopted in clinical practise.

Acknowledgements

The authors thank the Finnish Funding Agency for Technology and Innovation (TEKES Project Nos. 40430/05 and 40039/07), and the partners of the ImageRet project (`http://www.it.lut.fi/project/imageret/`) for their support.

References

1. Kauppi, T.: Eye Fundus Image Analysing for Automatic Detection of Diabetic Retinopathy. PhD Thesis, Lappeenranta University of Technology (2010)
2. 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. BMVC (2007)
3. Kauppi, T, Kalesnykiene, V., Kamarainen, J.-K., Lensu, L., Sorri, I., Raninen, A., Voutilainen, R., Pietilä, J., Kälviäinen, H., Uusitalo, H.: DIARETDB1 diabetic retinopathy database and evaluation protocol. In: Proc. MIUA (2007)

4. Thacker, N.A., Clark, A.F., Barron, J.L., Beveridge, J.R., Courtney, P., Crum, W.R., Ramesh, V., Clark, C.: Performance characterization in computer vision: A guide to best practices. CVIU 109, 305–334 (2008)
5. Zou, K.H.: Receiver operating characteristic (ROC) literature research, http:\/\/splweb.bwh.harvard.edu:8000\/pages\/ppl\/zou\/roc.html
6. Fawcett, T,: An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874 (2006)
7. STructured Analysis of the Retina (STARE), http://www.ces.clemson.edu/~ahoover/stare/
8. Digital Retinal Images for Vessel Extraction (DRIVE), http://www.isi.uu.nl/Research/Databases/DRIVE/
9. Methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology (MESSIDOR), http://messidor.crihan.fr
10. Collection of multispectral images of the fundus (CMIF), http://www.cs.bham.ac.uk/research/projects/fundus-multispectral/
11. Retinopathy Online Challenge (ROC), http://roc.healthcare.uiowa.edu/
12. REVIEW: Retinal Vessel Image set for Estimation of Widths (REVIEW), http://reviewdb.lincoln.ac.uk/
13. Application Protocol for Electronic Data Exchange in Healthcare Environments Version 2.5.1, ANSI Standard, http://www.hl7.org
14. Lasko, T.A., Bhagwat, J.G., Zou, K.H., Ohno-Machado, L.: The use of receiver operating characteristic curves in biomedical informatics. J of Biomedical Informatics 38, 404–415 (2005)
15. 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: Proc. SCIA (2009)
16. Phillips, P.J. Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET Evaluation Methodology for Face-Recognition Algorithms. IEEE TPAMI 22, 1090-1104 (2000)
17. Bailliere, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariethoz, J., Matas, J., Messer, K., Popovic, V., Poree, F., Ruiz, B., Thiran, J.P.: The BANCA Database and Evaluation Protocol. In: Proc. AVBPA (2003)
18. Everingham, M, Zisserman, A.: The Pascal Visual Object Classes Challenge VOC2006 Results. In: Proc. ECCV Workshop of VOC (2006)
19. Kittler, J., Hatef, M., R.,Duin, R.P.W., Matas, J.: On Combining Classfiers. IEEE TPAMI 20, 226–239 (1998)
20. Figueiredo, M.A.T., Jain, A.K.: Unsupervised learning of finite mixture models. IEEE TPAMI 24, 381–396 (2002)
21. Paalanen, P., Kamarainen, J.-K., Ilonen, J., Kälviäinen, H.: Feature Representation and Discrimination Based on Gaussian Mixture Model Probability Densities - Practices and Algorithms. Pattern Recognition 39, 1346–1358 (2006)
22. Diabetic retinopathy database and evaluation protocol (DIARETDB1), http://www.it.lut.fi/project/imageret/files/diaretdb_v02_01/
23. Image annotation tool (IMGANNOTOOL), http://www.it.lut.fi/project/imgannotool