DIARETDB1 diabetic retinopathy database and evaluation protocol

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

[1] Machine Vision and Pattern Recognition Research Group
Lappeenranta University of Technology, Finland

[2] Department of Ophthamology
Faculty of Medicine, University of Kuopio, Finland

[3]University of Tampere

[4] Perimetria Ltd., Helsinki, Finland

Abstract

For a particularly long time, automatic diagnosis of diabetic retinopathy from digital fundus images has been an active research topic in the medical image processing community. The research interest is justified by the excellent potential for new products in the medical industry and significant reductions in health care costs. However, the maturity of proposed algorithms cannot be judged due to the lack of commonly accepted and representative image database with a verified ground truth and strict evaluation protocol. In this study, an evaluation methodology is proposed and an image database with ground truth is described. The database is publicly available for benchmarking diagnosis algorithms. With the proposed database and protocol, it is possible to compare different algorithms, and correspondingly, analyse their maturity for technology transfer from the research laboratories to the medical practice.

1 Introduction

Diabetes has become one of the rapidly increasing health threats worldwide [21]. Only in Finland, there are 30 000 people diagnosed to the type 1 maturity onset diabetes in the young, and 200 000 people diagnosed to the type 2 latent autoimmune diabetes in adults [4]. In addition, the current estimate predicts that there are 50 000 undiagnosed patients [4]. Proper and early treatment of diabetes is cost effective since the implications of poor or late treatment are very expensive. In Finland, diabetes costs annually

505 million euros for the Finnish health care, and 90% of the care cost arises from treating the complications of diabetes [5]. These alarming facts promote the study of automatic diagnosis methods for screening over large populations.

Fundus imaging has an important role in diabetes monitoring since occurrences of retinal abnormalities are common and their consequences serious. However, since the eye fundus is sensitive to vascular diseases, fundus imaging is also considered as a candidate for non-invasive screening. The success of this type of screening approach depends on accurate fundus image capture, and especially on accurate and reliable image processing algorithms for detecting the abnormalities.

Numerous algorithms have been proposed for fundus image analysis by many research groups [13, 6, 25, 15, 18]. However, it is impossible to judge the accuracy and reliability of the approaches because there exists no commonly accepted and representative fundus image database and evaluation protocol. With a widely accepted protocol, it would be possible to evaluate the maturity and state-of-the-art of the current methods, i.e., produce the achieved sensitivity and selectivity rates.For example, commonly accepted strict guidelines for the evaluation of biometric authentication methods, such as the FERET and BANCA protocols for face recognition methods [16, 2], have enabled the rapid progress in that field, and the same can be expected in medical image processing related to diabetic retinopathy detection.

The main contribution of this work is to report a publicly available diabetic retinopathy database, DIARETDB1, containing the ground truth collected from several experts and a strict evaluation protocol. The protocol is demonstrated with a baseline method included to the available tool kit. This study provides the means for the reliable evaluation of automatic methods for detecting diabetic retinopathy.

2 Diabetic retinopathy

In the type 1 diabetes, the insulin production in the pancreas is permanently damaged, whereas in the type 2 diabetes, the person is suffering from increased resistance to insulin. The type 2 diabetes is a familial disease, but also related to limited physical activity and lifestyle [21]. The diabetes may cause abnormalities in the retina (diabetic retinopathy), kidneys (diabetic nefropathy), and nervous system (diabetic neuropathy) [14]. The diabetes is also a major risk factor in cardiovascular diseases [14].

The diabetic retinopathy is a microvascular complication of diabetes, causing abnormalities in the retina, and in the worst case, blindness. Typically there are no salient symptoms in the early stages of diabetic retinopathy, but the number and severity predominantly increase during the time. The diabetic retinopathy typically begins as small changes in the retinal capillaries. The first detectable abnormalities are mircroaneurysms (Ma) (Fig. 1(a)) which are local distensions of the retinal capillary and which cause intraretinal hemorrhage (H) (Fig. 1(b)) when ruptured. The disease severity is classified as mild non-proliferative diabetic retinopathy when the first apparent microaneurysms appear in the retina [24]. In time, the retinal edema and hard exudates (He) (Fig. 1(c)) are followed by the increased permeability of the capillary walls.The hard exudates are lipid formations leaking from these weakened blood vessels. This state of the retinopathy is called moderate non-proliferative diabetic retinopathy [24]. However, if the above-mentioned abnormalities appear in the central vision area (macula),

it is called diabetic maculopathy [21]. As the retinopathy advances, the blood vessels become obstructed which causes microinfarcts in the retina. These microinfarcts are called soft exudates (Se) (Fig. 1(d)). When a significant number of intraretinal hemorrhages, soft exudates, or intraretinal microvascular abnormalities are encountered, the state of the retinopathy is defined as severe nonproliferative diabetic retinopathy [24].

The severe nonproliferative diabetic retinopathy can quickly turn into proliferative diabetic retinopathy when extensive lack of oxygen causes the development of new fragile vessels [24]. This is called as neovascularisation (Fig. 1(e)) which is a serious eye sight threatening state. The proliferative diabetic retinopathy may cause sudden loss in visual acuity or even a permanent blindness due to vitreous hemorrhage or tractional detachment of the central retina. After diagnosis of diabetic retinopathy, regular monitoring is needed due to the progressive nature of the disease. However, broad screenings cannot be performed due to the fact that the fundus image examination requires attention of medical experts. For the screening, automatic image processing methods must be developed.

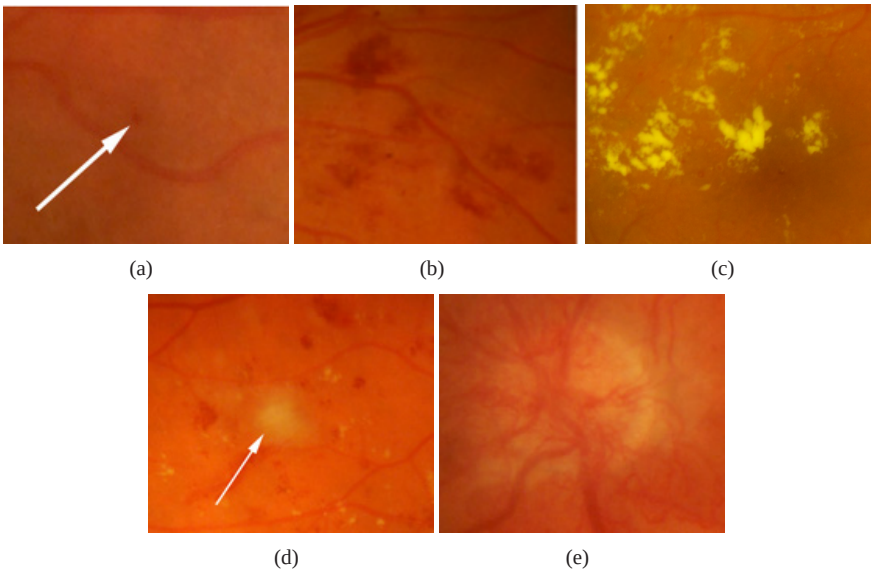


Figure 1: Abnormal findings in the eye fundus caused by the diabetic retinopathy: (a) microaneuryms (marked with an arrow), (b) hemorrhages, (c) hard exudates, (d) soft exudate (marked with an arrow), and (e) neovascularization.

2.1 Current evaluation practises

In medical diagnosis, the medical input data is usually classified into two classes, where the disease is either present or absent. The classification accuracy of the diagnosis is assessed using the sensitivity and specificity measures. Following the practises in the medical research, the fundus images related to the diabetic retinopathy are evaluated by using sensitivity and specificity per image basis. Sensitivity is the percentage of abnormal funduses classified as abnormal, and specificity is the percentage of normal fundus classified as normal by the screening. The higher the sensitivity and specificity

values, the better the diagnosis. Sensitivity and specificity can be computed as [22]:

$$\text{sensitivity (SN)} = \frac{T_P}{T_P + F_N}, \text{ specificity (SP)} = \frac{T_N}{T_N + F_P} \quad (1)$$

where T_P is the number of abnormal fundus images found as abnormal, T_N is the number of normal fundus images found as normal, F_P is the number of normal fundus images found as abnormal (false positives) and F_N is the number of abnormal fundus images found as normal (false negatives). Sensitivity and specificity are also referred to as the true positive rate (TPR) and true negative rate (TNR), respectively.

2.2 Automatic methods

As mentioned previously, the diagnosis of diabetic retinopathy can be divided into the following two categories:

1. Screening of the diabetic retinopathy
2. Monitoring of the diabetic retinopathy

Most automatic systems approach the detection directly using shape, color, and domain knowledge of diabetic retinopathy findings, but the abnormalities can also be found indirectly by detecting changes between two fundus images taken from the same eye in different time moment [11, 17]. The direct approach contributes to screening of the disease, where indirect approach contributes to both screening and monitoring of the diabetic retinopathy. Both approaches use roughly the following stages for finding abnormalities in fundus images: 1) image enhancement 2) candidate diabetic retinopathy finding detection 3) classification to correct diabetic retinopathy category (or hypothesis rejection).

Some of the main features distinguishing between the different findings and normal fundus parts are the color and brightness. The same features have been verified also by ophthalmologists. Unsurprisingly these features dominate in the automatic methods, and therefore will be shortly reviewed in our brief surveys for different type of findings in Section 2.2.1 and Section 2.2.2. Most of the automatic methods also detect normal fundus parts, such as optic disk, blood vessels, and macula. The automatic methods either use the vital domain information provided by the normal fundus parts or remove them due to their similar color and shape appearance with abnormal fundus findings. The detection of normal fundus parts is not considered in this study.

2.2.1 Microaneurysms and hemorrhages

Image enhancement methods: Niemeijer *et al.* [13] estimated non-uniform background intensity of fundus image by applying median filtering to the green channel of the fundus image. Shade correction was generated by subtracting the result from the original green channel.

Fleming *et al.* [6] had similar approach for microaneurysms, but the green channel of the original fundus image was divided with the background intensity image. In addition, the shade corrected image was normalized for global image contrast by dividing

with its standard deviation. Multiple local contrast enhancement methods were tested to improve detection accuracy.

In hemorrhage detection, Zhang and Chutape [25] used histogram specification applied to each individual RGB color component to normalize the colors between different fundus images.

Sinthayothin *et al.* [20] used local contrast enhancement to equalize the intensity variation in fundus images. The fundus images were transformed from RGB color model to IHS color model and the local contrast enhancement was applied to the intensity component of the image.

Detection and classification methods: Niemeijer *et al.* [13] extracted the candidate finding areas by assigning posterior probability of being red finding for every pixel using Gaussian filter and its derivates as features for k-nearest neighbor clustering. Shape and intensity properties of the candidate areas were used for more accurate abnormal red finding and normal red finding classification.

Fleming *et al.* [6] segmented candidate microaneurysm areas by applying region growing to image enhanced with morphological top-hat operation and thresholding. The result candidate areas were classified with k-nearest neighbor clustering using the shape and intensity information.

Zhang and Chutape [25, 26] used hemorrhage areas restricted by finite window in training images as input for support vector machine. To detect different sized hemorrhages a pyramid of images was generated by changing the resolution of fundus image. The local minima of the support vector machine provided evidence map were selected as hemorrhage locations. The principal component analysis was used to reduce the complexity of feature space.

Sinthanayothin *et al.* [20] sharpened the edges of red finding regions by applying moat operator to green channel of the contrast enhanced image. From the result image, red findings were extracted with recursive region growing and thresholding.

2.2.2 Hard and soft exudates

Image enhancement methods: Narasimha-iyer *et al.* [11] used normal retinal findings (vasculature, optic disk, fovea, and abnormal findings) to estimate the illumination component using iterative robust homographic surface fitting to compensate the non-uniform illumination in fundus images.

In detection of bright diabetic retinopathy areas from fundus images, Zhang and Chutape [27] applied adaptive local contrast enhancement to sub-image areas using the local mean and standard deviation of intensities. The same approach was used by Osareh *et al.* [15] after color normalization between fundus images using histogram specification.

Wang *et al.* [23] adjusted the image brightness using brightness transform function similar to gamma correction.

Detection and classification methods: Hsu *et al.* [8] determined abnormal and normal finding areas using intensity properties for dynamic clustering. From the result abnormal areas, hard exudates were separated from soft exudates and drusen using intensity contrast information between abnormal areas and immediate background. The domain knowledge of retinal blood vessels were used to remove false artifacts.

Walter *et al.* [22] eliminated the vessels by applying morphological closing to the luminance component of the fundus image. From the result, within a sliding window local standard variation image was calculated and thresholded into coarse exudate areas. More accurate countours were acquired by thresholding difference between original image and morphologically reconstructed image.

Sànchez *et al.* [18] used yellowish color and sharp edges to distinguish hard exudates from the fundus images. The image pixels were classified into background and yellowish objects using minimum distance discrimination, where the countour pixels of extracted optic disk were used as background color reference and pixels inside the contour were used as yellowish object color reference. The segmented yellowish areas and their edge information extracted with Kirsch's mask were combined to hard exudate areas using boolean operator.

Zhang and Chutape [27] located the bright abnormal regions in fundus images by applying fuzzy c-means clustering in LUV color space. The result areas were classified to hard exudates, soft exudates, and normal findings using support vector machine.

Osareh *et al.* [15] searched the coarse hard exudate areas using fuzzy c-means clustering with Gaussian smoothed histograms of each color band of the fundus image. The segmented areas were classified to exudate and non-exudate regions using neural networks. Color, region size, mean and standard deviation of intensity, and edge strength were used as features.

Li and Chutape [10] segmented exudates with combination of Canny edge detection and region growing in LUV color space. Gradient, mean pixel value, and seed pixel value were used as criteria in region growing.

Niemeijer *et al.* [12] used a similar approach for bright abnormal region detection as they used for finding abnormal red regions in [13]. In addition to the previous work, the prior knowledge of optic disk and vascular arch were used to improve detection accuracy.

Sinthanayothin *et al.* [20] clustered similar pixels using intensity difference as criteria for recursive region growing. The region with the most pixels were considered as background and defined the threshold value for hard exudate areas.

Wang *et al.* [23] used spherical color coordinates as features for the classification of fundus image pixels to background and bright abnormal findings using minimum distance discriminant. The abnormal findings were verified using local-window-based method.

3 Evaluation database

A necessary tool for reliable evaluations and comparisons of medical image processing algorithms is a database of dedicatedly selected high-quality medical images which are representatives of the problem and have been verified by experts. In addition, information about the medical findings, the ground truth, must accompany the image data. An accurate algorithm should take the images as input, and produce output which is consistent with the ground truth. In the evaluation, the consistency is measured, and algorithms can be compared based on these performance metrics. In the following, we describe the images and ground truth for the diabetic retinopathy database DIARETDB1.

3.1 Fundus images

The database consists of 89 colour fundus images of which 84 contain at least mild non-proliferative signs (Ma) of the diabetic retinopathy (two examples shown in Figs. 2(b) and 2(c)), and 5 are considered as normal which do not contain any signs of the diabetic retinopathy according to all experts participated in the evaluation (an example shown in Fig. 2(a)). The images were taken in the Kuopio university hospital. The images were selected by the medical experts, but their distribution does not correspond to any typical population, i.e., the data is biased and no a priori information can be devised from it. The diabetic retinopathy abnormalities in the database are relatively small, but they appear near the macula which is considered to threaten the eyesight. Images were captured with the same 50 degree field-of-view digital fundus camera with varying imaging settings (flash intensity, shutter speed, aperture, gain) controlled by the system. The images contain a varying amount of imaging noise, but the optical aberrations (dispersion, transverse and lateral chromatic, spherical, field curvature, coma, astigmatism, distortion) and photometric accuracy (colour or intensity) are the same. Therefore, the system induced photometric variance over the visual appearance of the different retinopathy findings can be considered as small. The data correspond to a good (not necessarily typical) practical situation, where the images are comparable, and can be used to evaluate the general performance of diagnostic methods. The general performance corresponds to the situation where no calibration is performed (actual physical measurement values cannot be recovered), but where the images correspond to commonly used imaging conditions, i.e., the conditions encountered in hospitals. This data set is referred to as "calibration level 1 fundus images". A data set taken with several fundus cameras containing different amounts imaging noise and optical aberrations is referred to as "calibration level 0 fundus images".

(a) (b) (c)

Figure 2: Examples of DIARETDB1 fundus images: (a) normal fundus, (b) abnormal fundus, and (c) abnormal fundus after treatment by photocoagulation.

3.2 Ground truth

The most important accuracy measures for medical diagnosis methods are *sensitivity* and *specificity* (see Sec. 2.1 for the definitions). Sensitivity and specificity are defined on the image basis – an image either contains a specific finding or not despite the fact that the diabetic retinopathy findings do have spatial locations in the fundus . For the computer vision researchers, it is important to ensure that the automatically extracted diabetic retinopathy findings also spatially correspond the findings marked by experts, that is, they appear at the same location in the image. Thus, the more detailed expert

groundtruth contains also the description of visual appearance of diabetic retinopathy findings.

3.2.1 Marking visual findings

The image groundtrtuh is based on expert-selected findings related to the diabetic retinopathy and normal fundus structures (see Fig. 3). A person with a medical education (M.D.) and specialization to ophthalmology is considered as an expert.

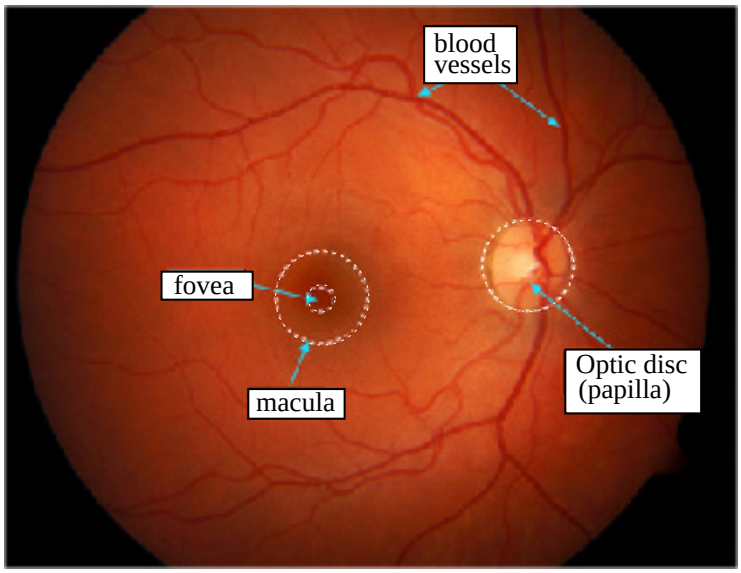


Figure 3: Structural elements of a normal fundus.

A special software tool was provided for the experts to inspect the fundus images and annotate the findings. The user interface of the current software version is shown in Fig. 4. It should be noted that the workstation displays were not calibrated. Therefore, the diabetic retinopathy findings were not equally visible on all displays. However, the situation corresponds to the current best practices.

The experts were asked to mark the areas related to the microaneurysms, hemorrhages, and hard and soft exudates. The experts were instructed to avoid marking the findings so that the borders of the marked areas contain any pixels belonging to the finding. The experts were further instructed to report their confidence and especially annotate the single most representative point for each finding. The ground truth confidence levels, $\{< 50\%, > 50\%, 100\%\}$, represented the certainty of the decision that a marked finding is correct. The experts were taught to use the image annotation tool, but they were not instructed how to make the annotations to prevent a biased scheme; the medical experts learnt their own best practises. Currently, Image annotation tool includes the following graphical directives

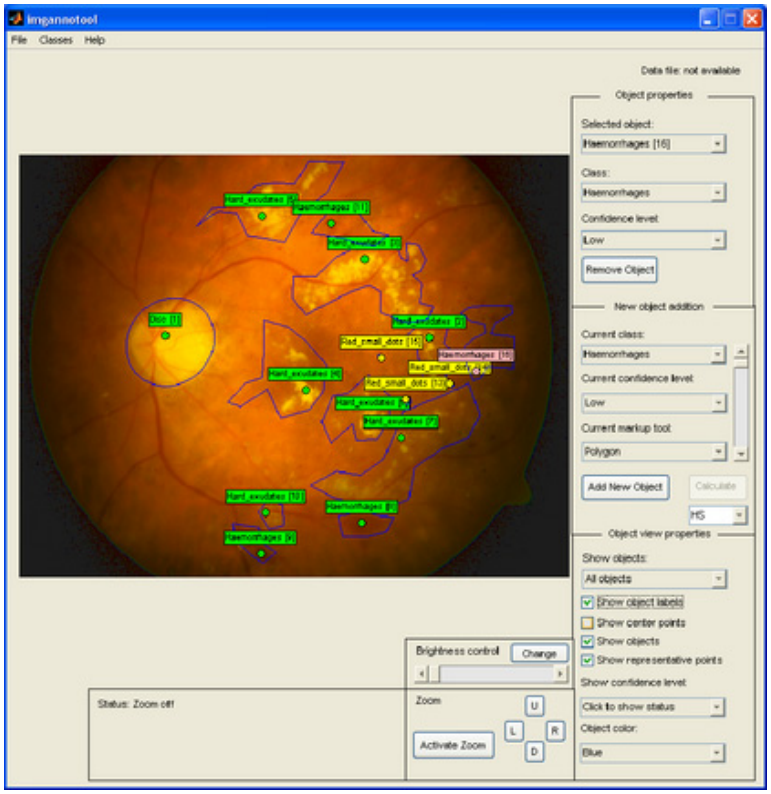
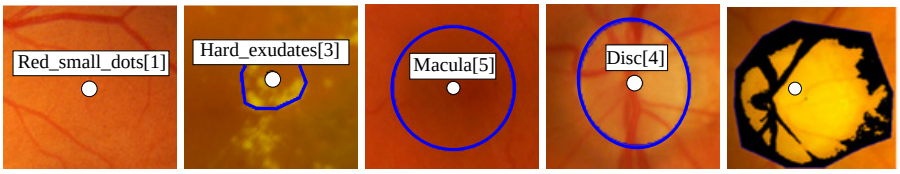


Figure 4: Graphical tool for gathering knowledge from medical experts.

1. Centroid (Fig. 5(a)),
2. Polygon region (Fig. 5(b)),
3. Circle region (Fig. 5(c)),
4. Ellipse region (Fig. 5(c)), and
5. Representative point (Fig. 5(e)).

In addition to the graphical directives, the image annotation tool provided a gamma correction tool and semi-automatic tool for more accurate definition of the finding areas (Fig. 5(e)). The semi-automatic tool used the color information provided by the representative point. In gathering the expert knowledge, the semi-automatic tool was not used.



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

Figure 5: Graphical directives for marking the visual findings: (a) centroid; (b) polygon region; and (c) circle region; d) semi-automatic region cropping tool and representative point.

3.2.2 Data format

The expert knowledge gathered with the groundtrtuh tool is stored to a text file. Each line in the text file corresponds to a visual finding marked with the groundtrtuh tool. The data format for visual finding is defined as

```
CENTROID [POLYGON_REGION] [CIRCLE/ELLIPSE_REGION] RP_POINT REGION_TOOL ...
CONFIDENCE_LEVEL FINDING_TYPE
```

Each marking in image is defined by the centroid and one other graphical directive. The same graphical directive should be consistently used for every medical finding type over the image set. The directive attributes are visualized in Fig. 6 and defined as follows:

```
CENTROID            X Y
POLYGON_REGION      [X_1 X_2 X_3 ...] [Y_1 Y_2 Y_3 ...]
CIRCLE_REGION       [RADIUS]
ELLIPSE REGION      [RADIUS1] [RADIUS2] [ANGLE]
RP_POINT            (X,Y)
REGION_TOOL         [METHOD,VALUE]
CONFIDENCE_LEVEL    HIGH, MEDIUM, LOW
FINDING_TYPE        RED_SMALL_DOTS, HARD_EXUDATES ...
```

where CONFIDENCE and FINDING TYPE values are preset by the experts and stored, but which the user can define.

An example file produced by the groundtrtuh tool based on the findings in Fig. 5 is as follows:

```
233 945 [242,251,245,213,206,] [921,949,971,967,938,] [] [] [] (231,944) [HS,25] High Haemorrhages
354 461 [] [] [102] [94] [117] (413,466) [HS,65] High Disc
543 104 [] [] [41] [] [] (543,104) [HS,39] High Soft_exudates
899 348 [] [] [] [] [] (899,348) [,] Medium Red_small_dots
```

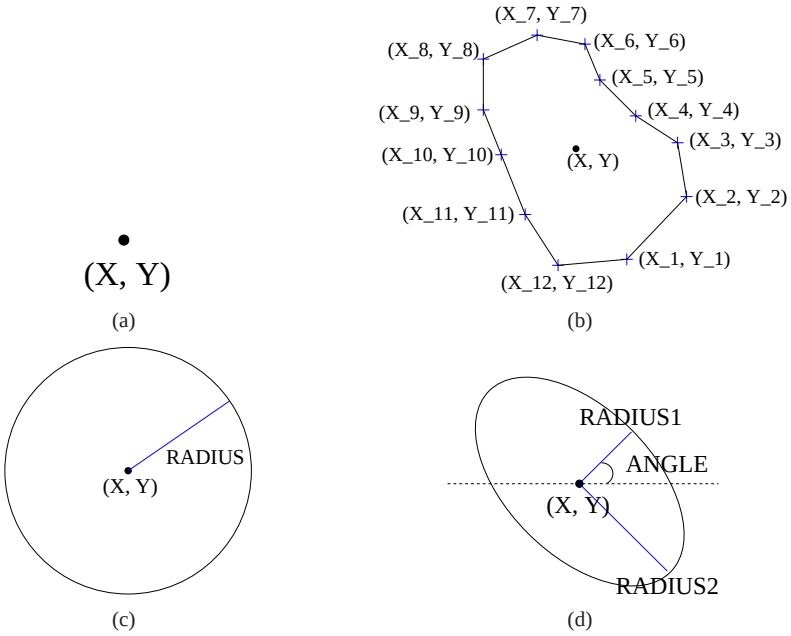
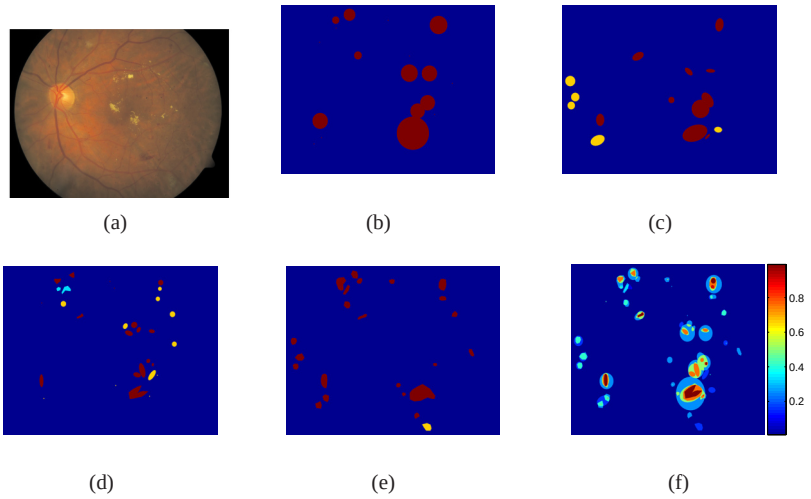


Figure 6: Graphical directives for marking visual findings: (a) a centroid defined by the coordinate pair; (b) a polygon defined by the point coordinates and centroid; (c) a circle defined by the radius and centroid; (d) an ellipse defined by the orientation(angle) and length of major-axis(radius1) and minor-axis(radius2) with centroid.

3.2.3 Refining expert knowledge

The uninstructed collection process caused significant differences between the medical experts as can be seen in Fig. 7. Therefore, it was not possible to use the expert information as such as the ground truth. However, using the original data the expert knowledge was fused for a better spatial accuracy and suppression of outliers. The fusion was performed on a pixel basis using the reported confidence levels.

Several different approaches for fusing the markings are possible, e.g., voting, minimum, maximum and the sum (average) of confidences. The first three provide binary classifications, but the normalised average provides values in the range $[0, 1]$ (see Fig. 7(f)). It should be noted that the markings do not provide any absolute ground truth of the findings, but reveal how the medical experts analyse and interpret the retinopathy from the digital fundus images. Not to discard any information, the approach using the average was selected since it provides a linear confidence scale, and in the evaluation, the confidence level can be fixed to one or several values. In DIARETDB1, the confidence level is fixed to $conf_{GT} = 0.75$.



(a) (b) (c)

(d) (e) (f)

Figure 7: Expert marked hemorrhages for the same image (colour decodes the ground truth confidence): (a) original; (b) expert 1; (c) 2; (d) 3 (d) 4; (e) mean.

3.3 Training and test set

The 89 images were manually assigned into categories representing the progressive states of retinopathy: normal, mild, moderate and severe non-proliferative, and proliferative. Using the categories, the images were divided into the representative training (28 images) and test sets (61 images).

In the training set with $conf_{GT} = 0.75$, 18 images contain hard exudates, 6 soft exudates, 19 microaneurysms, and 21 hemorrhages. In the test set with $conf_{GT} = 0.75$, 20 images contain hard exudates, 9 soft exudates, 20 microaneurysms, and 18 hemorrhages.

3.4 Security

The patient information from the database images were removed by converting the images into raw data. The raw data images were converted to lossless portable network images and marked to belong DIARETDB0 by redefining the comment field to "DIARETDB0 *image_filename* 1500x1152 DirectClassRGB".

4 Evaluation protocol

The training and test images, expert ground truth, a baseline method, Matlab functionality for computing performance measures and more detailed technical documentation are publicly available at the DIARETDB1 web page (`http://www.lut.it.fi/project/imageret`). Research groups investigating diabetic retinopathy detection

methods are encouraged to report their results according to the DIARETDB1 evaluation protocol. The performance measures for which the functionality is provided are defined next.

4.1 Performance measures

In the literature, the sensitivity and specificity values are typically reported since they correspond to the current medical practice and have straightforward interpretations in the medical terms. Sensitivity value depends on the diseased population and specificity on the healthy population (see Eq. 1). These values provide the means for analysing how many diseased and how many healthy patients are correctly diagnosed with a provided method. From the method comparison point of view, however, these two values are not feasible since the two distributions overlap and the true accuracy is always a trade-off.

In our evaluation protocol we have selected two evaluation principles: 1) the evaluation is image-based and 2) is done separately to different diabetic retinopathy findings (Sec. 2). The first principle is justified by the fact that it corresponds to the medical practice where decisions are "patient-based". Spatial area (pixel-wise) based evaluation can be useful in method development, but the problem itself is always image-based. The second principle is due to the practical fact that most researchers concentrate only on one or several finding types and a practically useful method does not necessarily need to detect all findings.

4.1.1 ROC

For a proper comparison the sensitivity and specificity values must be combined into a form which can describe the behavior over different combinations of the values. Receiving operating curve (ROC) is a natural selection due to its popularity and proven applicability in similar computer vision tasks, such as face recognition [7], object class recognition [3] and medical research [9]. The ROC provides a graphical representation for sensitivity (TPR) and 1-specificity (FPR) trade off. The ROC curve provides the means for the optimal analysis when the problem is to find the best method parameters for the task or compare performances irrespective to operating conditions.

In our evaluation we adapted the practises from [3], where each method is required to provide a score for each test image. A high score corresponds to a high probability that a finding is present in the image. By manipulating the provided scores the ROC curve can be automatically generated.

4.1.2 Weighted error rate (WER)

The ROC curve is a reliable method for method comparisons, but often method ranking is also needed, and then, single valued measures must be used. Single valued measures should be derived from the corresponding ROC curve, e.g. by computing an equal error rate (EER) ($TPR = TNR$) or a total area under roc curve (AUC). Here we prefer the more interpretable EER. The EER measure however assumes equal penalties for the both false positives and negatives, which is not typically the case in the medical

diagnosis. Therefore, we adapt a more versatile measure utilised in [16, 2], where the two measures, sensitivity (SN) and specificity (SP), are combined to a weighted error rate defined as

$$WER(R) = \frac{FPR + R \cdot FNR}{1 + R} = \frac{(1 - SP) + R \cdot (1 - SN)}{1 + R} \ . \qquad (2)$$

In (2) $R = \frac{C_{FNR}}{C_{FPR}}$ is the cost ratio between FPR and FNR ($R = 1$ corresponds to the equal penalty for the both). In the DIARETDB1 protocol the following measures are computed: $\mathbf{WER}(\mathbf{10^{-1}})$ (FNR is an order of magnitude less harmful), $\mathbf{WER}(\mathbf{1})$ (FPR and FNR are equally harmful) and $\mathbf{WER}(\mathbf{10})$ (FNR is an order of magnitude more harmful). These measures are computed from the nearest true points on the ROC without interpolation.

5 Evaluation example

In this section we present example evaluation according to the DIARETDB1 protocol. These reproducible results were computed using a simple baseline method which is included to the DIARETDB1 tool kit available at the web site.

5.1 Baseline method

Our method is based on the principle that different findings can be distinguished and detected based only on their photometric information, i.e. colour. We adapt the successful colour locus based face detection by Hadid et al. [1] to our multi-class diabetic retinopathy detection. The approach is justified by the fact that the photometric characteristics (illumination, camera and optics) remain the same in the DIARETDB1 images (calibration level 1). The colour variation should resemble the normal variation within the finding type and between different individuals.

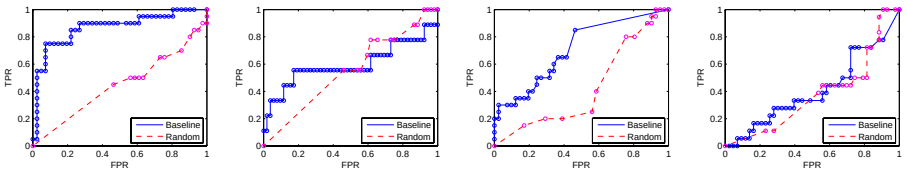
The method utilises two colour channels (e.g. R and G) without intensity component (e.g. normalisation by $R + G + B$). It should be noted that no particular improvement can be achieved by changing the colour space [1], and therefore, RGB was used. A colour locus for each finding type, F_i, is defined by forming their colour histograms $h_{F_i}(r, g)$. The histograms are computed from the intensity normalised pixel colours at the neighborhood (8x8) of the most representative points marked by the experts. By using the colour histograms of findings, $h_{F_i}(r, g)$, and a test image itself, $h_{total}(r, g)$, Schwerdt and Crowley [19] have derived a formula for the Bayesian decision rule to classify a pixel with color (r,g) to one of the finding classes. The formula reduces to the histogram ratio of finding and test image:

$$p(F_i|r, g) = \frac{h_{F_i}(r, g)}{h_{total}(r, g)} \qquad (3)$$

We manually select an optimal posterior threshold for every finding type and compute the sum of pixels having higher or equal posterior value as the image based score.

5.2 Results

Evaluation results, ROC and WER, for the baseline method are shown in Fig. 8 and Table 1, respectively. For a better visualisation also ROC curves of random classification (random score for each test image) are plotted. It is clear that the method performs moderately for the hard exudates while other findings are quite poorly detected. These results can however be used as the baseline which all reported methods should outperform. Only the WER values should be reported, but in Table 1 also the corresponding false positive and negative rates are given.



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

Figure 8: ROC curves for the XXXXXXXX baseline method: a) hard exudates; b) soft exudates; c) microaneurysms; d) hemorrhages.

Table 1: DIARETDB1 baseline performance measures.

Baseline method									
	FPR	FNR	**WER**			FPR	FNR	**WER**	
He	0.8049	0	0.0732	R = 0.1	Se	1.0000	0	0.0909	R = 0.1
	0.0732	0.2500	0.1616	R = 1		0.1731	0.4444	0.3088	R = 1
	0.0244	0.4500	0.0631	R = 10		0	0.8889	0.0808	R = 10
Ma	0.4634	0.1500	0.1785	R = 0.1	H	0.8837	0.2222	0.2824	R = 0.1
	0.4634	0.1500	0.3067	R = 1		0.1628	0.8333	0.4981	R = 1
	0	0.8000	0.0727	R = 10		0	1.0000	0.0909	R = 10

6 Discussion and future research

The development of medical image processing methods to a mature level where they are ready to be transferred from the research laboratories to medical practise requires properly designed benchmarking databases and protocols. The method testing must correspond to the strict regulations in the medical treatment and medicinal research. Medical image processing is not different from the medical practice in that sense.

We proposed a step towards standardized evaluation of methods for detecting findings of diabetic retinopathy by introducing the publicly available DIARETDB1 database and evaluation protocol. The quality and size of the database can be improved but already now the DIARETDB1 corresponds to the situation in practice very well. In the future, however, we will continue to develop the database and evaluation methodology. The following development steps will be taken:

1. A predefined set of instructions are defined for the experts to prevent the free form description, and thus, allow control over subjective interpretations and acquire spatially more accurate ground truth.

2. The effect of display calibration for the experts will be evaluated.

3. Location of normal findings will be added to the ground truth and their evaluation to the protocol

7 Conclusions

An image database, ground truth and evaluation methodology were proposed for evaluating and comparing methods for automatic detection of diabetic retinopathy. All data, a baseline method and evaluation functionality (tool kit) are publicly available at the DIARETDB1 web site (`http://www.it.lut.fi/project/imageret`). The DIAERETDB1 provides a unified framework for benchmarking the methods, but also points out clear deficiencies in the current practice in the method development. The work will continue and the research group's main objective is to publish an ultimate tool for the evaluation of diabetic retinopathy detection methods. The tool will provide accurate and reliable information of method performance to estimate their maturity before starting the technology transfer from the research laboratories to practice and industry.

References

[1] A. Hadid A, M. Pietikäinen, and B. Martinkauppi B. Color-based face detection using skin locus model and hierarchical filtering. In *Proc. 16th International Conference on Pattern Recognition*, pages 196–200.

[2] E. Bailliere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree, B. Ruiz, and J.P. Thiran. The BANCA database and evaluation protocol. In *Proc. of the Int. Conf. on Audio- and Video-Based Biometric Person Authentication*, pages 625–638, 2003.

[3] Mark Everingham and Andrew Zisserman. The pascal visual object classes challenge 2006 (voc2006) results. Workshop in ECCV06, May. Graz, Austria.

[4] Finnish Diabetes Association. Development programme for the prevention and care of diabetes, 2001. ISBN 952 5301-13-3.

[5] Finnish Diabetes Association. Programme for the prevention of type 2 diabetes in Finland, 2003. ISBN 952-5301-36-2.

[6] Alan D. Fleming, Sam Philip, Keith A. Goatman, John A. Olson, and Peter F. Sharp. Automated microaneurysm detection using local contrast normalization and local vessel detection. *IEEE Transactions in Medical Imaging*, 25(9):1223–1232, September 2006.

[7] P.J. Grother, R.J. Micheals, and P.J. Phillips. Face recognition vendor test 2002 performance metrics. In *Proc. of the. 4th Int. Conf. on Audio- and Video-based Biometric Person Authentication*, 2003.

[8] Wynne Hsu, P. M. D. S Pallawa, M. Li Lee, and Au Eong K.-G. The role of domain knowledege in the detection of retinal hard exudates. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 246–251, Kauai, Hi, USA, December 2001.

[9] Thomas A. Lasko, Jui G. Bhagwat, Kelly H. Zou, and Lucilla Ohno-Machado. The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*, 38:404–415, 2005.

[10] Huiqi Li and Opas Chutape. Automated feature extraction in color retinmal images by a model based approach. *IEEE Transactions on Biomedical Engineering*, 51(2):246–254, February 2004.

[11] Harihar Narasimha-Iyer, Ali Can, Bandrinath Roysam, Charles V. Stewart, Howard L.Tanenbau, Anna Majerovics, and Hanumant Singh. Robust detection and classfication of longitudinal changes in color retinalfundus images for monitoring diabetic retinopathy. *IEEE Transactions on Biomedical Engineering*, 53(6):1084–1098, June 2006.

[12] M. Niemeijer, M. D Abràmoff, and B. van Ginneken. Automatic detection of the presence of bright lesions in color fundus photgraphs. In *Proceedings of IFMBE the 3rd European Medical and Biological Engineering Confrence*, volume 11 of *1*, pages 1823–2839, Prague and Czech Republic, November 2005.

[13] M. Niemeijer, B. van Ginneken, J. Staal, M. S. A. Suttorp-Schulten, and N. D. Abramoff. Automatic detection of red lesion in digital color fundus photographs. *IEEE Transactions on Medical Imaging*, 24(5):584–592, May 2005.

[14] M. Niemi and K. Winell. Diabetes in Finland, prevalence and variation in quality of care. Kirjapaino Hermes Oy, Tampre, Finland, 2006.

[15] Alireza Osareh, Majid Mirmehdi, Barry Thomas, and Richard Markham. Classification and localization of diabetic-related eye disease. In *Proc. of 7th European Conference on Computer Vision (ECCV)*, pages 502–516, 2002.

[16] P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(10), 2000.

[17] Richard J. Radke, Srinivas Andra, Omar Al-Kofahi, and Badrinath Roysam. Image change detection algorithms: a systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307, March 2005.

[18] C. I. Sànchez, R. Hornero, M. I .Lòpez, and J. Poza. Retinal image analysis to detect and quantify lesions associated with diabetic retinopathy. In *Proceedings of the 26th Annual International Confrence of the IEEE Engineering in Medicine and Biology Society (EMBS)*, pages 1624–1627, San Francisco, CA, USA, September 2004.

[19] K. Schwerdt and J.L. Crowley. Robust face tracking using color. In *Proceedings of 4th IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[20] C. Sinthayothin, J. F. Boyce, T. H. Williamson, E. Mensah, S.Lal, and D. Usher. Automated detection of diabetic retinopathy on digital fundus images. *Diabetic Medicine*, 19:105–112, 2002.

[21] Gunvor von Wendt. *Screening for diabetic retinopathy: Aspects of photographic methods*. PhD thesis, Karolinska Institutet, 2005.

[22] T. Walter, J.-C. Klein, P. Massin, and A. Erginay. A contribution of image processing to the diagnosis of diabetic retinopathy - detection of exudates in color fundus images of the human retina. *IEEE Transactions on Medical Imaging*, 21:1236–1243, October 2002.

[23] Huan Wang, Wynne Hsu, Kheng Guan Goh, and Mong Li Lee. An effective approach to detect lesions in color retinal images. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 181–186, 2000.

[24] C. P. Wilkinson, Frederick L. Ferris, Ronald E. Klein, Paul P. Lee, Carl David Agardh, Matthew Davis, Diana Dills, Anselm Kampik, R. Pararajasegaram, and Juan T. Verdaguer. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 10(9):1677–1682, September 2003.

[25] Xiaohoui Zhang and Opas Chutape. A SVM approach for detection of hemorrhages in background diabetic retinopathy. In *Proceedings of International Joint Confrence on Neural Networks*, pages 2435–2440, Montreal and Canada, July 2005.

[26] Xiaohoui Zhang and Opas Chutape. Top-down and bottom-up strategies in lesion deection of background diabetic retinpathy. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 422–428, San diego, CA, USA, July 2005.

[27] Xiaohui Zhang and O. Chutatape. Detection and classification of bright lesions in colour fundus images. In *Proceedings of IEEE Internationl Confrence on Image Processing (ICIP)*, volume 1, pages 139–142, October 2004.