the DIARETDB1 diabetic retinopathy database and evaluation protocol

T. Kauppi[1] V. Kalesnykiene[2] J.-K. Kamarainen[1,3] L. Lensu[1] I. Sorri[2]
A. Raninen[2] R. Voutilainen[2] J. Pietilä[5]
H. Kälviäinen[1] H. Uusitalo[2]
[1]Machine Vision and Pattern Recognition Research Group,
Lappeenranta University of Technology
[2]Department of Ophthalmology, Faculty of Medicine,
University of Kuopio
[3]Centre for Vision, Speech and Signal Processing,
University of Surrey
[5]Perimetria Ltd., Helsinki, Finland

Abstract

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 [15]. 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 [9, 6, 18, 11, 13]. 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 [12, 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 toolkit. This study provides the means for the reliable evaluation of automatic methods for detecting diabetic retinopathy.

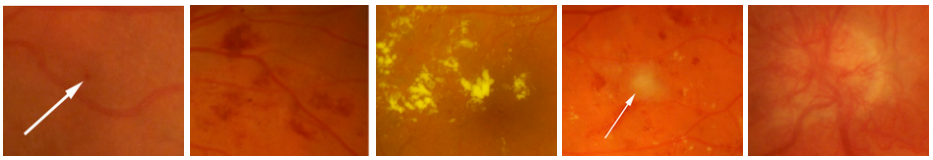
2 Diabetic retinopathy

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 [15]. The diabetes can cause abnormalities in the retina (diabetic retinopathy), kidneys (diabetic nefropathy), and nervous system (diabetic neuropathy) [10]. The diabetes is also a major risk factor in cardiovascular diseases [10].

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 their number and severity predominantly increase with time. The diabetic retinopathy typically begins as small changes in the retinal capillaries. The first detectable abnormalities are mircroaneurysms (Ma) shown in Fig. 1(a), which are local distensions of the retinal capillary and when ruptured, cause intraretinal hemorrhage (H) shown in Fig. 1(b). The disease severity is classified as mild non-proliferative diabetic retinopathy when the first apparent microaneurysms appear in the retina [17]. With time, the retinal edema and hard exudates (He) shown in Fig. 1(c) appear because of 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 [17]. However, if the above-mentioned abnormalities appear in the central vision area (macula), the condition is called diabetic maculopathy [15]. As the retinopathy advances, the blood vessels become obstructed which causes microinfarcts in the retina. These microinfarcts are called soft exudates (Se) shown in 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 non-proliferative diabetic retinopathy [17].

The severe non-proliferative diabetic retinopathy can quickly turn into proliferative di-

abetic retinopathy when extensive lack of oxygen causes the development of new fragile blood vessels [17]. This is called neovascularisation shown in Fig. 1(e), which is a serious eye sight threatening state. The proliferative diabetic retinopathy may cause sudden loss in visual acuity, or even permanent blindness due to vitreous hemorrhage or tractional detachment of the central retina. After the diabetic retinopathy has been diagnosed, regular monitoring is needed due to the progressive nature of the disease. However, broad pre-emptive screenings cannot be performed due to the fact that the fundus image examination requires attention of medical experts. For the screening, reliable automatic image processing methods must be developed.



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

Figure 1: Abnormal findings in the eye fundus caused by the diabetic retinopathy (best viewed in colour): (a) microaneuryms (marked with an arrow); (b) hemorrhages; (c) hard exudates; (d) soft exudate (marked with an arrow); (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 are computed as [16]:

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

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.

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 [1] with varying imaging settings and without preprocessing. The intensity of the camera flash was controlled by the photographer, but other settings such as shutter speed, aperture, and gain were controlled by the fundus camera. 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. In the absence of a well-founded procedure for camera characterisation, the data correspond to a good (not necessarily typical nor the most general) 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 a single hospital. This data set is referred to as "calibration level 1 fundus images". A data set taken with several fundus cameras containing different amounts of imaging noise and optical aberrations is referred to as "calibration level 0 fundus images". To publish an ultimate tool for the evaluation of diabetic retinopathy, which is the research group's main objective, it is necessary to study the different calibration levels. Therefore, the current database was not aimed to be statistically representative, but as one step in the development process.

3.2 Ground truth

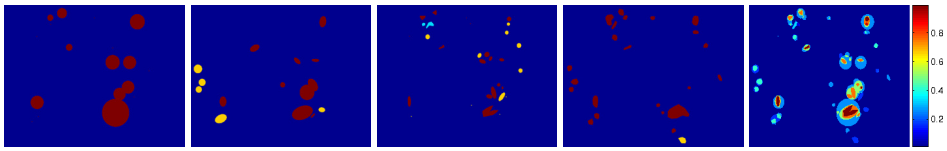
Independent markings from 4 medical experts were collected by using a software tool provided for image annotation. The computer displays used in the collection process were not calibrated. A person with medical education and solid experience in ophthalmology was considered as an expert. The experts were 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

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

(a) (b) (c)

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

especially annotate the single most representative point for each finding. The ground truth confidence levels defined by the medical experts, $\{< 50\%,\ > 50\%,\ 100\%\}$, represented the certainty of the decision that a marked finding is correct. Medical experts had varying certainty for less distinguishable findings and therefore inexact confidence intervals were approved for such findings. 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. The uninstructed collection process caused significant differences between the medical experts as can be seen in Fig. 3 for image in Fig. 2(b). 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



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

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

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. 3(e)). 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 the DIARETDB1, the confidence level is fixed to $conf_{GT} = 0.75$.

3.3 Training and test set

The 89 images were manually assigned into categories representing the progressive states of retinopathy: normal (27 images), mild (7 images), moderate and severe non-proliferative

(28 images), and profilerative (27 images). 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. Note that a single image may contain several finding types.

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.it.lut.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. The image acquisition process affects the evaluation protocol only through the ground truth, i.e., the marking accuracy of experts depends on the quality of image acquisition.

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 [8]. 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 [12, 2], where the two measures, sensitivity (SN) and specificity (SP), are combined to a weighted error rate defined as

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

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 we adopted the following measures [2]: $\mathbf{WER}(\mathbf{10^{-1}})$ (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 toolkit 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 [14] have derived a formula for the Bayesian decision rule to classify a pixel with

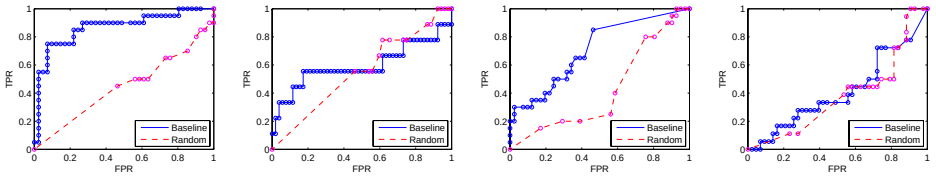
color (r,g) to one of the finding classes. The formula reduces to the histogram ratio of finding and test image:

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

An optimal posterior threshold for every finding type was defined by manual cross-selection and finally the sum of pixels having higher or equal value to the posterior threshold was used as the image based score.

5.2 Results

Evaluation results, ROC and WER, for the baseline method are shown in Fig. 4 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 4: ROC curves for the DIARETDB1 baseline method: a) hard exudates; b) soft exudates; c) microaneurysms; d) hemorrhages.

Table 1: The DIARETDB1 baseline performance measures.

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

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 (toolkit) are publicly available at the DIARETDB1 web site (IMAGERET[2], `http://www.it.lut.fi/project/imageret/`). DIARETDB1 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

[2]Supported from the FinnWell technology program (40430/05, 40039/07) of the Finnish Funding Agency for Technology and Innovation, Tekes.

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] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.