Information Fusion for Diabetic Retinopathy CAD in Digital Color Fundus Photographs

Meindert Niemeijer*, Michael D. Abràmoff, *Member, IEEE*, and Bram van Ginneken, *Member, IEEE*

***Abstract*—The purpose of computer-aided detection or diagnosis (CAD) technology has so far been to serve as a *second reader*. If, however, all relevant lesions in an image can be detected by CAD algorithms, use of CAD for automatic reading or prescreening may become feasible. This work addresses the question how to fuse information from multiple CAD algorithms, operating on multiple images that comprise an exam, to determine a likelihood that the exam is normal and would not require further inspection by human operators. We focus on retinal image screening for diabetic retinopathy, a common complication of diabetes. Current CAD systems are not designed to automatically evaluate complete exams consisting of multiple images for which several detection algorithm output sets are available. Information fusion will potentially play a crucial role in enabling the application of CAD technology to the automatic screening problem. Several different fusion methods are proposed and their effect on the performance of a complete comprehensive automatic diabetic retinopathy screening system is evaluated. Experiments show that the choice of fusion method can have a large impact on system performance. The complete system was evaluated on a set of 15 000 exams (60 000 images). The best performing fusion method obtained an area under the receiver operator characteristic curve of 0.881. This indicates that automated prescreening could be applied in diabetic retinopathy screening programs.**

***Index Terms*—Computer aided detection, computer aided diagnosis, diabetic retinopathy, fundus, information fusion, photographs, retina, screening.**

I. Introduction

MOST, if not all, CAD systems in the clinical arena today focus on the detection of a single type of lesion. Examples are a mass or a cluster of microcalcifications in a mammogram, a pulmonary nodule in a radiograph or computed tomography (CT) scans, or a polyp in an abdominal CT exam. In these cases, the D in CAD stands for detection. In other applications, not yet widely used in the clinic, the D is for diagnosis. Most often, however, the diagnosis is made using only a single lesion or region of interest in the image as input. An example is the estimation of the probability that a lung nodule is cancerous from morphological features. In all these scenarios, it is clear that the computer analysis cannot replace the clinician, because a clinician's diagnosis should be the integration of the presence of a large number of possible, possibly different, abnormalities. In [1], CAD is thus carefully defined as follows: "a diagnosis made by a radiologist who uses the output from a computerized analysis of medical images as a *second opinion* in detecting lesions, assessing extent of disease, and making diagnostic decisions. (...) With CAD, the final diagnosis is made by the radiologist."

In the last decade the research field of CAD has evolved rapidly and many different CAD systems have been developed in academia and industry. For some modalities, a range of abnormality detectors is now available. Theoretically, the paradigm of using CAD as a second opinion only, can now be abandoned. A computer can analyze a complete exam, integrating the output of various CAD algorithms, and establish a diagnosis. If a human expert never sees the exams the result is fully automated screening. If however, in this process, those exams that are flagged as probably abnormal are subsequently reviewed by a human expert, the process is referred to in the literature as *prescreening*, notably in the context of pap smear analysis. Prescreening is also currently performed by humans, either by experts that use only a limited amount of time [2] or by nonexperts, such as radiographers where this practice is called red dotting, named after the color and shape of stickers that are placed on suspect films [3].

In the clinic, where expert clinicians are present, automated screening or prescreening is not expected to have a major translational impact in the near future. In massive, large scale, screening applications however, where millions of exams are generated and have to be read, such systems can have a substantial effect on the practice of medicine, and the cost-effectiveness of healthcare. However, it also introduces a number of important issues that have not yet received a lot of attention in the literature. One way of thinking about this is to consider the scale on which decisions are made as a continuous variable. CAD systems, regardless of their application area, operate on a range of scales. This scale is limited at one end by the finite resolution of the imaging device (i.e., the pixel or voxel level), and at the other end, the amount of imaging that can be applied over a finite time (i.e., repeat imaging of the same subject). At

Manuscript received July 31, 2008; revised November 24, 2008. First published January 13, 2009; current version published April 29, 2009. This work was supported in part by the National Eye Institute under Grant R01 EY017066, in part by the Dutch Ministry of Economic Affairs under Grant IOP IBVA02016, in part by The Netherlands Organization for Scientific Research (NWO), in part by Research to Prevent Blindness, NY, and in part by the Wellmark Foundation. *Asterisk indicates corresponding author.*

*M. Niemeijer was with the Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242 USA, and also with the Department of Opthalmology and Visual Sciences, University of Iowa Hospitals and Clinics, Iowa City, IA 52242 USA. He is now with the Image Sciences Institute, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands (e-mail: meindert@isi.uu.nl).

M. D. Abràmoff is with the Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242 USA, and the Department of Opthalmology and Visual Sciences, University of Iowa Hospitals and Clinics, Iowa City, IA 52242 USA.

B. van Ginneken is with the Image Sciences Institute, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2008.2012029

the lowest level, CAD systems classify pixels (voxels for 3-D) and then continue with groups of pixels (lesions), areas (organs) in images, complete images, multiple images that make up an exam, and finally subjects. At each level the probability of abnormality is determined based on the findings at previous levels. At the highest scale the system is diagnosing a single patient based on the fused information from all the lower scales. Clearly the question how to effectively fuse all this information is nontrivial, and this is the focus of this study.

The fusion process should not only combine the output of multiple abnormality detectors, but also determine if it considers itself *qualified* to make a diagnosis. Even for human experts it is often impossible to make a reliable diagnosis because the quality of the data is insufficient. In an automated or prescreening process, it seems preferable to refer all exams that cannot be reliably analyzed—typically because image quality is deemed insufficient—to human operators for further inspection. Thus, the task of the system is two-fold: detect possibly abnormal exams and exams of insufficient quality. This poses unique challenges for the design of the information fusion process.

In this work the application will be the detection of diabetic retinopathy from color photographs of the retina. Diabetic retinopathy is a common complication of diabetes and the most frequent cause of blindness and vision loss in the working population of the western world [4]. Early detection and treatment through screening helps prevent blindness and vision loss [5]–[7]. Recently we and others have developed and evaluated comprehensive diabetic retinopathy screening systems [8], [9] and the use of CAD is considered a realistic option in this area [10]. For our system, information is gathered from four images, two for each eye, consisting of the position of multiple types of possible lesions, each associated with a degree of suspicion, and a quality indicator for each of the individual images.

The major goal of this work is to propose, describe, and evaluate different methods to fuse this information into a single number that indicates the likelihood that this exam should be inspected by a human operator. The presentation of a complete, general CAD system approach applied to the automatic large-scale screening for diabetic retinopathy is a second contribution. In contrast with much previous work, the proposed system functions on all scales—from pixels to exams. The third contribution is the presentation of an extensive evaluation of the complete system on a set of 15 000 unselected eye exams. To the best of our knowledge, this set of 60 000 images (four images per exam) represents the largest test set used in the retina literature to date. Using such a large and unselected set of exams offers a unique insight into the performance of the presented automated diabetic retinopathy screening system.

The structure of this paper is as follows. The data used in this study is described in Section II. In Section III the architecture of the entire system is presented. In Section IV different methods for combining the data produced by the CAD algorithms are proposed. Experiments and results are given in Section V. This is followed by discussion and conclusions in Section VI. In the Appendix, brief descriptions of each of the components of the automatic screening system are provided.

II. Data

A total of 15 000 exams of patients with diabetes without previously known diabetic retinopathy were included. The patients attended the EyeCheck project, a diabetic retinopathy screening program [11], over a two year period (2006–2008). To obtain a representative screening set, no selection of patients was performed. The only requirement was that four fundus photographs, two of each eye, were available. As a reference standard the judgement of the screening program ophthalmologists (trained retinal specialists, experienced in telediagnosis) was used. Every exam was graded by one ophthalmologist. A total of three ophthalmologists participate in the project. Each exam was assigned to one of three classes.

- *Not suspect*: No signs of referable diabetic retinopathy.
- *Suspect*: Signs of referable diabetic retinopathy present according to the protocol described in [11].
- *Ungradable*: The image quality was insufficient for grading as subjectively determined by the screening program ophthalmologists.

If an exam was considered ungradable another set of images of the same patient was acquired. When an exam was deemed suspect the patient was referred to an ophthalmologist for further diagnosis and treatment, this patient would then no longer attend the screening program. In total, 446 (3.0%) of the exams in our set were marked as "ungradable," 394 (2.6%) exams were marked as "suspect" and the remaining exams were marked as "not suspect."

For each exam four images were acquired, two of each eye, one optic disc centered and one macula centered. The total number of images thus was 60 000. All images and exams were anonymized and the study was performed according to the guidelines set forth in the Declaration of Helsinki. The image data were acquired at ten different sites. The image resolution varied from 768×576 to 2896×1944 pixels while the field of view coverage varied between 35° and 45°. Three different camera types were employed; the Topcon NW 100, the Topcon NW 200 (Topcon, Tokyo, Japan) and the Canon CR5-45NM (Canon, Tokyo, Japan). All images were JPEG compressed. All images had a circular or semi-circular field of view that ranged in diameter from around 747 pixels to around 2100 pixels.

III. Methods

Fig. 1(a) shows a flow diagram outlining the proposed approach to automated diabetic retinopathy screening in retinal images. At the top an exam, containing four images, is fed into the system and at the bottom a degree of suspicion is being assigned to the exam. This number can reflect different things, a high degree of suspicion can mean "a high likelihood of abnormality" or "a high likelihood of needing examination" depending on the way the information generated by the CAD system is fused. In an autonomous screening process, low quality (i.e., ungradable) exams need to be examined, therefore these exams should also be assigned a high degree of suspicion. Fig. 1(b) shows the processing pipeline for a single image throughout the CAD system. The first step after the image enters the system is preprocessing to reduce the differences in resolution between the images acquired at different screening

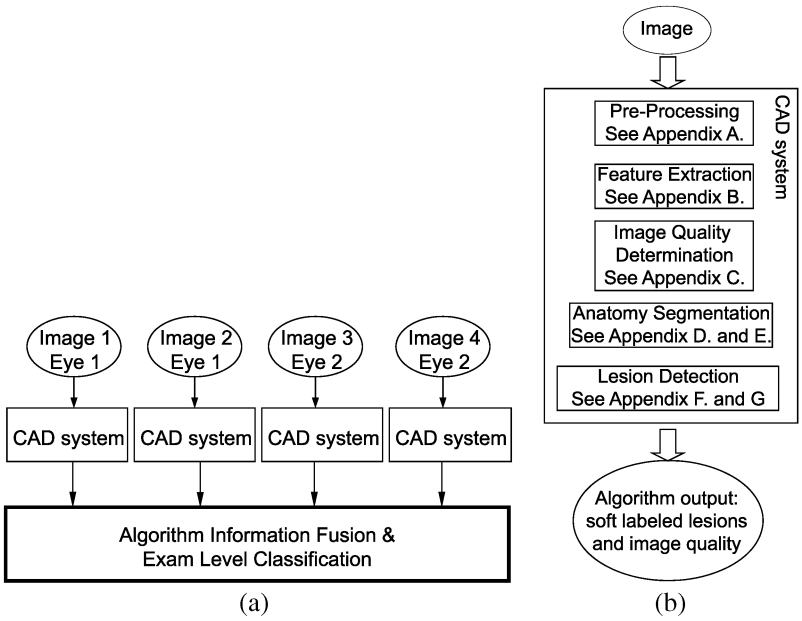

Fig. 1. (a) Diagram of the proposed automated screening system approach. In our application an exam consists of two images per eye, four images in total. After each image in the exam has been processed by a set of CAD algorithms, the algorithm results need to be combined into an outcome for the complete exam. This outcome is a single number indicating the likelihood that the exam needs to be examined by a human operator. (b) Diagram of the CAD system components as they are applied to each image. The image is first preprocessed and pixel features are extracted. Next, the normal anatomy, the image quality and finally two different types of abnormalities are detected. The output of the CAD system for a single image consists of two sets of soft-labeled candidate lesions and an image quality measure.

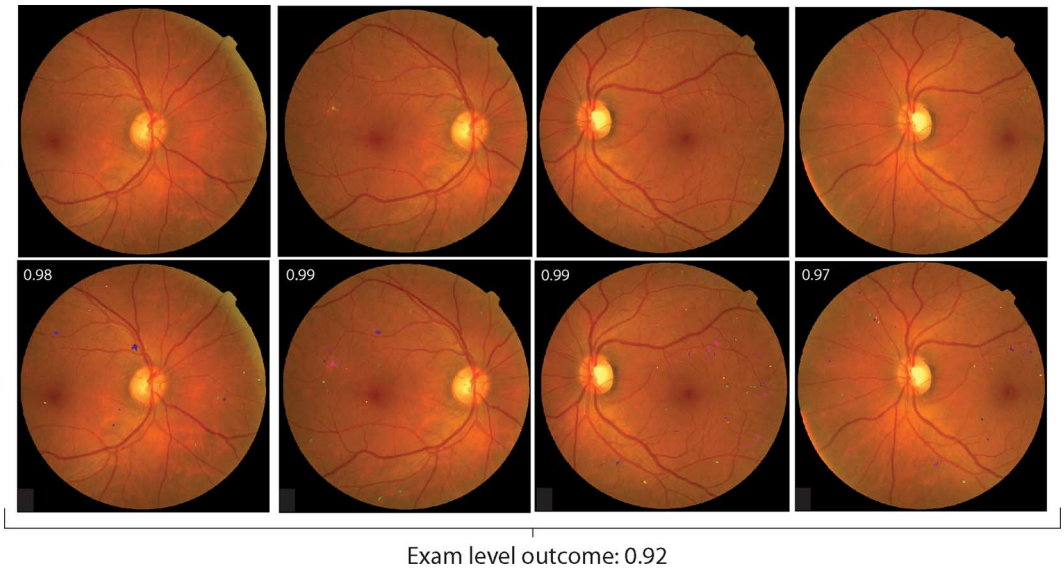

Fig. 2. The top four images comprise an example exam. Each of the four images is analyzed separately using the CAD system algorithms as outlined in Fig. 1(b). The results of these algorithms are visualized in the annotated exams in the bottom four images. All detections are color coded depending on the type of detection and the posterior probability assigned to the candidate to be a lesion. Red lesions are color coded from green (low posterior probability) to yellow (high posterior probability) and bright lesions are color coded in a similar fashion from blue to magenta. The image quality is shown as a number in the top left corner of each image, where a number closer to 1 denotes a higher quality image. These results (lesions with an associated PP, and a quality measure) are fused into one likelihood that the image is either abnormal or the quality is too low to warrant automatic analysis. For this exam, with obvious abnormalities, this likelihood, which ranges from 0 to 1, is 0.92.

locations. Next, the image is processed such that pixel level image structure descriptors (i.e., pixel features) that are useful in the detection of both the normal and abnormal structures in the image are extracted. The extracted features are fed into the different processing pipelines of the CAD screening system detecting image quality, abnormalities and normal anatomy. For our application to diabetic retinopathy detection we distinguish two types of abnormalities, red lesions (i.e., microaneurysms and hemorrhages) and bright lesions (i.e., exudates, cottonwool spots, and drusen). Even though drusen is not a sign of diabetic retinopathy the employed bright lesion detection component does not distinguish between it and the two diabetic retinopathy related bright lesion types. In Fig. 2, a visualization of the CAD system output is shown on an example exam.

Most components in the proposed automated screening system have been presented in previous work [12]–[16]. Short descriptions of each of the components can be found in Appendix I.

IV. Data Fusion

Fusion of the findings of the different algorithms in the different images at the exam level is a final and critical step of the screening system. The quality verification outputs a single number per image. The pathology detection components produce differing amounts of candidates with associated probabilities for being a lesion. The number of probabilities available depends on the number of detected red lesion and bright lesion objects in an image. In Fig. 1 the end result of the detection process is shown on four example images from the same exam. Each of the images has a color-coded quality indicator ([0, 1]) as well as a set of color-coded red and bright lesion objects with associated posterior probability ([0, 1]).

It is important to distinguish at this point between clinical ophthalmology and population wide screening. In a population based screening program the goal is to detect abnormalities early, and not to make a definitive diagnosis. Patients with (suspected) abnormalities should be referred to an ophthalmologist for further diagnosis and management. For automated screening, the system should thus assign a label, either soft or hard, to each exam in a way similar to what a screening ophthalmologist may do.

The manner in which the available data are fused, potentially has a large influence on the final labeling. How this fusion should be accomplished is a nontrivial and important question. The fusion is complicated by the fact that the input used to obtain the final labeling is not perfect as every individual detection component can and will occasionally produce erroneous results.

A. Features

In the data set used for this study, two overlapping images are obtained per eye. Consequently, lesions can be detected twice in a single eye and we are restricted to using a simple, inclusive fusion strategy for all findings in the complete exam. Below, whenever a set of lesions is mentioned, the combined set of lesions as detected in all four images in an exam together is implied. The fusion of the four image quality results with the two sets of candidate objects is complicated by the fact that their associated posterior probabilities are scaled differently and as such are not directly comparable. Although the range of posterior probability values is the same for each component, in practice the distribution of the values that are produced depends on the type of classifier used, the classifier parameters and the class prevalences in the training set. We define the following features to distinguish suspect from normal exams:

1) Highest red/bright posterior probability in the exam. The presence of a single lesion of high certainty can be an important factor in deciding if an exam is suspect.
2) The sum of all red/bright posterior probabilities as a measure for the total lesion load in the exam.
3) The $\sum_{l=1}^{N} a_l * p_l$ where N is the total number of red/bright lesions, a_l is the area in pixels of lesion l and p_l is the posterior probability associated with lesion l. This is a measure of the total lesion load weighted by the size of the detected lesions.
4) The average posterior probability $\bar{p} = (1/N)\sum_{l=1}^{N} p_l$ where only those lesions are included for which $p_l > 0$. This is an alternative method for measuring the lesion load.
5) The standard deviation of the posterior probability $\sigma_p = \sqrt{(1/N)\sum_{l=1}^{N} p_l - \bar{p}}$ where only those lesions are included for which $p_l > 0$.
6) A four bin histogram of the posterior probabilities of the red/bright lesion candidates. This histogram captures both the likelihood of the presence of any lesions as well as the distribution of them among the detected lesion objects.
7) A four bin histogram of the lesion area in pixels subdivided by posterior probability. This histogram captures the amounts of detected lesion pixels within a certain range of posterior probabilities.

The quality detection component already produces a single number per image that can directly be used as a feature.

- The four values of the image quality detection component. As ungradable exams should be detected these outputs must be included in the fusion scheme.

The fact that the abnormality detection components may be influenced in some systematic way by the image quality can possibly lead to more robust exam quality determination than by just using the four image quality outputs.

B. Fusion Techniques

1) Fusion Based on Single Features: A number of different techniques for using (a subset of) these features to assign a label to an exam can be identified. The most straightforward technique would be to select one of the features above and use the value of that feature to directly assign a hard label to the exam by thresholding. Especially features 1, 2, and 3 seem to be appropriate. It is not possible to directly incorporate information about the exam image quality in such a system. As the determination of exam image quality is an essential part of a practical screening system, we treat the exam image quality as pre-determined. Those exams that are of low quality will be assigned the feature value of 1. Note that the feature values are normalized to lie between 0 and 1 and thus assigning the value 1 flags these exams as always positive. To determine the exam image quality the four image quality feature values are combined as follows. The maximum quality output for each eye is taken, reducing the total number of outputs to two. Let $Q_{l1}, Q_{l2}, Q_{r1}, Q_{r2}$ be the first and second normalized quality output of the left and right eye, respectively. Then the combined quality output Q is given by

$$Q = \min\left(\max(Q_{l1}, Q_{l2}), \max(Q_{r1}, Q_{r2})\right). \qquad (1)$$

This combination strategy ensures there should be at least one good quality image of each eye. We determined the threshold on Q that eliminated all but 15% of true low quality exams and used this to determine the low quality exams beforehand. These simple, single feature, techniques will serve as benchmark systems to which the other fusion techniques can be compared.

2) Likelihood Distribution Normalization (PPDN)-Based Fusion: A major complication for fusion is the difference in

the distribution of the posterior probabilities as generated by the different detection components. The following method attempts to decrease this difference and thereby enable the use of standard pattern recognition posterior probability fusion techniques. First, cumulative histograms of the outputs of the quality detection component and the red and bright lesion detection components for a set of training exams are constructed. These three cumulative histograms now form lookup-tables that map the unnormalized output values to normalized values in the range [0, 1]. For example, if a red lesion has a posterior probability of 0.5 and the cumulative histogram shows that 67% of all objects detected by the red lesion detection component have a lower value than 0.5, the normalized value will be 0.67. Once all values are normalized they still need to be fused into an exam level output. The abnormality detection per image outputs are fused by putting them in one set and choosing the maximum. There are many existing techniques to combine posterior probabilities [17], [18] but the maximum rule is a logical, though not necessarily optimal, choice in this application. The four image quality features are combined according to (1). Finally, the soft label L_{soft} for the exam is generated by $L_{\text{soft}} = \max(1 - Q, A)$ where A represents the maximum posterior probability based on the abnormality detection.

3) Multithreshold Fusion: A technique that has been applied previously [19] enables the combination of the red lesion, bright lesion and image quality results by choosing individual thresholds. Three thresholds are selected, one for the image quality value, one for the red lesion posterior probability and one for the bright lesion posterior probability. These three thresholds determine at what point an image is considered ungradable and when an object detected by the red/bright lesion detection is considered a true lesion. Depending on the number of lesion detections and number of ungradable images in an exam one accepts as "normal" a sensitivity and specificity can be determined. By varying the number of acceptable lesions/ungradable images the sensitivity and specificity of the system can be varied. The label generated using this method is hard, either the exam is suspect or normal. Methods that use this type of technique usually report the sensitivity and specificity at a single setting of the system [19]. A disadvantage is that optimization of the three parameters (six if the posterior probability thresholds are included) of this method can be a time consuming task.

4) Supervised Fusion: The final fusion strategy we have investigated is a supervised technique. Here, a set of features based on the output from the different components are computed for each exam in the set. In this work we use the set of features as described in Section IV-A. These features form a feature vector which was labeled according to the reference standard. If the exam is suspect or ungradable the label 1 is assigned, otherwise the label 0 is assigned. Then, each of the exams is assigned a combined posterior probability by the classifier using leave one out classification. In this classification scheme the classifier is trained on the complete dataset except one sample, this one sample is then classified. This process is repeated for all samples. In preliminary experiments several classifiers were tested such as a linear discriminant classifier [18], a quadratic discriminant classifier [18], a support vector machine [20], and a *k*-nearest neighbor (kNN) classifier [21]. It was determined, on an independent training set not used for any other purpose in this research, that a kNN classifier with $k = 281$ provided the best results. A kNN classifier assigns a soft label to an exam by examining the labels of the nearest neighbors of the exam in the feature space based on the Euclidian distance. The soft label assigned to the exam then is $L_{\text{soft}} = n/k$ where n is the number of exams among the k nearest neighbors that have the label 1.

5) Combination of Supervised and PPDN System: In the final fusion scheme we combine the fusion results of the supervised method with the output of the posterior probability distribution normalization based system. This is done by calculating the soft labels as assigned by the PPDN based system and using this label as an addition feature value.

V. Experiments and Results

The complete screening system was applied to all 15 000 exams. The average processing time per image is about 5 min on a 2.4 Ghz Intel Core 2 (single core). This processing time can vary slightly based on the amount of possible lesion objects detected in the image. To decrease the time needed for the experiment (205 days on a single core) a cluster of five quad core machines (2.4 GHz Intel Core 2 Quad) combined with an advanced job distribution system was used to decrease the total runtime to approximately 10 days.

The previously described fusion strategies were applied to the resulting data. A receiver operator characteristic (ROC) analysis was performed by measuring the sensitivity and specificity of the screening system at different thresholds of the exam level probability. In the case of the multithreshold based fusion technique where no exam level probability is generated, sensitivity/specificity pairs were measured on a grid in the 3-D parameter space of the method. The ROC curve was then generated by selecting the set of points S_{roc} from the complete set of points S sorted by descending specificity for which the following relation held: for all points in S with a certain specificity, the one with the highest sensitivity was selected if there was no point in S_{roc} which had a higher sensitivity. The ROC curve was generated by linearly interpolating between the points in S_{roc}.

The resulting ROC curves are shown in Fig. 3. The curves in Fig. 3(a) are those of the single feature based reference techniques. Features 1, 2, and 3 were used for both the red and bright lesions separately resulting in six curves. In Fig. 3(b), the ROC curves of the PPDN based fusion, the multithreshold based fusion, the supervised fusion and the combination of the supervised and PPDN methods are shown. The area under the ROC curve (AUC) for each of the fusion strategies and the sensitivity at specificity 0.6 is shown in Table I. The best performing fusion system combines quality information and lesion information into a single outcome. The plot in Fig. 3(c) illustrates the relative amounts of abnormal and ungradable exams amongst the false negatives for each value of the specificity.

VI. Discussion and Conclusion

A number of different approaches to the exam level fusion of information collected from different images in the same exam by a prescreening system were presented. An overview of the

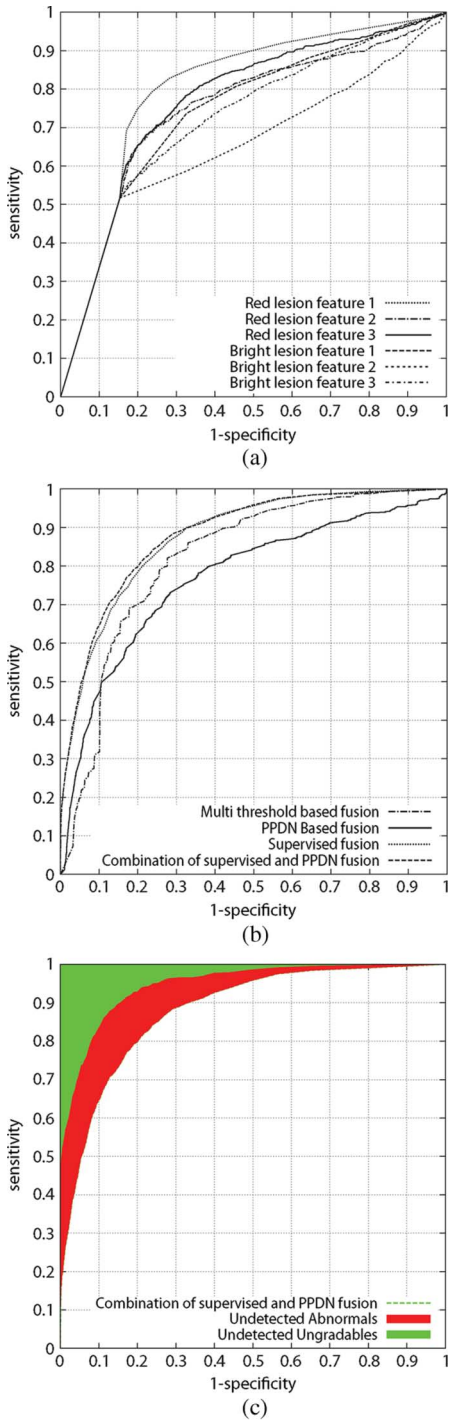

Fig. 3. Performance of various fusion systems in terms of ROC curves per exam evaluated on the test set of 15 000 exams. (a) Single feature reference systems. For each system the initial part of the curve (starting from the origin) is identical, because of the manner in which the image quality has been integrated into these simple fusion systems (see Section V for details). (b) Results of different methods for information fusion. (c) The best performing fusion system combines quality information and lesion information into a single outcome; this plot illustrates the breakdown of abnormal and ungradable exams among the false negatives for each value of the specificity.

comprehensive prescreening system was also provided. The results show that a supervised fusion system that uses a feature set extracted from the abnormality and image quality detection component outputs, augmented with the soft label assigned by the PPDN based fusion system has the highest AUC compared with the other methods. It is able to obtain an AUC of 0.881 on the unselected dataset of 15 000 exams. This is substantially better than results we have previously reported on a smaller test set with less advanced fusion techniques [9].

TABLE I
TABLE SHOWING THE AREAS UNDER THE ROC CURVE OF THE DIFFERENT FUSION STRATEGIES TOGETHER WITH THEIR SENSITIVITIES AT A FIXED SPECIFICITY OF 0.6

Method	AUC	Sensitivity
Red lesion feature 1	0.800	0.873
Red lesion feature 2	0.746	0.783
Red lesion feature 3	0.768	0.829
Bright lesion feature 1	0.741	0.777
Bright lesion feature 2	0.653	0.622
Bright lesion feature 3	0.725	0.736
Multi threshold based fusion	0.821	0.886
PPDN based fusion	0.770	0.804
Supervised fusion	0.877	0.930
Combination PPDN and supervised fusion	**0.881**	0.929

The curves of the single feature reference systems as shown in Fig. 3(a) are the same up to around 0.53 sensitivity. This is due to the manner in which the image quality has been integrated into these simple fusion systems. A certain set of images was assumed to be ungradable due to low image quality and this set was given beforehand, that first part of the curve represents these given images. The most interesting aspect in the variable part of the curves is the performance of the fusion system based on feature 1, the highest posterior probability of the red lesions in the exam. Essentially feature 1 represents a fusion strategy in itself by applying the maximum rule to the output of the red lesion detector. This system reaches a sensitivity of 0.873 at a specificity of 0.6 which is surprisingly close to the best performing system which reaches 0.929 at the same specificity. It shows that the red lesion based features are sufficient to detect the bulk of abnormal cases, an approach proposed in [8]. However, the compound fusion method that also takes other lesion types into account clearly performs better and the difference between the AUC of both approaches is significant $(p < 0.0001)$ [22].

In previous work on retinal screening the multithreshold fusion approach was used [19]. An advantage of using separate thresholds is that it is clear why a particular exam was assigned a certain label, for example due to the presence of a certain type of lesion. A disadvantage of this approach is that no soft labels for the exams are generated. This makes it much more difficult to dynamically adapt the sensitivity/specificity of the system to changing requirements. The proposed supervised fusion system has the advantage of being able to assign soft labels and straightforwardly combine data from sources that measure completely different things, e.g., lesion detection and quality verification. In the screening context it is less important why an exam is assigned a certain label as long as abnormal and ungradable exams can be separated from the normal exams.

The current set of features used in the supervised fusion scheme is by no means exhaustive. In the future more features could be added as they become available. One limitation of the current system is the fact that eye level information fusion is not possible as it is unknown how the two images acquired from a

single eye are positioned. Therefore, it is not possible to measure the exact lesion load per eye. Using a registration approach the information from both images of each eye could be fused, possibly enhancing the image quality at least in the overlapping part of the image. However, given the current feature set this is unlikely to yield significantly different results. Only if the added information derived from registration of the two images of one eye is integrated into the feature set can registration have an influence on the fusion algorithm output. A feature that could have a positive effect on the fusion result is location information. Even though Fig. 1 shows the establishment of a coordinate system on the retina, in the current implementation of the screening system only the position of the optic disc is detected. The fovea location is important as lesions close to the fovea are generally regarded as more serious [6]. So, an exam of a patient with two small exudates close to the fovea may be labeled as containing "referable diabetic retinopathy" while if these lesions would be located further away from the fovea this label would not be assigned. In future work, the output of a fovea detector may be integrated in the CAD system. We have described an algorithm to locate the fovea in [23]. If the position of the fovea is known, features that encode the spatial position of lesion candidates relative to the position of the fovea could be added. The spatial distribution of lesions could also be employed in more sophisticated fusion methods.

In general there are large differences in AUC between most of the different methods. An exception is the difference between the best performing fusion system, supervised + PPDN, and the second best performing system which was not significant ($p < 0.654$). The fact that there is a small difference does indicate that combination of different approaches or classification methods may further improve the presented results. Overall, the substantial differences in AUC between the other fusion methods illustrates the fact that the chosen fusion method has a large influence on the final performance of the automated screening system. As such, a move toward CAD systems for use in screening settings in general will likely benefit from additional research in this area.

There are a number of challenges in the development of methods for the fusion of information produced by different CAD algorithms. The output of the different algorithms cannot usually be directly compared. The algorithms might be measuring completely different things such as the detection of lesions or the quality of images. Some algorithms will produce a single output per image (e.g., image quality detection) others produce varying amounts of outputs, each with a different soft label (e.g., a lesion detector). In some cases the algorithms may produce erroneous results on one image but succeed on the remaining images in the exam. Finally, fusion can take place at multiple levels within the system. For our application this could be at the image level, the eye level, and the exam level. We have chosen to apply fusion only on the exam level by combining all lesion detections in the complete exam in a single set. However, as mentioned before, the system may benefit from fusing the information in multiple stages at different levels.

Fig. 3(c) shows the ROC of the best performing system with the sources of false negatives above the curve. The relative amount of quality related false negatives decreases more quickly than the abnormality related false negatives, with the majority of false negatives at 0.5 specificity being abnormal exams. At this setting, with the current test data, the system produces 25 false negative abnormal exams and 11 false negative ungradable exams. A second reading of these false negatives was performed by one of the retinal experts involved in the screening program (MDA). In 10 of 11 cases the second reader agreed with the reference standard as far as the ungradability of the exams was concerned. Interestingly, the second reader did agree with the quality assessment of the automated system of each of the individual images in 10 of the 11 exams. The main issue here seems to be in the fusion which fails to recognize these *exams* as low quality. As far as the abnormal exams were concerned the second reader agreed with the reference standard in 22 of 25 cases. The main causes of false negatives, 50%, are exams that contain a low number (i.e., up to two) of relatively large hemorrhages connected with the vasculature and no other abnormalities. The second major cause, 32%, are exams with a low number (i.e., up to four) of small, isolated exudates close to the fovea and no other abnormalities present. The remaining four cases are either subtle, i.e., a single microaneurysm near the fovea, contained laser scars or contain another abnormality not associated with diabetic retinopathy. Integration of fovea location information into the system is likely to help with detection of exams in which there is a small number of lesions close to the fovea. Note that, depending on requirements, in practice an operating point with higher specificity than the one chosen here would probably be used.

In addition to integrating a fovea detection component into the system another enhancement that could lead to improved performance is improved detection of large hemorrhages connected with the vasculature. This type of component should lead to the elimination of a large part of the remaining false negatives. Another potential source of errors is the presence of drusen that are not discriminated from the diabetic retinopathy related bright lesions in the current system. We estimate that about 1500 of all 15 000 exams contained some drusen. In our previous work [16] we have presented a method that was able to distinguish between the three different types of lesions successfully on a relatively small test dataset. For this final system we decided to take a more inclusive strategy to not miss any images due to misclassified bright lesions. As the presence of drusen likely causes high probability responses from the bright lesion detector in exams that are marked "not suspect" this may have lead to an increase in false positive exam-level responses.

A limitation of the current study is that the output of the screening system is only compared to a single read by one of the screening program ophthalmologists. Though our results in this study show that the performance is sufficient for deployment, a large scale safety study is still required for regulatory approval, which would need to consist of a comparison to the gold standard. We have previously proposed [9], [24] a prospective multicenter study, on populations with defined race and ethnicity distributions, involving comparison to standard multifield stereo photographs read according to the ETDRS [6] standard.

The major goal of this work is to propose, describe and evaluate different methods to fuse the information produced by a set of CAD algorithms into a single number that indicates the

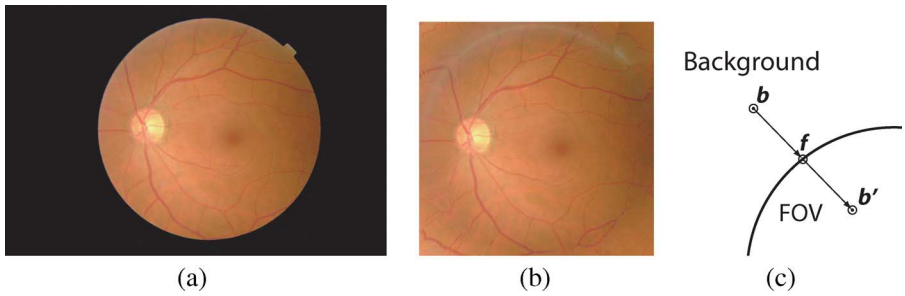

Fig. 4. (a) Example of an image as it is acquired on site. (b) The same image after preprocessing (i.e., resizing, clipping of black borders, and mirroring the inside of the FOV to the outside of the FOV). (c) Schematic illustrating the mirroring operation. A pixel value from within the FOV is mirrored to outside the FOV.

posterior probability that and exam should be inspected by a human operator. The presentation of a complete, general CAD system approach applied to the automatic large-scale screening for diabetic retinopathy is a second contribution. In contrast with much previous work, the proposed system functions on all scales—from pixels to exams. The third contribution is the presentation of an extensive evaluation of the complete system on a set of 15 000 unselected eye exams. To the best of our knowledge, this set of 60 000 images (four images per exam) represents the largest test set used in the retina literature to date. Using such a large and unselected set of exams offers a unique insight into the performance of the presented automated diabetic retinopathy screening system.

To summarize, we have developed and evaluated a compound computer-aided diagnosis system that takes into account abnormalities at multiple scales of multiple types as well as whether a reliable analysis can be produced. We have also shown that both the "supervised" as well as the "supervised plus PPDN" fusion method are superior over other fusion methods for the type of lesions detected by this system. Finally, we have performed the most extensive evaluation to date of a system for automated detection of diabetic retinopathy in retinal color images. The performance of the automatic screening system is such that we are considering applying it to screening practice, provided the remaining procedural, safety, and legal issues are resolved.

Appendix I
Automated Screening System

See Fig. 1 for a schematic overview of the proposed automatic screening system and the configuration of the components discussed below.

A. Preprocessing

The images that are fed to the system are acquired at different sites, by different operators using different cameras and camera settings. The first processing step of the system is aimed at making the difference in field-of-view (FOV) size between exams smaller, removing the sharp border between the FOV and the image background and clipping away unused background pixels. A typical image before and after processing is shown in Fig. 4(a) and (b), respectively. In normal quality images, the FOV may be segmented by thresholding one of the color planes to obtain a binary FOV mask. However, in many cases the use of a fixed threshold will fail due to differences in FOV brightness or the presence of local, underexposed areas in the image. Because there are different shapes of FOV the circularity cannot be used reliably to detect failed segmentations.

The proposed approach employs a set of template FOV segmentations that have been previously segmented manually. For the test set of 60 000 images used in this research, 20 template FOV segmentations were sufficient to describe the size and shape variations of all FOVs present in the test set. To match a template FOV segmentation to a query image, the gradient magnitude of the red color plane of the query image was used as a cost function image (I_{gm}). In a first selection step, only those templates with similar dimensions to the query image were selected as candidate FOV segmentations. A rough grid search for each of the candidates in which the template is translated in both the x and y direction is performed, for each variation of the two position parameters the mean value $\overline{gm}(i)$ and standard deviation $\sigma_{gm}(i)$ of I_{gm} for all border pixels i of the FOV of the template image were determined.

The values of x and y which maximized $\overline{gm}(i) - \sigma_{gm}(i)$ were chosen. This cost function ensures that the mean gradient magnitude under the template border is high combined with a low standard deviation. The standard deviation term was added to handle flash artifacts near the FOV edge. These artifact can result in a locally very high gradient magnitude which tends to disturb the fitting process. After the rough grid search a finer grid search is performed with the final selected FOV template. The location with the lowest cost function value is chosen as the final segmentation result. Following the FOV segmentation, the image is resized so that its FOV has a standardized diameter of 650 pixels.

The large difference in intensity at the edge of the FOV can be a problem when extracting image features near the border of the FOV. Therefore a "mirroring" operation is performed to remove the intensity gradient at the border. This operation is applied to every pixel outside the FOV. A line is projected from a background pixel $\mathbf{b}$ to the closest pixel on the FOV border $\mathbf{f}$ [see Fig. 4(c)]. The line is reflected in $\mathbf{f}$, the pixel value at $\mathbf{b}'$ inside the field of view is then assigned to $\mathbf{b}$. An example of the final result of this mirroring operation is shown in Fig. 4(b).

B. Feature Extraction

The algorithms used in the automated screening system are supervised. Many of them are based on pixel classification. In this technique a statistical classifier is trained using a set of labeled example pixels. For each of the training pixels, a feature vector is extracted from the training image. Using

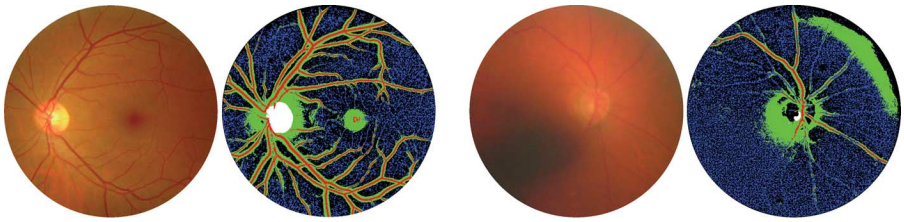

Fig. 5. (a) Example of the structure clusters in a good quality image. Each of the five clusters is represented by a different color. (b) The same but for a low quality image.

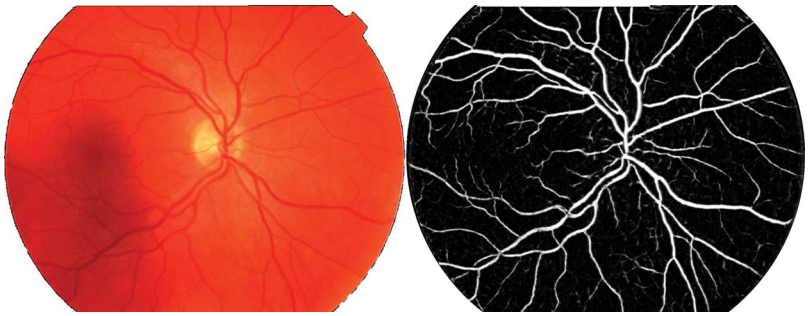

Fig. 6. An example result of the automatic vessel segmentation algorithm applied to a typical, normal quality fundus image. The pixel intensities of the segmentation image represent the posterior probabilities as produced by the classifier that performed the segmentation. The higher the intensity the higher the probability the pixel is inside a vessel.

the trained classifier, pixels in a test image are assigned a label based on the set of feature values extracted from the test image. In the proposed screening system, a similar set of basic pixel features was used by several algorithms and therefore these features were precomputed. The set consisted of Gaussian filterbank outputs up to and including second order derivatives $(G, G_x, G_y, G_{xx}, G_{xy}, G_{yy})$ at five different scales $\sigma = 1, 2, 4, 8, 16$ augmented with the intensity value from the original image. The total number of extracted features was $6 \times 5 + 1 = 31$. Some algorithms use a nonlinear combination of these features (see Section I-C) or use only a few selected features from this complete set.

C. Image Quality Verification

An important issue for automated screening is the danger of false negatives because of insufficient image quality. Image quality verification is therefore included in our approach. Details of the algorithm used here were described in [12].

We rely on the assumption that an image of sufficient quality should contain particular image structures according to a certain predefined distribution. Filterbank output response vectors are clustered to obtain a compact representation of the image structures found within an image. After preliminary experiments, five clusters was determined to provide good performance. Three of the five clusters corresponded to anatomical structures on the retina, two to the vasculature and one to the optic disk. The two remaining clusters represented mainly variations in the background of the image. An example showing the different clusters for a typical image is shown in Fig. 5. The image structures in the image can now be represented using a five bin histogram, one bin for each cluster. Using this compact representation together with histograms of the R, G, and B color planes as features, a statistical classifier is trained to distinguish normal from low quality images. A support vector machine achieved the best classification performance. The final output of the system consists of a single probability per image. This probability indicates the likelihood that the image has a normal quality.

D. Vessel Segmentation

The vasculature is one of the most important anatomical structures in retinal images and is used in the optic disc and both the red and bright lesion detection algorithms. The vessel segmentation method used in this work is based on the pixel classification method as described in [13]. A kNN classifier with $k = 31$ was trained using labeled example pixels to identify vessel pixels in previously unseen images. The training pixels were obtained from the publicly available DRIVE database [25]. All 40 images in the database were used for training. From each image 2500 random training samples in the FOV were collected. Each training sample was labeled as either vessel or non-vessel using the reference standard from the DRIVE database. The features outlined in Section I-B were used. Feature selection using the Sequential Forward Floating Selection algorithm [26] was performed on the training set to reduce the computational complexity of the algorithm. Ten features were selected: $I^{1,2,4,8}$, I^2_y, I^1_x, I^2_{yy}, $I^{2,4}_{xy}$, I^2_{xx}, where I^s_x represents the first derivative in the x direction at scale s.

After the training phase has finished, the trained classifier can be applied to any retina image to produce a vessel probability map that can be thresholded to obtain a binary vessel segmentation. An example vessel segmentation result is shown in Fig. 6.

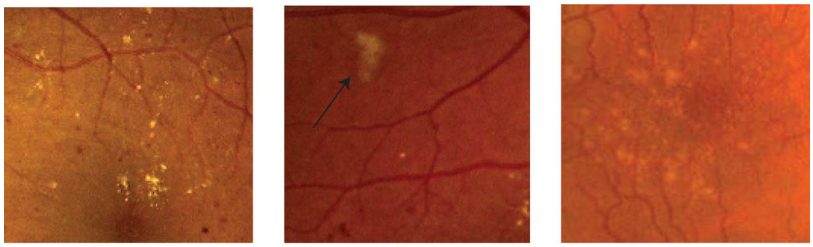

Fig. 7. (a) The dark, red spots are small hemorrhages and microaneurysms while the bright spots are exudates, both signs of the presence of diabetic retinopathy. (b) The arrow indicates a cottonwool spot, an abnormality associated with diabetic retinopathy. (c) Drusen, lesions not associated with diabetic retinopathy but not uncommonly encountered in typical diabetic retinopathy screening populations.

E. Optic Disk Detection

The optic disc is another important anatomical landmark. It usually appears as a bright yellow disc-like object on the retina. The optic disc can interfere with the detection of diabetic retinopathy related bright lesions as it locally appears very similar to a bright lesion. If its position is known, any bright lesion candidates on the optic disc can be masked out.

The optic disk detection algorithm used in this work was described in [14]. The method uses a kNN regressor to determine the location of the centerpoint of the optic disk. This regressor determines the relationship between a set of independent variables, a feature vector $\mathbf{v}$ measured around a circular template, and a dependent variable, the distance d between the center of the template and the true center of the optic disk. Using this approach it is relatively straightforward to combine the use of features based on the image intensities as well as features derived from the orientation and density of the vasculature at a certain location. The orientation of the vasculature around the optic disc for example tends to be similar across subjects, so the local orientation of the vessels at the template boundary provides information about the location of the template on the retina. Since the optic disc is a bright object the average image intensity under the template also provide important information. Combining these different cues has the advantage that the method can give robust results even in cases when the optic disc is not the brightest object in the image.

The regressor needs to be trained once using a set of examples that consist of $\mathbf{v}$ and d pairs. After training the kNN regressor, given a measured $\mathbf{v}$, provides an estimate of d, $\tilde{d}$. By locating the position in the image where $\tilde{d}$ is minimal the optic disk center is located. In the vast majority of cases, there are vessels running over the optic disk center so the search is constrained to only the vessel pixels as found by the method described in Section I-D. After determining $\tilde{d}$ for every vessel pixel the resulting image is blurred using a Gaussian kernel with standard deviation 16 pixels to handle noise and the location with the lowest value is taken as the optic disk center.

F. Red Lesion Detection

Red lesions (i.e., microaneurysms and hemorrhages) are important signs of diabetic retinopathy. Therefore the detection of red lesions is a crucial step for a diabetic retinopathy screening system. See Fig. 7(a) for example red lesions.

An automatic system which can detect the presence and the location of red lesions was described in [15]. The method uses a hybrid approach that combines two detection techniques which are complimentary. A mathematical morphology based detection technique [27]–[29], that works well for smaller red lesions is combined with a pixel classification based approach which also allows the detection of larger red lesions (i.e., hemorrhages). The supervised pixel classification based technique uses a kNN classifier and the filterbank features as described in Appendix I-B to detect both pixels in the vasculature as well as in red lesions in one step. After eliminating elongated structures (the vasculature) from the resulting posterior probability map the remaining objects are red lesion candidates. The two detection systems are both applied to an image and the resulting sets of candidate lesions are merged to arrive at the final set of candidate lesions. From each candidate lesion 69 features were extracted that describe shape, image structure, and color. A kNN classifier, trained using example lesions from an extensive training set, is used to assign each of the candidate lesions a posterior probability that it is a true red lesion and not a spurious object.

G. Bright Lesion Detection

In addition to red lesions some additional types of lesions are associated with diabetic retinopathy. Due to their appearance these lesions are named bright lesions. Of the three types of bright lesions frequently encountered in diabetic retinopathy screening populations, i.e., exudates [Fig. 7(a)], cottonwool spots [Fig. 7(b)], and drusen [Fig. 7(c)], only the first two are associated with diabetic retinopathy. Drusen is however a not uncommonly occuring lesion in the type of populations that attend diabetic retinopathy screening programs.

An automatic system capable of detecting and differentiating between the three types of bright lesions was presented in [16]. That system was used in this work, but without the component that distinguished between different types of lesions. Thus, we detect the likely presence of any of the three types of bright lesions in the image. In a first candidate detection step, pixel classification is used to find groups of pixels that are likely inside a bright lesion. This pixel classification approach used a kNN classifier and the features as described in Appendix I-B. The classifier was trained using a large set of images with available

ground truth. The detected lesion-like objects were then segmented by locally clustering the pixels in the object and those directly around the object into two classes, background and candidate. Any candidates that were closer than 60 pixels to the optic disk center were removed to prevent false positive detections on the optic disk. From the remaining candidates 83 features were extracted describing the shape, contrast, color, and distance to the nearest red lesion. A kNN classifier assigned each candidate a posterior probability.

REFERENCES

[1] M. L. Giger, N. Karssemeijer, and S. G. Armato, "Computer-aided diagnosis in medical imaging," *IEEE Trans. Med. Imag.*, vol. 20, no. 12, pp. 1205–1208, Dec. 2001.

[2] A. Djemli, K. Khetani, and M. Auger, "Rapid prescreening of papanicolaou smears: A practical and efficient quality control strategy," *Cancer*, vol. 108, no. 1, pp. 21–26, 2006.

[3] E. P. Sonnex, A. D. Tasker, and R. A. Coulden, "The role of preliminary interpretation of chest radiographs by radiographers in the management of acute medical problems within a cardiothoracic centre," *Br. J. Radiol.*, vol. 74, no. 879, pp. 230–233, 2001.

[4] D. Klonoff and D. Schwartz, "An economic analysis of interventions for diabetes," *Diabetes Care*, vol. 23, no. 3, pp. 390–404, 2000.

[5] J. Kinyoun, F. Barton, M. Fisher, L. Hubbard, L. Aiello, and F. Ferris, "Detection of diabetic macular edema. Ophthalmoscopy versus photography-Early Treatment Diabetic Retinopathy Study report number 5. The ETDRS research group," *Ophthalmology*, vol. 96, pp. 746–750, 1989.

[6] Early Treatment Diabetic Retinopathy Study Research Group, "Early Photocoagulation for Diabetic Retinopathy: ETDRS report 9," *Ophthalmology*, vol. 98, pp. 766–785, 1991.

[7] G. Bresnick, D. Mukamel, J. Dickinson, and D. Cole, "A screening approach to the surveillance of patients with diabetes for the presence of vision-threatening retinopathy," *Opthalmology*, vol. 107, no. 1, pp. 19–24, 2000.

[8] S. Philip, A. D. Fleming, K. A. Goatman, S. Fonseca, P. McNamee, G. S. Scotland, G. J. Prescott, P. F. Sharp, and J. A. Olson, "The efficacy of automated "disease/no disease" grading for diabetic retinopathy in a systematic screening programme," *Br. J. Ophthalmol.*, vol. 91, pp. 1512–1517, 2007.

[9] M. D. Abràmoff, M. Niemeijer, M. S. A. Suttorp-Schulten, M. A. Viergever, S. R. Russell, and B. van Ginneken, "Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes," *Diabetes Care*, vol. 31, no. 2, pp. 193–198, 2008.

[10] G. S. Scotland, P. McNamee, S. Philip, A. D. Fleming, K. A. Goatman, G. J. Prescott, S. Fonseca, P. F. Sharp, and O. A. John, "Cost-effectiveness of implementing automated grading within the national screening programme for diabetic retinopathy in scotland," *Br. J. Ophthalmol.*, vol. 91, pp. 1518–1523, 2007.

[11] M. D. Abràmoff and M. S. A. Suttorp-Schulten, "Web-based screening for diabetic retinopathy in a primary care population: The eyecheck project," *Telemed. J. E Health*, vol. 11, no. 6, pp. 668–674, 2005.

[12] M. Niemeijer, M. D. Abràmoff, and B. van Ginneken, "Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening," *Med. Image Anal.*, vol. 10, no. 6, pp. 888–898, 2006.

[13] M. Niemeijer, J. Staal, B. van Ginneken, M. Loog, and M. D. Abràmoff, "Comparative study of retinal vessel segmentation methods on a new publicly available database," in *Proc. SPIE: Med. Imag.*, 2004, vol. 5370, pp. 648–656.

[14] M. D. Abràmoff and M. Niemeijer, "The automatic detection of the optic disc location in retinal images using optic disc location regression," in *Eng. Med. Biol. Soc.*, 2006, pp. 4432–4435.

[15] M. Niemeijer, B. van Ginneken, J. Staal, M. Suttorp-Schulten, and M. D. Abràmoff, "Automatic detection of red lesions in digital color fundus photographs," *IEEE Trans. Med. Imag.*, vol. 24, no. 5, pp. 584–592, May 2005.

[16] M. Niemeijer, B. van Ginneken, S. Russel, M. Suttorp-Schulten, and M. D. Abràmoff, "Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis," *Investigative Ophthalmol. Vis. Sci.*, vol. 48, pp. 2260–2267, 2007.

[17] M. Loog, "Supervised dimensionality reduction and contextual pattern recognition in medical image Processing," Ph.D. dissertation, Image Sci. Inst., Utrecht, The Netherlands, Sep. 2004.

[18] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[19] D. Usher, M. Dumsky, M. Himaga, T. Williamson, S. Nussey, and J. Boyce, "Automated detection of diabetic retinopathy in digital retinal images: A tool for diabetic retinopathy screening," *Diabetic Medicine*, vol. 21, pp. 84–90, 2004.

[20] C. Chang and C. Lin, LIBSVM: A Library for Support Vector Machines 2001.

[21] S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu, "An optimal algorithm for approximate nearest neighbor searching in fixed dimensions," *J. ACM*, vol. 45, no. 6, pp. 891–923, 1998.

[22] J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases," *Radiology*, vol. 148, no. 3, pp. 839–843, Sep 1983.

[23] M. Niemeijer, B. van Ginneken, and M. Abràmoff, "Automated localization of the optic disc and the fovea," in *Proc. IEEE Int. Conf. EMBS*, 2008, pp. 3538–3541.

[24] M. Abràmoff, M. Niemeijer, M. Suttorp-Schulten, M. Viergever, S. Russell, and B. van Ginneken, "Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes: Response to Olson et al.," *Diabetes Care*, vol. 31, no. 8, p. e64, 2008.

[25] J. Staal, M. D. Abràmoff, M. Niemeijer, M. Viergever, and B. Van Ginneken, "Ridge based vessel segmentation in color image of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.

[26] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognit. Lett.*, vol. 15, no. 11, pp. 1119–1125, 1994.

[27] T. Spencer, J. Olson, K. McHardy, P. Sharp, and J. Forrester, "An image-processing strategy for the segmentation and quantification in fluorescein angiograms of the ocular fundus," *Comput. Biomed. Res.*, vol. 29, pp. 284–302, 1996.

[28] M. Cree, J. Olson, K. McHardy, P. Sharp, and J. Forrester, "A fully automated comparative microaneurysm digital detection system," *Eye*, vol. 11, pp. 622–628, 1997.

[29] A. Frame, P. Undrill, M. Cree, J. Olson, K. McHardy, P. Sharp, and J. Forrester, "A comparison of computer based classification methods applied to the detection of microaneurysms in ophthalmic fluorescein angiograms," *Comput. Biol. Med.*, vol. 28, pp. 225–238, 1998.