Automatic Detection of Retina Disease: Robustness to Image Quality and Localization of Anatomy Structure

T. P. Karnowski, *Member, IEEE,* D. Aykac, *Member, IEEE,* L. Giancardo, *Member, IEEE,* Y. Li,
Ph.D., *Member, IEEE,* T. Nichols, M.D., Ph.D., *Member, IEEE,* K.W. Tobin, Jr., Ph.D., *Senior Member, IEEE,* E. Chaum, M.D., Ph.D., *Member, IEEE*

Abstract— The automated detection of diabetic retinopathy and other eye diseases in images of the retina has great promise as a low-cost method for broad-based screening. Many systems in the literature which perform automated detection include a quality estimation step and physiological feature detection, including the vascular tree and the optic nerve / macula location. In this work, we study the robustness of an automated disease detection method with respect to the accuracy of the optic nerve location and the quality of the images obtained as judged by a quality estimation algorithm. The detection algorithm features microaneurysm and exudate detection followed by feature extraction on the detected population to describe the overall retina image. Labeled images of retinas ground-truthed to disease states are used to train a supervised learning algorithm to identify the disease state of the retina image and exam set. Under the restrictions of high confidence optic nerve detections and good quality imagery, the system achieves a sensitivity and specificity of 94.8% and 78.7% with area-under-curve of 95.3%. Analysis of the effect of constraining quality and the distinction between mild non-proliferative diabetic retinopathy, normal retina images, and more severe disease states is included.

I. INTRODUCTION

IN the United States of America, more than 25 million Americans have diabetes and this number is projected to increase to over 115 million by the year 2050 [1]. The growth of diabetes is not limited to the United States, as diabetic retinopathy (DR) is the leading cause of blindness in the industrialized world. In recent years, several machine vision algorithms for detecting DR and other eye diseases have been developed (see for example [2-4] among others). These systems and algorithms could lead to inexpensive, broad-based screening for DR. One example is the Telehealth Retinal Image Analysis and Diagnosis (TRIAD), a telemedicine network based in the Mid-South region of the

United States [5]. This system provides retina screening to diabetic patients in a primary care setting or walk-in clinics. State-of-the-art fundus camera available at the clinic are used to image the patient's retinas. These are submitted to a workflow process through a secure, HIPAA-compliant network and stored in a database. Notification is sent to an ophthalmologist who prepares a recommendation for the general practitioner delivered via secure protocols. While the system is currently functioning as a manual telemedicine network, greater levels of automation have been introduced. The initial automation was limited to a real-time quality measurement of the captured images [6]. Later automation included anatomical feature localization [7], and lesion detection [8-9], which will ultimately form the basis for a supervised disease stratification methodology based on content-based image retrieval (CBIR) [10]. We have also explored the impact of quality on disease detections using manually segmented lesions with our CBIR methodology [11], and investigated means of assigning a confidence to the optic nerve location using complementary methods [12-13]. Our belief is that system performance can be improved by setting a threshold T_Q on the quality of the images and optic nerve confidence T_C such that examinations that exceed these thresholds are reliably detected by automatic methods, while examinations below these thresholds are best dealt with by human screeners (see Figure 1 for an illustration of this process). In this work, we use the quality estimation algorithm and optic nerve location algorithms to determine the impact of these thresholds on disease detection using data collected from February 2009 to January 2011 with the TRIAD system. We first briefly discuss the quality metric and optic nerve detection confidence metric. We then show the system performance as we vary T_Q and restrict analysis based on T_C. We conclude with some observations on performance and future paths for development.
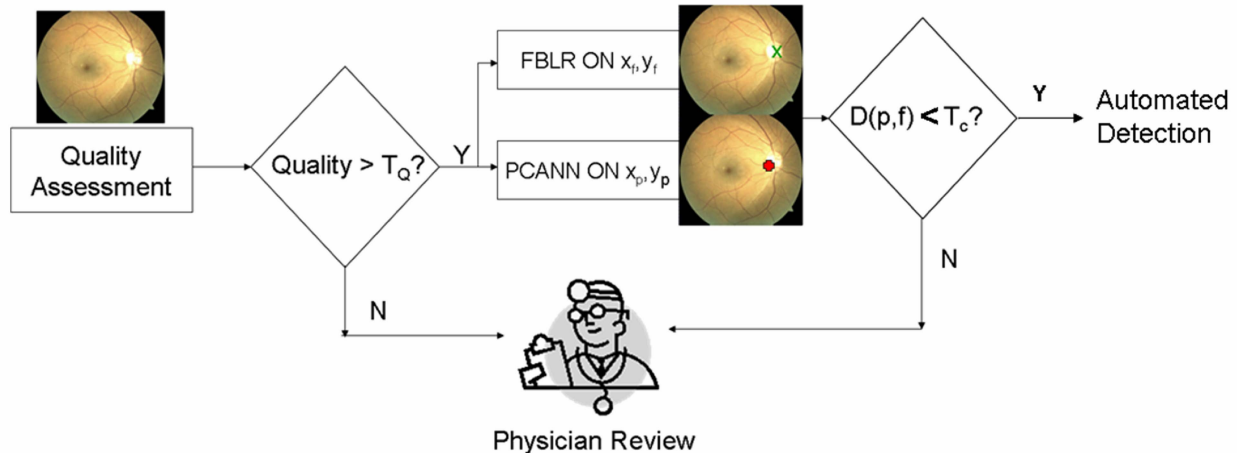
Figure 1. Overview of robustness study for automation in telemedicine network. After quality assessment, images which exceed a threshold T_Q undergo optic nerve locatlization using complementary methods. When the distance between the detected locations **p** and **f** is less than some threshold T_C, automated detection is performed. Any examinations failing these tests are reviewed by the ophthalmologist.

II. APPROACH

A. Quality Assessment

The image acquisition phase is currently performed by a technician in the clinical setting. After image capture, a measurement of the image quality is performed. The TRIAD system uses the elliptical vessel density or ELVD method [6]. This method was designed for computational efficiency and thus is well-suited for this application, giving quick feedback. In the method, the image undergoes a segmentation of the vascular tree based on the method of [14], which uses morphological processing to emphasize vessel-like structures. After the segmentation, which is preserved for additional processing, the image is divided into spatial regions. In each region a measurement of the density of the vessels is made, in addition to a measure of the overall color of the retina image. A supervised support vector machine classifier delivers a quality estimate ranging from 0 to 1 which was trained with example images labeled as "poor", "fair", and "good" quality. Any images which fall below a threshold (set empirically to 0.40) are judged insufficient quality. There is an empirical correlation between the quality level assigned by the currently trained ELVD method and the actual human perceptual quality, but the exact relationship is not well-defined. Thus, we have a fundamental question regarding the quality: what level is sufficient for accurate disease diagnosis? Intuitively, it would seem that higher quality thresholds would produce better results at the cost of more manual screening.

B. Optic Nerve Detection

The detection of anatomic structures, particularly the optic nerve and macula, are fundamental to the subsequent characterization of the normal or disease states that may exist in the retina. These structures establish a sort of coordinate system of the retina. Optic nerve detection is fairly mature field and there are many different methods that have been applied to a wide variety of data sets (see, for example, [7,18-21]). In our research we have developed a method which relies on a good quality segmentation of the optic nerve [7] which we will identify as the "feature based likelihood ratio" or FBLR. In this method, the segmented vascular tree is used and local measurements of the vessel density, orientation, and thickness are made. These features are combined with a brightness measure to compose a four-dimensional feature vector for each pixel. A library of training images is used with manually labeled ONs and maculae. Pixels within an average radius of the ON are used to compose an "optic nerve" training set with other pixels sampled to compose a "non-optic nerve" training set. A Gaussian model for both regions is formed with the training vectors. In addition, the manually labeled ON centers are used to form an *a priori* estimate of the ON center probability density function (pdf). The Gaussian parameters and the pdf are used to compute a likelihood ratio function using maximum *a posteriori* (MAP) estimation, and the pixel of highest likelihood is identified as the optic nerve. The macula is then found by modeling the vascular tree as a parabola, with the macula assumed to lie at a fixed distance from the ON at the angle indicated by the parabolic fit.

In earlier work [15] we compared the performance of the FBLR classifier with a model-based method [7] and explored the utility of fusing the methods to produce an optic nerve confidence metric [12-13]. To create the model, a group of optic nerves were manually segmented by non-clinical researchers. The images were resized to a common scale (the median size of the set), then resized to quarter size for computational efficiency. The images were rasterized into column vectors and projected using the small-sample size variation of principle component analysis (PCA) [16]. An additional data set taken from a later period of time was used to train a supervised classifier using neural networks to distinguish between optic nerve pixels and non-optic nerve pixels using the principle components as features. This differs somewhat from our earlier work which used a linear
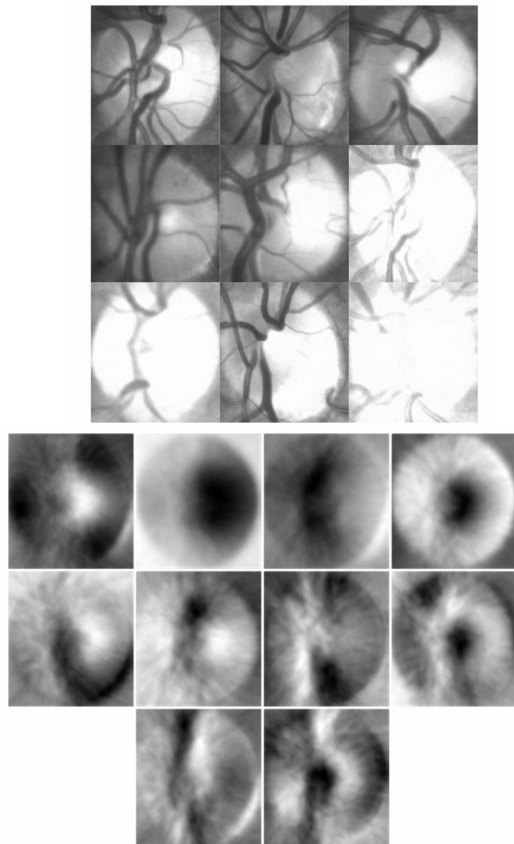
Figure 2. Top: Examples of cropped optic nerves. Bottom: Top 10 principle component eigendisks.

discriminate analysis (LDA) approach. The training samples were chosen from the optic nerve region and non-optic nerve regions at random, with the exception in the case of images where the FBLR method was shown to give an error; these erroneous regions were chosen as non-ON examples to help direct the training away from regions where the FBLR method had difficulty as discussed in [12]. The classifier was applied to test images by classifying each individual pixel and the neural network (NN) output was then convolved with a smoothing filter of the same size as the ON. The confidence metric was created by measuring the Euclidean distance between the locations determined by the two methods. When they were less than some threshold apart, a high confidence was assigned to the detection.

C. Disease Detection and Classification

For this work, we do not investigate the potential use of a confidence metric associated with the disease classification, but since this processing is our main means of evaluating the quality and optic nerve confidence metrics, we include a description of our processing in this area. Two main lesions or anomalies are currently detected with the system: exudates (small lipid deposits with a bright yellow color and distinctive spatial distribution) and microaneurysms (dilations of a small retinal capillary vessel, producing a small round dark spot in the vicinity of a capillary vessel.) The system uses specific detectors designed for these defects

as detailed in [8-9]. The detectors produce candidate regions of interest which are likely positive examples of each lesion. While these detectors were designed to provide good lesion detection without post-processing, we use supervised learning to separate the candidates into "true lesions" and "nuisance blobs". Both detectors generate candidate blobs which are filtered into "true lesions" and "nuisance blobs" using neural network classifiers trained on a set of ground truth lesions. The detected, filtered lesions create an overall fundus description that consists of a measurement of features related to the lesions detected (such as the number detected, histograms of the sharpness of the lesion edges, shape properties, etc.), the vascular density within the lesion population, population moments, and textural features of the macula region. This vector, which we refer to as the "population vector", is used to determine an index that can be used to locate similar images in an archive which then generates an automatic detection of disease by analysis of the ground-truth disease states of the similar images. We use PCA to reduce the dimensionality of the features space to a more efficient and effective search space, followed by a classification using a kNN classifier [17].

III. EXPERIMENTS

In this section we describe the experimental results using data from the telemedicine network. Some background on the data set is provided, and we then describe our validation process. We then present the results of the optic nerve confidence metrics on this data set, and finally we present the disease classification results on the quality confidence thresholds and optic nerve confidence methods.

A. Data Set

As of March 2011, there are five clinics in TRIAD based in Tennessee, North Carolina, and Mississippi. The examinations used span a roughly two-year period from February 2009 to January 2011. All images were collected with Zeiss VisuCam^{Pro} fundus cameras, in color with a roughly 6 microns per pixel resolution. Roughly 75% of the images are 45 degree field of view and 25% 30 degree field of view. The data set contains 5218 images obtained from 2378 patients over 2656 examinations; thus some patients are return visitors, with multiple examinations. Approximately 65% of the patients are African-American and 30% Caucasian.

The initial development of the optic nerve detectors and quality estimators, including testing, were performed on an early group of 1051 images from between February 2009 and March 2010. We refer to these as the "core set". The subsequent experiments described here do not include these images for classification studies, although they are used as part of the training dataset for the disease classification.
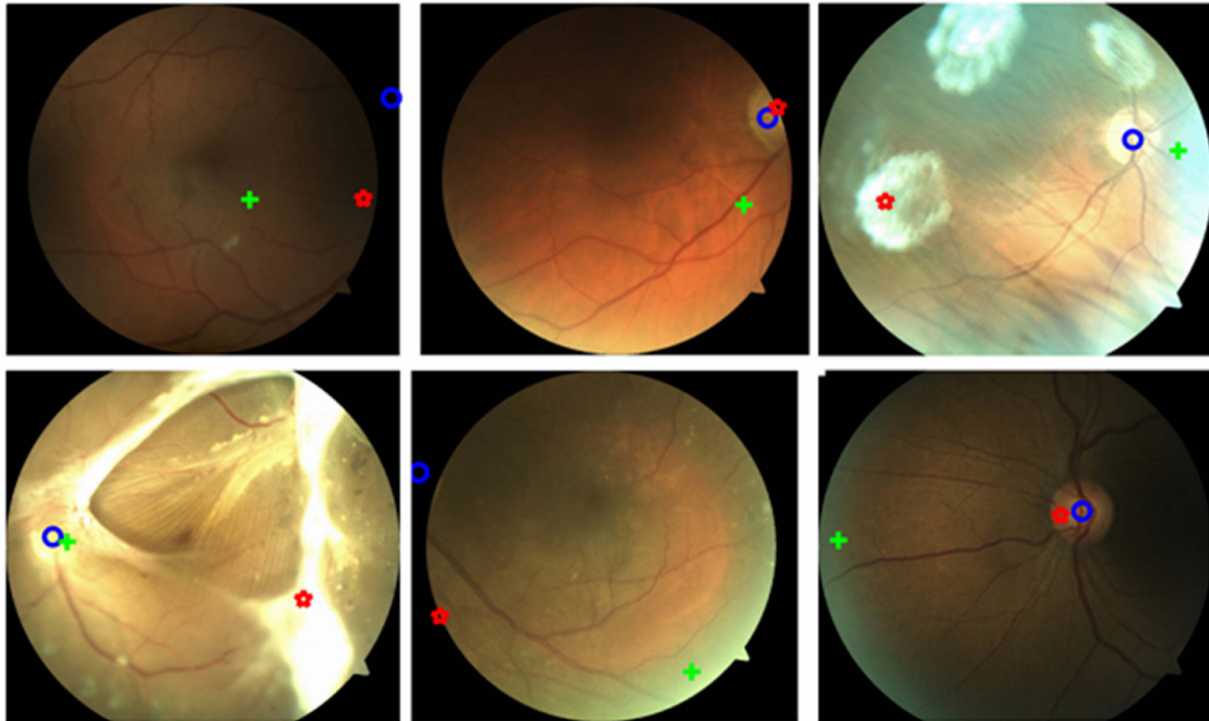
Figure 3. Examples of failed optic nerve detections. The ground truth value is in blue while f is the green cross and p is the red star. Top, left to right: missing optic nerve; partial optic nerve; flash artifacts. Bottom, left to right: Diseased retina; missing optic nerve; macula location incorrect (left eye identified as a right eye).

The disease state in the dataset is 83% normal / no diabetic retinopathy, 9% "mild DR without clinically significant macula edema (CSME)" and "mild Age-related Macular Degeneration (AMD)", and 8% consisting of more severe disease. These latter were grouped together into the "abnormal" or "positive" category for classification purposes, and the normal were grouped with mild DR and mild AMD into a "normal" or "negative" category. We wish to point out that this grouping may be problematic long term, but we justify this approach in this work since all these conditions share the same recommendations (6-month follow-up visit).

As a final note, our examinations are analyzed on a per-patient basis, using the physician diagnosis for the worse disease state of both eyes as ground-truth. In the case of patients who have been imaged by the system multiple times, we hold out all images of the patient when performing the classification of that patient, but only the earliest exam in the data set is classified. Finally, for each hold-one-out test, the training set consists of all images from the remaining patients, regardless of their frequency of examinations.

B. Optic Nerve Confidence Metric

The earliest 568 images from the "core set" were manually segmented by non-clinical researchers. The images were resized to the median size of 328 pixels square and the resulting images were resized to quarter size for computational efficiency. Some examples are shown in Figure 2. The PCA projection contained 90% of the energy in the first 10 components. We chose the 10 component option as our earlier work used the 90% threshold; however, we note that in this data set, we used much fewer components which may be due to the more normal appearance of most of these images as opposed to the data set of [15] which was from an ophthalmology practice as opposed to a screening environment. The eigendisks are also shown in Figure 2.

We then took the remaining images of the core set to create a training set for the ON detection neural network classifier. The optic nerve region was chosen and every fourth pixel was sampled as an example of an optic nerve. The remaining pixels were chosen by taking the region of lowest reconstruction error, choosing that pixel, removing pixels within one ON diameter of that pixel, and repeating using the next lowest reconstruction error. The data set thus consisted of two roughly equal numbers of ON pixel and non-ON pixel, each with a 10-dimensional feature vector. The resulting neural networks were trained using mean-square error and two hidden layers of 20 neurons each with the MATLAB© NN toolbox. After training the NNs, the optic nerve was located in a test image by classifying each individual pixel, then convolving the NN output for optic nerve targets with a smoothing filter of size 82 pixels square (the size of an ON). The position of peak response was chosen as the ON location. The two-fold validation test results were compared to the FBLR results and the Euclidean distance between the two estimates were measured. In the 2-fold test 99% accuracy was achieved
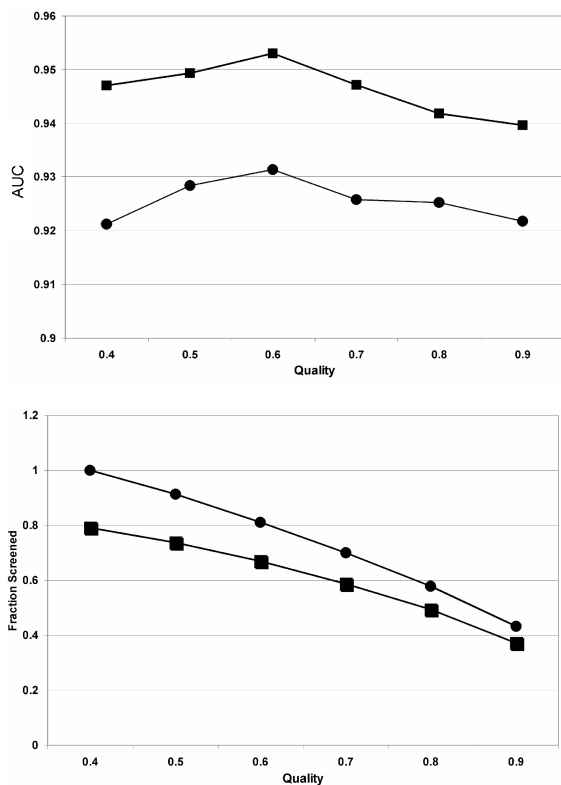
Figure 4 Top: Performance (based on AUC for disease detection) as T_Q is varied, for application with T_C (■) and without T_C (●). Bottom: The fraction of exams screened for these cases.

with ON differences of one ON diameter at a cost of 5% images rejected. A correct detection was interpreted as one where the estimated ON location was within one ON radii of the ground-truthed value. We therefore used this target as the ON detection threshold. We next tested the performance of the ON confidence metric on the remaining images after the core set. In this test, the accuracy at that distance threshold was slightly less at 98.6% but also with 5% images rejected. Some examples of images which are rejected are shown in Figure 3. In many cases artifacts or a missing ON are the primary issues.

C. Confidence Thresholds on Disease Detection

The quality confidence was evaluated by observing all image pairs (right eye / left eye) which passed a quality threshold T_Q varying from 0.4 to 0.9 in 0.1 increments. Each unique patient in the evaluation set was then evaluated on a "hold-one-out" basis, where all images from a patient were removed and the remaining images were used to form a training set. The feature dimension was reduced by using PCA with components selected to capture 80% of the signal energy, then the held-out exam was classified with a kNN classifier using K=67. A Receiver Operating Characteristics (ROC) curve was generated by increasing the weight of an abnormal vote, and the area under the ROC curve (AUC) was computed. At each threshold T_Q, the number of candidate exams (with two images over the quality

threshold) was computed as well and normalized by the maximum number possible (the number of exams with two images of quality 0.40 or greater). In addition, the optic nerve confidence metric was computed as described earlier, and an addition set of ROC curves were generated by applying this screening. In Figure 4, we show the AUC as a function of the T_Q and the number of exams screened as a function of T_Q. The fraction screened drops almost linearly for both non-T_C and T_C cases, from 100% screened to 40% for the former and 80% to just fewer than 40% for the latter. The T_C cases all perform better than no T_C by approximately 1%. However, as a function of quality, we see an increase in performance as T_Q increases up to 0.6 but then actually drops for higher values.

We examined the false negatives and false positive results for TQ = 0.6 with optic nerve confidence filtering. We selected the parameters which delivered a 94.8% sensitivity and 78.7% specificity. In this case there were a total of 610 exams, with 109 false positives and 5 false negatives. Two of the false negatives were AMD level 2, one was AMD level 3, one was mild DR with CSME, and one was moderate NPDR without CSME. We note that a drusen detector would likely improve the performance on the AMD images.

D. Discussion of Results

The overall detection performance is quite good from the standpoint of identifying serious disease conditions (those requiring immediate attention from an ophthalmologist.). We note however that the differentiation between normal / mild NPDR-CMSE, and mild NPDR-CMSE and more severe diagnosis is not as clear-cut. In a system which is being used with patients who are under regular screening, such as TRIAD, these distinctions may not be as important. If the system is intended to truly identify early disease, more importance should be attached to this issue. As a point of reference, if mild NPDR-CSME were treated as desired
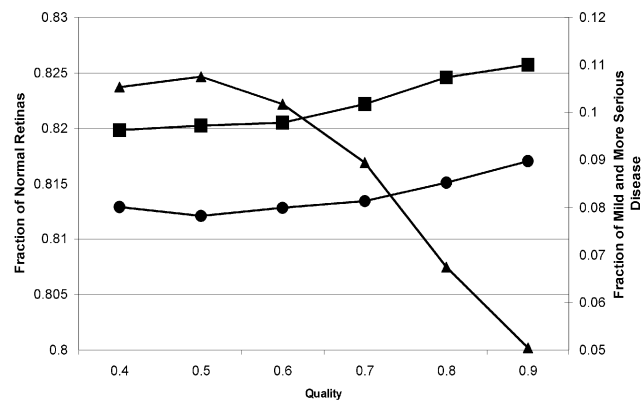

Figure 5. Prevalence of disease in database as quality changes: Normals (▲), Mild NPDR-CSME and Mild AMD (■), and more serious disease(●).

positive result, the system would perform with 77% sensitivity and 74% specificity; if mild NPDR-CSME were treated as a desired negative result, the performance is 91.8% sensitivity and 73.6% specificity. The latter result is greater most likely because the system was trained in this fashion with the kNN classifier.

Another rather unusual finding is the drop in performance as the quality threshold is raised, which is counter to intuition and our previous work [11]. Additional analysis of the images indicates that as the quality declines, a smaller percentage of the images in the datasets are of category normal as shown in Figure 5. This suggests that our methodology benefits more from the inclusion of more examples of normal retinas than abnormal, but further analysis is needed which is beyond the scope of this work. The result of the ON confidence filtering was more intuitive, as the use of the filter improves performance by removing more false positives. This is likely because images that satisfy the ON confidence metric are less likely to show false positive lesion detections from the typically brighter areas around the optic nerve.

IV. CONCLUSION

In this work we evaluate two confidence metrics, for quality and optic nerve detection to determine the effect of these constraints on overall classification of disease in retina images from a telemedical network. We note the optic nerve detection confidence metric could be applied to different optic nerve detection methods that use different or complimentary characteristics for the detection. Future work involves using larger data sets with hopefully more complete non-image data, as roughly 40% of the examinations did not have sufficient non-image data for inclusion. We hope to ultimately incorporate a more detailed disease stratification estimation which extends beyond the normal / abnormal type of detection described here. We would also like to investigate adding additional detectors to the lesion detection process, exploring other filtering techniques and thresholds, and the impact of additional non-image data as it is available.

REFERENCES

[1] Boyle, J., et. al., (2010) Projection of the year 2050 burden of diabetes in the US adult population: dynamic modeling of incidence, mortality, and prediabetes prevalence, Population Health Metrics 8:29

[2] Abramoff, M.D., Niemeijer, M., and Russell, S.R., (2010) Automated detection of diabetic retinopathy: barriers to translation into clinical practice, Expert Rev. Medical Devices 7:287-296

[3] Philip, S. et al., (2007) "The efficacy of automated 'disease / no disease' grading for diabetic retinopathy in a systematic sceening programme', British J. of Ophth., 91:1512

[4] Scotland, G.S., et al (2010) "Costs and consequences of automated algorithms versus manual grading for the detection of referable diabetic retinopathy", 94:712

[5] Li, Y., et.al., (2011) TRIAD: A HIPPAA-Compliant Ocular Telehealth Network for the Remote Diagnosis and Management of Diabetic Retinopathy, Telemedicine and E-Health (submitted)

[6] Giancardo, L., Abramoff, M.D., Chaum, E., Karnowski, T.P., Meriaudeau, F., and Tobin, K. W., (2008) Elliptical local vessel density: a fast and robust quality metric for fundus images. 30th Annual International Conf. of the IEEE EMBS Vancouver, Canada, 2008

[7] Tobin KW, Chaum E, Govindasamy VP, Karnowski TP. (2007) Detection of anatomic structures in human retinal imagery. IEEE Trans Med Imaging 26:1729-39

[8] Giancardo, L., et. al, (2010), Microaneurysms detection with the Radon Cliff operator in retinal fundus images, Proc. SPIE Medical Imaging, San Diego CA, 7623:7623

[9] Giancardo, L., et. al, (2011), Automatic retina exudates segmentation without a manually labeled training set, Proc. IEEE International Symposium on Biomedical Imaging

[10] Chaum, E., Karnowski, T.P., Govindasamy, V.P., Abdelrahamen M., Tobin K.W., (2008) Automated diagnosis of retinopathy by content-based image retrieval. Retina 28:1463-1477

[11] Tobin, K.W., Abramoff, M.D., Chaum, E., Giancardo, L., Govindasamy, V.P., Karnowski, T.P., Tennant, M.T.S., (2008) Using a patient image archive to diagnose retinopathy. 30th Annual International Conf. of the IEEE EMBS, Vancouver, Canada, August 2008

[12] Karnowski, T., et al (2009) Practical considerations for optic nerve location in telemedicine, 31st Annual International Conf. of the IEEE EMBS Minneapolis, M, 2009

[13] Karnowski, T., Aykac, D.; Chaum, E.; Giancardo, L.; Li, Y.; Tobin, K.W.; Abramoff, M.D. (2009) Evaluating the Accuracy of Optic Nerve Detections in Retina Imaging Using Complementary Methods, World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany

[14] Zana, F., Klein, J., (2001) Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation. IEEE Trans. Image Processing, 10:1010-1019

[15] Karnowski, T.P., Govindasamy, V.P., Tobin, K.W., Chaum, E., (2006) Locating the optic nerve in retinal images: comparing model-based and Bayesian decision methods. 28th Annual International Conf. of the IEEE EMBS New York, NY 2006

[16] Fukanaga, K., (1990) Introduction to Statistical Pattern Recognition, Academic Press

[17] Duda, R.O, Hart, P.E. and Stork, D.G, (2000) Pattern Classification, Wiley Interscience

[18] Hoover, A., and Goldbaum, M. (2003) Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels, IEEE Trans Med Imaging 22:951-958

[19] Foracchia, M., Grisan, E., and Ruggeri, A. (2004) Detection of optic disc in retinal images by means of a geometrical model of vessel structure, IEEE Trans Med Imaging 23:1189-1195

[20] Abramoff, M.D., and Neimeijer, M., (2006) The automatic detection of the optic disc location in retinal images using optic disc location regression, 30th Annual International Conf. of the IEEE EMBS, New York, August 2006

[21] Mishra, A., et. al., (2009) Intra-retinal layer segmentation in optical coherence tomography images, Optics Express 17:23719-23728