Automated image quality evaluation of retinal fundus photographs in diabetic retinopathy screening

Honggang Yu*[1,2], Carla Agurto[2,1], Simon Barriga[1,2], Sheila C. Nemeth[1], Peter Soliz[1], Gilberto Zamora[1]
[1]Visionquest Biomedical, Albuquerque, NM 87106, USA
[2]Dept. of Electrical & Computer Engineering, University of New Mexico, Albuquerque, NM 87131, USA
*email: hyu@visionquest-bio.com

Abstract— **This paper presents a system that can automatically determine whether the quality of a retinal image is sufficient for computer-based diabetic retinopathy (DR) screening. The system integrates global histogram features, textural features, and vessel density, as well as a local non-reference perceptual sharpness metric. A partial least square (PLS) classifier is trained to distinguish low quality images from normal quality images. The system was evaluated on a large, representative set of 1884 non-mydriatic retinal images from 412 subjects. An area under the ROC curve of 96% was achieved.**

Keywords - Retinal image; quality evaluation; diabetic retinopathy screening; non-reference image sharpness metric

I. Introduction

Digital fundus photography is a common procedure in ophthalmology and provides critical diagnostic information of retinal pathologies, such as diabetic retinopathy (DR), glaucoma, age-related macular degeneration, and vascular abnormalities. The research community has put forth a great effort towards the automation of a computer screening system able to promptly detect DR in fundus images [1,2]. An algorithm able to automatically assess the quality of the fundus image is an important preprocessing step for reliable lesion detection for a computer-based screening system.

In a DR screening system, an image is deemed as inadequate when it is difficult or impossible to make a reliable clinical judgment regarding the presence or absence of DR in the image [3]. Studies show that the percentage of images that are inadequate for screening systems is about 10% of the mydriatic (pupil dilation) images [4,5,6]. For single field non-mydriatic (no pupil dilation) images, the percentage of inadequate quality images has been reported at 20.8% [7]. Major causes of inadequate image quality in retinal screening images include illumination crescents due to small pupil size; loss of contrast due to poor focus, movement of the eye, or media opacity; and imaging of part of the eyelid and eyelash due to blinking; as well as insufficient illumination.

Several approaches to automatically determine the quality of retinal images have been developed. Reference-based methods were first developed by Lee and Lalonde [8,9]. However, a limited number of the good quality images might not be a “good” reference for the natural large variance encountered in retinal images acquired from screening.

Usher et al. [10] was one of the first authors to use vessel detection in the whole image for image quality assessment. To assess quality of macula-centered images, Fleming et al. measured vessel density in a region around the macula [11]. More recently, Giancardo et al. [12] split the field of view into sub-regions with different sizes of elliptical rings and different angles. The local vessel density and 5-bin color histogram features were used to determine image quality.

Other automatic retinal image quality methods do not require any segmentation of the image. Niemeijer et al. [3] clustered multiscale Gaussian derivative filterbank responses to obtain a compact representation of the image structures. Five bins of normalized histogram of image structure clusters (ISC) and 5 bins of normalized RGB histogram were used to classify the quality of each image. Image structure clusters (ISC), three Haralick features, and sharpness measures based on image gradient magnitudes were used by Paulus et al. [13] to classify inadequate images. Davis et al. used histogram and Haralick features in CIElab color space as additional information for classification [14].

In this work, we present an automated approach based on the classification of global and local features that correlate with the human perception of retinal image quality as assessed by eye care specialists. The overall image content, such as lightness homogeneity (detect crescents, etc.), brightness, and contrast are measured by global histogram and textural features. The sharpness of local structures, such as optic disc and vasculature network, is measured by a local perceptual sharpness metric and vessel density. In the next section we detail each of the components of the system.

II. Methodology

The retinal image quality evaluation system presented herein consists of two main processing phases: feature extraction and PLS classification. The first phase is further subdivided into four parts and described as follows.

A. Feature Extraction

Four categories of features are used to evaluate the image quality: histogram features, textural features, vessel density, and local sharpness metrics.

1) Vessel Density Features

Global vessel density measures are used in order to check the sharpness of dark vessel structures since the performance of vessel segmentation is sensitive to the blurriness of vasculature. To segment the retinal vasculature, a previously presented method [15] based on Hessian eigen system and the second order local entropy thresholding was developed. Vessel segmentation is performed after illumination correction and adaptive histogram equalization to remove

978-1-4673-1830-3/12/$31.00 ©2012 IEEE

uneven lightness and enhance contrast in the green channel of the retinal images. Fig. 1 shows an example of the vascular segmentation.

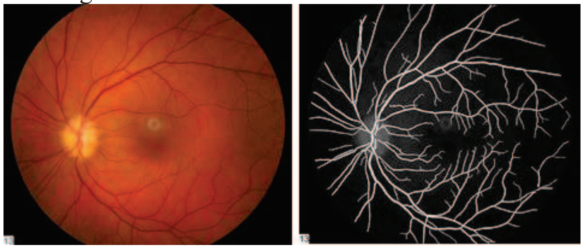

Figure 1. Retinal vasculature segmention. (a) retinal image (b) local entropy segmentation with Hessian enhancement

The vessel density is calculated as the ratio of the area of segmented vessels over the area of field of view (FOV) in an image.

2) Histogram Features

Seven histogram features are extracted from the RGB color spaces. They are mean, variance, skewness, kurtosis, and the first three CDF quartiles. These seven histogram features describe the overall image information such as brightness, contrast, and lightness homogeneity. We also compute the first order entropy and spatial frequency to detect the image complexity.

3) Textural Features

Texture is one of the important characteristics used in identifying objects in an image. The texture information can be obtained by computing the co-occurrence matrix of an image. We calculated five Haralick texture features: the second order entropy, contrast, correlation, energy and homogeneity [16]. Entropy measures the randomness of the elements of the matrix, when all elements of the matrix are maximally random, entropy has its highest value. Contrast measures the intensity difference between a pixel and its neighbor. The correlation feature measures the correlation between the elements of the matrix. When correlation is high, the image will be more complex than when correlation is low. The fourth feature, energy, describes the uniformity of the texture. In a homogeneous image there are very few dominant grey-tone transitions, hence, the co-occurrence matrix of this image will have fewer entries of large magnitude. Therefore, the energy of an image is high when the image is homogeneous. The last feature, homogeneity, also called inverse difference moment, has a relatively high value when the high values of the co-occurrence matrix are near the main diagonal. It is a measure of coarseness in an image.

4) Local Sharpness Features

A clear and sharp edge is important for a good quality image. A local sharpness metric, the cumulative probability of blur detection (CPBD), is applied on the green channel to measure the strength of sharp edges in the images.

Average edge width and gradient magnitude have been used to measure blurriness of images [13,17]. These have been found to be too simplistic to directly correspond to human visual perception of blurriness, which is a complicated process [18]. A metric, which uses the cumulative probability of blur detection (CPBD) at every edge, was therefore studied [19]. CPBD assumes that the blurriness around an edge is more or less noticeable depending on the local contrast around that edge. It derives a human perceptibility threshold called "Just Noticeable Blur" (JNB), which can be defined as the minimum amount of perceived blurriness around an edge given a contrast higher than the "Just Noticeable Difference" (JND). The perceived blurriness around an edge is counted only when its amount is larger than the "Just Noticeable Blur" under "Just Noticeable Difference" in contrast. It can be modeled as follows:

$$P_B(e_i) = 1 - e^{-\left|\frac{w(e_i)}{w_{JNB}(e_i)}\right|^{\beta}},$$

where $P_B(e_i)$ is the probability of blur detection at each edge e_i. If the actual width of the edge is the same as the JNB edge width, then $P_B(e_i)$ =63%, below which blur is said to be undetectable. β has a median value of 3.6 based on experimentally determined psychometric function [18]. The cumulative probability is based on the sum of the blur probabilities that are below 63%. The CPBD value is therefore negatively correlated to the edge blurriness. To measure the local edge blurriness, the image is divided into blocks of a size 64×64 for each block. And edge block ratio (EBR), in which the number of edge pixels is more than a certain threshold, over the number of all blocks, is used as another feature, since some hazy retinal images due to cataract may appear foggy and misty with sharp edge underneath.

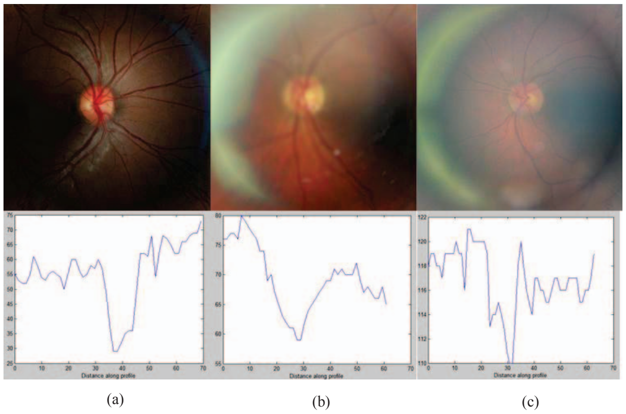

Figure 2. Local sharpness metrics of retinal images. (a) a good quality image with sharped edge, CPBD = 0.61, EBR = 0.69. (b) a defocus low quality image, CPBD = 0.26, EBR = 0.34. (c) a low quality image with cataract, CPBD = 0.75, EBR = 0.35.

Fig. 2 shows three example images (the first row) and their intensity profile across a vessel segment (the second row). The cataract image (c) has higher CPBD value than the good quality image (a) because of the edge sharpness. The edge block ratio (EBR) can capture the characteristic of hazy images as in Fig. 2(c).

B. Partial Least Squares Classification

Partial least squares (PLS) classifier was used to develop a predictor of image quality. PLS is a very powerful method for eliciting the relationship between one or more dependent variables and a set of independent (predictor) variables, especially when there is a high correlation between the input variables. In our case, the same feature when applied to different color channels tends to be highly correlated even if their magnitudes are quite different. PLS finds a lower dimensional orthogonal sub-space to the multi-dimensional feature space and provide robust prediction. We have found PLS is an effective tool in automated DR screening [1,20].

III. Data set

Digital fundus photographs from 412 subjects (1884 images) were captured with a Canon CR1 Mark II camera. The images are macula-centered or optic disc centered, 45 degree field of view, non-mydriatic images of both eyes with a dimension of 4752×3168 pixels. The image quality grading was provided by two optometrists. The graders assigned each image to one of four quality categories: high, medium, low, and reject based on overall illumination, image contrast, sharpness of fine structure, illumination on the macula, and grader's confidence on ability to identify lesions. The purpose of image quality evaluation is to detect the low quality images whose quality is inadequate for DR screening. We grouped reject and low quality into inadequate category, and medium and high quality into adequate category for providing the reference standard for classification.

IV. Results

To show the effectiveness of the proposed methodology, different feature sets have been tested. The isolated features (histogram, texture features, and blurriness features) and a combination of all features have been tested. All features were normalized to zero mean and unit standard deviation before classification.

In the experiment, 28% of 1884 images were marked as inadequate quality. External shots of the pupils were not excluded from the data set and were graded as inadequate quality. Statistical significance tests of the difference between ROC curves were performed. The sensitivity for classifying inadequate quality images at a specificity of 80% using leave-one-out validation is given in Table I.

TABLE I. Results of the classifier using different features for total 1884 images

	AUC (Std)	95% CI	Sensitivity (Specificity of 80%)	p value
Histogram	82.2% (1.06%)	(80.1%, 84.2%)	70.2%	0.0000
Haralick	72.3% (1.32%)	(70.0%, 75.1%)	53.5%	0.0000
Blurriness	92.6% (0.63%)	(91.4%, 93.9%)	89.3%	0.0000
All features	95.8% (0.44%)	(94.8%, 96.6%)	95.3%	--

The blurriness features (including vessel density and CPBD metrics) achieved the best performance of the selected subsets of features (92.6% of AUC). The highest classification performance is achieved by the final combination of all the features (95.8% of AUC). The tests of the difference between the areas under the ROC curves were performed by using Metz ROC Software [21]. The difference between all features and the isolated features was found to be significant with a p value less than 0.00005.

The performance of automated image quality evaluation is satisfactory considering the diversity of the images in the data set (shown in Fig. 3). We tested the system on 824 macula-centered images selected from the above data set. There are 97 images marked as inadequate quality, which is 12% of the 824 images. Results of this test are shown in Table II. The performance of the classifier using all features is significantly better than any isolated feature at the significant level of 95%. Because only macula-centered images were used in this experiment, the diversity of image contents was reduced; a higher AUC was achieved at 98.1%. But the standard deviation of AUC is slightly larger than that in the experiment where 1884 images were used.

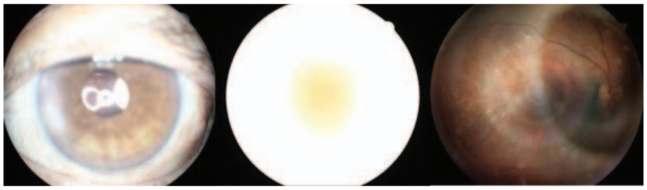

Figure 3. Diversity of the images in the data set. (a) external shot of the pupil, (b) over-exposured image, (c) retinal image with camera artifacts.

TABLE II. Results of the classifier using different features for 824 macula-centered images

	AUC (Std)	95% CI	Sensitivity (Specificity of 80%)	p value
Histogram	92.9% (1.43%)	(89.8%, 95.5%)	91.7%	0.0001
Haralick	81.6% (2.27%)	(77.0%, 85.9%)	65.3%	0.0000
Blurriness	96.4% (0.77%)	(94.7%, 97.8%)	95.8%	0.0017
All features	98.1% (0.46%)	(97.1%, 98.9%)	99.0%	--

To validate the system, a ten-folded cross-validation method was applied. Images were randomly chosen for each of the ten subsets. Each image was tested exactly once in the experiment. The average AUC for the ten runs for training and testing were 96.5% and 95.3%, respectively for 1884 images. For 824 macula-centered images, the average AUC of the ten runs for training and testing were 98.3% and 97.5%, respectively. We notice that the classification results are consistent when going through multiple rounds of cross-validation using difference data partitions. This indicates that the extracted features are effective in separating the images into two classes.

Fleiss' kappa correlation was calculated between the reference standard and the results generated by the automated method. We used the threshold that obtained the sensitivity of 95.3% at the specificity of 80%. The inter-observer correlation between the two different human observers is 0.25, while the correlation between the reference standard and the automated method is 0.72. We also calculate the kappa value by adding the automated method's results as a third observer's results to test the agreement. The initial inter-observer correlation (k =0.25) is increased to 0.43 for assuming the automated method to be an additional observer. The difference between the reference standard and the results of the proposed automated method is less than the inter-observer variability.

V. Discussion and Conclusion

The image quality evaluation system presented in this paper performs well by using global image appearance features and local sharpness measures. Due to the large variability presented in the data set, the content of the images is very heterogeneous; thus reducing the effectiveness of histogram and Haralick features in quality evaluation. Vessel density and local CPBD features enhance the system performance by used as effective measures of image structure blurriness.

We also used Wilcoxon rank sum and Ansari-Bradley test to determine whether the adequate and inadequate quality classes' median values and dispersions (e.g. variances) differ significantly. If neither the median nor the dispersion differ significantly between adequate and inadequate classes, the feature is unlikely to be useful for classification. We used the selected histogram features and Haralick features in classification. However, a decreased performance was presented. So no feature reduction was applied in the final classification.

Since the image quality may appear differently in localized areas of a retinal image, we performed an experiment in which all the features were extracted from three regions of interest (ROI): optic disc region, upper retinal and lower retinal hemispheres. However, the final performance of the system using ROIs was slightly worse measured in AUC than the system using global features with a p value of 0.4793. Using the bivariate binormal model, the shape of ROCs were not significant different with a p value of 0.1145.

In summary, we developed an automated method which integrates global histogram and texture features, as well as local sharpness features in retinal image quality evaluation. By introducing local sharpness criteria, the performance of the system is increased even if the image contents exhibit large heterogeneities. The discrimination of quality classes on the images which are on the borderline of normal and low quality categories remains a crucial task. Our method underlies a restriction that it is limited by the human graded reference standard. Nevertheless, it has the potential to provide an efficient, objective and robust tool in retinal image quality evaluation for broad screening applications.

References

[1] C. Agurto, V. Murray, E. Barriga, S. Murillo, M. Pattichis, H. Davis, S. Russell, M. Abramoff, and P. Soliz, "Multiscale AM-FM Methods for Diabetic Retinopathy Lesion Detection," Medical Imaging, IEEE Transactions on, vol. 29, pp. 502-512, 2010.

[2] Fleming AD, Goatman KA, Philip S, Prescott GJ, Sharp PF, Olson JA. Automated grading for diabetic retinopathy: a large-scale audit using arbitration by clinical experts. Br J Ophthalmol. 2010.

[3] Niemeijer M, Abràmoff MD, van Ginneken B. "Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening". Med Image Anal. 2006 Dec;10(6):888-98.

[4] Teng T., Lefley M., Claremont D., "Progress towards automated diabetic ocular screening: a review of image analysis and intelligent systems for diabetic retinopathy", Med. & Biological Engineering & Computing, vol. 40, pp2-13, 2002

[5] Liesenfeld B, Kohner E, Piehlmeier W et al. (2000): "A telemedical approach to the screening of diabetic retinopathy: digital fundus photography", diabetes care, 23, pp. 345–348

[6] Philip S, Cowie LM, Olson JA, The impact of the Health Technology Board for Scotland's grading model on referrals to ophthalmology, Br J Ophthalmol. 2005;89:891–896.

[7] Scanlon PH, Malhotra R, Greenwood RH, et al. Comparison of two reference standards in validating two field mydriatic digital photography as a method of screening for diabetic retinopathy. Br J Ophthalmol. 2003;87:1258–1263.

[8] Lee SC and Y Wang, "Automatic retinal image quality assessment and enhancement," Proceedings of SPIE Medical Imaging Processing, 3661:1581–1590. SPIE (Washington, DC 1999).

[9] Lalonde, M, L Gagon, and MC Boucher, "Automatic visual quality assessment in optical fundus images," Proceedings of Vision Interface 2001, Ottawa, 259-264, June 2001.

[10] Usher DB, Himaga M, Dumskyj MJ, et al. "Automated assessment of digital fundus image quality using detected vessel area. Proceedings of Medical Image Understanding and Analysis." 2003;81–84. British Machine Vision Association (BMVA) Sheffield, UK:

[11] Fleming AD, Philip S, Goatman K, Olson J, Sharp P. Automated assessment of diabetic retinal Image quality based on clarity and field definition. Invest Ophthalmol Vis Sci. 2006; 47:1120–1125.

[12] Giancardo L, Abràmoff MD,Chaum E, Karnowski TP, Meriaudeau F, Tobin KW Jr (2008) Elliptical local vessel density: a fast and robust quality metric for retinal images, Engineering in Medicine and Biology Society, 2008. EMBS 2008.

[13] Paulus J, Meier J, Bock R, Hornegger J, Michelson G. "Automated quality assessment of retinal fundus photos". International Journal of Computer Assisted Radiology and Surgery. 2010. vol 5(6), pp 557-564

[14] Davis H, Russell SR, Barriga ES, Abramoff M, soliz P, "Vision-based, real-time retinal image quality assessment". CBMS 2009: 1-6

[15] H Yu, S Barriga, C Agurto, G Zamora, W Bauman and P Soliz, Fast Vessel Segmentation in Retinal Images Using Multiscale Enhancement and Second-order Local Entropy, accepted by SPIE medical imaging, Feb, 2012, San Diego, USA

[16] Haralick, R.M., K. Shanmugan, and I. Dinstein, "Textural Features for Image Classification", IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-3, 1973, pp. 610-621.

[17] Winkler, S. et. al. No-reference Perceptual Blur Metric IEEE 2002 International Conference on Image Processing. 2002.

[18] Ferzli, R. Karam, L. No-reference Objective Image Sharpness Metric Based on the Notion of Just Noticeable Blur IEEE Transactions on Image Processing. pp. 717 728. 2009.

[19] N. D. Narvekar and L. J. Karam, "A No-Reference Image Blur Metric Based on the Cumulative Probability of Blur Detection (CPBD)," IEEE Transactions on Image Processing, Vol 20 (9), 2678-2683, Sep. 2011.

[20] Agurto Rios C., Barriga E S, Zamora G. , Murray V. , Murillo S., Yu H. , Bauman W.C. , and Soliz P., Automatic Screening of Eye Diseases using three-field Fundas Photographs, ARVO, Fort Lauderdale, FL, USA, May 1-5, 2011.

[21] http://metz-roc.uchicago.edu/MetzROC