REAL-TIME DIABETIC RETINOPATHY PATIENT SCREENING USING MULTISCALE AM-FM METHODS

Victor Murray[*,♯], *Carla Agurto*[*,♮], *Simon Barriga*[*,♮], *Marios S. Pattichis*[*], *and Peter Soliz*[♮]

[*]University of New Mexico, Department of Elect. and Comp. Eng., Albuquerque, New Mexico, USA.
[♯]Universidad de Ingeniería & Tecnología, Lima, Perú.
[♮]VisionQuest Biomedical, Albuquerque, New Mexico, USA.
E-mails: vmurray@ieee.org, pattichis@ece.unm.edu, psoliz@visionquest-bio.com

ABSTRACT

In this paper we present a robust and improved system for diabetic retinopathy (DR) screening. The goal of the system is to automatically screen out digital fundus photographs of diabetic patients who do not present signs of DR. This work is motivated by the large amount of diabetics in the world who do not receive their recommended eye exams, leading to widespread blindness as a complication of diabetes. The system is based on multiscale amplitude-modulation frequency-modulation (AM-FM) methods for feature extraction, and uses supervised and unsupervised methods to produce its final output, namely, a normal or abnormal grade. The most time-consuming processing routines of the system are implemented in C using a compute unified device architecture (CUDA) to produce results in real-time. The system was tested using 776 images from 388 patients (one macula-centered image from each eye). During the training phase of the system, the data was divided in 70% for training and 30% for testing. The system was tested using 20 random training/testing distributions, obtaining an average sensitivity of 89% and specificity of 59%. Analysis of sight-threatening conditions resulted in a sensitivity of 98% for these types of cases.

***Index Terms*—** amplitude-modulation frequency-modulation (AM-FM), multi-scale analysis, diabetic retinopathy.

1. INTRODUCTION

The National Institutes of Health (NIH) states that the leading cause of new blindness among adults of 20-74 years old is diabetes [1]. In 2005-2008, in the United States only, 4.2 million people (28.5%) with diabetes ages 40 years or older had diabetic retinopathy, and of these, 4.4% had advanced diabetic retinopathy (DR) that could lead to severe vision loss. It is estimated that over 10 million diabetics do not receive the recommended annual eye examinations, significantly increasing their risk of vision loss.

A system for automatic diabetic retinopathy (DR) screening using multi-scale amplitude-modulation frequency-modulation (AM-FM) methods has been first presented in [2, 3]. In this paper, we present a new version of the system that is patient-based. In other words, instead of classifying individual images, we are now producing a classification for two retinal images taken from the same patient. Patient classification is better suited for screening because it allows us to integrate several components of the exam into a final result. It also improves the sensitivity and specificity of the system depending on which strategy is used for combination of grades. Finally, in an automatic screening environment an assessment of the whole case needs to be provided as the final result of the examination. Furthermore, the new system is validated on a diverse set of images that were collected from 3 different DR screening centers, with new image normalization pre-processing, using extended AM-FM decompositions over 5 scales (instead of 4), with strict cross-validation (robust regression), that is also re-implemented to run in real-time.

Related research for DR screening approaches can be found in [4] where the authors compared k-Nearest Neighbour (kNN) versus the use of random Forest. Also, in recent years, many authors have proposed automatic DR screening systems based on the detection of microaneurysms, blot haemorrhages and exudates ('bottom-up' approaches, see for example [5]). Gabor based methods are also applied for example in [6]. Although great progress has been made in the development of automatic systems for DR screening, there are still several open problems. Thus, AM-FM has been applied to all types of DR diseases, while other approaches are problem specific. Our system is based on multi-scale AM-FM decompositions that allows us to estimate the instantaneous amplitude (IA) and the instantaneous frequency (IF) at every pixel over 5 different scales. This approach allows us to provide a top-down approach that can be applied to several types of eye diseases [7].

We describe the methodology in section 2. Then, we present the results in section 3 and the discussion in section 4. Finally, conclusions are presented in section 5.

2. METHODOLOGY

We present the block diagram of the system in Fig. 1. For each patient, a digital retinal image is taken for each eye. The mean image intensity is normalized so as to ensure that all images have the same mean. Then, the AM-FM texture features are computed using 13 different combination of scales. Next, we apply k-means clustering to group extracted regions of interest according to their characteristics. After that, partial least squares is used to reduce the high dimensionality of the data and to classify each image. Finally, a combiner block is used to produce the final patient grade from the individual image grades. The following subsections provide more details for the most significant steps.

2.1. AM-FM demodulation

For each normalized image, we consider a multi-scale AM-FM representation of digital images given by [8]

$$I(k_1, k_2) \simeq \sum_{n=1}^{M} a_n(k_1, k_2) \cos \varphi_n(k_1, k_2), \qquad (1)$$

This work was supported in part by the National Eye Institute under Grants EY018280, EY020015 and RC3 EY020749.

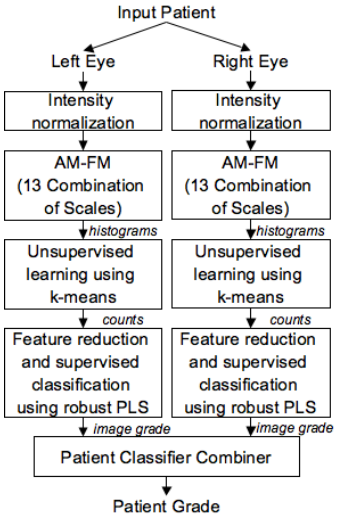

Fig. 1. Block diagram of the patient DR screening system. The digital images from both eyes of each patient are processed and used to produce the patient grade.

where $n = 1, 2, \ldots, M$ denote different frequency scales, a_n denotes the instantaneous amplitude (IA) functions, and φ_n denotes the instantaneous phase (IP) functions. Each scale is defined in terms of a set of separable bandpass filters with similar frequency magnitude ranges. It is assumed that the IA functions represent the slow varying over the image; a high IA value means a strong presence of the corresponding frequency scale. Associated with each AM-FM component, the instantaneous frequency (IF) is defined as $\nabla\varphi = (\varphi_x, \varphi_y)$.

The input images are first filtered using an extended 2-D Hilbert filter (see [8] for more details). Then, they are processed through a dyadic filterbank using separable filters. Some examples of similar AM-FM applications can be found in [2, 8]. Next, the AM-FM demodulation is applied at the output of each bandpass filter. For each pixel, a dominant component analysis (DCA, [9]) is applied within each frequency scale.

To compute the AM-FM estimates, from the output of each bandpass filter, we estimate the IA and IP using: $\hat{a}(k_1, k_2) = |\hat{I}_{AS}(k_1, k_2)|$ and $\hat{\varphi}(k_1, k_2) = \arctan(\text{imag}(\hat{I}_{AS}(k_1, k_2))/\,\text{real}(\hat{I}_{AS}(k_1, k_2))$, respectively, with $\hat{I}_{AS}(k_1, k_2) = I(k_1, k_2) + j\mathcal{H}_{2d}[I(k_1, k_2)]$, where $\mathcal{H}_{2d}$ denotes a two-dimensional extension of the one-dimensional Hilbert transform operator. We use the variable spacing, local linear phase (VS-LLP) method for robust IA-IF estimation (described in [8]). To estimate the first IF component, we use $\hat{\varphi}_x(k_1, k_2) = \frac{1}{n_1}\arccos(\gamma(n_1))$, where $\bar{I}_{AS}(k_1, k_2) = \hat{I}_{AS}(k_1, k_2)/|\hat{I}_{AS}(k_1, k_2)|$ and $\gamma(n_1) = \left(\bar{I}_{AS}(k_1 + n_1, k_2) + \bar{I}_{AS}(k_1 - n_1, k_2)\right) / \left(2\bar{I}_{AS}(k_1, k_2)\right)$. We perform a similar approach for the second IF component $\hat{\varphi}_2$.

As described in [8], n_1 represents a variable displacement, from 1 to 4 formulated from the optimization problem

$$\begin{aligned} \underset{n_1}{\text{minimize}} \quad & |\gamma_{\arccos}(n_1)| \\ \text{subject to} \quad & \hat{\varphi}_x \in [w_{p_{1x}}, w_{p_{2x}}], \end{aligned} \tag{2}$$

where $w_{p_{1x}}$ and $w_{p_{2x}}$ represents the limits to the x-direction projection of the bandpass used.

In the DR system, we use 3 AM-FM estimates: (i) IA, (ii) IF magnitude, and (iii) IF angle. We use a 5-scale filterbank with the space of frequency scales defined as: (i) H: high, (ii) M: medium, (iii) L: low, (iv) V: very low (half of L), (v) U: ultra low (half of V), and (vi) F: lowpass filter. Then, we compute the AM-FM estimates (3 in total) using 13 combinations of the frequency scales (CoS) mentioned before. The goal of each CoS is to analyze the AM-FM estimation at independent scales and at neighbors of scales (see [3] for more details).

Since the AM-FM estimation is time demanding, we use a C/CUDA (compute unified device architecture) implementation as stated in [10].

Thus, an input image is divided into 39 (3 AM-FM estimates and 13 CoS) new images.

2.2. Unsupervised learning

Each image from the multiscale AM-FM estimation (39 in total) is divided into a fixed number of regions of interest (ROI) with fixed dimensions. Then, a histogram of the values of each AM-FM estimate for each ROI is produced to approximate its probability density function. These histograms become the descriptors of each ROI. Thus, assuming that we have N_R ROIs per image and that we are using b variables for each histogram, each of the 39 images (observations) is now represented using $N_R \times b$ variables.

Next, to reduce the number of variables to represent each image, we apply k-means as an unsupervised learning method to cluster the different types of ROIs into K groups [2, 11]. By this, we reduce the number of variables to represent each of the 39 images from $N_R \times b$ to K.

Due to the high dimensionality of the data in the training set (we describe the images used in subsection 3.1), different random initial cluster centroid positions (seeds) can lead to different final centroids. To make the system more robust and to have more probabilities to reproduce each clustering with the same final centroids, we repeat the clustering 3 times. Thus, we use as answer the repetition with the minimum sum of point-to-centroid distances.

For the testing images, the centroids computed here are used to produce the feature vectors with K variables.

2.3. Feature reduction and supervised classification using partial least squares and determination of patient grade

Up to this point, each input fundus image has been described by using 39 features vectors of K variables each. Thus, to create a DR system model, given N_{Tr} images for the training, with their corresponding grades, we use Partial Least Squares (PLS, [12]) first to reduce the dimensionality from high correlation observations at each CoS to robust and uncorrelated observations.

PLS is a linear regression method formulated as $y = X\beta + \varepsilon$, where y is a $n \times 1$ vector of the classification variables, X is a $n \times p$ matrix of the extracted AM-FM features, β is a $p \times 1$ vector of regression weights, and ε is a $n \times 1$ vector of residuals. The least squares solution to estimating β is given by the normal equations $\beta = (X^T X)^{-1}(X^T y)$ (see more details in [2, 7]).

In most of classification applications, there are much more observations than images (variables, $p < n$), and AM-FM features in X can be highly correlated. Thus, $X^T X$ can be singular or nearly singular and a unique solution to the normal equations could not exist. PLS reduces X to a lower dimensional subspace ($k \ll p$, where k represent the number of factors used). The first step is to factor X as $X = TL$, where T is an orthogonal $n \times k$ matrix of T-scores and L is a $k \times p$ matrix of factor loadings. The T-scores matrix are used to find a threshold for classification as outlined in [13].

For the training stage, using all the images in the training database, we compute the T-scores matrices for each CoS. The dimensionality of the input matrix will be reduced from $N_{Tr} \times K$ to $N_{Tr} \times k$, with $k \leq K$. It is important to mention that the $\hat{y}$ produced for each input training image using the recursive method SIMPLS [12] will be different than the $\hat{y}$ value produced when the loading matrix L are used. Let's define the T-scores and the loading factors L computed using the recursive method SIMPLS as T_{SIMPLS} and L with $X \approx T_{SIMPLS}L$. If we compute the T-scores using L and X using regular matrix operations: $X \approx TL$ then $T_{\{\text{using}L\}} = T = \left(XL^T\right)\left(LL^T\right)^{-1}$, there are differences between the recursive T_{SIMPLS} and $T_{\{\text{using}L\}}$ computed using the loading factors. Thus, we use SIMPLS to compute the optimum number of factors k and the loading matrix L for each CoS. Then, we compute the T-scores by matrix multiplications using X and L.

We compute the optimum number of factor k based on the mean-squared errors (MSE) for the matrix of the extracted AM-FM features X and the MSE for the classification variables y.

Now that we have T-scores for each CoS, $T_i = \{T_1, T_2, \cdots, T_{39}\}$, they are combined in a new matrix M such that $M = [T_1 T_2 \cdots T_{39}]$. Then, we use PLS againbut as a classifier to solve the equation $y = M\beta$. Finally, we keep the regression parameters β for classification of new images. At this stage, we select a threshold to reach our target sensitivity and specificity values.

For the testing stage, the generalized inverse of L for each CoS is used first to reduce the input features. Then, the reduced features are combined to compute the estimated y value given the regression parameters β found during the training.

The final grade (normal or abnormal) per image is estimated using the threshold selected during the training stage. To produce the grade per patient, we use either an OR rule (patient is abnormal if either ayer is abnormal) between the 2 eyes, or an AVERAGE rule ($\hat{y}_{patient} = \frac{1}{2}\hat{y}_{LeftEye} + \frac{1}{2}\hat{y}_{RightEye}$) given the selected threshold.

3. RESULTS

3.1. Images Used

Digital fundus photographs were collected in three different DR screening centers: (i) Project HOPE in Albuquerque, NM, (ii) Communicare Clinics, and (iii) the Retina Institute of South Texas, in San Antonio, TX. Macula-centered (field 2), non-mydriatic images of both eyes were captured with a Canon CR1 Mark II camera. The field of view (FOV) of the images is 45°. The images were originally 4752×3168 pixels of resolution with 72 dots per inch (dpi). The images were resized to 2224×1888 pixels due to memory and processing time constraints. Ground truth was provided independently by two optometrists. Any discrepancies were adjudicated by a certified retinal grader. Cases without enough image quality for retinal evaluation were manually eliminated from our dataset.

A total of 776 images, corresponding to 388 patients, were collected. For the training, we used 544 images corresponding to 393 normals and 151 abnormals. Note that some patients might have early signs of DR only in one eye. For the testing, the distribution was 167 normals and 65 abnormals.

3.2. Parameters used in the DR system

We use a 5-scale filterbank as described in subsection 2.1. For the k-means clustering, we have used $K = 30$ clusters. This K value was selected empirically after experimenting with different number of clusters.

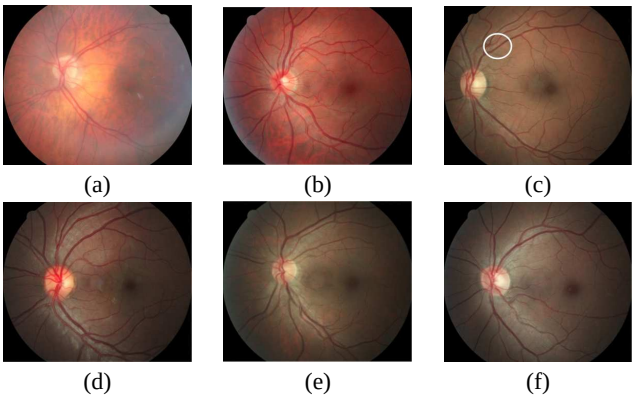

Fig. 2. Analysis of misclassified images during all tests. See text for the possible reasons for each failure.

The ROIs were selected to be square, non-overlapping regions of 140×140 pixels. We cover the full retina picture using a total of 202 ROIs per image.

To avoid the computation of a model which can be producing extreme results (either good or bad), we produce 20 different runs. For each run, we randomly change the training and testing sets. The constraints are that a patient must have his or her 2 photographs in the same set either training or testing and that the ratio of normal and abnormal cases in the training and testing sets remains constant. The original images were normalized to have mean intensity value equal to 75 in the green channel.

In Table 1 we present a summary of the results in terms of sensitivity/specificity (Sen/Spc) for the 20 models created. We show each individual result for each model. Also, we show the individuals area under the receiver operating characteristic (ROC) curve (AUC) for training and testing. The final Sen/Spc value is given by the average of the 20 models.

In Fig. 2 we present 6 images that were always misclassified under different tests for discussion purposes (next section).

4. DISCUSSION

AM-FM methods are robust under intensity variations in the IF features. However, IA estimates will be different under cases with similar FM textures but with higher or lower intensities. By normalizing them so that they have the same mean intensity values we reduced the effect of this problem.

This bring us to the discussion of the need for pre-processing methods that can normalize any input image under different characteristics. For example, images with non-constant illumination need to be normalized. The current problem of pre-processing methods is that they introduce noise or small image artifacts that can be confused with small lesions in a DR problem. We are currently working on developing pre-processing methods that can improve the quality of the image without the creation of artifacts.

In terms of the results from Table 1, we can see that $AUC_{Tr} \geq AUC_{Tt}$ (note that only in the models 8 and 15 $AUC_{Tr} = AUC_{Tt}$). When the T-scores from the recursive methods (like SIMPLS, see subsection 2.3) are used, the difference between AUC_{Tr} and AUC_{Tt} was bigger. This is due to the fact that optimum and unrealistic values that cannot be generalized to new sets when there are many variables such as high dimensionality and correlated data were computed. The use of the computed T-scores based on the loading

matrices L represent the behavior of new data with more generalization. Also, related with PLS, the number of factors k needs to as small as possible. Different error metrics can be used (we use the MSE) keeping in mind that the bigger the k the less generalization of the system.

We can see that there are models with very high Sen/Spc, for example, model 17 using the OR rule (92%/65%) or model 13 using the AVERAGE rule (92%/64%). Recall that the Sen/Spc values were calculated using a threshold from the training selected automatically. Given the high AUC values in the testing we can get different values for Sen/Spc according to the medical requirements. In this manuscript we show the results given the automatically selected threshold to analyze the generalization of the problem.

In terms of *failure analysis* (see Fig. 2), most of the misclassified images were at very early stages of DR. Fig. 2a presents some microaneurysms, hemorrhages and exudates out of the optic disc. However, the image is blurry, which reduces the quality of the image and smooths these lesions. Thus, for the system, this image was graded as normal. Fig. 2b presents some microaneurysms but the image is graded as normal due to the small size and quantities of this lesion. Fig. 2c presents some microaneurysms and hemorrhages indicated by the circle but the retinal is graded as normal since the hemorrhages are small and too close to the vessels. Fig. 2d and e presents the same lesions and problems of Fig. 2c. Finally, Fig. 2f is an image with some microaneurysms and some symptoms of drusen. The quantity of microaneurysms is small and the contrast of the drusen is low which helps explain why the system mis-classified these cases.

Table 1. Summary of the results in terms sensitivity (Sen.) and specificity (Spc.) in the testing results give a selected threshold from the training. The results are per patient given the OR rule or the AVERAGE rule (see subsection 2.3). We also show the individual AUC for each model for training (Tr) and testing (Tt).

	OR rule		AVERAGE rule		AUC	
Run #	**Sen.**	**Spc.**	**Sen.**	**Spc.**	**Tr.**	**Tt.**
1	0.83	0.60	0.78	0.73	0.92	0.79
2	0.89	0.51	0.89	0.63	0.94	0.77
3	0.89	0.76	0.78	0.89	0.92	0.87
4	0.86	0.49	0.81	0.75	0.91	0.81
5	0.92	0.66	0.72	0.78	0.91	0.82
6	0.86	0.51	0.78	0.78	0.91	0.83
7	0.92	0.51	0.83	0.60	0.93	0.77
8	0.89	0.80	0.78	0.90	0.87	0.87
9	0.86	0.45	0.81	0.60	0.94	0.76
10	0.94	0.54	0.81	0.65	0.91	0.81
11	0.83	0.46	0.72	0.69	0.92	0.72
12	0.89	0.68	0.72	0.74	0.92	0.82
13	0.89	0.53	0.92	0.64	0.91	0.81
14	0.92	0.59	0.83	0.73	0.88	0.82
15	0.83	0.75	0.81	0.89	0.87	0.87
16	0.92	0.59	0.81	0.69	0.89	0.80
17	0.92	0.65	0.89	0.78	0.91	0.84
18	0.86	0.69	0.75	0.76	0.89	0.84
19	0.89	0.50	0.86	0.73	0.89	0.84
20	0.92	0.61	0.86	0.68	0.89	0.81
Average	**0.89**	**0.59**	**0.81**	**0.73**	**0.91**	**0.81**
Minimum	0.83	0.45	0.72	0.60	0.87	0.72
Maximum	0.94	0.80	0.92	0.90	0.94	0.87
Standard Dev.	0.03	0.10	0.06	0.09	0.02	0.04
Median	0.89	0.59	0.81	0.73	0.91	0.82

5. CONCLUSIONS

We have presented an improved version of a DR screening system based on multiscale AM-FM features. We used fundus images collected from 338 patients from different screening centers. Given the estimated DR population, we need to get a performance of about Sen/Spc = 90%/60% for applying the system to population screening. Some of the models presented here produce results better than that requirement.

As future work, we will develop methods for reducing the misclassification of abnormal images due to the image quality of the fundus images (blurry, out of focus, under- or over-exposured, etc.) and due to some patterns such as choroidal vessels.

6. REFERENCES

[1] U.S. Department of Health and Human Services, "National diabetes statistics," February 2011, NIH Publication No. 11-3892.

[2] 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," *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 502–512, 2010.

[3] C. Agurto, S. Barriga, V. Murray, S. Murillo, G. Zamora, W. Bauman, M. Pattichis, and P. Soliz, "Toward comprehensive detection of sight threatening retinal disease using a multiscale AM-FM methodology," in *Medical Imaging 2011: Computer-Aided Diagnosis*, Ronald M. Summers and Bram van Ginneken, Eds. 2011, vol. 7963, p. 796316, SPIE.

[4] M. Niemeijer, M.D. Abramoff, N. Joshi, and M. Brady, "Comparison of classifier performance for information fusion in automated diabetic retinopathy screening," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, 30 2011-april 2 2011, pp. 685 –688.

[5] Keith Goatman, Amanda Charnley, Laura Webster, and Stephen Nussey, "Assessment of automated disease detection in diabetic retinopathy screening using two-field photography," *PLoS ONE*, vol. 6, no. 12, pp. e27524, 12 2011.

[6] Meindert Niemeijer, Bram van Ginneken, Stephen R. Russell, Maria S. A. Suttorp-Schulten, and Michael D. Abrmoff, "Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis," *Investigative Ophthalmology & Visual Science*, vol. 48, no. 5, pp. 2260–2267, 2007.

[7] Carla Agurto, E. Simon Barriga, Victor Murray, Sheila Nemeth, Robert Crammer, Wendall Bauman, Gilberto Zamora, Marios S. Pattichis, and Peter Soliz, "Automatic detection of diabetic retinopathy and age-related macular degeneration in digital fundus images," *Investigative Ophthalmology & Visual Science*, vol. 52, no. 8, pp. 5862–5871, 2011.

[8] V. Murray, P. Rodriguez, and M. Pattichis, "Multi-scale AM-FM demodulation and image reconstruction methods with improved accuracy," *IEEE Transactions on Image Processing*, vol. 19, no. 5, pp. 1138 –1152, May 2010.

[9] J. P. Havlicek, *AM-FM Image Models*, Ph.D. thesis, The University of Texas at Austin, 1996.

[10] Cesar Carranza, Victor Murray, Marios Pattichis, and E.S. Barriga, "Multiscale AM-FM decompositions with gpu acceleration for diabetic retinopathy screening," in *IEEE Southwest Symposium on Image Analysis and Interpretation*, 2012.

[11] Greg Hamerly and Charles Elkan, "Alternatives to the k-means algorithm that find better clusterings," in *Proceedings of the eleventh international conference on Information and knowledge management*, New York, NY, USA, 2002, CIKM '02, pp. 600–607, ACM.

[12] Sijmen de Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 251 – 263, 1993.

[13] Matthew Barker and William Rayens, "Partial least squares for discrimination," *Journal of Chemometrics*, vol. 17, no. 3, pp. 166–173, 2003.