35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Quality Control and Multi-lesion Detection in Automated Retinopathy Classification using a Visual Words Dictionary

Herbert F. Jelinek, *Member, IEEE*, Ramon Pires, *Member, IEEE*, Rafael Padilha, Siome Goldenstein, *Member, IEEE*, Jacques Wainer, and Anderson Rocha, *Member, IEEE*

***Abstract* — Automated identification of diabetic retinopathy has followed a number of paths. For accurate automated lesion detection, the first step is to identify images that are not suitable for automated assessment. This is followed by lesion or disease classification. For appropriate referral of individuals, the accurate recognition of a lesion and the extent of the lesion is important. The extent of a specific lesion and whether there are more than one type of lesion typify disease progression. In this sense, the visual word dictionary analysis allows automated image analysis where no pre-processing of images for specific lesions is required; training can be from different image resolutions, cameras and allows for some noise in the image. We demonstrate that the method when combined with machine learning allows feature fusion that incorporates identification of one or multiple lesions within one image at an appropriate accuracy level for clinical referral. In addition, the same general visual dictionary methodology can be applied to identify blurred images that are not suitable for automated assessment triggering alternative actions from the operator such as acquiring new images or referring to a qualified reader. For the detection of hard exudates, our approach has achieved an area under the curve (AUC) of 94.7%, while for superficial hemorrhages detection, the AUC achieved is 83.2%. The best performance reached for quality analysis is 87.4%.**

I. Introduction

Machine learning methods and automated data mining are important for health informatics and have been actively investigated in automated classification of disease, including diabetic retinopathy [1-5]. Quality control is an important part of automated image analysis [6, 7] as is the detection of multiple lesions in images of different resolution and ethnic background. This requires algorithms that unify image quality assessment and do not require preprocessing for each type of lesion separately, have a high accuracy for each type of lesion and, if possible, improve the classification when lesion types are combined in the classification framework. In this context, we have previously shown that visual word dictionaries have good accuracy with training of the classifier on different images to the test images and no preprocessing of the test images used in the research [8]. This paper now describes further developments using visual word dictionaries by considering a means of identifying poor quality images, identifying multiple lesions and detector fusion to optimize classification. Developing a framework that can identify discontinuities associated with retinal lesions and combining these results is an important prerequisite, as the number of visual word sets increases with the number of lesions and associated discontinuities present in the retinal image. Detector fusion has been applied in some areas of pattern recognition including multi-lesion detection associated with diabetic retinopathy (DR) [9-11].

The rest of the paper is organized as follows. Section II presents our methods of visual word dictionaries for detecting diabetic retinopathy lesions in images as well as its adaptation to determine the quality of an input image. Section III presents the results for the proposed approach both in terms of lesion detection as well as image quality analysis. Finally, Section IV concludes the paper and discusses the next steps on our research.

II. Methods

In this paper, we have two main contributions: first, we present the visual words dictionary approach for detecting diabetic retinopathy lesions in fundus images. After the detection using one or more detectors, we present a simple approach for combining such detectors to yield a final answer whether or not diabetic retinopathy is present in an image. The visual words methodology for detecting different lesions can be retained with only the set of training images associated with each specific lesion being selected. The second contribution of this paper is the adaptation of the visual words dictionary methodology to classify whether or not an input image meets the quality standard required for automatic assessment. Although image quality analysis can have innumerable ramifications before arbitrating on the quality of an image, in this paper we focus on a very common problem during image acquisition: blurring.

First, we will describe the general methodology for detecting multiple lesions, one at a time, and for combining such detectors towards deciding the diagnosis for an image.

A. Lesion Detectors and Fusion

In the last few years, the computer vision literature has reported successfully using invariant points for image matching, and for panorama creation as they are good reference points to match image portions with very similar properties. For image registration purposes, each image is characterized by finding stable points across multiple scales that capture image discontinuities. Points that are consistent, among other properties, across scales are selected and called points of interest (PoI). However, in retinal image analysis, matching images to each other is not the primary target. Instead, we are interested in characterizing an image in order to capture any inconsistencies/discontinuities it might have (e.g., lesion) in order to classify the image correctly. In this sense, exact matching techniques are not readily suitable. For automated retinal image analysis, the points of interest in a set of training images, that is the points that are most

H.F. Jelinek is with the Centre of Research of Complex Systems, Charles Sturt University, Albury, Australia (phone: +61-2-60519219; fax: +61-2-60519219; e-mail: hjelinek@csu.edu.au).

The other authors are with the Institute of Computing, University of Campinas, Campinas, SP, Brazil. Contact A. Rocha (phone: +55-19-3521-5854; e-mail: anderson.rocha@ic.unicamp.br).

important to the image classification problem, are determined. This approach is known as visual dictionaries, which is a robust representation methodology as each image is represented by a collection of regions.

To build a visual dictionary and detect a specific retinal lesion, training images classified as normal (no lesions) by a specialist as well images associated with the lesion of interest (e.g., hard exudates) are required. Regions of interest are marked on the training images by the specialist, which contain the lesion(s) rather than identifying the exact location and extent of the lesion.

After collecting the training images, the next step consists of finding the PoIs in all training images within the regions highlighted by the specialist. At this point, there are several choices as to which feature detector to use. Here we describe the Speeded Up Robust Features (SURF) approach [12] as it is a good feature detector with reasonable speed. SURF is a detection and description algorithm which is based on an approximation of the Hessian matrix. The descriptor uses a distribution of Haar-wavelet components based on the image intensity pattern within the neighborhood of the points of interest. The standard descriptor vector has a length of 64 floating-point numbers but other configurations are possible if the number of basic orientations is changed. In our experiments, we used 128 orientations. During training, the regions of interest marked by specialist are used to filter the points of interest that will be evaluated further. All other points not within the regions of interest will be discarded. For a normal image during training, however, all points of interest are retained since the entire retinal view is considered a region of interest. Figure 1 depicts these first steps for one image with hard exudates and superficial hemorrhages.

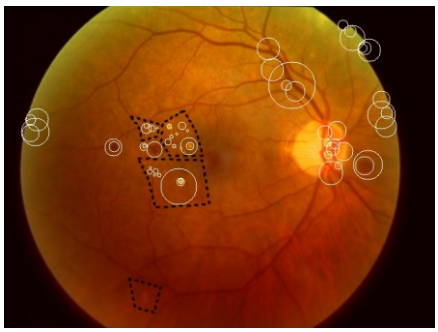

Figure 1. An input image, its points of interest as found by SURF (white circles) and the specialist selection of important regions of interest (dashed black regions).

From the points of interest obtained during training and representing the images with lesion(s) as well as images with no lesions, a clustering approach is applied to select the most representative PoIs for each group. At this stage, the number of PoIs (k) to be retained as representative of the lesion is decided. First, we use K-Means to find the $k/2$ most representative points of interest associated with a normal image. Then, the process is repeated to find the most representative points associated with images with the specific type of lesion of interest. At this point, we have $k/2 + k/2 = k$ representative points of interest for the classification problem. These k points of interest form the visual dictionary.

In order to use any machine learning method, the next stage is to map the points of interest within each image to the k most representative points in the dictionary. Hence, for each point in the training image, the closest PoIs to the most representative PoI within the k dictionary is mapped. Finally, each image is represented by a histogram of size k bins, with each bin representing the number of times one representative PoI is associated with the best PoI in the image. For instance, an image with 10 PoIs and all of them close to the first representative out of the k calculated PoIs will have its first bin equal to 10 and all other $k-1$ bins equal to zero in the histogram. This step provides feature vector histograms. Using the feature vectors of the normal and pathology images of interest, a machine learning classifier can be trained to find the classification model able to classify a new input image as normal or having a specific lesion.

For classifying a new input image, its SURF points of interest are identified and each PoI mapped onto its closest (in Euclidean distance) word in the pre-computed dictionary, resulting in the feature vector of the new image. Then, the machine learning classifier will be fed with the feature vectors to classify the new image into having (yes) or not having (no) the lesion. In the final classification, the Support Vector Machines algorithm was applied, which is a two-class machine learning classifier [13]. The kernel type explored was the radial basis function. This procedure is repeated for every lesion of interest resulting in a different yes/no answer for each kind of lesion.

According to this methodology, the only aspect that changes from one type of lesion to another is the training images associated with the lesion(s). The normal images can be the same. In this paper, we demonstrate this methodology applied to detection of hard exudates and superficial hemorrhages and also their fusion using a simple logical OR.

B. Quality selection

The same general methodology described in Section II-A for detection of lesions can be used to evaluate whether or not an image is adequate for automated analysis. In this case, before any action, the image quality needs to be assessed.

Among all types of problems associated with the image acquisition process, one of particular interest is the detection of blurred images. This paper, as a preliminary study of visual dictionaries for image quality analysis, focuses on classifying the quality of an image based on blurring.

For this intent, the general visual words methodology needs to be adapted with two small differences. The first difference is in the set of training images. In the case of image quality analysis for automated assessment all that is required is a set of images considered as good quality and a set of images considered low quality. The detection of low quality images using visual words is based on blurred images normally having less high-frequency information than good quality images. In particular, for retinal images, high-frequency information is more pronounced in the border regions associated with the venous branching pattern.

To capture behavior such as blurring, the edge map of each training image is first calculated using the Canny algorithm [14]. Next, the representative patches for the image are centered using the edge map. Fifty non-overlapping patches (each one with 50x50 pixels) in the edge map are centered in order to capture the differences of such regions.

We analyzed several sizes and quantity of patches and noted that 50 patches of 50*x*50 pixels were satisfactory to cover the edges of the blood vessels. The use of patches is the second difference with respect to the general methodology described in Section II-A. SURF is therefore not used directly on the image, rather it is directed to regions on the edge map that are more important to differentiate blur and non-blur artifacts, namely regions with edges.

After calculating the points of interest within the selected 50 regions, the most representative PoIs, according to Section II-A, have to be found for each training image. For that, K-Means is applied to select a specialized visual dictionary for image quality analysis. In this case, are selected *k/2* regions that represent good quality images and *k/2* regions for low quality images. Figure 2 depicts an example of a retinal image and its Canny edge map with the 50 patches centered on the localized edges. After generating each image feature vector, it is normalized using the traditional term-frequency (divide the entries by total sum of the bins).

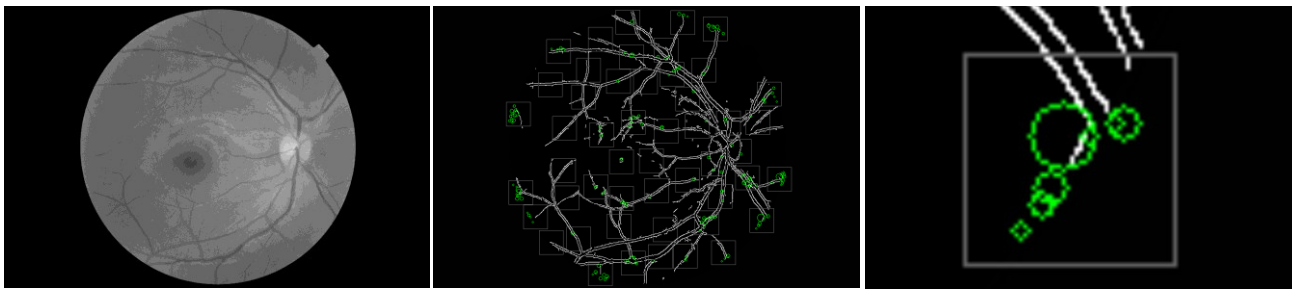

Figure 2. Input image with its Canny edge map as well as the 50 50*x*50-selected image regions centered on the edges (small squared regions) and the calculated SURF PoIs within each region (green circles), followed by a highlighted patch.

III. RESULTS

This section shows the results for evaluating the quality of an image for automatic screening and also for detecting the presence/absence of multiple diabetic retinopathy lesions in fundus images. All the experiments reported consider a 5-fold cross-validation protocol in which the data set is divided into five parts, train with four parts and test on the fifth, repeating the process five times each time changing the training and test sets.

A. Data sets

All experiments for lesion detection and fusion were conducted using the DR1 dataset from the Ophthalmology Department of the Federal University of São Paulo, collected during 2010. The DR1 dataset comprises 1,014 images with an average resolution of 640 × 480 pixels, (687 are normal retinas, 245 images contain bright lesions, 191 contain red lesions and 109 contain signs of both bright and red lesions). Three medical specialists manually annotated all the images in this dataset. The images were captured using a TRC50X (Topcon Inc., Tokyo, Japan) mydriatic camera with maximum resolution of one megapixel and a field of view of 45 degrees. The experiments for quality analysis were conducted on the more recent DR2 dataset for which we have quality assessment grading performed by one medical specialist. DR2 comprises 660 12.2MP images decimated to 867 × 575 for speed purposes divided into 466 good and 194 low quality images captured using a TRC-NW8 mydriatic camera with a D90 camera for image capture. For more details and for downloading both data sets, please refer to http://www.recod.ic.unicamp.br/site/asdr.

B. Image Quality

Figure 3 depicts the results for image quality analysis. In this case, a good-quality image is one with no blurring.

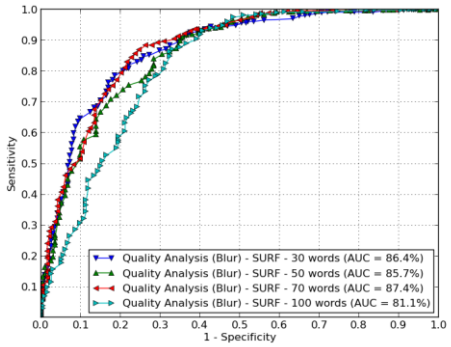

Figure 3. Image quality analysis considering 50, 50*x*50-regions per image from DR2 data set and various dictionary (most representative regions) sizes. Note that the dictionary needs 70 words for a reasonable performance resulting in an AUC of 87.4%, in this case. For a dictionary with 30 words, the AUC is 86.4% while for 50 words the AUC is 85.7% and for 100 it is 81.1%. These are promising results for a first attempt for solving image quality assessment.

C. Multi-lesion detection

After assessing the quality of an image, images of good quality can be considered for lesion analysis. In this paper, however, we decided to validate the two contributions separately since our approach for quality analysis is still preliminary. Therefore, the methodology we presented in Section II-A to detect hard exudates and superficial hemorrhages with no pre-analysis regarding image quality is validated.

Figure 4 shows the ROC classification results for images with hard exudates and superficial hemorrhages. The best-performing dictionary size for hard exudates consisted of 500 words (250 for hard exudates plus 250 normal-based words) with a corresponding AUC = 94.7%. The best-performing visual dictionary for superficial hemorrhages also consisted of 500 words (250 for superficial hemorrhage and 250 normal-based words) with a corresponding AUC = 83.2%.

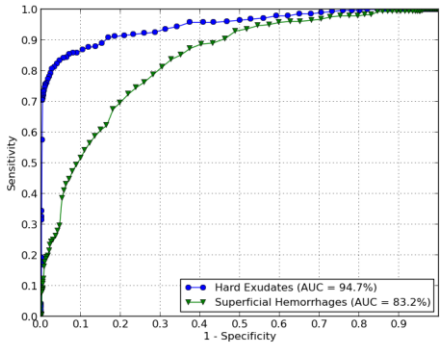

Figure 4. ROC results for detecting hard exudates (HE) and superficial hemorrhages (SH) considering the DR1 data set and a 5-fold cross-validation protocol. The area under the curve for HEs is approximately 94.7% while for SH is 83.2%.

D. Detector fusion

The most interesting aspect of the visual word methodology for assessing image quality and for lesion detection is in its unified modus operandi.

This unification allows easy implementation of a combination method for the different lesion detectors towards a final classification of an unknown input image.

Using a simple OR rule on top of the results in the previous section for a specific point of the curve, namely specificity = 90% for both lesions, a final sensitivity of 93% was obtained. Note, however, the data set considered herein has images that are either normal or have at least one lesion.

IV. Discussion and Conclusion

Many feature descriptors have been proposed in the literature: Gaussian derivatives [15], complex features [16], SIFT [17], and SURF [12]. Such methods need to capture sufficient image details, whilst being robust to small deformations or localization errors [12]. Using the Hessian approximation within the visual word dictionary framework is comparable to and, in some instances, better than current state-of-the-art interest point detectors. SURF´s advantage relies on its robustness against rotation, scale change, image noise, change in brightness across the image and change of view being suitable for adaptation for a classification framework instead of its usual image matching form.

Current microaneurysm detectors in theory should detect one MA if it is present but may not due to image quality or closeness to a blood vessel as well as differences in retinal background [18, 19]. In addition, when classifying disease presence the detection sensitivity for microaneurysms is usually turned down to reduce the number of false positives (classifying into disease present/disease absent).

In essence, the visual words based detector indicates which lesion is present regardless of image quality and with no pre- or post-processing required. In addition, the visual words dictionary is able to identify multiple lesions in an image, which can be easily combined with a logical OR as all detectors have a unified framework. The results shown here improve on previous solutions, which are based on correctly identifying a lesion such as a microaneurysm and its correct location. These methods cannot identify multiple lesions unless a number of machine learning algorithms are combined with specific preprocessing for each type of lesion and classification algorithms.

Finally, the general visual words methodology used to pinpoint the presence of specific lesions in fundus images can also provide information on image quality for automatic screening. The initial results discussed here are promising.

Acknowledgments

We thank Dr. Eduardo Dib for technical assistance with image acquisition. We also thank the financial support of FAPESP, CNPq, CAPES, and Microsoft.

References

[1] G. Gardner, D. Keating, T. Williamson, and A. Elliott, "Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool," *British Journal of Ophthalmology,* vol. 80, pp. 940-944, 1996.

[2] M. H. Goldbaum, P. A. Sample, K. Chan, J. Williams, T.-W. Lee, E. Blumenthal, C. A. Girkin, L. M. Zangwill, C. Bowd, T. Sejnowski, and R. N. Weinreb, "Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry," in IOVS*,* vol. 43, pp. 162-169, 2002.

[3] D. J. Cornforth, H. F. Jelinek, M. C. Teich, and S. B. Lowen, "Wrapper subset evaluation facilitates the automated detection of diabetes from heart rate variability measures," in *CIMCA,* 2004, pp. 446-455.

[4] J. V. B. Soares, J. J. G. Leandro, R. M. Cesar-Jr, H. F. Jelinek, and M. J. Cree, "Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification," *IEEE TMI,* vol. 25, pp. 1214-1222, 2006.

[5] G. Quellec, M. Lamard, M. D. Abràmoff, E. Decencière, B. Lay, A. Erginay, B. Cochener, and G. Cazuguel, "A multiple-instance learning framework for diabetic retinopathy screening," in Medical Image Analysis, vol. 16, no. 6, pp. 1228-1240, 2012.

[6] A. D. Fleming, S. Philip, K. A. Goatman, J. A. Olson, and P. F. Sharp, "Automated Assessment of Diabetic Retinal Image Quality Based on Clarity and Field Definition," in *IOVS,* vol. 47, pp. 1120-1125, 2006.

[7] R. Pires, H. F. Jelinek, J. Wainer, and A. Rocha, "Retinal Image Quality Analysis for Automatic Diabetic Retinopathy Detection," in SIBGRAPI, 2012, pp. 1-8.

[8] H. F. Jelinek, A. Rocha, T. Carvalho, S. Goldenstein, and J. Wainer, "Machine learning and pattern classification in identification of indigenous retinal pathology," in *IEEE EMBS,* 2011, pp. 5951-5954.

[9] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade Object Detection with Deformable Part Models," in *IEEE CVPR*, 2010, pp. 2241-2248.

[10] U. R. Acharya, C. M. Lim, E. Y. K. Ng, C. Chee, and T. Tamura, "Computer-based detection of diabetes retinopathy stages using digital fundus images," *Journal of Engineering in Medicine,* pp. 545-553, 2009.

[11] H. F. Jelinek, R. Pires, R. Padilha, S. Goldenstein, J. Wainer, T. Bossomaier, and A. Rocha, "Data fusion for multi-lesion diabetic retinopathy detection," in IEEE EMBS, 2012, pp. 1-4.

[12] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," *in ECCV,* pp. 404-417, 2006.

[13] M. Bishop, "Pattern Recognition and Machine Learning, 1st ed.," 2006.

[14] R. Gonzalez and R. Woods, "Digital Image Processing, 3rd ed.," 2007.

[15] L. M. J. Florack, B. M. ter Haar Romeny, J. J. Koenderink, and M. A. Viergever, "General intensity transformations and differential invariants," in *JMIV*. vol. 4, 1994, pp. 171–187.

[16] A. Baumberg, "Reliable feature matching across widely separated views," in IEEE *CVPR*, 2000, pp. 774-781.

[17] D. Lowe, "Distinctive image features from scale-invariant keypoints, cascade filtering approach," *IJCV,* vol. 60, pp. 91-110, 2004.

[18] M. J. Cree, E. Gamble, and D. J. Cornforth, "Colour normalisation to reduce inter-patient and intra-patient variability in microaneurysm detection in colour retinal images," in *Workshop on Digital Image Computing*, 2005, pp. 163-169.

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