Automatic Diabetic Retinopathy Detection using BossaNova Representation*

Ramon Pires[1], Sandra Avila[2], Herbert F. Jelinek[3], Jacques Wainer[1], Eduardo Valle[2], and Anderson Rocha[1]

***Abstract*— The biomedical community has shown a continued interest in automated detection of Diabetic Retinopathy (DR), with new imaging techniques, evolving diagnostic criteria, and advancing computing methods. Existing state of the art for detecting DR-related lesions tends to emphasize different, specific approaches for each type of lesion. However, recent research has aimed at general frameworks adaptable for large classes of lesions. In this paper, we follow this latter trend by exploring a very flexible framework, based upon two-tiered feature extraction (low-level and mid-level) from images and Support Vector Machines. The main contribution of this work is the evaluation of *BossaNova*, a recent and powerful mid-level image characterization technique, which we contrast with previous art based upon classical Bag of Visual Words (BoVW). The new technique using BossaNova achieves a detection performance (measured by area under the curve — AUC) of 96.4% for hard exudates, and 93.5% for red lesions using a cross-dataset training/testing protocol.**

I. INTRODUCTION

Diabetic Retinopathy (DR) is the most common eye disease in American adults. It is estimated that 4.1 million persons age 40 and older in the US general population have DR and approximately 25,000 people go blind every year due to diabetic retinopathy [1], [2]. The accelerated increase of the number of people with diabetes complications creates the need for systems that screen larger sections of the community effectively for early signs of diabetic retinopathy.

Over the last three decades, several computer-aided image processing methods, computer vision and machine learning tools have been proposed in order to diagnose fundus images, increasing the efficiency of DR detection in clinical environments and also to reach remote and rural communities [3]–[8]. In general, the approaches based restrictively on classical image processing methods explore the characteristics of individual DR lesions in order to develop a detection method for the lesion of interest [3], [4]. This practice results in effective highly accurate outcomes, but is in general limited to the detection of single lesions only.

More recently, some techniques which do not require any pre- or post-processing of retinal images have been proposed [7], [8], and provided algorithms that are able to automatically detect individual DR lesions. These algorithms also have the advantage of being able to be easily adjusted to multiple DR anomalies and they rely on a characterization scheme known as Bags of Visual Words (BoVW). Methods for diabetic retinopathy progression analysis, which try not only to examine the presence/absence of DR lesions, but also to indicate the stage of retinopathy (mild, moderate or severe non-proliferative, or proliferative retinopathy), have also been proposed [9], [10]. Another useful and interesting topic related to automated DR detection and classification is the referral assessment [11].

This paper explores a new form of characterizing retinal images in order to obtain powerful lesion classifiers without the requirement of additional pre- or post-processing operations. We evaluate *BossaNova*, a powerful tool for mid-level feature characterization and image representation that allows us to create a highly discriminative two-tiered feature extraction (low- and mid-level) approach amenable for designing different lesion detectors with a unified technique.

II. METHODS

This section provides a description of the extraction of low-level local features from retinal images, the aggregation of those local features into mid-level BossaNova features, and then the classification of those BossaNova features by a Support Vector Machine (SVM) classifier [12].

A. Low-level Local Feature Extraction

Typically, local feature extraction includes two steps: feature detection and feature description. The former aims at finding a set of interest points, or salient regions in the image that are invariant to a range of image transformations. The latter step aims at obtaining robust local descriptors from the detected features. In this work, we extract Speeded-Up Robust Features (SURF) local descriptors [13].

Two types of local feature extraction can be distinguished [14]: (i) *sparse*, based upon the detection of salient regions or interest points, or (ii) *dense*, where patches of fixed size are placed on a regular grid over multiple scales. For sparse feature detection, the SURF [13] is used. SURF sensitivity parameters are pre-tuned to detect, on average, 400 interest points per retinal image.

For dense features, patches are selected on a dense grid using radii of 12, 19, 31, 50, 80, 128 pixels. These radii are used both as scale and as the vertical/horizontal sampling steps of the grid.

SURF is used to create a feature vector for each detected point of interest. The algorithm is parameterized to

*Funded by the Samsung, Microsoft Research, FAPESP, and CNPq.

[1]Institute of Computing, University of Campinas, Campinas, SP, Brazil, 13083-852. Corresponding author: Anderson Rocha, e-mail: anderson.rocha@ic.unicamp.br.

[2]School of Electrical and Computing Engineering, University of Campinas, Campinas 13083-852, Brazil

[3]Centre of Research in Complex Systems, Charles Sturt University, Albury, Australia, 2640 and also with the Australian School of Advanced Medicine, Macquarie University, North Ryde, N.S.W. 2113, Australia.

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

operate on twice the image resolution and to extract 128-dimensional extended feature vectors instead of the default 64-dimensional feature vectors.

Once extracted, these points need to be analyzed and filtered so as to select the ones most appropriate for detecting DR lesions. For that, we transform the low level features into mid-level ones creating a two-tiered representation scheme.

B. Mid-level Feature Extraction: BossaNova Representation

BossaNova is a recent mid-level representation for multi-tier ensemble data mining for image classification [15], [16].

Let $\mathscr{X}$ be an unordered set of local descriptors extracted from an image. $\mathscr{X} = \{\mathbf{x}_j\}$, $j \in \{1,\ldots,N\}$, where $\mathbf{x}_j \in \mathbb{R}^D$ is a local descriptor vector and N is the number of local descriptors of the image. Let $\mathscr{C}$ be a visual codebook obtained by the k-means algorithm (the codebook can be defined by the set of codewords corresponding to the centroids of clusters). $\mathscr{C} = \{\mathbf{c}_m\}$, $m \in \{1,\ldots,M\}$, where $\mathbf{c}_m \in \mathbb{R}^D$ is a codeword and M is the number of visual codewords. $\mathbf{z}$ is the final vectorial BossaNova representation of the image used for classification.

The BossaNova approach follows the BoVW formalism [17], but proposes an image representation which keeps more information than BoVW during the pooling step, by computing a histogram of distances between the descriptors found in the image and each codeword. Recall that in a BoVW model, pooling is the step responsible for aggregating different features activating the same visual word onto a final summarized feature vector.

The BossaNova pooling function g estimates the probability density function of α_m: $g(\alpha_m) = \mathrm{pdf}(\alpha_\mathbf{m})$, by computing the following histogram of distances $z_{m,b}$:

$$\begin{aligned} g : \mathbb{R}^N &\longrightarrow \mathbb{R}^B, \\ \alpha_\mathbf{m} &\longrightarrow g(\alpha_m) = z_m, \\ z_{m,b} &= \mathrm{card}\left(\mathbf{x}_j \mid \alpha_{m,j} \in \left[\frac{b}{B}; \frac{b+1}{B}\right]\right), \\ & \frac{b}{B} \geq \alpha_m^{min} \quad \text{and} \quad \frac{b+1}{B} \leq \alpha_m^{max}, \end{aligned} \tag{1}$$

where B denotes the number of bins of each histogram z_m, $\alpha_{m,j}$ represents a dissimilarity (i.e., a distance) between $\mathbf{c}_m$ and $\mathbf{x}_j$, and $[\alpha_m^{min}; \alpha_m^{max}]$ limits the range of distances for the descriptors considered in the histogram computation.

After computing the local histograms z_m for all the $\mathbf{c}_m$ centers, the BossaNova vector $\mathbf{z}$ [16] can be written as:

$$\mathbf{z} = \left[\left[z_{m,b}\right], st_m\right]^{\mathrm{T}}, \quad (m,b) \in \{1,...,M\} \times \{1,...,B\}, \tag{2}$$

where $\mathbf{z}$ is a vector of size $M \times (B+1)$, s is a nonnegative constant and t_m is a scalar value for each codeword, counting the number of local descriptors $\mathbf{x}_j$ close to that codeword.

In summary, by using a histogram of distances to capture the relevant information, the BossaNova approach remains very flexible and keeps the representation compact. In comparison to the BoVW representation, BossaNova significantly outperforms BoVW on many challenging image classification benchmarks [16]. Considering those results, we chose the BossaNova approach for mid-level features given that it takes into account some spatial relationship between features which we believe would be important for DR lesions.

In our experiments, we kept the default BossaNova parameter values the same as in [16] ($B = 2$, $\alpha_m^{min} = 0.4\sigma_m$, $\alpha_m^{max} = 2.0\sigma_m$, $s = 10^{-3}$), except for the number of visual codewords M, where we considered $\{1{,}000,\ 4{,}000\}$.

C. Class-based Scheme vs. Global Dictionary

In our previous work, we have used the *class-based* scheme for image representation, which performs well for retinal images. The class-based scheme, proposed by Rocha et al. [7], creates two independent codebooks, one from descriptors extracted from retinal images with the lesion present, and one from descriptors extracted from images of healthy retinas. Then, two independent k-means clustering methods are performed, each with k corresponding to half the size of the desired codebook. After the clustering process, the two sets of centroids are concatenated, generating a codebook of the desired size.

This class-based scheme is compared to the global dictionary scheme when applying the BossaNova approach in which the clustering is performed only once with the desired codebook size.

III. RESULTS

All the experiments were performed using a cross-dataset protocol, which consists of distinct training and test datasets preferably collected in very different settings with different cameras. The cross-dataset protocol is an important precaution of the design, since in clinical practice the images that need to be classified have rarely the same image specification (camera, resolution, FOV) than the images used for training.

To quantify performance as a single scalar, all the results are reported as the area under the receiver operating characteristic curve (AUC-ROC). The AUC gives a better overall performance measure than any particular point of the specificity-sensitivity metrics.

A. Datasets

The experiments were performed using two different and freely-available[1] retinal image datasets:

- **DR1 dataset**, provided by the Federal University of São Paulo. Each image was manually annotated by three medical specialists. The images were captured using a TRC-50X (Topcon Inc., Tokyo, Japan) mydriatic camera with maximum resolution of one megapixel (640 × 480 pixels) and a field of view (FOV) of 45°. The DR1 dataset comprises 1,077 images, of which 595 are normal (no lesions) and 482 contain lesions (234 contain hard exudates, and 180 contain red lesions)[2].
- **DR2 dataset**, also provided by the Federal University of São Paulo (Unifesp). Each image was manually annotated by two medical specialists, who did not work on

[1] Available at http://dx.doi.org/10.6084/m9.figshare.953671.

[2] The number of images with lesions do not match the total because, in this paper, we considered only two types of lesions.

the DR1 dataset. The dataset was captured using a TRC-NW8 retinograph with a Nikon D90 camera, creating 12.2 megapixel images, which were then reduced to 867 $\times$ 575 pixels for accelerating computation. The DR2 dataset comprises 520 images, of which 300 are normal and 220 contain lesions (79 contain hard exudates, and 98 contain red lesions)[2].

The DR1 dataset was used as training set, whilst the DR2 dataset was used as the test set in our experiments.

B. Diabetic Retinopathy Lesion Detection

In this paper, we demonstrate the methodology described in Sec. II for the detection of hard exudates and red lesions. The experiments were performed with two distinct codebook sizes: 1,000 and 4,000. Note that, in this paper, we use larger codebooks than in previous work. That is why, for BossaNova, larger codebooks lead to higher accuracy [16]. For BoVW, however, small codebooks have provided more satisfactory results for DR detection [7].

Fig. 1 shows the ROC curves with their respective AUCs for the detection of hard exudates employing the class-based and the global dictionary approaches. For comparison purposes, Fig. 1 presents also the results reached in [11] using the same datasets but describing the retinal images with BoVW. Initial results using the sparse low-level feature extraction but employing larger codebooks, showed that the BossaNova provides a better representation and a better accuracy. The proposed new method provides an AUC of 96.0% with a global dictionary and 95.9% with the class-based approach, compared to an AUC of 95.6% obtained by the BoVW with soft-max coding/pooling. When the dense extraction step is used, the difference between BossaNova and BoVW increases: the best result was obtained using class-based scheme with 4,000 codewords, with an AUC of 96.4% compared to the BoVW with an AUC of 95.6%, which represents an error reduction of over 18%.

For the detection of red lesions, we present the ROC curves and their respective AUCs in Fig. 1. BossaNova with the sparse low-level feature extraction technique did not provide a significant advantage over the Bag of Visual Words approach based on the hard-sum coding/pooling. AUCs of 91.9% and 92.4% were achieved using the class-based scheme and a global dictionary, respectively, against 92.3% obtained with BoVW [11]. However, once again, the dense extraction shows its superiority when applied with the BossaNova mid-level feature extraction, presenting an AUC of 93.5% using a codebook of size 4,000 in the class-based scheme, compared to the BoVW with an AUC of 92.3%, which represents an error reduction of over 15%.

IV. CONCLUSIONS

The increasing prevalence of diabetes and the subsequent rise in the number of diabetes-related complications are a considerable challenge for health care. The lack of specialists for classification of diabetic retinopathy has stimulated the scientific community to investigate and develop more effective solutions to the screening of diabetic retinopathy.

In previous work, we have explored machine learning and computer vision methods for the development of a framework that examines retinal image quality [18], detects DR-related lesions [8], [11], and assesses the need for referral within one year of screening [11]. For diabetic retinopathy detection, we have shown that Bag of Visual Words methodology has performed very well, providing results comparable to the literature and achieving promising outcomes for some hard-to-detect lesions such as drusen and cotton-wool spots (see [11] for more details). One important strategy of our current design is the use of the class-based scheme, that consists of creating not only a unique codebook, but one codebook per class as this methodology performs well for two-class problems (e.g., normal/abnormal).

In the current paper, we explored the BossaNova [15], a new mid-level feature extraction technique that consists of an improvement in the pooling stage. Due to the use of a new representation method, we considered again the global dictionary scheme for comparison purposes.

Both for hard exudates and for red lesions detection, the results outperformed previous methods showing the importance of preserving some relationship between the detected features in retinal images instead of just throwing them away as previous BoVW-based solutions have done.

As the proposed two-tier BossaNova mid-level representation presented interesting results for both white and red lesions with a unified framework, as future work we will investigate this methodology for detecting other lesions such as drusen and cotton-wool spots. We also propose a deeper analysis on key aspects of the parametric space of BossaNova (e.g., number of bins B and range of distances $[\alpha_m^{min}; \alpha_m^{max}]$) in the visual recognition task. Another future work we are interested in consists of combining a coding technique that has shown to be more suitable for DR detection (semi-soft) [19] with the BossaNova pooling approach.

REFERENCES

[1] E. D. P. R. Group, "The prevalence of diabetic retinopathy among adults in the united states," *Archives of Ophthalmology*, vol. 122, no. 4, pp. 552–563, 2004.

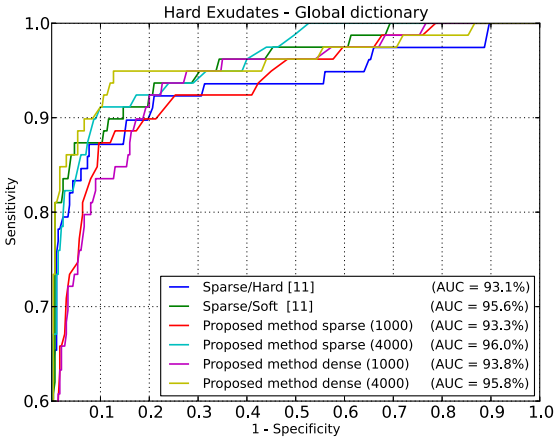
[2] M. D. Abràmoff, M. Niemeijer, M. S. Suttorp-Schulten, M. A. Viergever, S. R. Russell, and B. van Ginneken, "Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes," *Diabetes care*, vol. 31, no. 2, pp. 193–198,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

[3] C. Sinthanayothin, J. Boyce, T. Williamson, H. Cook, E. Mensah, S. Lal, and D. Usher, "Automated detection of diabetic retinopathy on digital fundus images," *Diabetic Medicine*, vol. 19, no. 2, pp. 105–112, 2002.

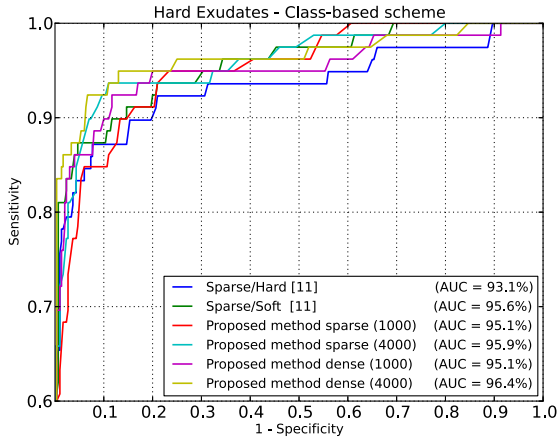
[4] M. Niemeijer, B. van Ginneken, S. R. Russell, M. S. Suttorp-Schulten, and M. D. Abràmoff, "Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis," *Investigative ophthalmology & visual science*, vol. 48, no. 5, pp. 2260–2267, 2007.

[5] E. S. Barriga, V. Murray, C. Agurto, M. Pattichis, W. Bauman, G. Zamora, and P. Soliz, "Automatic system for diabetic retinopathy screening based on am-fm, partial least squares, and support vector machines," in *International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1349–1352, IEEE, 2010.

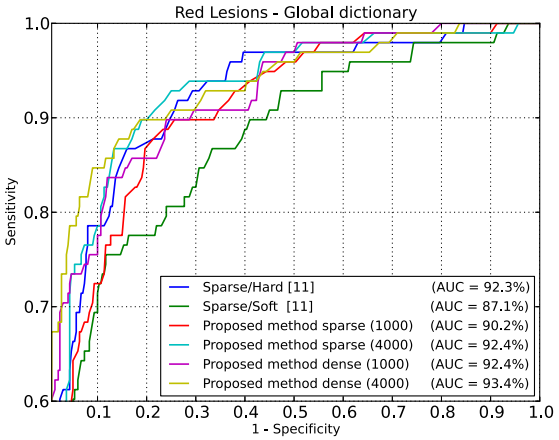
[6] C. Agurto, E. S. Barriga, V. Murray, S. Nemeth, R. Crammer, W. Bauman, G. Zamora, M. S. Pattichis, and P. Soliz, "Automatic detection of diabetic retinopathy and age-related macular degeneration in digital fundus images," *Investigative ophthalmology & visual science*, vol. 52, no. 8, pp. 5862–5871, 2011.



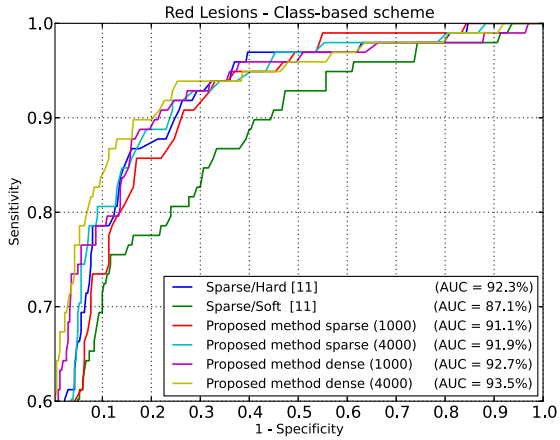
(a) All mid-level features built on global codebooks



(b) All mid-level features built on class-based sampling of codebooks



(c) All mid-level features built on global codebooks



(d) All mid-level features built on class-based sampling of codebooks

Fig. 1. ROC results for hard exudates and red lesions detection using in the global codebooks (a) (c) and class-based sampling (b) (d), for codebooks of sizes 1,000 and 4,000. The results are contrasted with those obtained in [11] using the BoVW approach with 500 codewords for sparse low-level technique and class-based scheme (the mismatch in codebook size is due to the fact the previous art performed better with smaller codebooks [7]). As the work [11] already comprises the best results reported in [7], we opted for not repeating them here. The best configuration is the new technique, using BossaNova, dense low-level features, and large codebooks (4,000) with class-based sampling. For hard exudates and red lesions, the error reduction compared to the prior art is over 15% and 18%, respectively, for the best proposed method.

[7] A. Rocha, T. Carvalho, H. F. Jelinek, S. Goldenstein, and J. Wainer, "Points of interest and visual dictionaries for automatic retinal lesion detection," *Transactions on Biomedical Engineering*, vol. 59, no. 8, pp. 2244–2253, 2012.

[8] H. F. Jelinek, R. Pires, R. Padilha, S. Goldenstein, J. Wainer, and A. Rocha, "Quality control and multi-lesion detection in automated retinopathy classification using a visual words dictionary," in *IEEE Engineering in Medicine and Biology Society*, vol. 2013, pp. 5857–5860, 2013.

[9] J. Nayak, P. S. Bhat, R. Acharya, C. Lim, and M. Kagathi, "Automated identification of diabetic retinopathy stages using digital fundus images," *Journal of Medical Systems*, vol. 32, no. 2, pp. 107–115, 2008.

[10] H. F. Jelinek, K. Al-Saedi, and L. B. Bäcklund, "Computer assisted top-down assesment of diabetic retinopathy," in *World Congress on Medical Physics and Biomedical Engineering*, pp. 127–130, 2009.

[11] R. Pires, H. Jelinek, J. Wainer, S. Goldenstein, E. Valle, and A. Rocha, "Assessing the need for referral in automatic diabetic retinopathy detection," *Transactions on Biomedical Engineering*, vol. 60, no. 12, pp. 3391–3398, 2013.

[12] V. Vapnik, *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

[13] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[14] T. Tuytelaars, "Dense interest points," in *Computer Vision and Pattern Recognition*, pp. 2281–2288, 2010.

[15] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A Araújo, "Bossa: Extended bow formalism for image classification," in *International Conference on Image Processing*, pp. 2909–2912, IEEE, 2011.

[16] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A Araújo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 453–465, 2013.

[17] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, vol. 2, 2003.

[18] R. Pires, H. F. Jelinek, J. Wainer, and A. Rocha, "Retinal image quality analysis for automatic diabetic retinopathy detection," in *Conference on Graphics, Patterns and Images*, pp. 229–236, IEEE, 2012.

[19] R. Pires, H. F. Jelinek, J. Wainer, E. Valle, and A. Rocha, "Advancing bag-of-visual-words representations for lesion classification in retinal images," *PLoS ONE*, vol. 9, p. e96814, 06 2014.