Beyond Lesion-based Diabetic Retinopathy: a Direct Approach for Referral

Ramon Pires, *Member, IEEE,* Sandra Avila, *Member, IEEE,* Herbert F. Jelinek, *Member, IEEE,* Jacques Wainer, Eduardo Valle, and Anderson Rocha, *Senior Member, IEEE*

***Abstract*—Diabetic retinopathy (DR) is the leading cause of blindness in adults, but can be managed if detected early. Automated DR screening helps by indicating which patients should be referred to the doctor. However, current techniques of automated screening still depend too much on the detection of individual lesions. In this work we bypass lesion detection, and directly train a classifier for DR referral. Additional novelties are the use of state-of-the-art mid-level features for the retinal images: BossaNova and Fisher Vector. Those features extend the classical Bags of Visual Words and greatly improve the accuracy of complex classification tasks. The proposed technique for direct referral is promising, achieving an area under the curve (AUC) of 96.4%, thus reducing the classification error by almost 40% over the current state of the art, held by lesion-based techniques.**

***Index Terms*—Diabetic Retinopathy, Referral, Referability, Direct Referral, Bag of Visual Words, BossaNova, Fisher Vector.**

I. Introduction

DIABETIC Retinopathy (DR) is the leading cause of blindness in adults worldwide, with 93 million people affected in 2010 [1]. Just considering 40+ year-old Americans, we have 7.7 million people with DR [2], most of which in counties with a shortage of ophthalmologists and optometrists [3].

Optimal treatment of DR requires early diagnostic and, thus, regular eye examinations. In practice, however, many communities cannot offer the frequent consultations and continuous follow-up required for screening [4]. Therefore, DR management requires reaching underserviced populations. Recent research addresses the issue with Computer-Aided Diagnosis [5]–[14]. Most existing art focuses on the detection of DR lesions using visual characteristics specific to each type of lesion [5]–[10]. More recently, a few unified DR-lesion detectors have been proposed [14]–[20].

It is much harder (and polemic) to decide automatically to refer or not the patient to the ophthalmologist [18], [20]–[22]. Automated referral is a hot topic, because DR risk assessment is complex, based not only on the presence of lesions and their evolution, but also on subtle hints revealed during examination and anamnesis. We argue, however, that automated or semi-automated decision of referable cases can have a huge impact on the management of care, reducing the specialist's workload while still attending to the patients in need. While agreeing that face-to-face consultations with a specialist are always ideal, we stress that many communities simply lack the luxury of offering them to every suspect case. For poor, isolated, or rural communities, it is critical to prioritize the high-risk cases, in order to better serve the patients [23].

Currently, automated referral decisions combine separate DR-lesion classifiers into a final decision. Those models are complex, cumbersome to implement, and often have limited accuracy. We take a different approach, with an effective method for directly assessing the referability of patients. That method is based upon the Bags of Visual Words (BoVW) model [24] and maximum-margin support vector machine (SVM) classifiers. We also improve the referral assessment with advanced mid-level features (BossaNova [25] and Fisher Vector [26]) that considerably extend the BoVW model. Because image classification is very sensitive to the choice of parameters, we employ a rigorous experimental design to investigate the significance of our choices.

We have organized the remainder of this paper into six sections. In Section II, we describe related work. In Section III, we overview the lesion-based referable methodology, while in Section IV we explain our methodology for direct automated referral decision. In Sections V and VI, we present, respectively, the experimental protocol and the results for the proposed method, using the 5×2-fold cross-validation protocol. Finally, in Section VII, we conclude the paper and discuss future possibilities.

II. State of the Art

The simple presence of DR lesions is insufficient to warrant referral. For instance, a small number of microaneurysms in a safe region of the retina might be considered mild nonproliferative DR, without need of referral. Much existing art addresses that problem [18], [20]–[22], [27], but still those approaches depend strongly on lesion detection, which is often very specific to each type of lesion. More recently, unified models for detecting any kind of DR lesion appeared [15]–[19], but even those require an extra fusion step to combine the lesion scores into a single referability decision.

R. Pires is with the Institute of Computing, University of Campinas (Unicamp), Campinas 13083-852, Brazil (e-mail pires.ramon@ic.unicamp.br).

S. Avila and E. Valle are with the School of Electrical and Computing Engineering, University of Campinas (UNICAMP), Campinas 13083-852, Brazil (e-mail sandra@dca.fee.unicamp.br; dovalle@dca.fee.unicamp.br).

H. F. Jelinek is with the Australian School of Advanced Medicine, Macquarie University, N.S.W. 2113, Australia, and also with the School of Community Health, Charles Sturt University, Albury, Australia (e-mail hjelinek@csu.edu.au).

J. Wainer and A. Rocha are with the Institute of Computing, University of Campinas (Unicamp), Campinas 13083-852, Brazil (e-mail wainer@ic.unicamp.br; anderson.rocha@ic.unicamp.br).

Manuscript received Month XX, YYYY; revised Month XX, YYYY.

Fleming et al. [27] compared automatic and human grading of DR, to assess the effectiveness of the former in a screening program. Their procedure started with quality assessment, to ensure that the retina images had the recommended field definition [28]. If the image had the required quality, it was then assessed by detecting microaneurysms and hemorrhages. The screening classified an image as positive if it was low-quality or if it had lesions. The discrepancies among the automatic and the manual grading were evaluated by seven senior ophthalmologists. The authors showed that, in most conflicting cases, the grading software result was correct. The study indicated that for 45.7% of patients, manual grading could be avoided.

Soto-Pedre et al. [22] investigated the advantages of using DR detectors to reduce specialists' workload. Their automated grading system assessed image quality, and counted microaneurysms [27], [28]. For comparison, the retinal images were graded by a specialist who classified the gradable (with enough quality) images using the International Clinical Diabetic Retinopathy (ICDR)[1] severity scale [29]. When an image reached the threshold level of DR (e.g., a few microaneurysms), the patient was referred to an ophthalmologist. The method achieved 94.5% sensitivity, and 68.8% specificity. The automated system had a classification accuracy of 72.3% and was able to reduce human workload by 44%. However, the work focused only on microaneurysms, ignoring exudates, deep hemorrhages, and cotton-wool spots from the analysis. It is worth mentioning that the simple presence of a few microaneurysms in one quadrant of the retina only characterizes a mild nonproliferative DR.

Abràmoff et al. [21] combined individual DR-lesion detectors into a referability decision. The Iowa Detection Program incorporated image quality assessment and detectors of exudates, microaneurysms, hemorrhages, cotton-wool spots, and neovascularization. After analyzing the retinal images, a fusion algorithm combined the extracted information and generated a score that expressed the likelihood for referral [30]. Iowa Detection Program's AUC was 93.7% after adjudication and consensus by three experts.

Existing DR-lesion detectors tend to be specific for each type of lesion. However, recent research has stressed general frameworks adaptable for large classes of lesions [15]–[17], [19]. Pires et al. [18], [20] employed this methodology for detecting the commonest DR lesions: hard exudates, red lesions, superficial hemorrhages, deep hemorrhages, cotton-wool spots, and drusen. The authors employed a general model composed of BoVW mid-level features, and SVM classifiers. They created a high-level feature vector with the decision scores of all lesion classifiers, and used that vector as input for the referral classifier. Their referability decision had an AUC of 93.4% [20], which was then improved to 94.2% by enhancing the lesion detectors with better mid-level image features [18].

[1]ICDR is a simplification of the Early Treatment Diabetic Retinopathy Study (ETDRS), formulated by a consensus of international experts.

III. Lesion-based Referable

We now overview the classical approach for detecting referable DR, which, in general, follows three steps: (1) detection of individual DR lesions; (2) fusion of the lesion responses; and (3) referability decision. We will, in particular, follow how that methodology was implemented by Pires et al. [18], [20], based on the responses of six distinct lesion detectors. We will refer to that particular work, in the remainder of this paper, simply as "the lesion-based approach" (Figure 1).

For low-level feature extraction, the lesion-based approach used the Speeded-Up Robust Features (SURF) algorithm [31] pre-tuned to detect and describe a pre-determined number of points of interest (PoIs), e.g., 400. The PoIs were described by a 128-dimensional SURF which had previously proved more effective than the 64-orientation versions [15].

Codebook learning was performed by a k-means clustering over features randomly chosen from a training set of images. The codebooks were learned using a class-aware policy, avoiding a domination by codewords representing healthy regions. The class-aware scheme worked by creating two independent codebooks: one from descriptors outside the regions with lesions (including images from healthy patients), and one from descriptors sampled from regions marked by the specialist as having lesions.

The lesion-based approach employed BoVW with semi-soft coding as mid-level features. For comparison purposes, we will also use those same mid-level features in this work. Because they turn out to be so important, we will discuss mid-level features in detail in this section, and revisit them in Section IV.

The lesion-based approach had two distinct classifiers (and thus, two training phases): one for the lesion detectors, and one for the referability decision.

Mid-Level Feature Extraction: Bags of Visual Words

Mid-level feature extraction aims at transforming low-level local descriptors (e.g., SURF [31]) into a global and richer image representation of intermediate complexity [32].

BoVW [24], the most popular mid-level representation in Computer Vision, describes an image as a histogram of quantized local descriptors. It can be understood as the application of two steps [32]: *coding*, which transforms the local descriptors into a code adapted to the task, and *pooling*, which summarizes the codes obtained into a single feature vector.

In the standard BoVW, the coding step associates the low-level local descriptors to the closest codeword in the codebook[2] (*hard-assignment coding*), and the pooling step averages those codes over the entire image (*average pooling*).

Many alternatives to this standard scheme have been developed. For instance, to attenuate the effect of coding errors introduced by quantization, hard-assignment can be replaced by *soft-assignment coding* [33], [34]. On the other hand, soft assignment results in dense code-vectors, which is undesirable as it leads to ambiguities in the pooling of all codes present in the image. Therefore, a *semi-soft* scheme is often more appropriate. Pires et al. designed a semi-soft for retinal

[2]The codebook is usually obtained by clustering a sample of local descriptors from the training data.

images analysis that makes a partial (soft) assignment only to the codeword closest to the local descriptor, attaining, thus, sparsity [18].

Also, to overcome the "blurring" effect due to the averaging of the codes of all elements in the image, average pooling can be replaced by max pooling [35].

IV. Direct Referral Analysis

The automated referral assessment proposed in this paper dispenses with the intermediate stage of detecting DR lesions. That direct approach has both theoretical and practical motivations.

Lesion-based referral decisions loses critical information on the interface between the lesion-specific classifiers and the referability classifier. Often, the referral classifier receives just a vector of classification scores, one per lesion classifiers. This is unfortunate, because cogent information is lost, like the number, intensity, and even position of the lesions in the retina. One can create ad hoc schemes to transfer those data between the classifiers, but, it may be simpler to forgo the lesion classifiers altogether, and just provide the retinal images, with all cogent information, directly to the referability classifier.

Giving the whole image to the referability classifier has also practical advantages. The classical scheme involves implementing, debugging, training, and testing several lesion classifiers, and then one additional layer to combine the results and make the referral decision. Although using a unified approach for the lesion detectors [18]–[20] simplifies the task, it remains much more complex than implementing, training, and testing a single model. The streamlined technique for direct decision on referability is illustrated in Figure 2.

For low-level feature extraction, the method selects patches on a dense grid using diameters of 12, 19, 31, 50, 80, 128 pixels. The selected patches are described with SURF [31] in 128 dimensions. The low-level features are then integrated into a single feature vector using mid-level features.

For the mid-level features, we employ simple BoVW as a baseline and explore two recent alternatives: BossaNova [25] and Fisher Vector [26] (Section IV-A).

In the codebook learning step, while the lesion-based approach uses pre-computed codebooks for each individual lesion detector [18], the direct methodology uses k-means with Euclidean distance over a sample of low-level features. For BoVW or BossaNova [25], the codewords are the centroids of k-means, keeping the class-aware scheme proposed in [18] (half of the codebook from descriptors sampled from referable images, and half from nonreferable images). In Fisher Vector [26], the codebook learning, using (class-agnostic) Gaussian Mixture Models, is intrinsic to the representation.

The method trains just one decision model for referability.

A. Alternative Mid-Level Representations

In the following, we overview two recent mid-level representation methods: BossaNova and Fisher Vector.

1) BossaNova: In order to keep more information than the BoVW during the pooling step, BossaNova [25] introduces a density-based pooling strategy, which computes the histogram of distances between the local descriptors and the codewords. More formally, BossaNova pooling function g estimates the probability density function of α_m: $g(\alpha_m) = \text{pdf}(\alpha_{\mathbf{m}})$, by computing the following histogram of distances $z_{m,b}$:

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

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

In addition to that pooling strategy, Avila et al. [25] also proposed a localized soft-assignment coding that considers only the k-nearest codewords for coding a local descriptor, and keeps the representation compact.

Pires et al. [19] have shown that the BossaNova approach performed very well for detecting DR-related lesions. For white and red lesions detection, the results improved the hitherto best in literature (see [19] for more details).

In the current paper, we evaluated BossaNova representation in a direct approach for referable DR detection.

2) Fisher Vector: Fisher Vector [26] is the mid-level image representations with consistently best results in computer vision literature [36], [37]. Based upon the idea of Fisher information vectors [38] in the parametric space of Gaussian Mixture Models (GMM) estimated over the whole set of images, it extends the BoVW paradigm by encoding first- and second-order average differences between the descriptors and codewords. Furthermore, Fisher Vector is a compact representation, since much smaller codebooks are required in order to achieve a good classification performance in general vision problems.

Formally, given a GMM with N Gaussians, let us denote its parameters by $\lambda = \{w_i, \mu_i, \sigma_i, i = 1 \ldots N\}$, where w_i, μ_i and σ_i are respectively the mixture weight, mean vector and diagonal covariance matrix of Gaussian i. In the Fisher Vector framework, the D-dimensional descriptor $\mathbf{x}_j$ is encoding with a function $\Phi(\mathbf{x}_j) = [\varphi_1(\mathbf{x}_j), \ldots, \varphi_N(\mathbf{x}_j)]$ into a $2ND$-dimensional space where each function $\varphi_i(\mathbf{x}_j)$ is defined by:

$$\begin{aligned} \varphi_i(\mathbf{x}_j) &: \mathbb{R}^D \longrightarrow \mathbb{R}^{2D}, \\ \varphi_i(\mathbf{x}_j) &= \left[\frac{\gamma_j(i)}{\sqrt{w_i}}\left(\frac{\mathbf{x}_j - \mu_i}{\sigma_i}\right), \frac{\gamma_j(i)}{\sqrt{2w_i}}\left(\frac{(\mathbf{x}_j - \mu_i)^2}{{\sigma_i}^2} - 1\right)\right], \end{aligned} \tag{2}$$

where $\gamma_j(i)$ denotes the soft assignment of descriptor $\mathbf{x}_j$ to Gaussian i.

We evaluate the Fisher Vector mid-level features given that it offers a more complete representation of the sample set, which we believe would be important for DR referable assessment.

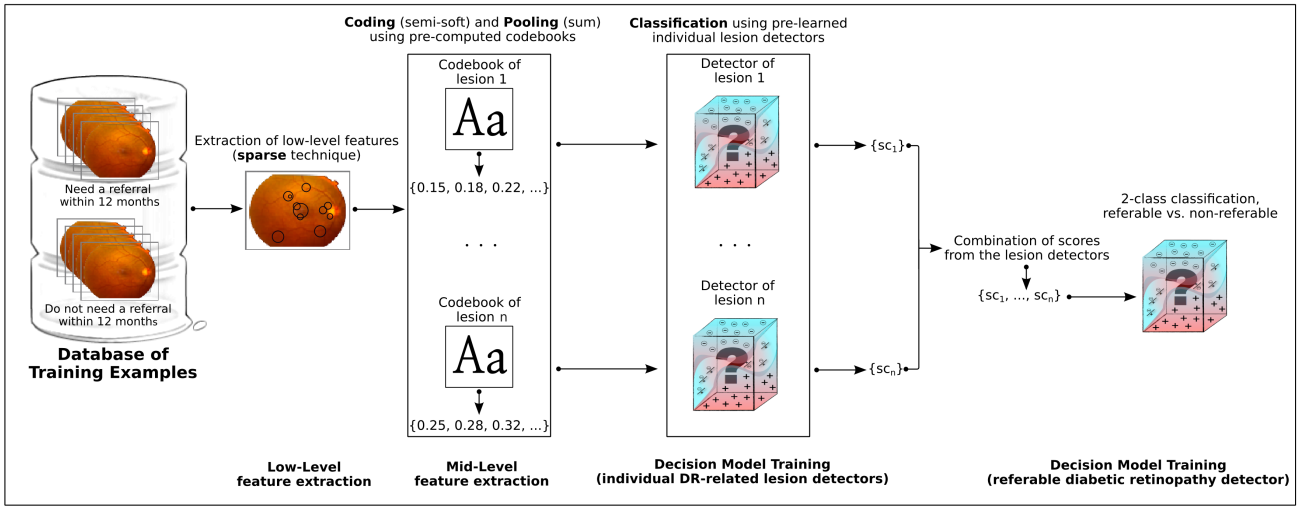

Fig. 1. Pipeline of the lesion-based methodology for detection of referable diabetic retinopathy as described in [18].

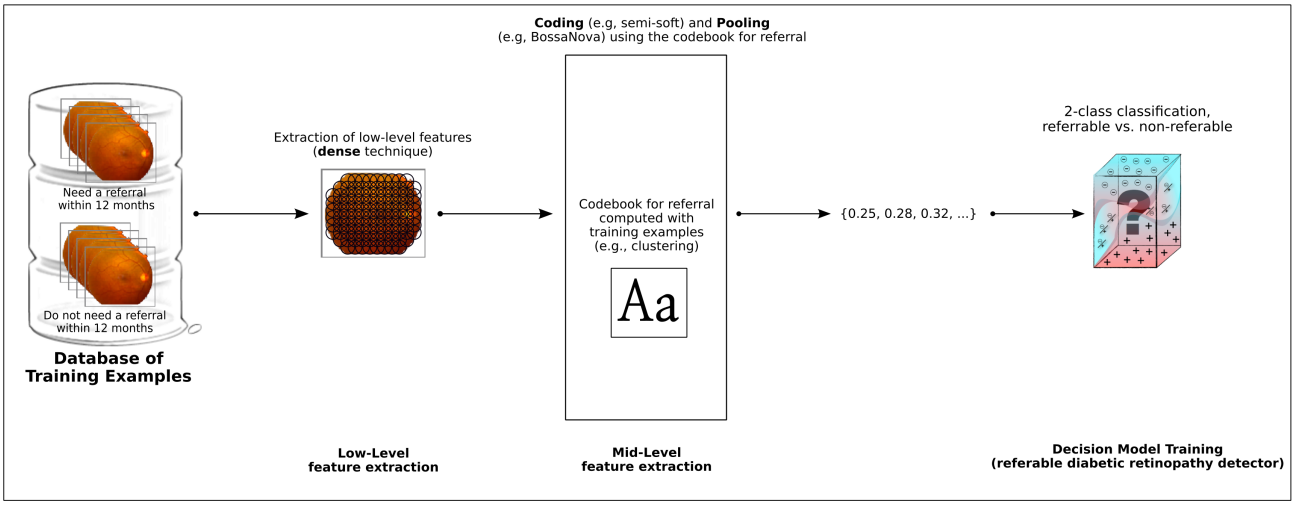

Fig. 2. Pipeline of the methodology for direct diabetic retinopathy referral assessment proposed in this work.

To the best of our knowledge, it has never been applied to this problem. More details on Fisher Vector representation can be found in [26], [37].

V. Experimental Protocol

Here we evaluate whether the proposed direct referal leads to better decisions than the previous lesion-based schemes. We also address the limitations of BoVW for retinal images (specifically the semi-soft coding proposed for eye-fundus images), and explore sophisticated approaches for mid-level description, such as Fisher Vector and BossaNova, detailed in Section IV-A.

In this section, we describe the datasets used in the development of the system for referable diabetic retinopathy, the validation protocol, and a description of the experiments.

A. Datasets

Two different datasets annotated by medical specialists were considered in this work: DR2 and MESSIDOR.

The DR2 dataset[3], from the Department of Ophthalmology, Federal University of São Paulo, comprises 520 images captured using a TRC-NW8 (Topcon Inc., Tokyo, Japan) nonmydriatic retinal camera with a Nikon D90 camera. To increase the processing speed, the images with 12.2 megapixels were cropped to 867×575 pixels. According to the annotations related to referable DR, 435 images were manually categorized by two independent specialists, whose mean intergrader κ is 0.77. Of these, 98 images were graded by at least one expert as requiring referral (56 images graded as positive by both experts), while 337 images were annotated by both experts as not requiring referral within one year. Although all patients in the DR2 dataset are diabetic, the specialists were asked to tag an image as referable or nonreferable based on any reason they considered relevant, not just the severity of a particular lesion.

[3]publicly available under accession number 10.6084 and URL http://dx.doi.org/10.6084/m9.figshare.953671.

TABLE I
RETINOPATHY GRADE CRITERION USED FOR MESSIDOR ANNOTATION.

Grade	Criterion
0	(μA = 0) AND (H = 0)
1	(0 < μA <= 5) AND (H = 0)
2	((5 < μA < 15) OR (0 < H < 5)) AND (NV = 0)
3	(μA >= 15) OR (H >= 5) OR (NV = 1)

μA: number of microaneurysms
H: number of hemorrhages
NV = 1: neovascularization - NV = 0: no neovascularization

The MESSIDOR dataset[4] was acquired in three French ophthalmologic departments using a color video 3CCD camera on a Topcon TRC NW6 non-mydriatic retinograph with a 45 degree field of view. It comprises 1,200 eye-fundus images captured with three distinct resolutions: 1,440×960, 2,240×1,488 or 2,304×1,536. The images have been cropped in order to establish that the relevant circular area of the retina has a radius similar to the DR2 database. Although it has not been graded for referable diabetic retinopathy, the dataset was annotated with two significant criteria: retinopathy grade (see Table I) and risk of macular edema.

The severity of the diabetic retinopathy is largely used as guideline for the frequency of examinations [21], [39]. Although a frequent consultation is recommended for patients with moderate or severe nonproliferative DR [39], for those without DR-related lesion or with just microaneurysms the annual incidence of progression is low [40]. In these situations, longer intervals between examinations may be recommended (one year for diabetics). Hence, based on the original annotations about DR severity, the guidelines of periodic referrals and the opinion of an expert about the criterion employed in the MESSIDOR dataset, we switched the grading into referable or nonreferable. Given that the presence of just microaneurysms does not suggest a referral in less than one year, the grades 0 and 1 (including also no risk of macular edema) are considered as nonreferable, while grades 2 and 3 (and also apparent macular edema) as referable, resulting in 688 negative and 512 positive images.

B. Validation Protocol

We employed a 5×2-fold cross-validation protocol [41], the same used in previous lesion-based assessment works [18], [20]. This protocol consists of repeating the process of two-fold cross validation five times. In each of the five steps, we randomly separate the samples in two groups, balanced by class, and use one of the groups for training and the other for testing. We perform two experiments per step, with the groups switching roles.

C. Experiments

The experiments were divided into three parts, to answer the following questions:

1) **Question 1:** Can we forgo the detection of individual DR lesions and still have an effective referral decision? In these experiments, we employ the DR2 dataset, divided in the same configuration used in the work used as baseline [18].
2) **Question 2:** Can sophisticated mid-level features — BossaNova [25] and Fisher Vector [26] — improve referral decision? The protocol and baseline are the same as in Question 1, but here we also contrast our different direct referral techniques (BoVW vs. BossaNova vs. Fisher).
3) **Question 3:** Can we confirm the suitability of direct referral in a second, independent, dataset? These experiments are performed in the reputable MESSIDOR dataset, using both the direct and the lesion-based approaches for comparison purposes. As explained in V-A, the MESSIDOR dataset was obtained in very different circumstances (time, country, equipment, people) than the DR2 dataset used in Question 1, reinforcing the independence of the results.

[4]kindly provided by the MESSIDOR program partners (see http://messidor.crihan.fr)

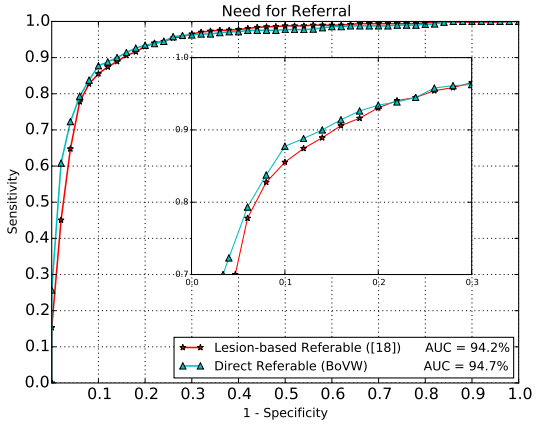

Fig. 3. ROC results for direct assessment for need of referral using BoVW mid-level characterization approach. The experiments were performed for hypothesis validation.

In our experiments, we extracted visual codebooks of {1,000, 2,000} visual codewords. Except for the number of visual codewords, we kept the default BossaNova parameter values the same as in [25]. For Fisher Vector, we used GMM with {128, 256} Gaussians after reducing the dimensionality of the SURF descriptors to 64 by applying Principal Component Analysis (PCA), as suggested in [37].

For the binary classification (referable vs. nonreferable), we used Support Vector Machines (SVM). We searched for the best parameters during the training with the standard LIBSVM's built-in grid search algorithm [42].

VI. RESULTS

In order to investigate the hypothesis that *lesion detection is nonessential for an effective referral assessment* and answer Question #1 in Section V-C, we use exactly the same mid-level features both for the traditional lesion-based method and the current approach. Both employ the BoVW with semi-soft coding explained in [18].

Figure 3 shows the results on the DR2 dataset for both methodologies: lesion-based [18] (AUC = 94.2%) and the

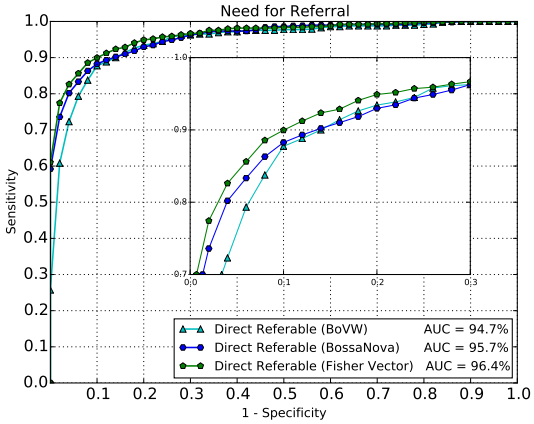

Fig. 4. ROC results for direct assessment for need of referral using advanced mid-level characterization approaches. The experiments were performed for improvement of the method.

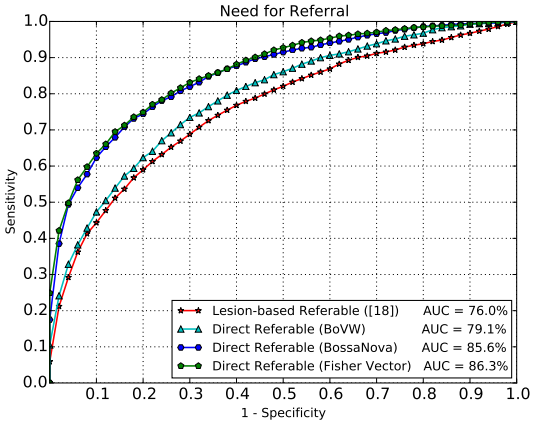

Fig. 5. ROC results for direct assessment for need of referral using advanced mid-level characterization approaches. The experiments were performed for emphasizing the fitness of the method in the MESSIDOR dataset.

best direct-referral. Direct referral performs better with 2,000 codewords, reaching an AUC of 94.7%.

The results obtained with BoVW, shown in Figure 3, validate our hypothesis that the detection of individual DR-related lesions is not necessary to provide effective referral decisions.

The second part of the experiments aims at exploring advanced mid-level features for answering Question #2 in Section V-C , where the results obtained in the first part will act as a baseline. As previously explained, we evaluate two recent mid-level features: BossaNova [25] and Fisher Vector [26].

Figure 4 shows the best results achieved with each mid-level feature. While BoVW achieved its best result with 2,000 codewords (AUC = 94.7%), BossaNova reached the best AUC using 1,000 codewords (AUC = 95.7%). Finally, Fisher Vector obtained the best result using just 128 Gaussians (AUC = 96.4%).

The results presented in Figure 4 express how accurate are the referral decisions by direct assessment, emphasizing that richer representation approaches yield better results for referable DR detection. We highlight that the Fisher Vector approach outperforms the traditional BoVW, reducing the classification error by over 30% (5.3% to 3.6%).

The third part of the experiments aims at reinforcing, in a reputable dataset, the direct methodology for referral and answering Question #3 in Section V-C. We use the MESSIDOR dataset as benchmark.

Once again, we perform this experiment with both the current direct referral and the previous lesion-based approaches. Figure 5 depicts the results reached with the lesion-based method, as well as the best results achieved with the direct approach for each mid-level representation. For BoVW, the best result was obtained with a codebook of size 2,000 (AUC = 79.1%). BossaNova reached its best AUC using 2,000 codewords (AUC = 85.6%). For Fisher Vector, the best result was achieved using 256 Gaussians (AUC = 86.3%).

While the lesion-based method obtained an AUC of 76.0%, we achieve the promising result of 86.3% using the current method that does not depend of lesion detection. The results reveal MESSIDOR as much more challenging dataset than DR2. Nevertheless, the relative performance of the techniques confirm the interest of the direct referral choice, which, in this case, appear prominently better.

To the best of our knowledge, only Sánchez et al. [43] used MESSIDOR for referral decisions. They reported an AUC of 91.0% using, however, ad hoc pre-processing techniques such as quality analysis, vessel segmentation, optic disc detection, and lesion detection.

Additionally, we attempt a new spatial pooling in concentric circular regions, in order to explore the location of low-level descriptors in the retinal image. Since the position of the lesions in the retina is well-known by specialists as one important factor for identifying referable retinopathy [44], we propose to divide the image on concentric circular regions emphasizing the lesions in the macular area. Nonetheless, the results we obtained so far contradict the expectations. For a more detailed discussion on this topic, please refer to supplementary material available along with this paper.

A. Statistical Analysis

In this section we explore the significance of the previous results.

Each single experiment requires picking a large number of parameters (mid-level representation, codebook size), that may have interactions. In order to investigate if the choice of mid-level feature would still appear significant when considering the totality of experiments performed, we applied a factorial analysis of variance (ANOVA) [45, chap. 22], with a block design using the folds as blocks, on the DR2 dataset. Since the AUC is a rate and behaves very non-linearly at the extremes of the [0-1] scale, we employ the more linear “log odds” scale (logit). We lessened the nuisance effect of the choice of the training set, by subtracting the global average of each fold from the results relative to that fold. In Table II, the statistical results reinforce the importance of the choice mid-level representation (p-value < 0.001). Note that the mid-level is responsible for more than 40% of the variation (see the column ‘Sum of squares’).

TABLE II
PARTIAL VIEW OF THE ANOVA TABLE. WE OMIT THE SECOND-ORDER INTERACTIONS SINCE NONE OF THEM WERE SIGNIFICANT. THE CHOICE OF MID-LEVEL REPRESENTATION EXPLAINS THE NON-RANDOM VARIATION, AS SEEN IN THE SUM OF SQUARES COLUMN.

Parameter	Degrees of freedom	Sum of squares	Mean square	F value	p-value	
mid-level	2	59.77	29.885	20.787	2.02×10^{-7}	***
codebook	1	2.57	2.568	1.786	0.187	
residuals	54	77.64	1.438			
total	59	143.86				

Significance codes: *** p-value < 0.001; ** p-value < 0.01; * p-value < 0.05

To further investigate the difference between the two competing advanced mid-level features, BossaNova and Fisher Vector, we perform a t-test [45, chap. 13] on paired samples obtained by 10 folds on the DR2 dataset (recall that we employed a 5×2-fold cross-validation protocol). For a confidence level of 95%, the difference between Fisher Vector and BossaNova is not significant (p-value = 0.3569), showing that both Fisher Vector and BossaNova play important roles in direct referral assessment.

VII. CONCLUSION

We proposed a novel approach to decide, directly from the retinal images and without preliminary DR-lesion detection, whether or not a patient will need to be referred to an ophthalmic specialist within a year. This decision to forgo specific DR-lesion detection has both theoretical motivations (making the referral decision using all information present in the image instead of just lesion scores) and practical advantages (much simpler to implement, test, and deploy). We highlight that direct assessment is new for referable diabetic retinopathy, and has not been developed before.

The experiments show that direct analysis for referral is not only feasible, but also advantageous. The direct methodology provided conclusive and promising results that outperformed the traditional lesion-based methodology in a strict comparative investigation. Direct referral assessment is possible because the rich mid-level representations we employ capture all the cogent information from the image, allowing the classifier to make complex decisions without losing important information.

Our experimental results contradicts previous beliefs showing no advantage in using two levels of classification: on the contrary, our experiments suggest that the loss of information in the interface between the classifiers is detrimental to accurate classification. Direct referral assessment is possible and effective using appropriate choices of low- and mid-level representations of the retina images.

We emphasize the novelty of using cutting-edge mid-level representations (BossaNova and Fisher Vector), over the traditional BoVW approach. The experiments and statistical analysis confirm that the choice of the mid-level representation is critical. The best result for direct referral, reached by the Fisher Vector approach, clearly outperforms the traditional lesion-based method by more than two percentage points, reducing the classification error by almost 40% (from 5.8% to 3.6%). We also conclude that, the richer the image representations, the more accurate the diagnosis of need of referral.

As the proposed direct methodology presented interesting results for referral assessment, in future work we aim to investigate automated decision with respect to DR progression. Other future work consists of studying alternative techniques using very recent approaches such as convolutional neural networks.

ACKNOWLEDGMENT

The authors would like to thank the medical team for helping us to collect and tag the ocular-fundus images. This work was supported in part by Microsoft Research, São Paulo Research Foundation (Fapesp) under the grants MSR-Fapesp 2008/54443-2 and Fapesp 2010/05647-4, Amazon Web Services, and Samsung Electronics of Amazon.

REFERENCES

[1] J. W. Yau, S. L. Rogers, R. Kawasaki, E. L. Lamoureux, J. W. Kowalski, T. Bek, S.-J. Chen, J. M. Dekker, A. Fletcher, J. Grauslund *et al.*, "Global prevalence and major risk factors of diabetic retinopathy," *Diabetes care*, vol. 35, no. 3, pp. 556–564, 2012.

[2] Vision Problems in the U.S., "Prevalence of adult vision impairment and age-related eye disease in America," accessed: 2015-02-12. [Online]. Available: http://www.visionproblemsus.org/

[3] D. M. Gibson, "The geographic distribution of eye care providers in the united states: Implications for a national strategy to improve vision health," *Preventive Medicine*, vol. 73, pp. 30–36, 2015.

[4] R. Hazin, M. Colyer, F. Lum, and M. K. Barazi, "Revisiting diabetes 2000: challenges in establishing nationwide diabetic retinopathy prevention programs," *American Journal of Ophthalmology*, vol. 152, no. 5, pp. 723–729, 2011.

[5] A. D. Fleming, S. Philip, K. A. Goatman, J. A. Olson, and P. F. Sharp, "Automated microaneurysm detection using local contrast normalization and local vessel detection," *IEEE Transactions Medical Imaging*, vol. 25, pp. 1223–1232, 2006.

[6] A. D. Fleming, S. Philip, K. A. Goatman, G. J. Williams, J. A. Olson, and P. F. Sharp, "Automated detection of exudates for diabetic retinopathy screening," *Physics in Medicine and Biology*, vol. 52, no. 24, pp. 7385–7396, 2007.

[7] A. D. Fleming, K. A. Goatman, S. Philip, G. J. Williams, G. J. Prescott, G. S. Scotland, P. McNamee, G. P. Leese, W. N. Wykes, P. F. Sharp, and J. A. Olson, "The role of haemorrhage and exudate detection in automated grading of diabetic retinopathy," *British Journal of Ophthalmology*, vol. 94, no. 6, pp. 706–711, 2010.

[8] L. Giancardo, F. Meriaudeau, T. Karnowski, Y. Li, K. Tobin, and E. Chaum, "Microaneurysm detection with radon transform-based classification on retina images," in *Intl. Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 5939–5942.

[9] M. Niemeijer, B. van Ginneken, M. J. Cree, A. Mizutani, G. Quellec, C. I. Sanchez, B. Zhang, R. Hornero, M. Lamard, C. Muramatsu, X. Wu, G. Cazuguel, J. You, A. Mayo, L. Qin, Y. Hatanaka, B. Cochener, C. Roux, F. Karray, M. Garcia, H. Fujita, and M. D. Abràmoff, "Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 185–195, 2010.

[10] M. Niemeijer, B. van Ginneken, S. R. Russell, M. S. A. 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.

[11] A. Sopharak, B. Uyyanonvara, S. Barman, and T. H. Williamson, "Automatic detection of diabetic retinopathy exudates from non-dilated retinal images using mathematical morphology methods," *Computerized Medical Imaging and Graphics*, vol. 32, p. 8, 2008.

[12] M. D. Abràmoff and M. S. A. Suttorp-Schulten, "Web-based screening for diabetic retinopathy in a primary care population: the eyecheck project," *Telemed J E Health*, vol. 11, pp. 668–674, 2005.

[13] K. Ganesan, R. J. Martis, U. R. Acharya, C. K. Chua, L. C. Min, E. Ng, and A. Laude, "Computer-aided diabetic retinopathy detection using trace transforms on digital fundus images," *Medical & biological engineering & computing*, vol. 52, no. 8, pp. 663–672, 2014.

[14] D. Sidibé, I. Sadek, and F. Mériaudeau, "Discrimination of retinal images containing bright lesions using sparse coded features and svm," *Computers in biology and medicine*, vol. 62, pp. 175–184, 2015.

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

[16] H. Jelinek, R. Pires, R. Padilha, S. Goldenstein, J. Wainer, and A. Rocha, "Data fusion for multi-lesion diabetic retinopathy detection," in *IEEE Computer-Based Medical Systems*, 2012, pp. 1–4.

[17] 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 *Intl. Conference of the IEEE Engineering in Medicine and Biology Society*, 2013, pp. 5857–5860.

[18] 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, no. 6, p. e96814, 06 2014.

[19] R. Pires, S. Avila, H. F. Jelinek, J. Wainer, E. Valle, and A. Rocha, "Automatic diabetic retinopathy detection using BossaNova representation," in *Intl. Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 146–149.

[20] 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.

[21] M. D. Abràmoff, J. C. Folk, D. P. Han, J. D. Walker, D. F. Williams, S. R. Russell, P. Massin, B. Cochener, P. Gain, L. Tang *et al.*, "Automated analysis of retinal images for detection of referable diabetic retinopathy," *JAMA Ophthalmology*, vol. 131, no. 3, pp. 351–357, 2013.

[22] E. Soto-Pedre, A. Navea, S. Millan, M. C. Hernaez-Ortega, J. Morales, M. C. Desco, and P. Pérez, "Evaluation of automated image analysis software for the detection of diabetic retinopathy to reduce the ophthalmologists' workload," *Acta ophthalmologica*, vol. 93, no. 1, pp. e52–e56, 2015.

[23] R. Pires, T. Carvalho, G. Spurling, S. Goldenstein, J. Wainer, A. Luckie, H. F. Jelinek, and A. Rocha, "Automated multi-lesion detection for referable diabetic retinopathy in indigenous health care," *PLoS ONE*, vol. 10, no. 6, p. e0127664, 06 2015.

[24] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," in *IEEE Intl. Conference on Computer Vision*, 2003, pp. 1470–1477.

[25] 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.

[26] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision*. Springer, 2010, pp. 143–156.

[27] A. D. Fleming, K. A. Goatman, S. Philip, G. J. Prescott, P. F. Sharp, and J. A. Olson, "Automated grading for diabetic retinopathy: a large-scale audit using arbitration by clinical experts," *British Journal of Ophthalmology*, vol. 94, no. 12, pp. 1606–1610, 2010.

[28] 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," *Investigative Ophthalmology & Visual Science*, vol. 47, no. 3, pp. 1120–1125, 2006.

[29] C. Wilkinson, F. L. Ferris III, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, and J. T. Verdaguer, "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, 2003.

[30] M. Niemeijer, M. D. Abramoff, and B. van Ginneken, "Information fusion for diabetic retinopathy cad in digital color fundus photographs," *Medical Imaging, IEEE Transactions on*, vol. 28, no. 5, pp. 775–785, 2009.

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

[32] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2559–2566.

[33] J. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.

[34] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *IEEE Intl. Conference on Computer Vision*, 2011, pp. 2486–2493.

[35] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition*, 2009, pp. 1794–1801.

[36] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *British Machine Vision Conference (BMVC)*, 2011, pp. 76.1–76.12.

[37] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision (IJCV)*, vol. 105, no. 3, pp. 222–245, 2013.

[38] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 1999, pp. 487–493. [Online]. Available: http://dl.acm.org/citation.cfm?id=340534.340715

[39] D. S. Fong, L. Aiello, T. W. Gardner, G. L. King, G. Blankenship, J. D. Cavallerano, F. L. Ferris, and R. Klein, "Retinopathy in diabetes," *Diabetes care*, vol. 27, no. suppl 1, pp. s84–s87, 2004.

[40] T. Batchelder and M. Barricks, "The wisconsin epidemiologic study of diabetic retinopathy," *Archives of ophthalmology*, vol. 113, no. 6, pp. 702–703, 1995.

[41] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, pp. 1895–1923, 1998.

[42] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[43] C. I. Sánchez, M. Niemeijer, A. V. Dumitrescu, M. Suttorp-Schulten, M. D. Abramoff, and B. v. Ginneken, "Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data," *Investigative Ophthalmology & Visual Science*, vol. 52, no. 7, pp. 4866–4871, June 2011.

[44] T. V. Litvin, G. Y. Ozawa, G. H. Bresnick, J. A. Cuadros, M. S. Muller, A. E. Elsner, and T. J. Gast, "Utility of hard exudates for the screening of macular edema," *Optometry and Vision Science*, vol. 91, no. 4, pp. 370–375, 2014.

[45] R. Jain, *The art of computer systems performance analysis - techniques for experimental design, measurement, simulation, and modeling.*, ser. Wiley professional computing. Wiley, 1991.