Assessing the Need for Referral in Automatic Diabetic Retinopathy Detection

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

***Abstract*—Emerging technologies in health care aim at reducing unnecessary visits to medical specialists, minimizing overall cost of treatment and optimizing the number of patients seen by each doctor. This paper explores image recognition for the screening of diabetic retinopathy, a complication of diabetes that can lead to blindness if not discovered in its initial stages. Many previous reports on DR imaging focus on the segmentation of the retinal image, on quality assessment, and on the analysis of presence of DR-related lesions. Although this study has advanced the detection of individual DR lesions from retinal images, the simple presence of any lesion is not enough to decide on the need for referral of a patient. Deciding if a patient should be referred to a doctor is an essential requirement for the deployment of an automated screening tool for rural and remote communities. We introduce an algorithm to make that decision based on the fusion of results by metaclassification. The input of the metaclassifier is the output of several lesion detectors, creating a powerful high-level feature representation for the retinal images. We explore alternatives for the bag-of-visual-words (BoVW)-based lesion detectors, which critically depends on the choices of coding and pooling the low-level local descriptors. The final classification approach achieved an area under the curve of 93.4% using SOFT–MAX BoVW (soft-assignment coding/max pooling), without the need of normalizing the high-level feature vector of scores.**

***Index Terms*—Bag-of-visual-words (BoVW), diabetic retinopathy, lesion detectors, metaclassification, referral, visual dictionaries.**

I. Introduction

THE development of computational systems that support specialists in diverse areas of health care has been the focus of several studies [1]–[6]. The use of computational methods that aid in the diagnosis of disease has contributed significantly to improve the quality of life of patients. In this context, several computational systems have been proposed (see, e.g., [1]–[3] and [7]) for dealing with complications related to *Diabetes Mellitus*.

According to the International Diabetes Federation,[1] diabetes will nearly double to 552 million people by 2030 [8]. Diabetes-related complications are also increasing in prevalence including diabetic retinopathy, which currently affects 2–4% of people with diabetes [9], [10] and is the main cause of blindness in the 20–74 age group in developed countries [11].

The development of a unified screening system that simultaneously identifies several different DR-related lesions has been described using a bag-of-visual-words (BoVW) model based upon visual dictionaries [1], [2], [12]. This model needs a visual dictionary for each type of lesion, and hence, a specific classifier is required for each type of lesion. To decide on the level of DR progression (from mild to severe), or the need for referral, one must combine the separate classifiers into a unified model.

In this paper, we propose a method that recommends referring a patient with diabetes for diabetic retinopathy assessment based on the image classification outcome, which is especially useful in remote and rural areas. The method captures retinal images from nonmydriatic or mydriatic cameras, evaluates the images in real time, and suggests whether or not the patient requires a review by an ophthalmic specialist within one year after the screening. The method consists of 1) detecting individual retinal anomalies and extracting the appropriate assessment scores, and 2) classifying the image as *referable versus nonreferable* by means of metaclassification techniques built upon the output of several lesion detectors. Different from [1], [2], and [12], we explore alternatives for the BoVW lesion detectors because the performance of BoVW depends critically on the choices of coding and pooling the low-level local descriptors and aim at characterizing the properties and signs related to each kind of lesion of interest.

We organized the remainder of this paper in four sections. Section II describes the related work, while Section III explains our methodology for referral versus nonreferral classification. Section IV presents the experimental results for the proposed methods, and, finally, Section V concludes this paper.

II. State of the Art

The existence of a DR-related lesion in a retinal image does not necessarily indicate a vision-threatening sign that requires

Manuscript received February 15, 2013; accepted August 4, 2013. Date of publication August 16, 2013; date of current version November 18, 2013. This work was supported in part by São Paulo Research Foundation FAPESP under Grant 2010/05647-4 and Grant 2011/15349-3, National Counsel of Technological and Scientific Development (CNPq) under Grant 307018/2010-5 and Grant 304352/2012-8), and Microsoft. *Asterisk indicates corresponding author.*

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

H. F. Jelinek is with the Department of Biomedical Engineering, Khalifa University, Abu Dhabi 127788, UAE, and also with the Australian School of Advanced Medicine, Macquarie University, North Ryde, N.S.W. 2113, Australia (e-mail: hjelinek@csu.edu.au).

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

E. Valle is with the School of Electrical and Computing Engineering, University of Campinas, Campinas 13083-852, Brazil (e-mail: dovalle@dca.fee.unicamp.br).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2013.2278845

[1] http://www.idf.org/diabetesatlas/5e/diabetes

a referral. The presence of microaneurysms, that characterize a moderate nonproliferative DR type, does not indicate an urgent consultation, but an indication of a follow-up between three months and 12 months depending on the number and location of the microaneurysms. Conversely, the presence of neovascularization indicates proliferative retinopathy and, if not under treatment, needs urgent referral for management by an ophthalmologist [13]. Other important retinal lesions are cotton wool spots, especially if there are more than five [14].

A nurse-managed primary care clinic is an essential step toward ensuring a satisfactory cost reduction as well as the opportunity of screening, assessment, and treatment reaching remote communities. Nurse-led screening programs are designed to verify the presence of any DR-related lesion, as well as to identify the lesion and whether referral is required.

Screening programs for DR have been developed in many countries such as the Netherlands [3], U.K. [4], and Australia [5]. Two well-known examples are the EyeCheck project [3] and the Challenge2009 [15]. In [16], the authors reported that both programs have statistically similar results (Areas under the receiver operating characteristic (ROC) curve (AUCs) of 82.0% for Challenge2009 [15] and 84.0% for EyeCheck [3]). Both programs focus on the detection of specific lesions and require pre- and postprocessing.

In the U.K., the National Screening Committee (UK NSC) recommends a systematic population screening program to be offered annually to all people with type 1 and type 2 diabetes aged 12 or over [17]. In 2010–2011, 79% of people in England aged 12 and over diagnosed with diabetes actually attended a retinopathy screening. In the U.K., automated DR screening has been investigated with software for image quality assessment and microaneurysm/dot hemorrhage detection. The outcome of this study for two community health screening clinics suggests a reduction of 36.6% in the workload of manual graders [18].

There are several other studies aimed at providing automated DR screening for the use by primary health care providers in rural communities. Luckie *et al.* [5] have proposed the identification of proliferative retinopathy (characterized by new vessel growth). The authors have explored wave transformation, mathematical morphology operations, and fractal analysis to provide an automated assessment of images to detect vascular proliferation from one image of the macular (posterior pole) region. However, the proposed technique requires an extensive stage of preprocessing.

Decencière *et al.* [19] have developed a strategy for combining a set of heterogeneous information for classification regarding the referral of patients. The authors have considered pathological information per lesion (microaneurysms, exudates, and hemorrhages), a signature-based pathological score (a proposed solution which relies on wavelet-based image characterization to detect the DR signs), image quality metrics, demographic information, and diabetes-related information (age, weight, diabetes type, etc.). The heterogeneous information is later combined with an algorithm for association rule mining.

III. Proposed Methodology

This section presents the method employed to decide if a patient is to be referred to an ophthalmologist within one year after the screening. The solution consists of 1) training detectors for individual DR-related lesions, and 2) using the decision scores from those detectors for training a metaclassifier that labels the retinal images as *referable* or *nonreferable*. The individual detectors are based on the BoVW model, for which several possibilities of coding and pooling [20] are investigated. The metaclassification can be interpreted as the creation of a high-level feature vector of scores, for which we test three possibilities of normalization.

A. Detection of Individual DR-Related Lesions

Individual DR lesions are detected using a BoVW model [20]–[22]. These models work by detecting a large number of feature vectors in the images [usually around points of interest (POIs)] and assigning these vectors to "visual words" using a dictionary of visual appearances. Six lesion detectors are considered: hard exudates, red lesions, superficial hemorrhages, deep hemorrhages, cotton-wool spots, and drusen.

Speeded-Up Robust Features (SURF) [23] is employed to detect the PoIs in the images and extract the local feature vectors based on local patches around the PoIs. The PoIs are detected based on approximations of the Hessian matrix in a scale-space, and the feature vectors are based upon the Haar-wavelet responses around the PoIs [23].

An identical protocol is applied to create the visual dictionary associated with the individual detectors. Using a training set of images, two sets of SURF feature vectors are extracted, one coming from regions annotated by medical specialists as containing lesions (positive) and one coming from healthy image regions (negative). A k-means clustering algorithm [24] then finds $k/2$ clusters from the vectors obtained using SURF and associated with images containing lesions and $k/2$ clusters associated with those images not containing lesions. The centroids of the clusters of the two sets are used as the visual words in a visual dictionary of size $M = k/2 + k/2$ with a total dictionary size of 500 words. We have set the dictionary size to 500 words (250 visual words for each class of interest) although more automated schemes could be employed to find the best size for each lesion detector [25].

The solution devised in this paper differs from traditional methods (see, e.g., [21] and [22]), where the visual dictionary is obtained from an indiscriminate sample of local features, without considering the classes (positive and negative) by creating a visual dictionary for each lesion/nonlesion. We create a visual dictionary for each lesion/nonlesion case containing specific visual words for each class.

The visual dictionaries are used to transform the low-level local feature vectors extracted by SURF onto mid-level BoVW feature vectors. This transformation requires first a coding step in which the low-level feature vectors are given a representation based on the dictionary. Then, for each image, all encoded vectors are aggregated in a *pooling* step (using operators such as sum, average, and max).

An outline of the mathematical description of the coding and pooling step follows. Once the visual dictionary is created, the visual dictionary can be represented as $\mathcal{C} = \{\mathbf{c}_i\}$, $i \in \{1, \ldots, M\}$, where $\mathbf{c}_i \in \mathbb{R}^d$ is a visual word in d-dimensional space. Then, for a given image, we start with the set of local descriptors $\mathcal{X} = \{\mathbf{x}_j\}$, $j \in \{1, \ldots, N\}$, where $\mathbf{x}_j \in \mathbb{R}^d$ is a local feature and N is the number POIs. Let $\mathbf{z}$ be the final BoVW vector representation [20].

The coding step transforms the low-level descriptors onto a representation based upon the codewords. The coding step can be modeled by a function $f\colon \mathbb{R}^d \to \mathbb{R}^M$, $f(\mathbf{x}_j) = \boldsymbol{\alpha}_j$ that takes the individual local descriptors $\mathbf{x}_j$ and maps them onto individual codes $\boldsymbol{\alpha}_j$. The classical BoVW model employs the "hard assignment" of a low-level descriptor to the closest codeword:

$$\alpha_{m,j} = 1 \;\; \text{if } m = \arg\min_k \|\mathbf{c}_k - \mathbf{x}_j\|_2^2 \quad \text{else} \quad 0 \tag{1}$$

where $\boldsymbol{\alpha}_{m,j}$ is the mth component of the encoded descriptor.

As a local descriptor can be roughly equidistant to several visual words, the method described in this paper proposes the use of a soft-assignment coding (degrees of association between the low-level descriptors and the elements of the codebook are allowed while avoiding the boundary effects of hard assignment) [26], [27]. To achieve this, a *codeword uncertainty* algorithm is employed [27]:

$$\alpha_{m,j} = \frac{G_\sigma(\|\mathbf{c}_m - \mathbf{x}_j\|\|_2)}{\sum_{\mathbf{c}\in\mathcal{C}} K_\sigma(\|\mathbf{c} - \mathbf{x}_j\|\|_2)} \tag{2}$$

where G_σ is the Gaussian kernel with $\sigma = 45$ [27]. Soft treatment has never been explored in the retinopathy-related literature before.

The pooling step takes place after the coding and can also be represented by a function $g\colon \{\boldsymbol{\alpha}_j\}_{j\in 1,\ldots,N} \to \mathbb{R}^M$, $g(\{\boldsymbol{\alpha}_j\}) = \mathbf{z}$. The classical BoVW corresponds to a "counting of words" (called sum-pooling):

$$g(\{\boldsymbol{\alpha}_j\}) = \mathbf{z} : \; \forall m, \; z_m = \sum_{j=1}^{N} \alpha_{m,j}. \tag{3}$$

While previous approaches in DR-related lesion detection using BoVW have used the sum-pooling approach [1], [2], we use the max-pooling approach that is applied here for creating the lesion detectors, taking the maximum activation of each codeword:

$$g(\{\boldsymbol{\alpha}_j\}) = \mathbf{z} : \; \forall m, \; z_m = \max_{j\in\{1,\ldots,N\}} \alpha_{m,j}. \tag{4}$$

The vector $\mathbf{z} \in \mathbb{R}^M$ obtained from pooling is a mid-level feature vector ready for use in a lesion/nonlesion classifier.

The common HARD–SUM (hard-assignment coding/sum pooling) and the more recent SOFT–MAX (soft-assignment coding/max pooling) approaches are applied here for creating the lesion detectors.

The final classification step for the individual lesion detectors is based upon a two-class *support vector machine* (SVM) [24] classifier, which employs the mid-level BoVW feature vectors for training and classification. In this step, a binary classifier is trained for each individual lesion (see Fig. 1). Although other classifiers could be used, SVM classifiers are preferred for a number of desirable traits: their solutions are global and unique; they have a simple geometric interpretation; and they are robust to overfitting, even when the input space is very high dimensional.

B. High-Level Feature Extraction and Referral Classifier

The information provided by each individual lesion detector is insufficient for deciding on whether a referral is necessary based on the lesion detector output because the lesions can be minor, just a few, they may not indicate there will be future deterioration of visual function, and their location might also be important (a few lesions present in the fovea may need referral whereas lesions in the periphery may not).

Our aim is to combine the results of the individual detectors in a metaclassification step that indicates whether or not a patient should be referred to an ophthalmologist for further review. This step can be interpreted as the creation of a high-level feature vector from the decision scores of all the lesion detectors. The metaclassification is made possible by an annotated dataset (not used in the training of any lesion detector), with images from patients tagged as referable versus nonreferable by two independent experts.

The goal is to have a high sensitivity (very few false negatives), while also keeping high specificity (few false positives): the former is important to ensure that no patient in need stays without care, the latter is important to avoid swamping the health care professionals with unneeded referrals.

The high-level referral classifier works as follows: 1) extraction of the low-level SURF feature vectors from the training images; 2) creation of the visual dictionaries for the lesions using annotated images of the lesion training sets; 3) extraction of the mid-level BoVW feature vectors using the visual dictionaries; 4) training of the independent SVM lesion detectors. After training individual lesion detectors, a referral classifier needs to be trained. For that, we first 5) extract the high-level feature vectors from the scores of the SVM lesion detectors on a training set of referable versus nonreferable tagged images; and finally, 6) we train a metaclassifier with the high-level feature vectors from the referral/nonreferral image training set. Fig. 1 illustrates these steps.

To analyze retinal images from a patient, we 1) extract the low-level SURF feature vectors from the retinograph images of this patient; 2) extract the mid-level BoVW feature vectors; 3) extract of the high-level feature vectors from the scores of the individual lesion detectors; and 4) perform the final decision based on the high-level feature vector (outcomes of the individual lesion detectors). As we consider six lesion detectors, each high-level feature vector is formed with six decision scores. Fig. 2 illustrates these steps.

To obtain the decision score for a particular feature vector representing an image for one detector, the distance of the feature vector representing such image to the decision hyperplane representing the detector is calculated. Alternatively, the decision confidence for an input rather than just the binary outcome could be obtained.

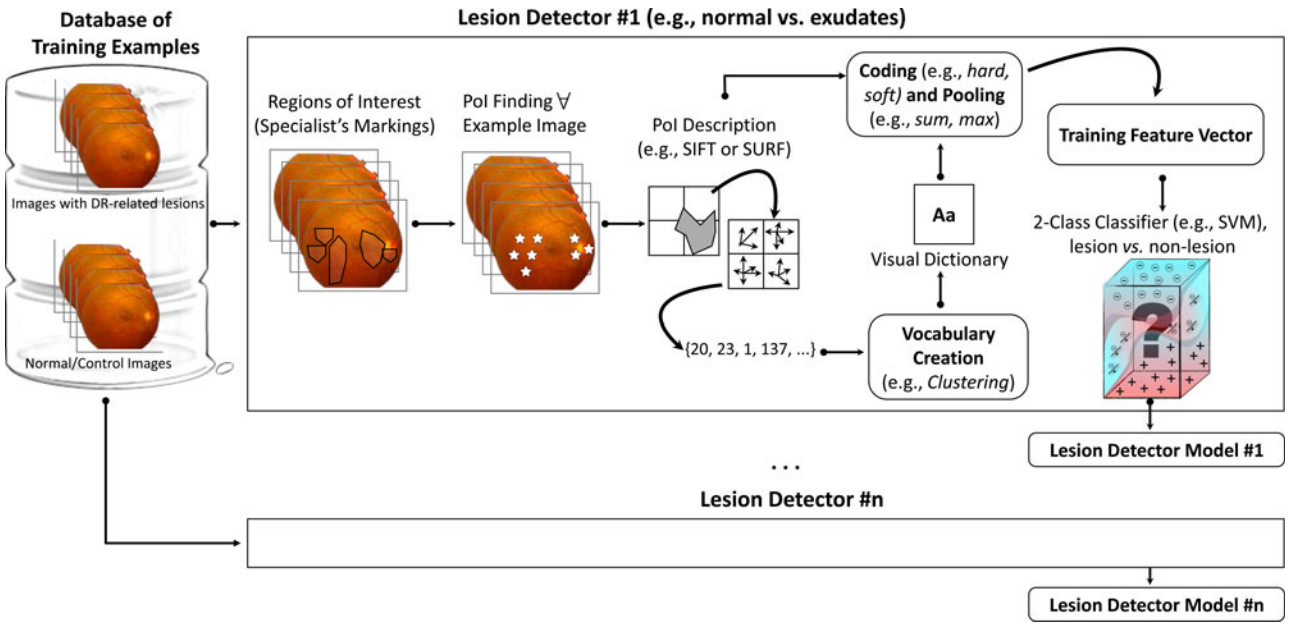
Database of Training Examples
Images with DR-related lesions
Normal/Control Images
Lesion Detector #1 (e.g., normal vs. exudates)
Regions of Interest (Specialist's Markings)
PoI Finding ∀ Example Image
PoI Description (e.g., SIFT or SURF)
Coding (e.g., hard, soft) and Pooling (e.g., sum, max)
Aa
Visual Dictionary
{20, 23, 1, 137, ...}
Vocabulary Creation (e.g., Clustering)
Training Feature Vector
2-Class Classifier (e.g., SVM), lesion vs. non-lesion
Lesion Detector Model #1
. . .
Lesion Detector #n
Lesion Detector Model #n

Fig. 1. Overview of the individual lesion detectors. Using regions marked by specialists for each kind of lesion, we extract POIs (low-level features) in such regions as well as in normal images and describe them using SURF. After that, we use clustering for creating a visual dictionary representing the normal images and the images with lesions. After creating the dictionary, we use coding (e.g., soft) and pooling (e.g., max) to create the training feature vectors (mid-level features) which feed a two-class classifier (lesion versus nonlesion). We end up with one classifier model for each considered lesion.

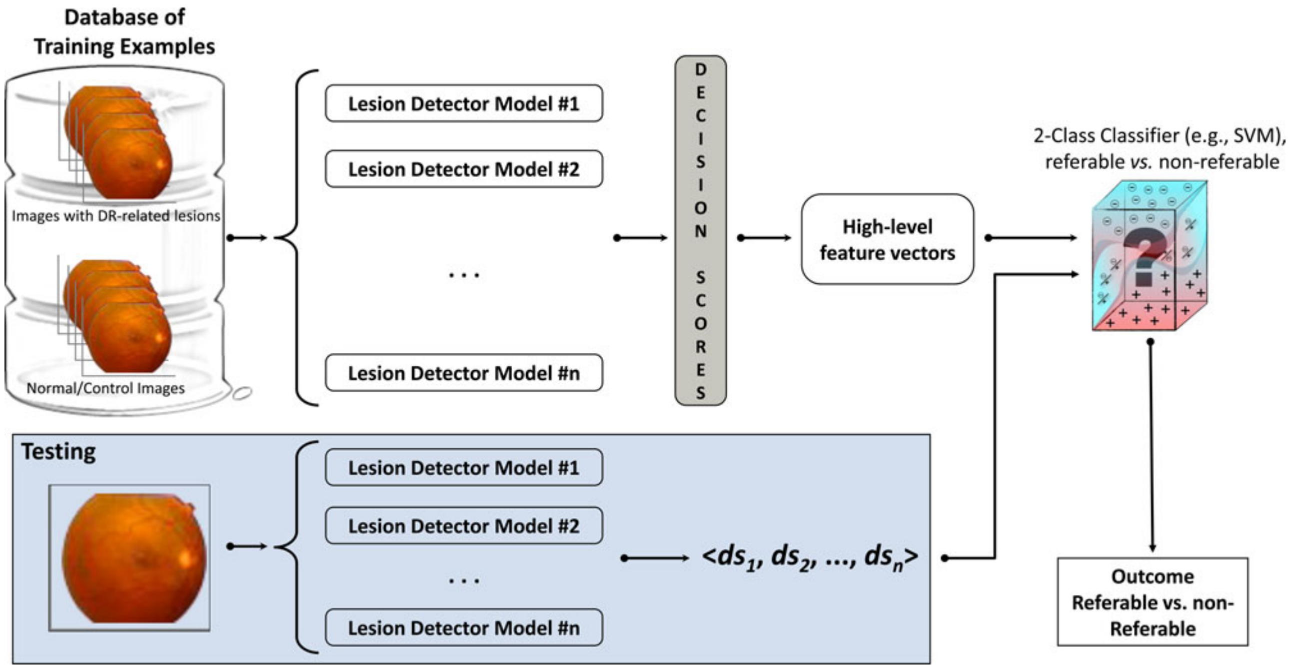

Fig. 2. Overview of the referral approach. Using a training set of examples, we apply the specific lesion detectors as depicted in Fig. 1. For each training image, we have n decision scores, one for each lesion detector. We use such decision scores as high-level feature vectors and train a referral two-class classifier using additional info regarding the patients (should they be referred to a doctor or not). For deciding on the referral for one particular patient (testing), we apply the same lesion detectors for the image of this particular patient obtaining the decision scores which represent the high-level feature vector of this image and feed it to the trained referral classifier which will issue the final decision.

C. Normalization

Normally, choosing the most appropriate normalization technique for the obtained decision scores before combining them in a high-level feature descriptor is a difficult task due to the heterogeneity of the score distributions [28].

Two normalization techniques are considered to verify whether the normalization methods improve on the classification outcome regarding the referral classifier: *term frequency* (TF), and z-*scores*, or z-norm. TF is widely used in text retrieval, where each document is represented by a vector of word frequencies [21]. It is expressed as the division of the number of occurrences of word i in document d (n_{id}) by the total number of words in the document d (n_d). *Z*-norm is calculated by taking the feature vector's arithmetic mean μ and standard deviation σ, subtracting μ from each score and dividing the result by σ [28].

IV. Validation and Experiments

In this section, we describe the datasets used in the development of the system, the validation protocol employed in the metaclassification, as well as the experimental results.

A. Datasets

Two different datasets tagged by medical specialists, DR1 and DR2, were used to perform the experiments.

DR1 is used for creating the DR-related lesion detectors. The dataset was created by the Department of Ophthalmology, Federal University of São Paulo. It has 1077 retinal images with an average resolution of 640 × 480 pixels, of which 595 images are normal and 482 images have at least one disease (234 images contain hard exudates, 102 images contain superficial hemorrhages, 146 images contain deep hemorrhages, 139 images contain drusen, 73 images contain cotton wool spots, and 180 images present signals of either superficial or deep hemorrhages, red lesions). Each image was manually annotated for DR-related lesion (presence/absence) by three medical specialists, and only the images for which the three specialists agree were kept in the final dataset. The images were captured using a TRC-50X (Topcon Inc., Tokyo, Japan) mydriatic camera with maximum resolution of one megapixel and a 45° field of view.

DR2 comprises 520 images with 12.2 megapixels cropped to 867 × 575 pixels to increase the processing speed. The dataset has been collected in the same ophthalmology department as DR1. Among the 520 images, 300 are normal and 149 have at least one lesion (79 images contain hard exudates, 98 images contain red lesions, 50 images contain drusen, and 17 images contain cotton wool spots). Ignoring the specific lesion that can be present, 337 images have been manually categorized by two independent specialists with over ten years of clinical experience as not requiring referral and 98 images require referral within one year to a specialist. Although all patients in the DR2 dataset are diabetic, the specialists were asked to tag an image as referrable or nonreferrable based on any reason they considered relevant, not just the severity of a particular DR lesion. The images were captured using a TRC-NW8 (Topcon Inc., Tokyo, Japan) nonmydriatic retinal camera with a Nikon D90 camera. We have released both datasets at http://www.recod.ic.unicamp.br/site/asdr.

B. Validation Protocol

A cross-dataset training/testing protocol for the individual lesion detectors is used. For that, we train each individual lesion detector using the DR1 dataset and test them on DR2.

The training of a referral classifier uses the individual lesion detectors created with DR1 dataset to obtain decision scores for the images in DR2 selected as training along with their annotated referable information. The testing uses the appropriate part of the DR2 dataset. In this case, a (5 × 2)-fold cross-validation protocol [29] is used which consists of repeating the process of two-fold cross validation five times. In each step, the dataset is randomly divided in two groups. The first group is used for training and the second group for testing. Then, the groups are switched.

TABLE I
AUCs for the Six Individual Detectors We Consider Here

DR-related Lesion	**HARD–SUM**	**SOFT–MAX**
Hard Exudates	93.1%	**95.5%**
Superficial Hemorrhages	88.8%	88.7%
Deep Hemorrhages	**90.0%**	86.5%
Red Lesions	**92.3%**	87.1%
Cotton Wool Spots	82.1%	**84.9%**
Drusen	**66.5%**	62.6%

C. Experiments

AUC is used to quantify the experimental results.

The experiments are divided into two parts:

1) Part #1. Experiments for detecting individual anomalies using a cross-dataset validation (training the classifiers with DR1 and testing with DR2).
2) Part #2. Experiments for determining the need of referral using the lesion classifiers trained on DR1 dataset and image scores from the DR2 dataset.

1) Experiments—Part #1: Previous research on individual lesion detection [1], [2] has explored only the HARD–SUM (hard-assignment coding/sum pooling) as mid-level representation. In this paper, we address the limitations of the HARD–SUM approach as discussed in Section III-A and explore alternatives to the mid-level feature extraction such as the SOFT–MAX (soft-assignment coding/max pooling).

For the binary classification technique, we have used the SVM classifier. We searched for the best SVM parameters during training using the standard LibSVM's grid search algorithm [30].

Table I shows the results obtained by the individual detectors. The results indicate that the SOFT–MAX technique has a considerable advantage compared to HARD–SUM for detecting hard exudates and cotton wool spots. On the other hand, HARD–SUM performed better for detecting deep hemorrhages. This complementary results goes in line with recent studies in the Computer Vision literature which hints at the interesting properties and results of SOFT–MAX techniques [20], [26], [27]

2) Experiments—Part #2: This part evaluates the referral metaclassifier. For each testing image, we extracted the decision scores from the individual DR-related lesion detectors generating a high-level description and feed the high-level metaclassifier, which produces the final referral decision. We explore how this metaclassifier performs when fed with normalized and nonnormalized decision scores.

• *Without normalization:* We use six lesion detectors to characterize each image producing a raw high-level feature vector (no normalization) of decision scores in six dimensions which goes to the referral classifier for the final decision making. Fig. 3

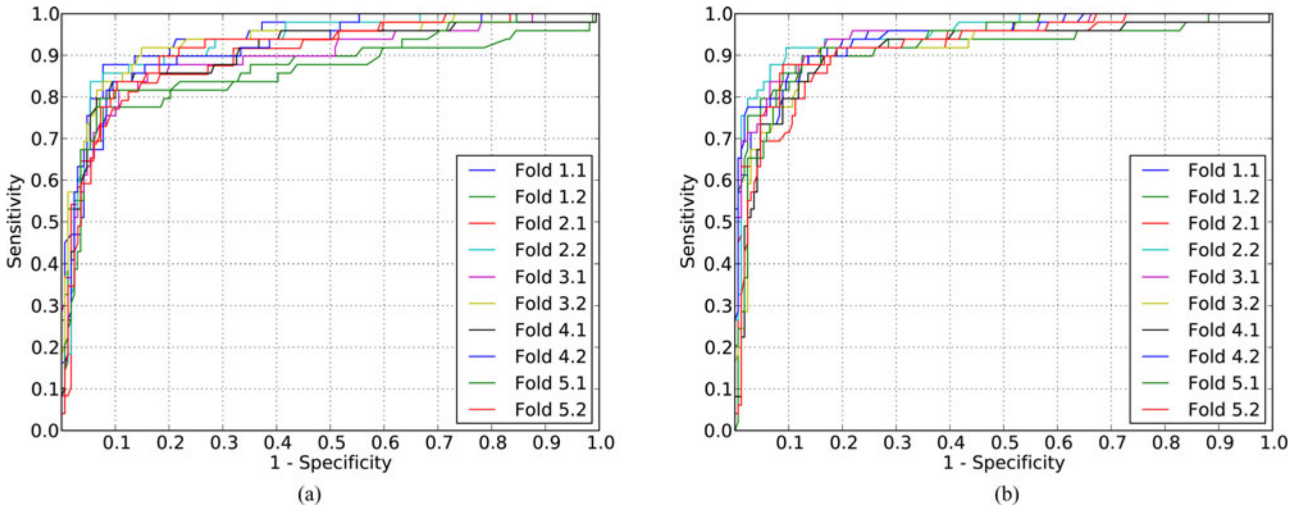

Fig. 3. Sensitivity and specificity for the final referral decision, when using HARD–SUM and SOFT–MAX mid-level BoVW feature vectors, and no normalization for the high-level feature vector of decision scores. Referral validation using a 5 × 2-fold cross-validation protocol. (a) Hard-assignment coding and sum pooling. (b) Soft-assignment coding and max pooling.

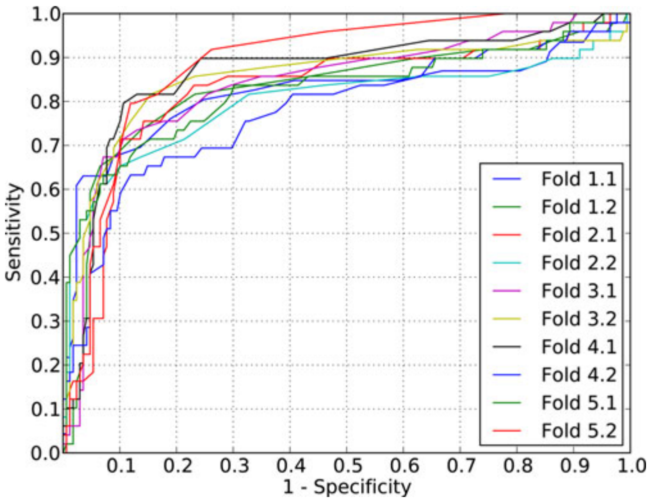

Fig. 4. Sensitivity and specificity for the final referral decision, when using SOFT–MAX mid-level BoVW feature vectors, and term-frequency normalization for the high-level feature vector of decision scores. Referral validation using a 5 × 2-fold cross-validation protocol.

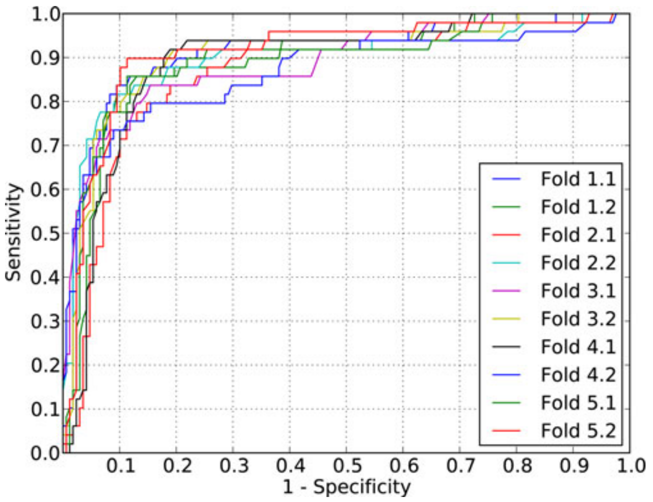

Fig. 5. Sensitivity and specificity for the final referral decision, when using SOFT–MAX mid-level BoVW feature vectors, and *z*-scores (standard normalization) for the high-level feature vector of scores. Referral validation using a 5 × 2-fold cross-validation runs.

shows the results for HARD–SUM and SOFT–MAX respectively, with SOFT–MAX yielding the best results.

- *Normalization with term frequency:* These experiments explore the normalization of high-level feature vectors (decision scores) using term frequency. Once the normalization process is complete, the images are classified. Fig. 4 depicts the ROC classification results for SOFT–MAX considering term-frequency normalization. Similar results are obtained when considering HARD–SUM.

- *Normalization with z-score:* The high-level feature vectors are this time normalized using the z-score approach followed by the classification step. Fig. 5 depicts the AUCs obtained for SOFT–MAX. Similar experiments with HARD–SUM do not show classification improvement over SOFT–MAX.

Table II summarizes all the results obtained for referral classification, presenting the arithmetic mean and standard deviation across the 5 × 2 cross-validation protocol. Considering the HARD–SUM technique, the term-frequency normalization does not help (AUC = 82.5%) and was more than eight percentage points worse than the method without normalization. On the other hand, the *z*-score technique reached an AUC equal to 91.7%, higher than the result without normalization. However, the difference is not statistically significant.

TABLE II
AUCs FOR REFERRAL

Technique	HARD–SUM	SOFT–MAX
Without normalization	90.8%±3.1%	**93.4%±2.1%**
Term-frequency	82.5%±4.7%	83.4%±4.6%
Z-score	**91.7%±2.2%**	89.4%±3.0%

Fig. 6(a) depicts average AUCs along with standard deviation for the HARD–SUM approach considering the 5 × 2 cross-validation protocol. Note that *z*-score is also equivalent to not using normalization.

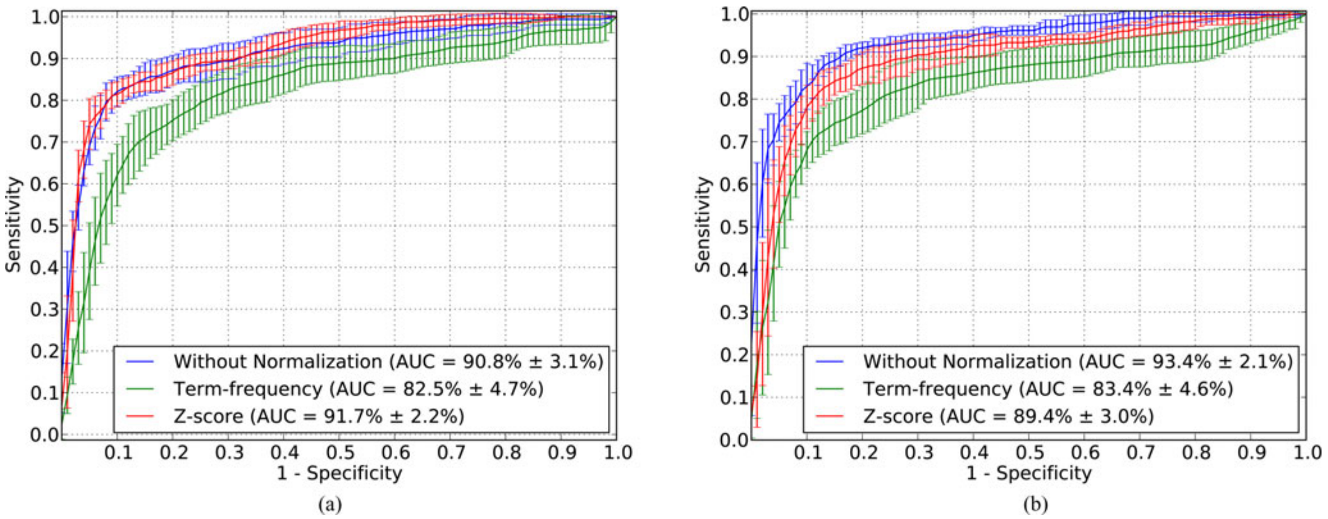

Fig. 6. Average behavior along with standard deviation of the ROC curves for the final referral decision across the 5 × 2 cross-validation protocol, when using HARD–SUM and SOFT–MAX mid-level BoVW features considering term-frequency and *z*-score normalization techniques as well as no-normalization. (a) Hard-assignment coding and sum pooling. (b) Soft-assignment coding and max pooling.

The metaclassifier without normalization performs better than with normalization using term frequency because term frequency ends up reducing the importance of some activations in the high-level feature vector when subtracting all entries by the sum of the feature vector decision scores. In the high-level descriptor, a high activation is actually interesting once the higher the decision score the more confident is the lesion detector regarding the outcome of that image. On the other hand, there is no difference when using the metaclassifier without normalization with respect to *z*-norm. The reason is that the SVM implementation used already performs a scaling of the input data so as to standardize it.

As for the SOFT–MAX approach, the disadvantage of the term-frequency technique is maintained. The result without normalization, 93.4% ± 2.1%, is significantly higher than the result with *z*-score normalization, 89.4 ± 3.0%. Fig. 6(b) depicts the average curves along with standard deviation for the SOFT–MAX approach considering the 5 × 2 cross-validation protocol.

The best results for HARD–SUM and SOFT–MAX, highlighted in Table II, indicate that normalization does not improve the final classification accuracy. Although the normalization with *z*-score provides an approximate outcome, the extra computation burden for computing the mean and standard deviation makes the method unnecessarily computationally more expensive than its counterpart with no normalization.

D. Comparison With the State of the Art

In a strategy similar to ours, Decencière *et al.* [19] combined heterogeneous information of individuals for devising a referral classifier. It is important to emphasize some crucial differences to our work: 1) the authors use parameters of quality assessment as descriptors, while we perform the analysis in a previous step (no quality analysis needed) nor any form of pre- or postprocessing; 2) we do not explore contextual information other than training images with annotated referral information; and 3) we perform a cross-dataset training/testing, a setup closer to a real scenario.

Considering the decision scores obtained with SOFT–MAX lesion detectors [see Fig. 6(b)], the results of the current classification experiments (sensitivity 90% and specificity 85%) were better compared to the results published by Decencière *et al.* (sensitivity 90% and specificity 50%), although caution is required as the testing scenarios and datasets were different. This translates to a greater reduction in unnecessary referral using the current methods.

Some recently published research, for example, [31] and [32], present higher AUCs. Barriga *et al.* [31] report an AUC of 98%, while Deepak and Sivaswamy [32] report a sensitivity of 100% and specificity of 97% for "cases needing immediate referral." However, in both cases, their definition of referral is detection of a specific DR lesion with high severity; for [32] the detection of images with risk 2 of macular edema in the Messidor dataset, and for [31], the detection of images with retinopathy grade 3 or risk of macular edema 2 also in Messidor.

Our definition of referable is more specific; it refers to the opinion of two experts that the patient should see a specialist within one year, *for whatever reason* not just the severity of a specific lesion.

V. Conclusion

Previous work has applied machine learning and computer vision techniques to detect DR-related lesions such as hard exudates, superficial hemorrhages, deep hemorrhages, cotton wool spots, drusen, and red lesions [2]. However, only the most traditional BoVW coding/pooling technique, hard-assignment coding, and sum pooling (HARD–SUM) was investigated. In this paper, we expand upon our previous work developing DR-related lesion detectors based on soft-assignment coding and max pooling. Our experiments show that, although the more recent SOFT–MAX scheme performs better for some of the lesions, the traditional HARD–SUM scheme is better for others, indicating that aided-diagnostic datasets have specificities unlike the ones of generic image recognition and evaluated Boureau *et al.* [20].

In addition to the new approach for characterizing DR-related lesions, the main novelty of this paper is the proposition of a fusion by metaclassification as a powerful tool for deciding whether a retinal image warrants a referral of the patient to a specialist within one year after screening. This metaclassification approach is based on the development of a new metaclassifier trained and tested with the decision scores generated by the individual lesion detectors. The scheme may be interpreted as the extraction of a high-level feature, composed of the scores of the individual lesion detectors, which is then used in the training and the decision of the metaclassifier.

Contrary to the expectation from previous results, normalization of the high-level feature vector scores did not improve on classification accuracy. Thus, future work could investigate more robust normalization techniques such as the w-score [28].

The best result achieved by our approach reached an AUC of 93.4% using the SOFT–MAX BoVW (soft-assignment coding/max pooling). These results indicate that our proposed method, applying a unified framework for lesion detection using soft-assignment coding and max pooling, has good potential for being used as a referral classification algorithm. This is especially the case as a cross-training/testing protocol was used for developing the methodology.

A unified framework is one where any new detector can be simply connected to the framework by creating its appropriate visual dictionary without the need for any pre- or postprocessing operation. The cross-training/testing protocol allows the use of retinal images to train individual lesion detectors which are different to the retinal images presented for classification that may have a different resolution, quality, or captured using different retinal cameras.

Classification whether a patient needs to be referred or not based on individual lesion detectors may lead to over-referral, which is unlikely using the proposed metaclassifier.

REFERENCES

[1] A. Rocha, T. Carvalho, H. Jelinek, S. Goldenstein, and J. Wainer, "Points of interest and visual dictionaries for automatic retinal lesion detection," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 8, pp. 2244–2253, Aug. 2012.

[2] H. Jelinek, R. Pires, R. Padilha, S. Goldenstein, J. Wainer, and A. Rocha, "Data fusion for multi-lesion diabetic retinopathy detection," in *Proc. IEEE Comput.-Based Med. Syst.*, 2012, pp. 1–4.

[3] 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.

[4] T. Peto and C. Tadros, "Screening for diabetic retinopathy and diabetic macular edema in the United Kingdom," *Curr. Diab. Rep.*, vol. 12, no. 4, pp. 338–345, 2012.

[5] A. Luckie, H. Jelinek, M. Cree, R. Cesar, J. Leandro, C. McQuellin, and P. Mitchell, "Identification and follow-up of diabetic retinopathy in rural health in Australia: An automated screening model," *Investigat. Ophtalmol. Visual Sci.*, vol. 45, no. 5, p. 5245, 2004.

[6] R. Pires, H. F. Jelinek, J. Wainer, and A. Rocha, "Retinal image quality analysis for automatic diabetic retinopathy detection," in *Proc. IEEE Conf. Graph., Patterns Images (SIBGRAPI)*, 2012, pp. 229–236.

[7] 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 *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2013, pp. 5857–5860.

[8] World Health Organization. (2012, Sep.). "Diabetes programme," [Online]. Available: http://www.who.int/diabetes/en

[9] N. Younis, D. M. Broadbent, S. P. Harding, and J. R. Vora, "Prevalence of diabetic eye disease in patients entering a systematic primary care-based eye screening programme," *Diabet. Med.*, vol. 19, pp. 1014–1021, 2002.

[10] H. Taylor, J. Xie, S. Fox, R. Dunn, A. Arnold, and J. Keeffe, "The prevalence and causes of vision loss in indigenous australians: The national indigenous eye health survey," *Med. J. Aust.*, vol. 192, no. 6, pp. 312–318, 2010.

[11] D. J. Pettitt, A. Okada Wollitzer, L. Jovanovic, G. He, and E. Ipp, "Decreasing the risk of diabetic retinopathy in a study of case management: The california medical type 2 diabetes study," *Diabetes Care*, vol. 28, pp. 2819–2822, 2005.

[12] H. F. Jelinek, A. Rocha, T. Carvalho, S. Goldenstein, and J. Wainer, "Machine learning and pattern classification in identification of indigenous retinal pathology," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2011, pp. 5951–5954.

[13] H. Jelinek and M. Cree, *Automated Image Detection of Retinal Pathology*. Boca Raton, FL, USA: CRC Press, 2010.

[14] P. J. Watkins, "ABC of diabetes: Retinopathy," *BMJ: Brit. Med. J.*, vol. 326, no. 7395, pp. 924–926, 2003.

[15] G. Quellec, M. Lamard, P. Josselin, G. Cazuguel, B. Cochener, and C. Roux, "Optimal wavelet transform for the detection of microaneurysms in retina photographs," *IEEE Trans. Med. Imag.*, vol. 27, no. 9, pp. 1230–1241, Sep. 2008.

[16] M. D. Abràmoff, J. M. Reinhardt, S. R. Russell, J. C. Folk, V. B. Mahajan, M. Niemeijer, and G. Quellec, "Automated early detection of diabetic retinopathy," *Ophthalmology*, vol. 117, no. 6, pp. 1147–1154, 2010.

[17] (2013, Feb.). NHS Diabetic Eye Screening Programme, [Online]. Available: http://diabeticeye.screening.nhs.uk

[18] 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," *Brit. J. Ophthalmol.*, vol. 94, no. 12, pp. 1606–1610, 2010.

[19] E. Decencière, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, and R. Danno, "Teleophta: Machine learning and image processing methods for teleophthalmology," *Ingénierie et Recherche Biomédicale*, vol. 34, pp. 196–203, 2013.

[20] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2559–2566.

[21] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1470–1477.

[22] E. A. do Valle Jr., "Local-descriptor matching for image identification systems," Ph.D. dissertation, Univ. Cergy-Pontoise École Doctorale Sciences et Ingénierie, Cergy-Pontoise, France, Jun. 2008.

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

[24] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. New York, NY, USA: Springer, 2006.

[25] J. P. Papa and A. Rocha, "Image categorization through optimum path forest and visual words," in *Proc. Int. Conf. Image Process.*, 2011, pp. 3586–3589.

[26] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2486–2493.

[27] J. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.

[28] W. Scheirer, A. Rocha, R. Micheals, and T. Boult, "Robust fusion: Extreme value theory for recognition score normalization," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 481–495.

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

[30] C.-C. Chang and C.-J. Lin. (2001). *LIBSVM: A Library for Support Vector Machines*, [Online]. Available: Software at http://www.csie.ntu.edu.tw/∼cjlin/libsvm

[31] 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 *Proc. IEEE Int. Symp. Biomed. Imag.: Nano Macro*, 2010, pp. 1349–1352.

[32] K. Deepak and J. Sivaswamy, "Automatic assessment of macular edema from color retinal images," *IEEE Trans. Med. Imag.*, vol. 31, no. 3, pp. 766–776, Mar. 2012.

Authors' photographs and biographies not available at the time of publication.