Retinal Image Quality Analysis for Automatic Diabetic Retinopathy Detection

Ramon Pires*, Herbert F. Jelinek†, Jacques Wainer* and Anderson Rocha*
*Institute of Computing, University of Campinas, UNICAMP, Campinas, Brazil
Email: pires.ramon@gmail.com, wainer@ic.unicamp.br, anderson.rocha@ic.unicamp.br
†Centre of Research of Complex Systems, Charles Sturt University, CSU, Albury, Australia
Email: hjelinek@csu.edu.au

***Abstract*—Sufficient image quality is a necessary prerequisite for reliable automatic detection systems in several healthcare environments. Specifically for Diabetic Retinopathy (DR) detection, poor quality fundus makes more difficult the analysis of discontinuities that characterize lesions, as well as to generate evidence that can incorrectly diagnose the presence of anomalies. Several methods have been applied for classification of image quality and recently, have shown satisfactory results. However, most of the authors have focused only on the visibility of blood vessels through detection of blurring. Furthermore, these studies frequently only used fundus images from specific cameras which are not validated on datasets obtained from different retinographers. In this paper, we propose an approach to verify essential requirements of retinal image quality for DR screening: field definition and blur detection. The methods were developed and validated on two large, representative datasets collected by different cameras. The first dataset comprises 5,776 images and the second, 920 images. For field definition, the method yields a performance close to optimal with an area under the Receiver Operating Characteristic curve (ROC) of 96.0%. For blur detection, the method achieves an area under the ROC curve of 95.5%.**

***Keywords*-Retinal Quality Assessment; Field Definition; Blur Detection.**

I. Introduction

Diabetes and associated complications including diabetic retinopathy (DR) is increasing with a predicted prevalence tripling by 2050 in the United States [1]. Developing countries and Indigenous populations are likely to exceed this percentage [2]. In addition, DR is the leading cause of blindness in developed countries and therefore screening and targeted case management programs that are economically viable and identify and implement early treatment are required [3].

Mobile screening of high-risk populations, especially in rural and remote locations is an effective means of increasing the screening coverage of DR prevention programs [4]. Two-field photography in the hands of photographers with diverse skill levels and irrespective of using mydriatic or nonmydriatic photography compares favorably to ophthalmic investigations by specialists in metropolitan clinics [5].

To further enhance rural and remote area screening, automated image analysis programs have been developed and are now in use as a first line screening for microaneurysms in Scotland [6]. Several algorithms have been proposed for detecting parts of the retina, the presence/absence of retinopathy as well as specific lesions from mild nonproliferative to proliferative retinopathy and maculopathy (see [7] and references therein). An important aspect of automated image analysis and the factor that successful image analysis relies on is *image quality*.

Assessing image quality has been discussed in the literature by a number of authors [8]–[11] and represents an important limiting factor for automated DR screening [12]. Image quality is reduced by artifacts in the image such as eye lashes or dust specs on the lens, only part of the retina is seen, the image is out-of-focus or the image is badly illuminated or blurred, among others. Image compression is often included with current software packages, which affects quality as does the resolution, field of view and type of camera [8]. Not directly related to image quality is retinal epithelial background, which often makes microaneurysm detection more difficult if the classifier is not trained for the specific ethnic group [13].

Furthermore, to ensure that automatic screening will be able to identify lesions like deep and superficial hemorrhages, it is necessary that the retinal images cover the appropriate portion of the retina, making the blood vessels visible. According to [14], the photographs should be centered on the macular region (See Fig. 2). Some authors have analyzed this aspect of image quality, known as *field definition* [15].

This paper proposes methods to verify these important factors of retinal image quality: *field definition* and *blur detection*. We aim at finding approaches that work well especially when trained with one type and tested with other types of retinal images. By introducing and adapting techniques such as visual words, quality analysis by similarity measures and classifier fusion to this context, we achieve promising classification results. In particular, for the field definition, our method is able to accurately distinguish between appropriate and inappropriate retinal images for automated DR screening.

II. Related work

Several methods for retinal image quality analysis are based on edge intensity histograms or luminosity to characterize the sharpness of the image [10]. In both approaches, the quality of a given image is determined through the difference between its histogram and the mean histogram of a small set of good-quality images used as reference.

Retinal morphology-based methods such as detection of blurring and its correlation to vessel visibility and retinal field definition have been applied for automatic detection of retinal image quality [9], [15]. The method of image assessment proposed by Fleming et al. [15], similarly to our work, involves two aspects: (1) image clarity and (2) field definition. The clarity analysis is based upon the vasculature of a circular area around the macula. The authors concluded whether or not a given image has enough quality using the presence/absence of small vessels in the selected circular area as evidence. The approach proposed by Fleming et al. requires a segmentation step to find the region of interest. However, for low-quality images, detecting segmentation failures is trivial.

Niemeijer et al. [11] proposed a method for image quality verification that is comparable to the well-known visual words dictionary classification technique, used extensively in pattern recognition tasks [16] and also one of the methods we rely upon in this paper. The purpose of Niemeijer et al. was to identify image structures that were present in a set of images. Local image structure at each pixel is described using the outputs of a set of 25 filters. Because the raw features are too numerous to be used directly in the classification process, a clustering algorithm is used to express the features in a compact way creating a visual dictionary. Once the visual dictionary is built, the features of each pixel are mapped to words and a histogram of word frequencies for each image is created. These histograms are used to feed a classifier.

Visual words dictionaries constitute one of the approaches proposed to analyze image quality in this work. However, different to [11] we utilize visual words in the space of features representing discontinuities in the retina and not directly on every pixel. Second, our method is based on points of interest which are reasonably robust to some image distortions (e.g., rotation) and exhibit high repeatability, which allows us to easily find similar discontinuities in different images. Third, we have used the same method to detect lesions associated with DR in previous work of ours [17]. Finally, the visual words dictionary calculated on the space of features exploits the benefits of an all-in-one classification algorithm which does not require any pre- or post-processing of the image.

Although good results for the assessment of diabetic retinal image quality have been obtained previously, the authors have not paid attention to one crucial factor needed for an acceptable screening of diabetic retinopathy. The image has to encompass the correct portion of the retina [14]. An analysis of DR images can fail because of inadequate field definition. As one exception, Fleming et al. [15] reported retinal image field definition in their work. In the viewpoint of the authors, an image is defined as having adequate field definition if it satisfies a series of constraints, that aim at verifying distances between important elements of the anatomy of the retina, such as the optic disc and fovea (top left of Fig. 2).

III. Technique for Field Definition

Here, we discuss a simple method to verify the field definition. In this problem, a good retinal image for further DR analysis is one image centered on the macula (See Fig. 2).

The method we discuss here operates based on the methodology of full-reference comparison. In this methodology, a reference image with assured quality is assumed to be known and quantitative measures of quality for any image are extracted by comparisons with the reference [18]. Given that the macular region has a distinguishable contrast in comparison with the remaining regions, and we are interested in the content of the center of retinal images, metrics of similarity have shown to be highly suitable for this objective.

We selected a set of images centered on the macular region as well as a set of images not centered on the macular region (centered on the optic disc or in any other location on the retina). Then, we calculated similarities between a given image and the reference images (positive and negative), with respect to their central regions and created a feature vector for later classification. In the next section, we explain the method employed for the feature extraction as well as the learning step of the technique for field definition.

A. Characterization

Wang et al. [18] proposed a new philosophy for comparison of images that considers image degradation as perceived changes in structural information instead of perceived errors (visibility of errors). The method, known as Structural Similarity (SSIM) [18] is calculated according to Eq. 3 which we shall define later.

Given that we are interested in assessing if the macula is present in the center of the image and it is clearly different from other regions of the retina, we use one region of interest (RoI) of pre-defined size (121×121) on the center of the retinal image. Fig. 1 depicts some positive (centered on the macular region) and negative (centered on the optic disc or in other region) RoIs.

To characterize each retinal image, we measure the structural similarity between the RoI of the image of interest and the RoIs of a set of reference images and calculate their average. We selected a set of 40 retinal images for reference (20 represent the retina with good field definition and 20 that would be discarded for not being centered on the macula). For the group not centered on the macula, we selected 12 RoIs centered on the optic disc and eight in any other area. The reference images are not used further neither for training nor for testing.

As we are comparing pixels directly, we investigated if a simple contrast normalization technique helps to boost classification results. For that, we tested the use of the images in grayscale as well as in RGB color space with and without the normalization considering contrast limited adaptive histogram equalization (CLAHE) [19]. CLAHE is suitable to improve the local contrast of an image.

After comparing each image with the references, its feature vector considering color images comprises 18 features: three comparison functions from SSIM $\times$ three color channels (RGB) $\times$ two sets of reference patches (positive and negative).

SSIM was calculated breaking Eq. 3 to three terms: luminance, contrast, and structure according to [18].

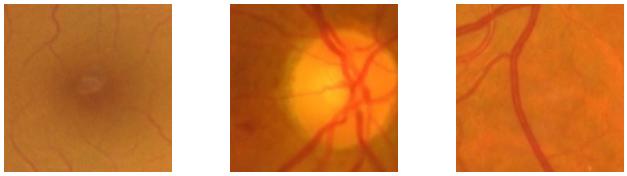

Fig. 1. Examples of RoIs whose images are centered on the macula (left), centered on the optic disc (middle), and non-representative (right).

B. Learning

At the end of the characterization process, we have a set of feature vectors representing the structural similarities with positive and negative reference images. The final classification procedure is performed using the *Support Vector Machine* (SVM) algorithm [20]. We train the classifier with feature vectors calculated from training images containing positive (images centered on the macular region) and negative (images centered on any other region of the retina) examples. When training the SVM, we use "grid search" for fine tuning the SVM parameters based only on the training examples [20].

IV. Technique for Blur Detection

Although image quality analysis can have several ramifications before arbitrating on the quality of an image, we focus on two very common problems during image acquisition: blurring and out-of-focus capture.

A. Characterization

The method involves a series of different blurring classifiers and classifier fusion to optimize the classification. Next, we present the details of the methods we use for blurring classification. Basically, we rely upon four descriptors: vessel area, visual dictionaries, progressive blurring and progressive sharpening. We also explore combinations of them.

Area Descriptor: Given that blurring affects the visibility of the blood vessels, our first descriptor consists of the measurement of the area occupied by the retinal vessels. For that, we calculate the image's edge map using the Canny algorithm [21]. Next, we measure the area occupied by the vessels counting the quantity of pixels on the edges and dividing it by the retina's total number of pixels. Fig. 2 depicts retinal images followed by their respective Canny edge maps.

In the end of the characterization phase, we have an 1-d feature vector whose area descriptor is the unique feature.

Visual Dictionary Descriptor: In this descriptor, each image is characterized by finding stable points of interest (PoIs) across multiple image scales that capture image discontinuities. We are interested in characterizing an image in order to capture any inconsistencies/discontinuities it might have (e.g., blood vessels) in order to classify it.

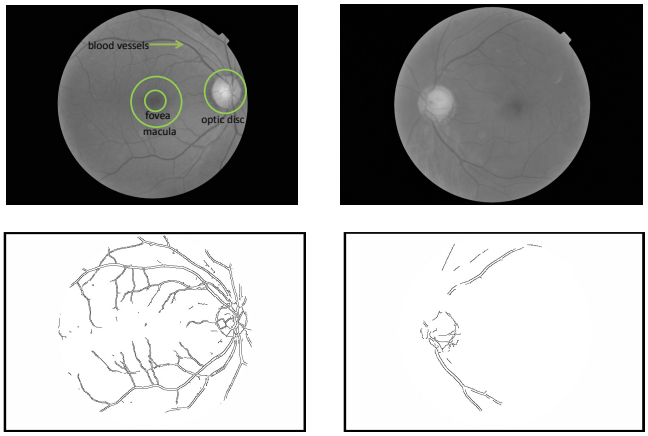

Fig. 2. Retina with enough quality (left) and with blurring (right) with their respective Canny edge maps (inverted for visualization purposes). The top left image also shows the typical elements present in the retina.

To build a visual dictionary and define whether a specific retinal image has enough quality, training images tagged as having quality (no blur) by a medical specialist as well as images associated with blurring are required. After collecting the training images, the next step consists of finding the points of interest in all training images. To detect the points, we use the *Speeded Up Robust Features* (SURF) [22] as it is a good feature detector with reasonable speed.

From the points of interest representing the images with quality as well as blurred images during training, we randomly select a set of PoIs for each group. At this stage the number of PoIs (k) to be retained as representative of the quality or non-quality images is decided. We find the $k/2$ points of interest associated with a high-quality image and repeat the process to find the $k/2$ points associated with images with blurring. We refer to these k points of interest as a visual dictionary. Note that this is different from other approaches in the literature (e.g., [23], [24]) which normally find a global unique dictionary and not one per class. In our experience, class-based dictionaries are more appropriate for retinal images.

In order to use any machine learning method, the next step is to map the PoIs within each image to the most representative points in the dictionary. For each image, we associate each one of its PoIs to the closest word in the dictionary using Euclidean distance. In the end, each training image is represented by a histogram of k bins which counts the number of times each PoI in the image was mapped to that word in the dictionary. We used such histogram as the image's feature vector. During testing, the process is simple: we extract the points of interest of the test image and map its PoIs to the dictionary creating its k dimensional feature vector.

Determining the optimal number of clusters for any given set is still an open problem and is therefore best determined empirically. In our experiments, we evaluated the performance of the visual dictionary descriptor with k = 30, 50, 70, 100 and 150. We avoided bigger dictionaries in order to keep the classification process fast and accurate.

Blurring, Sharpening, Blurring + Sharpening descriptors: We propose a variation of the traditional method widely employed in the literature to quantify the visibility of errors: full-reference method for assessment of quality [18]. In our variation, the reference image is not defined previously, but each image under analysis is elected as a reference and compared to progressive transformations of itself.

For the blurring descriptor, we progressively blur the input image with different intensities and measure how much the image can lose the discontinuities that characterize the blood vessels. It is expected that an image with poor quality be more similar to its transformed version than a good-quality image in comparison with its transformed version.

For the sharpening descriptor, we employ different unsharpening filters that enhance edges and provide higher similarity values for good-quality images than for blurred images. The unsharpening filter is a simple sharpening operator which enhances edges (and other high frequency components in an image) via a procedure which subtracts a smoothed version of an image from the input image.

To explore simultaneously the two features, we investigated a Blurring + Sharpening descriptor to represent retinal images.

Each input retinal image is considered as a reference image and is compared with its filtered images. For that, we define a filter-bank as a set of rotationally symmetric Gaussian lowpass filters $G_\sigma(i,j)$. The set comprises 12 filters with kernel sizes $k_s \times k_s$ where $k_s \in \{3,5,7\}$, and standard deviations $\sigma \in \{0.5, 1.5, 3.0, 4.5\}$.

For the blurring descriptor, each resulting image $f^i_{smooth}(x,y)$ is a filtered version of the original image $f(x,y)$, denoted as

$$f^i_{smooth}(x,y) = \sum_{i,j}^{k_s} G_\sigma(i,j) f(x+i, y+j) \quad (1)$$

For the sharpening descriptor, each resulting image $f^i_{sharp}(x,y)$ is calculated as

$$f^i_{sharp}(x,y) = f(x,y) + \lambda(f(x,y) - f^i_{smooth}(x,y)) \quad (2)$$

where λ is a scaling constant $\in [0.0, 1.0]$. Here, we fixed the constant, $\lambda = 0.7$ without any further analysis.

For each retinal image, we measured the similarity between the input image $f(x,y)$ (considered as reference) and each response image $f^i(x,y)$ blurred or sharpened according to the descriptor of interest. We calculated the similarity $sim(f(x,y), f^i(x,y))$ using three different metrics:

- SSIM: the structural similarity index between two images can be viewed as a quality measure of one of the images being compared, provided the other image is regarded as of good quality. We calculated SSIM for 11×11 windows centered on every pixel. The result is a matrix with the same dimensions as the compared images. We report the final similarity value as the average of such matrix. The $SSIM(R,S)$ where R and S are two 11×11 windows centered on a pixel (x,y) is given by

$$SSIM(R,S) = (2\mu_R\mu_S + c_1)(2\sigma_{RS} + c_2) \times 1/[(\mu_R^2 + \mu_S^2 + c_1)(\sigma_R^2 + \sigma_S^2 + c_2)] \quad (3)$$

  where μ_R and μ_S are the average of R and S regions, σ_R^2 and σ_S^2 their variances, σ_{RS} their covariance, c_1 and c_2 are two variables to stabilize the division with weak denominator. These variables depend upon two constants $k \ll 1$ ($k_1 = 0.01$ and $k_2 = 0.03$) and the image's dynamic range L which is 255 in our case. The final values, for $c = (k * L)^2$, are: $c_1 = 6.5$ and $c_2 = 58.5$.

- SSD: the sum of squared differences is calculated by subtracting pixels between the reference image $f(x,y)$ and the target image $f^i(x,y)$. The differences are squared.

$$SSD(f(x,y), f^i(x,y)) = \frac{1}{MN} \sum_{x,y} [f(x,y) - f^i(x,y)]^2, \quad (4)$$

  where M and N are the number of rows and columns.

- NCC: the normalized cross correlation is defined as

$$NCC(f(x,y), f^i(x,y)) = \frac{1}{MN} \sum_{x,y} \frac{f(x,y) f^i(x,y)}{\sqrt{f(x,y)^2}\sqrt{f^i(x,y)^2}}, \quad (5)$$

  where M and N are the number of rows and columns.

For each image, the blurring and the sharpening descriptors have feature vectors with 108 similarity measures: 12 gaussian filters $\times$ 3 metrics of similarity $\times$ 3 color channels (RGB). The blurring + sharpening descriptor is the concatenation of the feature vectors extracted by the blurring and the sharpening descriptors leading to a 216-d feature vector.

B. Learning

In the end, for each retinal image, we have a set of five feature vectors considering the area descriptor, visual dictionary descriptor, blurring and sharpening descriptors and their concatenation. The final classification procedure is performed using the SVM algorithm [20]. We trained the classifier with feature vectors calculated from training images containing positive (images tagged by a medical specialist as good quality) and negative (images tagged by a medical specialist as containing blur) examples. When training the SVM, we use "grid search" for fine tuning the SVM parameters based only on the training examples [20].

C. Fusion

It is possible that a series of complementary classifiers are more suited to accurately assess the quality of retinal images operating over several instances observed in the two classes of images. For example, analyzing not only one characteristic, but a series as the area occupied by visible blood vessels, the distributions of positive/negative visual words, similarities with blurred images and similarities with sharpened images provide a higher probability of correctly evaluating any retinal image from any camera.

We evaluated two approaches for fusion: at feature-level combining the feature vectors directly by concatenation and at classifier level by creating a Meta-SVM classifier trained over the outputs of individual classifiers, in this case, the marginal distances to the decision hyperplane produced by the SVMs.

V. Experiments and validation

This section shows the results for evaluating the quality of an image with respect to field definition and blurring artifacts as an effective pre-processing before using any classifier for detecting diabetic retinopathy lesions.

There are many metrics to measure the success of a detection/classification algorithm. For the purposes of this project, we are interested in *per image* metrics, such as sensitivity (number of images tagged as having enough quality over the total number of images with quality), and specificity (number of images tagged as blurred over the total number of blurred images). However, for quantifying the performance of the proposed methods, we calculated the area under the receiver operating characteristic curve (ROC). The area under the curve (AUC) is an accuracy measurement that explores how well the classifier is based on its ROC curve. An AUC of 100% represents a perfect test while an area of 50% represents a worthless test.

We organized the experiments in four rounds:

- **Round #1 – Single results for field definition.** Field definition approach using single classifiers. We performed all tests on single datasets using 5-fold cross-validation.
- **Round #2 – Cross-dataset results for field definition.** Cross-dataset approach, in which we trained the field definition classifiers in one dataset and test in another. We evaluated the ability of the field definition system to operate over images from different acquisition conditions.
- **Round #3 – Single results for blur detection.** Blur classification using single classifiers. We also evaluated fusion methods to check if they improved the classification results. We performed all tests on single datasets using 5-fold cross-validation.
- **Round #4 – Cross-dataset results for blur detection.** Cross-dataset approach, in which we trained the blur classifiers in one dataset and tested in another. We evaluated the ability of the blur classifiers to operate over images from different acquisition conditions.

In the 5-fold cross-validation protocol, we split the dataset into five parts, train with four parts and test on the fifth, repeating the process five times each time changing the training and testing sets.

A. Datasets

We performed the experiments for quality analysis using the DR1 and DR2 datasets annotated by medical specialists.

The DR1 dataset is from the ophthalmology department of Federal University of São Paulo (Unifesp), collected during 2010. It comprises 5,776 images with an average resolution of 640 $\times$ 480 pixels. 1,300 images have good quality (do not contain blur and are correctly centered on the macula), 1,392 represent poor quality (blur) and 3,084 are diagnosed as images of the periphery (not centered on the macula). Three medical specialists manually annotated all of the images. The images were captured using a TRC-50X (Topcon Inc., Tokyo, Japan) mydriatic camera with maximum resolution of one megapixel and a field of view of 45 degrees.

The DR2 dataset is from the same ophthalmology department, collected during 2011. One medical specialist graded the images. DR2 comprises 920 12.2MP images decimated to 867 $\times$ 575 for speed purposes and containing 260 images not centered on the macula (146 centered on the optic disc and 114 not centered on any interesting region) and 660 images centered on the macula (466 good and 194 low quality). The images were captured using a TRC-NW8 retinographer with a Nikon D90 camera.

For more details and for downloading the datasets, please refer to http://www.recod.ic.unicamp.br/site/asdr.

B. Round #1: Single Results for Field Definition

Here, we explore the measures of structural similarity in order to create a classifier able to analyze a retinal image and evaluate if it comprises the correct portion for diabetic retinopathy screening (centered on the macula).

We performed four experiments for field definition. In the first experiment, the images were analyzed in grayscale. The second experiment also was performed with the images in grayscale, but after an adaptive histogram equalization (CLAHE). Next, we considered the case of color images with and without histogram equalization.

For all experiments of field definition, we used 40 reference images. All of them were not considered further for training and for testing.

Fig. 3 and Fig. 4 depict the ROC curves for the field definition approach using 5-fold cross-validation protocol of the DR1 and DR2 datasets, respectively.

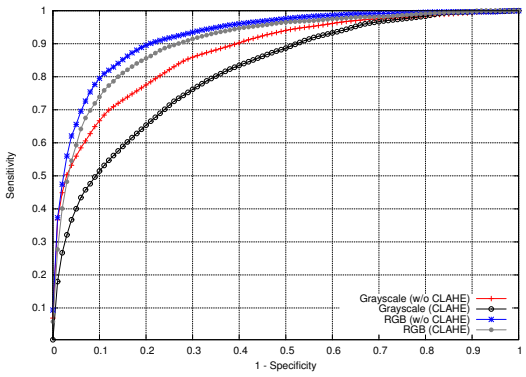

Fig. 3. DR1 field definition using 5-fold cross-validation.

As we can observe in Fig. 4, the method achieves reasonably successful results for field definition. The experiments using the DR2 dataset present even better results. The experiment with color images considering histogram equalization provides the best result, but this result in not statistically different to the others in DR2. However, in the experiments using the DR1 dataset (Fig. 3), that comprises a larger quantity of images (1,300 positives and 3,084 negatives), we can note a great difference of AUCs between the different techniques. The method that uses the color images without requiring an adaptive histogram equalization is the highlight.

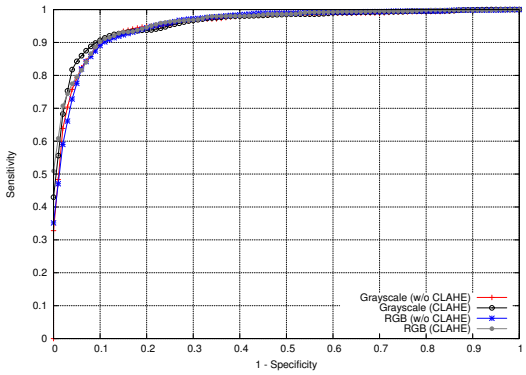

Fig. 4. DR2 field definition using 5-fold cross-validation.

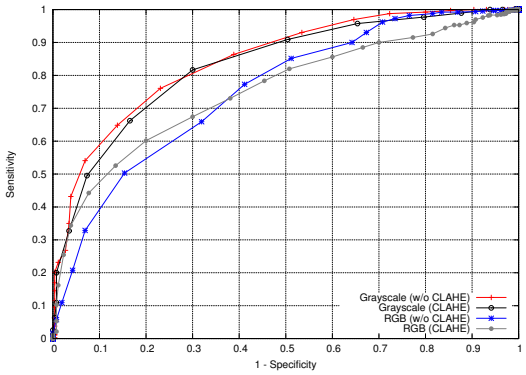

Fig. 5. Cross-dataset validation for field definition using DR1 as training and DR2 as testing sets.

As mentioned, there is not a considerable difference between the experiments with and without adaptive histogram equalization using the DR2 dataset. The reason is that the images from DR2 present small variations in illumination. The images from DR1 dataset exhibit a high variation of illumination making the CLAHE insufficient to distinguish them and improve classification.

C. Round #2: Cross-dataset Results for Field Definition

Conventional detectors usually build a classifier from labeled examples and assume the testing samples are generated from the same distribution. When a new dataset has a different distribution from the training dataset (e.g., different acquisition conditions), the performance may not be as expected.

In this round, we validated the field definition approaches considering the problem of cross-dataset field definition testing, which aims at generalizing field definition models built from a source dataset to a target dataset. We refer the DR1 as the source dataset (training), and the DR2 as the target dataset (testing). We emphasize that the two datasets were collected in very different environments with different cameras, at least one year apart and in different hospitals.

For this round, we trained the classifiers with DR1 dataset (3,064 images located on the periphery of the retina, 1,280 images centered on the macula and 40 images removed and used as reference), and tested with DR2 dataset (260 images not centered on the interest region and 660 images centered on the macular region).

Fig. 5 presents the ROC curves achieved by the method under the cross-dataset validation.

As discussed in the previous section, the high variation of the illumination in DR1 in comparison with DR2 makes the histogram equalization technique unable to improve the results. Table I summarizes the results for field definition for the single and cross-dataset tests.

Comparison with state-of-the-art: in a previous work, Fleming et al. [15] introduced the first automatic field definition study. The authors obtained 95.3% for sensitivity and 96.4% for specificity. Our results for field definition are somewhat comparable to the previous results (96% AUC, and 93% sensitivity and 92% specificity using DR2 and RGB-CLAHE). However, Fleming at al. used a different dataset with 1,039 retinal images and did not evaluate the algorithms in a cross-dataset scenario.

TABLE I
FIELD DEFINITION: AUC FOR THE EXPERIMENTS.

Method	DR1	DR2	Cross
Grayscale	87.6%±0.7%	95.5%±1.3%	84.7%
Grayscale (CLAHE)	81.5%±0.6%	95.9%±1.2%	83.2%
RGB	92.5%±0.7%	95.4%±1.1%	75.5%
RGB (CLAHE)	90.5%±0.9%	96.0%±0.8%	75.5%

D. Round #3: Single Results for Blur Detection

In the third round, we performed experiments to verify the descriptors and classifiers to separate good-quality images from blurred ones. We explored several descriptors, each one trying to take full advantage of the differences observed between poor and good-quality images, aimed at providing a series of blur classifiers. In this experiment, we developed classifiers that work in parallel, assuming competitive operation and contributing equally to the final decision.

Fig. 6 and Fig. 7 depict the results for DR1 and DR2 datasets.

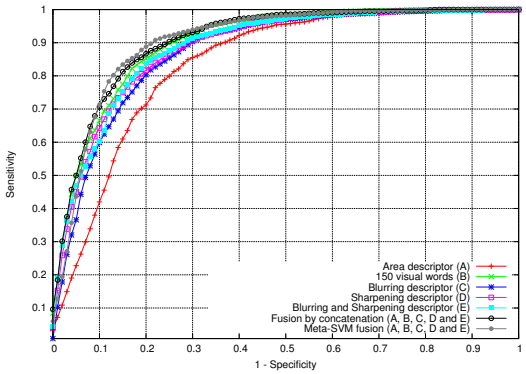

Fig. 6. DR1 blur classification using 5-fold cross-validation.

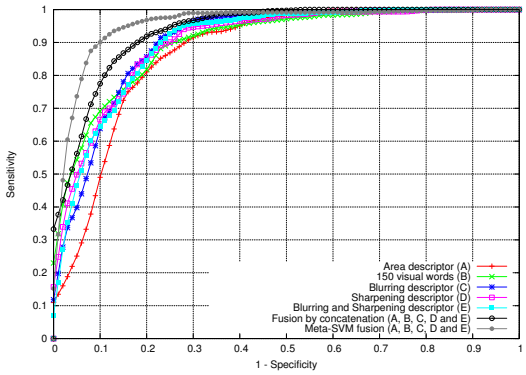

Fig. 7. DR2 blur classification using 5-fold cross-validation

Table II summarizes the results. The ROC curves as well as the areas under the curves reflect that interesting results are obtained for blur classification. We observe in the table that, for single classifiers, the best result using the DR1 dataset was achieved by the visual words approach (a dictionary size of 150 words was previously defined as the best number of words for the dictionary and not shown here). For the DR2 dataset, the visual words approach also presents good results but are outperformed by classifiers trained with the blurring and sharpening descriptors. The blurring, sharpening and the blurring + sharpening descriptors provide acceptable results in both datasets.

As expected, the more exciting results were provided by the fusion methods. As discussed before, exploring not only one evidence of incoherence, but several complementary information of poor and good-quality images, gives more chances of obtaining better results. In our case, the ensemble method that uses only the concatenation of the feature vectors provides the highest result for DR1 (AUC = 90.8%), followed closely by the Meta-SVM fusion method (AUC = 90.7%).

Here, it is important to emphasize that the ensemble by concatenation operates on large feature vectors making the method highly sensitive to the curse of dimensionality, and presents limitations for classification for specific classifiers and specific machines [25]. In addition, it is often necessary to deal with complicated normalization techniques to put different features in the same domain [25]. Conversely, the Meta-SVM fusion method is less subject to such limitations, since it only adds a new level of classification on a response vector composed of five classification scores (distances to the decision hyperplane) provided by the individual classifiers.

For the DR2 dataset, the highest AUC was obtained with a large difference using the Meta-SVM fusion method (AUC = 95.5%), followed by the fusion by concatenation technique (AUC = 93.4%).

E. Round #4: Cross-dataset Results for Blur Detection

The last round of experiments explored the cross-dataset validation to evaluate how the classifier models built from a source dataset (DR1) to a target dataset (DR2) generalize.

TABLE II
BLUR DETECTION: AUC FOR THE EXPERIMENTS.

Descriptor/Fusion	DR1	DR2	Cross
Area	83.9%±2.4%	87.2%±2.6%	87.1%
Visual words	**90.3%±1.2%**	**90.3%±2.3%**	**85.6%**
Blurring	87.6%±1.3%	90.3%±2.6%	60.8%
Sharpening	88.8%±1.4%	90.3%±3.9%	83.9%
Blurring and Sharpening	89.0%±0.9%	90.2%±3.0%	69.0%
Fusion by Concatenation	**90.8%±0.9%**	**93.4%±1.4%**	**87.0%**
Fusion by Meta-SVM	**90.7%±2.3%**	**95.5%±1.6%**	**87.5%**

For this round, we trained the classifiers with DR1 (1,392 images with poor quality and 1,300 images with good quality) and tested the classifiers with DR2 dataset (194 retinal images with enough quality and 660 images with no quality). Fig. 8 depicts the resulting ROC curves.

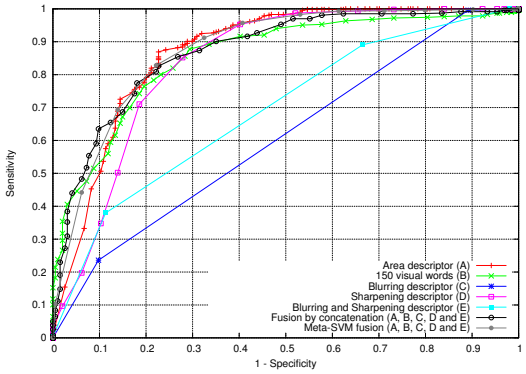

Fig. 8. Cross-dataset validation for blur classification using DR1 as training and DR2 as testing sets.

Observing the AUCs in Fig. 8 and summarized in Table II, we note that the visual words descriptor presents satisfactory results using the cross-dataset protocol. However, the simple area descriptor is the highlight in this experiment, showing that the density of blood vessels may be considered as an acceptable approach to assess the quality of retinal images.

Fortunately, with this experiment we can show the importance of a cross-dataset validation protocol. Although the blurring descriptor showed interesting results in the validation with single datasets, here it failed along with blurring + sharpening combination. With them, a large number of images from the DR2 dataset was classified at the same distance to the SVM decision hyperplane. This fact happens because the DR1 has greater contrast and illumination variation than DR2 dataset and, therefore, the descriptions of the DR2 match to approximate scores given by a classifier trained with DR1. Consequently, a small amount of operating points are available, as we can see in Fig. 8. This effect might be reverted using image normalization techniques more complex than CLAHE but we did not investigate this in this paper.

As we expected, detector fusion with the Meta-SVM method provides the best AUC with the caveat that in this analysis the Meta-SVM results are not statistically better than the single classifier using the single area descriptor.

Comparison with state-of-the-art: our results are comparable to several prior results. The approach proposed by Niemeijer et. al. [11] and explained in Sec. II provided an AUC of 99.6% operating over a dataset comprising 1,000 images. Davis et. al. [8] achieved a sensitivity of 100.0% and a specificity of 96.0% using a dataset comprising 2,000 images. However, no conclusion can be drawed observing only the final results, since we must consider that the datasets are different (camera model, acquisition conditions) and the methodologies employed are distinct. We emphasize that only one dataset is not enough as a validation protocol for a reliable system.

VI. Conclusions and Future Work

The assessment of diabetic retinal image quality presented in this paper shows promising results. Several studies have obtained satisfactory results for image quality verification in the literature. However, these have only focused on image quality as a generalized approach and have not paid attention to field definition, which is one crucial factor for an effective automatic screening of diabetic retinopathy. In addition, cross-dataset validation is hardly performed.

In the approach we discuss in this paper, image quality was defined by two aspects: field definition and blur analysis. For field definition, we proposed the use of structural similarity measures to evaluate the quality of retinal images. We obtained an AUC of 96.0% using color images and the DR2 dataset.

For blur analysis, we explored several descriptors, each one taking full advantage of the specific variations between poor and good-quality images. Furthermore, we aimed at providing a series of blur classifiers that work in parallel, assuming competitive operations and contributing equally to the final decision. We also evaluated the use of fusion techniques and the best result was reached with the Meta-SVM fusion method (AUC = 95.5% on DR2 dataset).

With the proposed methods for assessment of diabetic retinal images, it is possible to devise and deploy a system capable of robustly identifying images with low quality and, afterwards, discard them. A retinal camera equipped with quality assessment methods would be adequate to analyze fundus images taken in real-time, preventing misdiagnosis and posterior retake.

Our future works include building lesion-based classifiers specialized in the detection of single anomalies, and investigating methods to combine the single detectors, providing a final high-level classifier able to label a retinal image according the presence/absence of any DR lesion.

Acknowledgment

We would like to thank Microsoft Research and the São Paulo Research Foundation (FAPESP) for the financial support. We also thank Dr. Eduardo Dib for technical assistance with image acquisition.

References

[1] J. Saaddine, A. Honeycutt, K. Narayan, X. Zhang, R. Klein, and J. Boyle, "Projection of diabetic retinopathy and other major eye diseases among people with diabetes mellitus: United states, 2005-2050," *Arch Ophthalmol.*, vol. 126, no. 12, pp. 1740–1747, 2008.

[2] G. Spurling, D. Askew, N. H. N. Hansar, A. Cooney, and C. Jackson, "Retinal photography for diabetic retinopathy screening in indigenous primary health care: the inala experience," *Australian and New Zealand Journal of Public Health*, vol. 34, pp. S30–S33, 2010.

[3] D. Pettitt, A. Wollitzer, L. Jovanovic, H. Guozhong, and I. Eli, "Decreasing the risk of diabetic retinopathy in a study of case management: the california medical type 2 diabetes study," *Diabetes Care*, vol. 28, no. 12, pp. 2819–2822, 2005.

[4] P. Bragge, R. Gruen, M. Chau, A. Forbes, and H. Taylor, "Screening for Presence or Absence of Diabetic Retinopathy: A Meta-analysis," *Arch Ophthalmol.*, vol. 129, no. 4, pp. 435–444, 2011.

[5] D. Maberley, A. Morris, D. Hay, A. Chang, L. Hall, and N. Mandava, "A comparison of digital retinal image quality among photographers with different levels of training using a non-mydriatic fundus camera," *Ophthalmic Epidemiology*, vol. 11, no. 3, pp. 191–197, 2004.

[6] S. Philip, A. Fleming, K. Goatman, S. Fonseca, P. Mcnamee, G. Scotland, G. Prescott, P. Sharp, and J. Olson, "The efficacy of automated disease/no disease grading for diabetic retinopathy in a systematic screening programme," *British Journal of Ophthalmology*, vol. 91, no. 11, pp. 1512–1517, 2007.

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

[8] H. Davis, S. Russell, E. Barriga, M. Abramoff, and P. Soliz, "Vision-based, real-time retinal image quality assessment," in *IEEE CMBS*, 2009, pp. 1–6.

[9] L. Giancardo, F. Meriaudeau, T. Karnowski, E. Chaum, and K. Tobin, *New Developments in Biomedical Engineering*. InTech, 2010, ch. Quality Assessment of Retinal Fundus Images using Elliptical Local Vessel Density, pp. 201–224.

[10] M. Lalonde, L. Gagnon, and M.-C. Boucher, "Automatic visual quality assessment in optical fundus images," *Vision Interface*, pp. 259–264, 2001.

[11] M. Niemeijer, M. Abramoff, and B. van Ginneken, "Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening." *IEEE Med. Image Analysis*, vol. 10, no. 6, pp. 888–898, 2006.

[12] N. Patton, T. Aslam, T. MacGillivray, I. Deary, B. Dhillon, R. Eikelboom, K. Yogesan, and I. Constable, "Retinal image analysis: concepts, applications and potential," *Progress in Retinal and Eye Research*, vol. 25, no. 1, pp. 99–127, 2006.

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

[14] K. Facey, *Health Tech. Assessment: Organisation of services for diabetic retinopathy screening*. Health Tech. Board for Scotland, 2002.

[15] A. Fleming, S. Philip, K. Goatman, J. Olson, and P. 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.

[16] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *IEEE ICCV*, 2005, pp. 1800–1807.

[17] J. Herbert, R. Pires, R. Padilha, S. Goldenstein, J. Wainer, T. Bossomaier, and A. Rocha, "Data fusion for multi-lesion diabetic retinopathy detection," in *IEEE EMBS*, 2012.

[18] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[19] S. Pizer, E. Amburn, J. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. Romeny, and J. Zimmerman, "Adaptive histogram equalization and its variations," *Comput. Vision Graph. Image Process.*, vol. 39, no. 3, pp. 355–368, Sep. 1987.

[20] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. on Intelligent Systems and Tech.*, vol. 2, pp. 27:1–27:27, 2011.

[21] R. Gonzalez and R. Woods, *Digital Image Processing (3rd Ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006.

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

[23] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *IEEE ICCV*, 2003, pp. 1470–1477.

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

[25] A. Rocha, J. Papa, and L. Meira, "How far do we get using machine learning black-boxes?" *Intl. Journal of Pattern Recognition and Artificial Intelligence*, pp. 1–8, 2012.