A Function for Quality Evaluation of Retinal Vessel Segmentations

Manuel Emilio Gegúndez-Arias, Arturo Aquino, José Manuel Bravo, Diego Marín

M.E. Gegúndez-Arias is with the Department of Mathematics, "La Rábida" High Technical School of Engineering, University of Huelva, Spain (e-mail: gegundez@uhu.es).

A. Aquino, J.M. Bravo and D. Marín are with the Department of Electronic, Computer Science and Automatic Engineering, "La Rábida" High Technical School of Engineering, University of Huelva, Spain (e-mail: {arturo.aquino, caro, diego.marin}@diesia.uhu.es).

Abstract

Retinal blood vessel assessment plays an important role in the diagnosis of ophthalmic pathologies. The use of digital images for this purpose enables the application of a computerized approach and has fostered the development of multiple methods for automated vascular tree segmentation. Metrics based on contingency tables for binary classification have been widely used for evaluating the performance of these algorithms. Metrics from this family are based on the measurement of a success or failure rate in the detected pixels, obtained by means of pixel- to-pixel comparison between the automated segmentation and a manually-labeled reference image. Therefore, vessel pixels are not considered as a part of a vascular structure with specific features. This paper contributes a function for the evaluation of global quality in retinal vessel segmentations. This function is based on the characterization of vascular structures as connected segments with measurable area and length. Thus, its design is meant to be sensitive to anatomical vascularity features. Comparison of results between the proposed function and other general quality evaluation functions shows that this proposal renders a high matching degree with human quality perception. Therefore, it can be used to enhance quality evaluation in retinal vessel segmentations, supplementing the existing functions. On the other hand, from a general point of view, the applied concept of measuring descriptive properties may be used to design specialized functions aimed at segmentation quality evaluation in other complex structures.

Index terms: Ophthalmic pathologies diagnosis, retinal vessel segmentation, image segmentation quality evaluation, quality evaluation function.

1. INTRODUCTION

Digital eye-fundus images are widely used nowadays for computerized detection of ophthalmic pathologies. Specifically, blood vessel assessment through segmentation into retinal images is an important diagnosis key for automatic detection and evaluation of multiple pathologies leading to vascular anomalies. Some of the main applications reported for vessel segmentation include: location of other fundus features such as the optic disc [1]–[3] and fovea [4], reduction of the number of false positives in the detection of microaneurysms and haemorrhages [5], [6], extraction of reference vascula- ture points for image registration [7], [8], evaluation of the retinopathy of prematurity [9], arteriolar narrowing [10], [11], vessel tortuosity to characterize hypertensive retinopathy [12], vessel diameter measurement for the diagnosis of hypertension and cardiovascular diseases [13], [14], and computer-assisted laser surgery [15], [16].

As a result of this interest, many automated methods for vascular tree segmentation have been reported over the last years [9], [17]–[35]. Quality evaluation in the resulting seg- mentations is an important issue. The difficulties involved by this task have already been pointed out and discussed by Niemeijer et al. [35]. The methods of algorithmic performance on a fundus image are usually quantified by measuring metrics based on contingency tables for binary classification [36]. The most commonly-used metrics from this family are sensitivity (Se), specificity (Sp) and accuracy (Acc) [9], [18], [22], [24], [26]–[33], [35]. While Se and Sp metrics are the ratio of well-classified vessel and non-vessel (background) pixels, respectively, Acc is a measure that provides the ratio of total (both vessel and non-vessel) well-classified pixels (see [36] for a detailed description of these metrics). These evaluation functions provide global information on segmentation quality. They are obtained through pixel-to-pixel comparison between the automated segmentation and a reference-standard image, without taking into account that detected pixels are part of a vascular structure with specific features. This reference image is a manually-labeled image made by a medical expert. This method for quality measurement also faces the problem that human expert delineations of medical images are not exact, since the exact location of the real boundaries of the objects is unknown for experts (see Bioux et al. [37] for a comprehensive discussion on this issue). Therefore, differences between manual vessel segmentations performed by different specialists on the same fundus image are fairly common. To this respect, the attempts of different researches to produce more accurate reference-standard images are worth mentioning. Probabilistic adjustment of manually-segmented images to compensate possible generation-related differences is an example of this [38], [39].

This paper proposes a function based on the evaluation of measurable features describing vasculature. Specifically, this proposal enables vascular structure assessment through its characterization as connected segments with measurable area and length. Thus, this function is sensitive to vasculature features such as connectivity, area and length, and supplements widely-used metrics based on contingency tables. On the other hand, this function has shown a high matching degree with human quality perception when compared to other quality evaluation functions reported in literature. Therefore, it may be considered a useful tool for performance measurement in automated methodologies for blood vessel detection when the morphology of vascular structures is taken into account.

The rest of the paper is organized as follows. Section II explains the proposed quality function for retinal vessel segmentation evaluation and shows examples of its application. Section III offers the results of an experimentation aimed at measuring the matching degree of both the proposed function and other general quality evaluation functions with human quality perception. Finally, section IV contributes the main conclusions of this work.

2. FUNCTION FOR QUALITY ASSESSMENT OF RETINAL VESSEL SEGMENTATION

A quality evaluation function (QEF hereafter) for vessel segmentation assessment is described in this section. More- over, the methodology applied for parameter setting, as well as evaluation examples which show some of its properties, are also contributed.

A. Description

The aim of this paper is to design a QEF of vessel segmentations able to measure vascular tree descriptive features.

This QEF is based on three functions that evaluate connectivity, area and length in vessel segmentations with respect to their corresponding reference-standard images. Denoting S as the segmentation to be evaluated and S_G as the reference image, these functions are defined within the [0,1] interval as follows:

- Connectivity (C): This factor assesses the fragmentation degree between S and S_G. Since the vascular tree is a connected structure, proper vascular segmentation is expected to have only a few connected components (ideally one). This factor penalizes fragmented segmentations by comparing the number of connected components in S and S_G with regard to the total number of vessel pixels in S_G. Mathematically:

$$C(S, S_G) = 1 - min\left(1, \frac{|\#_C(S_G) - \#_C(S)|}{\#(S_G)}\right) \quad (1)$$

  where *min* is the minimum function, $\#_C(S_G)$ and $\#_C(S)$ stand for the number of connected components in S_G and S, respectively, and $\#(S_G)$ denotes the cardinality of S_G. Note that, for the sake of simplicity, segmentation is referred to the set of vessel pixels exclusively, thus excluding the set of background pixels.

- Area (A): This factor, based on the Jaccard coefficient [40], evaluates the degree of overlapping areas between S and S_G and is defined as:

$$A(S, S_G) = \frac{\#((\delta_\alpha(S) \cap S_G) \cup (S \cap \delta_\alpha(S_G)))}{\#(S \cup S_G)} \quad (2)$$

  Function δ_α is a morphological dilation using a disc of α pixels in radius. The introduction of this operator provides tolerance to slight differences in vessel width. The magnitude of this tolerance is controlled through α.

- Length (L): This factor measures the degree of coincidence between S and S_G in terms of total length and is formally expressed as:

$$L(S,S_G)=\frac{\#((\varphi(S)\cap\delta_\beta(S_G))\cup(\delta_\beta(S)\cap\varphi(S_G)))}{\#(\varphi(S)\cup\varphi(S_G))} \quad (3)$$

  where ϕ is an homotopic keletonization [41] and δ_β is a morphological dilation with a disc of β pixels in radius to reduce the impact of slight differences in vessel tracing. The value of β controls sensitivity degree to these differences.

According to these features, a function f is defined to be monotonically increasing as:

$$\begin{aligned} f &: \Re^3 \to \Re \\ (C, A, L) &\mapsto [0,1] \subseteq \Re \end{aligned} \quad (4)$$

where

$$x_i \geq y_i,\ i = 1,2,3 \ \Rightarrow f\,(x_1, x_2, x_3) \geq f\,(y_1, y_2, y_3) \quad (5)$$

Thus, f is dependent on the set of descriptive features C, A and L, presents an monotonically increasing behavior with respect to them, and takes values within the [0, 1] interval. The extreme values 0 and 1 denote the worst and perfect segmentations, respectively.

In this work, the product of C, A and L is proposed as a QEF for global quality assessment in retinal vessel segmentations (equation 6). This function will be referred to as CAL hereafter.

$$f(C, A, L) = C \times A \times L \equiv CAL \quad (6)$$

Note that any function fulfilling (4) and (5) can be considered. The product of C, A and L was selected because it tends to preserve equal quality in all features. On the other hand, this choice also allows the interpretation of segmentation results from the evaluation of important vascularity features.

B. Parameters Settings

Before using CAL, the values of the α and β parameters defined in equations (2) and (3) must be set. Note that A and L are monotonically increasing functions with respect to their α and β parameters, respectively. Low values in these parameters would make A and L very rigorous regarding differences between images. On the contrary, high values in these parameters would make functions very tolerant to such differences, thus reducing their descriptive potential.

An experimental study tried to determine the tolerance mar- gin between segmentations made by different human experts. The test set of the DRIVE database was used (see [42] for a detailed description of this retinal image database) for this purpose. This set provides in each of its 20 eye-fundus color images two manual segmentations generated by two different specialists. Thus, comparison between human observer-labeled images is possible. Since the manual images made by the first observer are commonly accepted as reference standard in literature, in our experimentation this set is considered as a reference and the set generated by the second observer is considered as segmentations to be evaluated. Taking this approach into account, for a given α parameter value, area A was obtained in the last 15 of the 20 images segmented manually by the second observer. These 15 values of A were then averaged to obtain a mean value linked to the selected α value. The same procedure was applied for length L and parameter β. The A and L values corresponding to the manual segmentations of the first five fundus images in the DRIVE database test set were not considered in average calculations. These five manual segmentations are part of the set of images used in the experimentation of this paper, section III. In this section, vessel segmentation evaluations of CAL and other QEFs are compared. Therefore, to avoid bias in the obtained results, these manual segmentations were excluded in this process of CAL-parameter setting.

Figure 1, image (a), shows the evolution of the mean values of A and L as functions of their α and β parameters, re- spectively. Figure 1, image (b), presents the forward-difference functions of A and L. Our aim is finding the value of α and β from which increase in functions A and L is low and almost constant. This fact can be observed to occur independently for both functions when α and β are equal to 2. Therefore, for all the experimentation described hereafter, α and β values were set to 2.

C. Application Examples

This subsection contributes examples of different cases of CAL-assessed vessel segmentation. These examples illustrate and highlight certain outstanding properties of this QEF. On the other hand, the results provided by sensitivity (Se), specificity (Sp) and accuracy (Acc) are also presented as a reference of the evaluations rendered by other commonly-used metrics.

1) Dependence on vascular tree structure features: The images used in this example are shown in Figure 2. Three synthetic segmentations were generated from manual vessel segmentation in image (a) according to the following criteria:

- Figure 2, image (b): Firstly, N true vessel pixels from wider vessels are labeled as background; secondly, N true background pixels located at the edges of narrower vessels are labeled as vessel.
- Figure 2, image (c): Firstly, N true vessel pixels from thinner vessels are labeled as background; secondly, N true background pixels located at the edges of wider vessels are labeled as vessel.
- Figure 2, image (d): Firstly, N true vessel pixels from thinner vessels are labeled as background; secondly, N true random background pixels are labeled as vessel.

The functions C, A, L, and CAL, expressed in equations (1), (2), (3), and (6), respectively, were calculated for images (b), (c) and (d), taking image (a) as reference standard. The obtained values are presented, together with the evaluations provided by Se, Sp and Acc, in TABLE I. Se, Sp and Acc can be observed to indicate equal quality for these three segmentations (the images were generated with this purpose). However, CAL quality evaluations show differences between tested images. Image (b) is very similar to reference image (a), since both images were generated to have only small differences at vessel edges. This results in the maximum possible CAL value, because all of its functions are 1.0 (A is also 1 due to the morphological dilation applied in its formulation). Although image (c) keeps all vessel pixels connected (C = 1.0), it presents narrow vessels shorter than in image (a). This fact is specially observable in the L value (0.8372). On the other hand, image (d) contains many isolated noisy pixels that result in the lowest CAL value (0.5286), because the whole set of measured features (connectivity, area and length) have been penalized. Therefore, CAL evaluates some vascular tree features in segmented images and thus enables interpreting its results within this framework.

2) Tolerance to small tracing differences in expert-labeled images: Quality assessment of automated vessel segmentations is usually performed by mathematical evaluation of the distortion between these segmentations and reference- standard images. These images are manually performed by human observers, thus including a subjective factor in their generation. Consequently, differences between expert-labeled images generated by different observers on the same image are fairly common, especially when tracing vessel borders or narrow vessels. As an example, Figure 3, image (c), shows coincidence (colored gray) and disagreement (colored black) in vessel tracing between two manual segmentations (images (a) and (b)) of a single fundus image. Therefore, these human- made images cannot be considered absolute ground truths [37]. A QEF for evaluating retinal vessel segmentation quality should minimize the

impact of this fact.

In this example, the influence of slight variations in vessel tracing on CAL is analyzed. Figure 3, image (d), shows automated-vessel segmentation on the fundus image whose expert-labeled images are shown in Figure 3. Segmentation was generated by the recently-published approach by Marín et al. [18]. Segmentation quality was measured by CAL taking images (a) and (b) as reference standards. The results are contributed in TABLE II (Individual Measures). Sp, Se and Acc values are also shown as a reference of evaluations by other metrics. CAL values can be observed to be very close, thus indicating the low dependence of this QEF on the expert- labeled image taken as reference. The same calculations were completed for each of the 20 fundus images available in the test set of the DRIVE database. The mean and standard deviation (std) of the QEFs are also shown in TABLE II (Averaged Measures). CAL averages for both reference standards differ in a small amount, thus corroborating the tolerance of CAL to differences in expert-labeled images. The same conclusions can be drawn for the Acc metric. However, their different scales should be taken into account when comparing both QEFs: while CAL varies from 0 to 1 for a black or background image (no vessel pixel detected) and the perfect segmentation, respectively, Acc varies within a smaller range. The average Acc of a null vessel-detected image measured with respect to the 20 manually-segmented images of the DRIVE test set was 0.8727 and 0.8774 for both sets of labeled images, respectively.

3) Correspondence with human perception: The first example shown in this section can be considered as an indicator of the existing correlation between CAL and human quality evaluations. Going back to the images of Figure 2, 20 human observers[1] were

[1] [1]Lecturers at departments of Mathematics as well as Electronic, Computer Science and Automatic Engineering, from the University of Huelva, Spain.

asked to rank quality of images (b), (c) and (d) with respect to the reference standard image (a). All of them qualified image (b) as the best segmentation. This may be explained by the fact that segmentation of image (b) preserves most of the vessels present in reference-standard image (a) (only with some slight differences in vessel width). On the other hand, segmentation in image (c) was considered more valuable than segmentation in image (d). Although both segmentations detected a similar amount of vascularity, noise in segmentation (d) degrades the visual perception of quality more than deficiencies in (c). As shown by the values in TABLE I, this interpretation matches CAL-rendered quality evaluation results.

3. EXPERIMENTATION

This last above-presented example seems to suggest that CAL-computed vessel segmentation evaluations are correlated with human-perceived quality. Next, this QEF is compared from this perspective to commonly-used metrics (Se, Sp and Acc) and other general QEFs that, to the best of our knowledge, have not been extensively used in retinal vessel segmentation. The aim is the subjective evaluation of the behavior of these functions in terms of correspondence with human perception. This section is divided into three subsections. The first subsection briefly describes the QEFs used in this experimentation, while the second presents the procedure and materials used. The third subsection firstly introduces the comparison methodologies applied to measure matching between the analyzed QEFs and human perception. Finally, the obtained results are contributed and discussed.

A. Description of the New QEFs Evaluated in this Section

The following general QEFs are included in the metrics evaluation study presented in this section: the Jaccard [40], Dice [43] and Kappa [44] coefficients, the average symmetric contour distance, the root mean square symmetric contour *distance,* and the *maximum* *symmetric contour distance*. The last three mentioned QEFs are 2-D adaptations of the *average symmetric surface distance* [45], the *root mean square symmetric surface distance* [45], and the *maximum symmetric surface distance* [46], respectively, as they were originally defined for quality evaluation in 3-D segmentations.

In order to compare these QEFs with *CAL* under the same conditions, their original formulations were modified to introduce tolerance on spatial overlap. This tolerance was implemented according to the same approach used in *CAL* (i.e., performing morphological dilations on the evaluated segmentations). Note that if the value of disc radius in such dilations is set to 0, the QEFs formulation described below corresponds to their original versions. Next, the mathematical formulation of each QEF evaluating a vessel segmentation, denoted by *S*, with respect to its corresponding reference-standard image, denoted by S_G, is contributed.

1) Jaccard coefficient: The Jaccard coefficient is defined as the ratio between the intersection and union of both images:

$$J_\gamma(S, S_G) = \frac{\#((\delta_\gamma(S) \cap S_G) \cup (S \cap \delta_\gamma(S_G)))}{\#(S \cup S_G)} \quad (7)$$

where δ_γ is a morphological dilation using a disc of γ pixels in radius. This operator is introduced in the modified versions of the following QEFs described below.

2) Dice coefficient: The Dice coefficient is defined as the size of the intersection of the two images divided by their average size:

$$D_\gamma(S, S_G) = \frac{\#((\delta_\gamma(S) \cap S_G) \cup (S \cap \delta_\gamma(S_G)))}{\frac{1}{2}((\#(S)) + (\#(S_G)))} \quad (8)$$

3) Kappa Coefficient: This coefficient can be expressed as:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (9)$$

where Pr(a) is the relative observed agreement and Pr(e) is the hypothetical probability of chance agreement, both calculated using segmentation S and reference standard S_G.

4) Average Symmetric Contour Distance: It is based on the edge points of S and S_G. For each edge point of S, the Euclidean distance to the nearest edge point of S_G is calculated. In order to provide symmetry, the same process is applied for each edge point of S_G. Average symmetric contour distance is then defined as the average of all stored distances (0 for perfect segmentation).

Let P (S) denote the set of edge points of S. The shortest distance of an arbitrary point p to P(S) is defined as $d(p, P(S)) = \min_{p_S \in P(S)} \|p - p_S\|$ where $\|.\|$ denotes the Euclidean distance. Average symmetric contour distance is then given by

$$ASD_\gamma(S, S_G) = \frac{1}{|P(S)|+|P(S_G)|} \times \tag{10}$$
$$\times \left(\sum_{p_S \in P(S)} d(p_S, P(\delta_\gamma(S_G))) + \sum_{p_{S_G} \in P(S_G)} d(p_{S_G}, P(\delta_\gamma(S))) \right)$$

5) Root Mean Square Symmetric Contour Distance: It is also based on contour distances, as it is calculated as the average symmetric contour distance described above. However, Euclidean distances between edge points are now squared before storing. The root of averaged squared distances then yields the root mean square symmetric contour distance (0 for a perfect segmentation):

$$RMSD_\gamma(S, S_G) = \sqrt{\frac{1}{|P(S)|+|P(S_G)|}} \times \tag{11}$$
$$\times \sqrt{\sum_{p_S \in P(S)} d^2(p_S, P(\delta_\gamma(S_G))) + \sum_{p_{S_G} \in P(S_G)} d^2(p_{S_G}, P(\delta_\gamma(S)))}$$

6) Maximum Symmetric Contour Distance: It is also known as Hausdorff distance and determined similarly to the two previous QEFs. Differences between both sets of edge points are determined using Euclidean distances, and the maximum value

yields the maximum symmetric

$$MSD_{\gamma}(S,S_G)=\max\left\{\max_{p_S\in P(S)} d(p_S,P(\delta_{\gamma}(S_G))),\quad \max_{p_{S_G}\in P(S_G)} d(p_{S_G},P(\delta_{\gamma}(S)))\right\} \quad (12)$$

With the aim of comparing the range of distances ASDγ, $RMSD_{\gamma}$ and MSD_{γ} to the remaining QEFs considered in this study (range [0, 1], 1 corresponding to perfect segmentation), these distance measures were normalized according to the following equation:

$$N(M)=\frac{1}{1+M},\ \text{with } M = ASD_{\gamma}, RMSD_{\gamma}, MSD_{\gamma} \quad (13)$$

B. Procedure and Material

The procedure applied in this study was aimed at comparing the matching degree of certain QEFs with human perception and can be summarized as follows. Different vessel segmentations of the first five eye-fundus images from the DRIVE database test set were selected. Then, 20 human observers (see footnote 1, page 4) were asked to score quality between the segmentations and the reference standards. In this process, the original color retinal images were also shown to the observers. Thus, they could overlay the segmentations on the color image and check segmentation goodness of fit. Scores were real numbers within the [0, 10] interval, where 0 and 10 denote the worst and best quality cases, respectively. A subjective human perception of quality can be then obtained. In addition to this, the values of the QEFs under study were also calculated for the same images. Thus, human- and functions-provided evaluations are compared to measure the correspondence degree between them through some statistical approach.

As an example, Figure 4 shows the set of vessel segmentations corresponding to one of the five fundus images (marked O in Figure 4) used for experimentation. The reference-standard image (marked G) is the corresponding image hand- labeled by the first observer.

The remaining images can be divided into two sets: a set of synthetic images (marked S) and a set of real algorithm images (marked M). The set of synthetic images is composed by 5 segmentations (images S_1-S_5) created by distorting G with the aim of disposing of varied-quality images. On the other hand, the set of real algorithm images is composed by the image manually labeled by the second observer (image M_1) and 8 segmentations generated by real vessel segmentation algorithms present in literature (images M_2-M_9). Concretely, these images were rendered by the following methods: Niemeijer et al., 2004 [35] (image M_2), Staal et al., 2004 [28] (image M_3), Zana and Klein, 2001 [21] (image M_4), Soares et al., 2006 [29] (image M_5), Chaudhuri et al., 1989 [23] (image M_6), Jiang and Mojon, 2003 [33] (image M_7) and M. E. Martínez-Pérez et al., 1999 [34] (image M_8) and Marín et al., 2011 [18](image M 9). The images were downloaded from www.isi.uu.nl/Research/Databases, except those taken from the methodologies by Soares et al., 2006 [29] and Marín et al., 2011 [18], that were provided by the authors.

The results of this survey are summarized in TABLE III. This table shows the averages ($\overline{HO}$) and standard deviations (HO_σ) of the scores divided by 10 (largest possible score) given by the 20 human observers for each image of both synthetic and real algorithms sets. Moreover, CAL, sensitivity (Se), specificity (Sp) and accuracy (Acc), as well as the metrics presented in the previous subsection for γ values 0 and 2, were also calculated for the segmentations of each selected fundus image. Their averaged values are shown in the last rows in TABLE III.

C. Comparison Methodologies and Results

The dataset of human-perceived and QEF-computed quality evaluation in TABLE III was analyzed according to the following statistical methodologies:

1) Consistence-based methodology: This methodology, proposed by Paglieroni (see [47] for a comprehensive description), consists on evaluating disparity between

human-perceived and QEF-computed quality. It computes two measures of inconsistency: mean inconsistency, denoted by Δ_{R1}, and standard inconsistency, denoted by Δ_{R2}. Both measures reflect rank disparity in units from 0 to n−1, n being the number of segmentations. Value 0 denotes perfect matching between perceived and computed quality. It is important to point out that, while Δ_{R2} is only sensitive to changes in the or- der of the compared scores, Δ_{R1} is also sensitive to the real distance between the compared scores. Therefore, this measure provides further sensitivity to differences than Δ_{R2}.

1. Correlation-based methodology: This is based on statistical correlation between subjective and objective quality evaluation. The random variables selected to represent the subjective and objective evaluation are, respectively, the average of human observer scores (defined above and denoted by $\overline{HO}$) and the average of QEF-generated scores (denoted by $\overline{S}_{QEF}$). For n segmentations, n samples of each variable can be obtained and the correlation between both datasets can be calculated as:

$$\rho = \frac{\sigma_{\overline{HO},\overline{S}_{QEF}}}{\sigma_{\overline{HO}}\ \sigma_{\overline{S}_{QEF}}} \qquad (14)$$

where σ denotes standard deviation and $\sigma_{\overline{HO},\overline{S}_{QEF}}$ is the covariance of $\overline{HO}$ and $\overline{S}_{QEF}$. Correlation ρ measures the linear relationship of both variables, which in this case is an indicator of the similarity degree of the behavior of subjective and objective evaluation.

TABLE IV shows the values of Δ_{R1}, Δ_{R2} and ρ for the QEFs in TABLE III. All sets of synthetic and real algorithm images were jointly considered. Regarding the set of tested QEFs, CAL is observed to render the lowest Δ_{R1} and Δ_{R2}-values, J_2 providing the second lowest value in both inconsistency measures. To this respect, note that J_2 is an equivalent metric to the A-area function in CAL expressed in equation (2). Regarding correlation ρ, CAL is the best- correlated QEF with human scores, followed again by J_2. This connection between CAL and human quality evaluations can be visually checked in Figure 5. This figure represents human- averaged scores $\overline{\mathrm{HO}}$ and averaged CAL-rendered measures for each of the considered images. This representation includes evaluations of other representative QEFs to compare the behavior of different functions.

Therefore, the analysis of the results provided by the matching degree indicators used in this experimentation concludes that CAL provides the best correspondence with human perception when compared to the remaining tested QEFs.

4. CONCLUSIONS

Blood vessel segmentation in retinal digital images plays an important role in the computerized detection of different ophthalmic pathologies leading to vascular anomalies. This applicability has led to the publication of numerous automated methods designed for this purpose over last years. As far as our understanding, most quality evaluation functions (QEFs) applied for measuring the performance of these methods do not consider vascularity as a tree-like connected structure with specific anatomical features. They are based on the individual pixel-to-pixel comparison of the resulting segmentation with an image labeled by a medical expert (reference-standard image). This paper proposes a function for vessel segmentation assessment based on vascular tree descriptive features with the aim of supplementing the existing QEFs. Specifically, this new proposed function, denoted by CAL, evaluates vessel connectivity, area and length in a segmented image in comparison with those in a reference-standard image.

Section II focused on the description of CAL, as well as on examples of its evaluation on different vessel-segmented images. These examples show evidence that CAL is sensitive to the anatomical features under evaluation (connectivity, area and length), thus allowing the interpretation of results from this point of view. In addition, they also suggest that CAL presents tolerance to small tracing differences in reference- standard images, as well as correspondence with human perception. With the aim of analyzing CAL behavior from this perspective, a comparison between CAL-provided values and human subjective evaluations was carried out on different vessel segmentations of five eye-fundus images. Other general QEFs were also included in the comparison. This experimentation is explained in detail in Section III. Two methodologies based on consistency assessment and statistical correlation respectively, were applied to measure correspondence between QEF-computed and human-perceived quality evaluations. The results obtained with both methodologies conclude that CAL renders a higher matching degree with human quality per- ception than the remaining tested QEFs. Anyway, it is worth mentioning that some others of these QEFs such as J and ASD also seem to correspond reasonably well to the quality assessment of human observers.

Due to these properties, the QEF proposed in this paper can be used as a good supplement of the information provided by other QEFs. However, it is important to notice that this proposal was designed for segmentation quality measurement in a specific structure (retinal vascular tree) and is not therefore applicable to general cases. Anyway, the applied concept of measuring descriptive features may be useful to design other specialized QEFs aimed at enhancing segmentation quality assessment of other complex shapes.

ACKNOWLEDGMENT

The authors would like to thank J. J. Staal and colleagues for making their database publicly available.

References

[1] M. Foracchia, E. Grisan, and A. Ruggeri, “Detection of optic disc in retinal images by means of a geometrical model of vessel structure,” IEEE Trans. Med. Imag., vol. 23, pp. 1189-1195, 2004.

[2] A. A. H. A. R. Youssif, A. Z. Ghalwash, and A. R Ghoneim, “Optic disc detection from normalized digital fundus images by means of a vessels’ direction matched filter,” IEEE Trans. Med. Imag, vol. 27, pp. 11-18, 2008.

[3] A. Aquino, M. E. Gegúndez-Arias, D. Marín “Detecting the Optic Disc Boundary in Digital Fundus Images Using Morphological, Edge Detection and Feature Extraction Techniques,” IEEE Trans. Med. Imag, doi: 10.1109/TMI.2010.2053042, 2010.

[4] H. Li and O. Chutatape, “Automated Feature Extraction in Color Retinal Images by a Model Based Approach,” IEEE Trans. Biomed. Eng., vol. 51, pp. 246-254, 2004.

[5] M. Larsen, J. Godt, N. Larsen, H. Lund-Andersen, A. K. Sjølie, E. Agardh, H. Kalm, M. Grunkin, and D. R. Owens, “Automated detection of fundus photographic red lesions in diabetic retinopathy,” Investigat. Opht. Vis. Sci., vol. 44, no. 2, pp. 761-766, 2003.

[6] M. Niemeijer, B. van Ginneken, J. J. Staal, M. S. A. Suttorp-Schulten, and M. D. Abramoff, “Automatic detection of red lesions in digital color fundus photographs,” IEEE Trans. Med. Imag., vol. 24, no. 5, pp. 584- 592, 2005.

[7] F. Zana and J. C. Klein, “A multimodal registration algorithm of eye fundus images using vessels detection and Hough transform,” IEEE Trans. Med. Imag., vol. 18, pp. 419-428, 1999.

[8] G. K. Matsopoulos, P. A. Asvestas, N. A. Mouravliansky, and K. K. Delibasis, “Multimodal Registration of Retinal Images Using Self Organizing Maps,” IEEE Trans. Med. Imag., vol. 23, pp. 1557-1563, 2004.

[9] C. Heneghan, J. Flynn, M. OKeefe, and M. Cahill, “Characterization of changes in

blood vessel width and tortuosity in retinopathy of prematurity using image analysis," Med. Image Anal., vol. 6, pp. 407- 429, 2002.

[10] E. Grisan and A. Ruggeri, "A divide and impera strategy for the automatic classification of retinal vessels into arteries and veins," in Proc. 25th Int. Conf. IEEE Eng. Med. Biol. Soc., 2003, pp. 890-893.

[11] Y. Hatanaka, H. Fujita, M. Aoyama, H. Uchida, and T. Yamamoto, "Automated analysis of the distribuitions and geometries of blood vessels on retinal fundus images," Proc. SPIE Med. Imag. 2004: Image Process., vol. 5370, pp. 1621-1628, 2004.

[12] M. Foracchia, E. Grisan, and A. Ruggeri, "Extraction and quantitative description of vessel features in hypertensive retinopathy fundus im- ages," in Book Abstracts 2nd Int. Workshop Comput. Asst. Fundus Image Anal., 2001, p. 6.

[13] M. E. Martínez-Pérez, A. D. Hughes, A. V. Stanton, S. A. Thom, N. Chapman, A. A. Bharath, and K. H. Parker, "Retinal vascular tree morphology: A semiautomatic quantification," IEEE Trans. Biomed. Eng., vol. 49, no. 8, pp. 912-917, 2002.

[14] J. Lowell, A. Hunter, D. Steel, A. Basu, R. Ryder, and R. L. Kennedy, "Measurement of retinal vessel widths from fundus images based on 2- D modeling," IEEE Trans. Med. Imag., vol. 23, no. 10, pp. 1196-1204, 2004.

[15] D. E. Becker, A. Can, J. N. Turner, H. L. Tanenbaum, and B. Roysam, "Image processing algorithms for retinal montage, synthesis, mapping and real-time location determination," IEEE Trans. Biomed. Eng., vol. 45, pp. 115-118, 1998.

[16] H. Shen, B. Roysam, C. V. Stewart, J. N. Turner, and H. L. Tanenbaum, "Optimal scheduling of tracing computations for real-time vascular landmark extraction from retinal fundus images," IEEE Trans. Inf. Technol. Biomed., vol. 5, pp. 77-91, 2001.

[17] Y. A. Tolias and S. M. Panas, "A fuzzy vessel tracking algorithm for retinal images based on fuzzy clustering," IEEE Trans. Med. Imag., vol. 17, pp. 263-273, 1998.

[18] D. Marín, A. Aquino, M. E. Gegúndez-Arias, and J. M. Bravo, “A New Supervised Method for Blood Vessel Segmentation in Retinal Images by Using Grey-Level and Moment Invariants-Based Features ,” IEEE Trans. Med. Imag, vol. 30, no. 1, pp. 146-158, 2011.

[19] I. Liu and Y. Sun, “Recursive tracking of vascular networks in angiograms based on the detection-deletion scheme,” IEEE Trans. Med. Imag., vol. 12, , pp. 334-341, 1993.

[20] L. Zhou, M. S. Rzeszotarski, L. J. Singerman, and J. M. Chokreff, “The detection and quantification of retinopathy using digital angiograms,” IEEE Trans. Med. Imag., vol. 13, pp. 619-626, 1994.

[21] F. Zana and J. C. Klein, “Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation,” IEEE Trans. Image Processing, vol. 10, pp. 1010-1019, 2001.

[22] A. M. Mendonça and A. Campilho, “Segmentation of Retinal Blood Vessels by Combining the Detection of Centerlines and Morphological Reconstruction,” IEEE Trans. Med. Imag., vol. 25, pp. 1200-1213, 2006.

[23] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum, “Detection of blood vessels in retinal images using two-dimensional matched filters,” IEEE Trans. Med. Imag., vol. 8, pp. 263-269, 1989.

[24] A. Hoover, V. Kouznetsova, and M. Goldbaum, “Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response,” IEEE Trans. Med. Imag., vol. 19, pp. 203-210, 2000.

[25] L. Gang, O. Chutatape, and S. M. Krishnan, “Detection and measurement of retinal vessels in fundus images using amplitude modified second-order Gaussian filter,” IEEE Trans. Biomed. Eng., vol. 49, pp. 168-172, 2002.

[26] G. G. Gardner, D. Keating, T. H. Williamson, and A. T. Elliott, “Automatic detection

of diabetic retinopathy using an artificial neural network: A screening tool," Br. J. Ophthalmol., vol. 80, pp. 940-944, 1996.

[27] C. Sinthanayothin, J. F. Boyce, H. L. Cook, and T. H. Williamson, "Automated localisation of the optic disc, fovea and retinal blood vessels from digital colour fundus images," Br. J. Ophtalmol., vol. 83, pp. 902- 910, 1999.

[28] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge based vessel segmentation in color images of the retina," IEEE Trans. Med. Imag., vol. 23, pp. 501-509, 2004.

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

[30] E. Ricci and R. Perfetti, "Retinal Blood Vessel Segmentation Using Line Operators and Support Vector Classification," IEEE Trans. Med. Imag., vol. 26, pp. 1357-1365, 2007.

[31] M. E. Martínez-Pérez, A. D. Hughes, S. A. Thom, A. A. Bharath, and K. H. Parker, "Segmentation of blood vessels from red-free and fluorescein retinal images," Med. Imag. Anal. vol. 11, pp. 47-61, 2007.

[32] B. S. Y. Lam and H. Yan, "A Novel Vessel Segmentation Algorithm for Pathological Retina Images Based on the Divergence of Vector Fields," IEEE Trans. Med. Imag., vol. 27, pp. 237-246, 2008.

[33] X. Jiang and D. Mojon, "Adaptive Local Thresholding by Verification- Based Multithreshold Probing with Application to Vessel Detection in Retinal Images," IEEE Trans. Pattern Analy. Mach. Intell., vol. 254, pp. 131-137, 2003.

[34] M. E. Martínez-Pérez, A. D. Hughes, A. V. Stanton, S. A. Thom, A. A. Bharath, and K. H. Parker, "Scale-space analysis for the characterisation of retinal blood vessels," in

Medical Image Computing and Computer- Assisted Intervention - MICCAI99, C. Taylor and A. Colchester, Eds.: 1999, pp. 90-97.

[35] M. Niemeijer, J. Staal, B. van Ginneken, M. Loog, and M. D. Abramoff, “Comparative study of retinal vessel segmentation methods on a new publicly available database,” in SPIE Med. Imag., J.M. Fitzpatrick and M. Sonka, Eds., 2004, vol. 5370, pp. 648-656.

[36] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition Letters, vol. 27, pp. 861-874, 2006.

[37] S. Bioux, M. Martin-Fernandez, L. Ungar, M. Nakamura, M. -S. Koo, R. W. McCarley, and M. E. Shenton, “On evaluating brain tissue classifiers without a ground truth,” NeuroImage, vol. 36, no. 4, pp. 1207- 1224, 2007.

[38] N. J. Pizzi, “Fuzzy Preprocessing of Gold Standards As Applied to Biomedical Spectra Classification,” Artificial Intelligence in Medicine, vol. 16, pp. 171-182, 1999.

[39] A Martina, H Laanaya, and A Arnold-Bosa, “Evaluation for uncertain image classification and segmentation, ”Pattern Recognition, vol. 39, pp. 1187-1195, 2006.

[40] P. Jaccard, “Etude comparative de la distribution florale dans une portion des alpes et des jura,” Bull. Soc. Vaudoise Sci. Nat., vol. 37, pp. 547-579, 1901.

[41] J. Serra, Image Analysis and Mathematical Morphology. London: Academic Press, 1982, vol. 1.

[42] Research section, digital retinal image for vessel extraction (DRIVE) database University Medical Center Utrecht, Image Sciences Institute, Utrecht, The Netherlands [Online].

Available: http://www.isi.uu.nl/Research/Databases/DRIVE/

[43] L. R. Dice, “Measures of the amount of ecologic association between species,” J. Ecol., vol. 26, pp. 297-302, 1945.

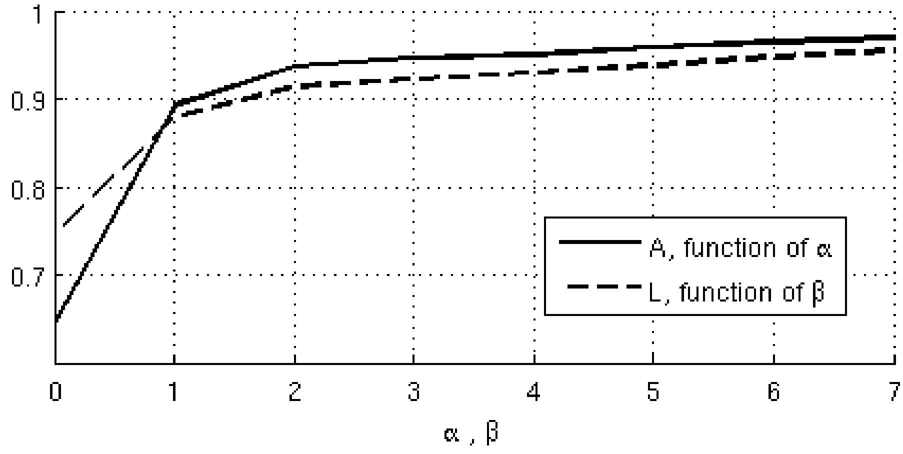
[44] J. A. Cohen, “A coefficient of agreement for nominal scales,” Educ. Psychol. Meas., vol. 20, pp. 37-46, 1960.

[45] H. Lamecker, T. Lange, and M. Seebass, “Segmentation of the Liver using a 3D Statistical Shape Model,” Technical report, Zuse Institute, Berlin, 2004.

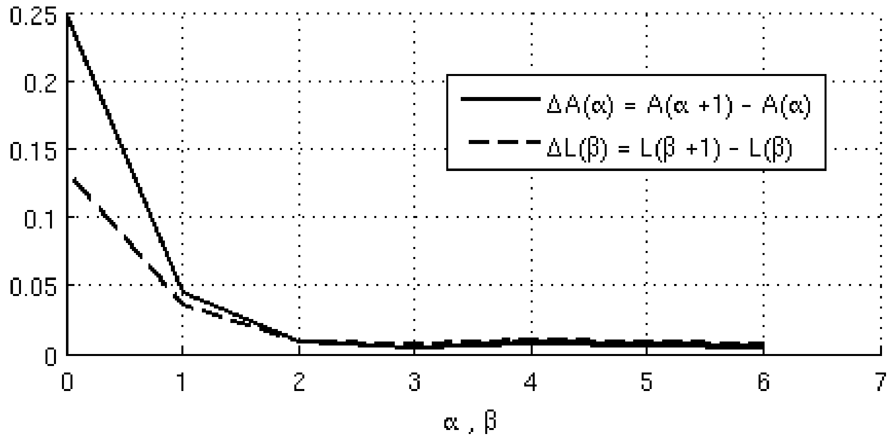
[46] D. Huttenlocher, D. Klanderman, and A. Rucklige, “Comparing images using the Hausdorff distance,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 9, pp. 850-863, 1993.

[47] D. W. Paglieroni, “Design considerations for image segmentation quality assessment measures,” Pattern Recognition, vol. 37, pp. 1607-1617, 2004.

Figures



(a)



(b)

Fig. 1. Study to determine optimal α and β values: (a) Evolutions of A and L mean values as functions of their α and β parameters; (b) Forward differences of A and L as functions of their α and β parameters.

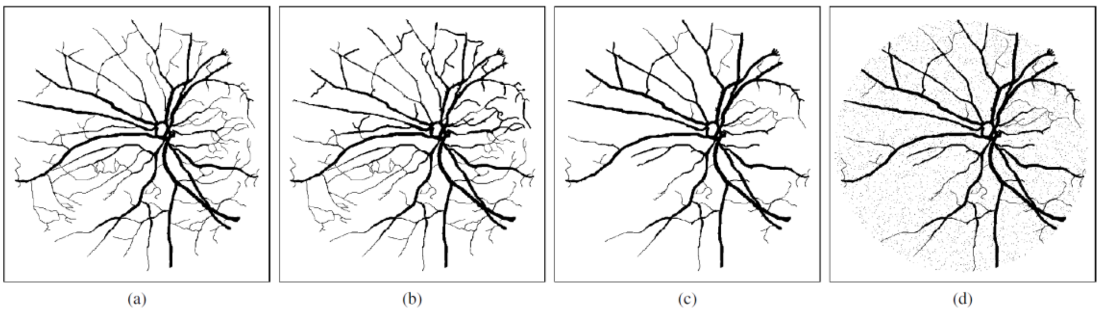


Fig. 2. Images used to show the dependence of CAL on vascularity features: (a) Reference-standard image; (b)-(d) Images created by distorting (a).

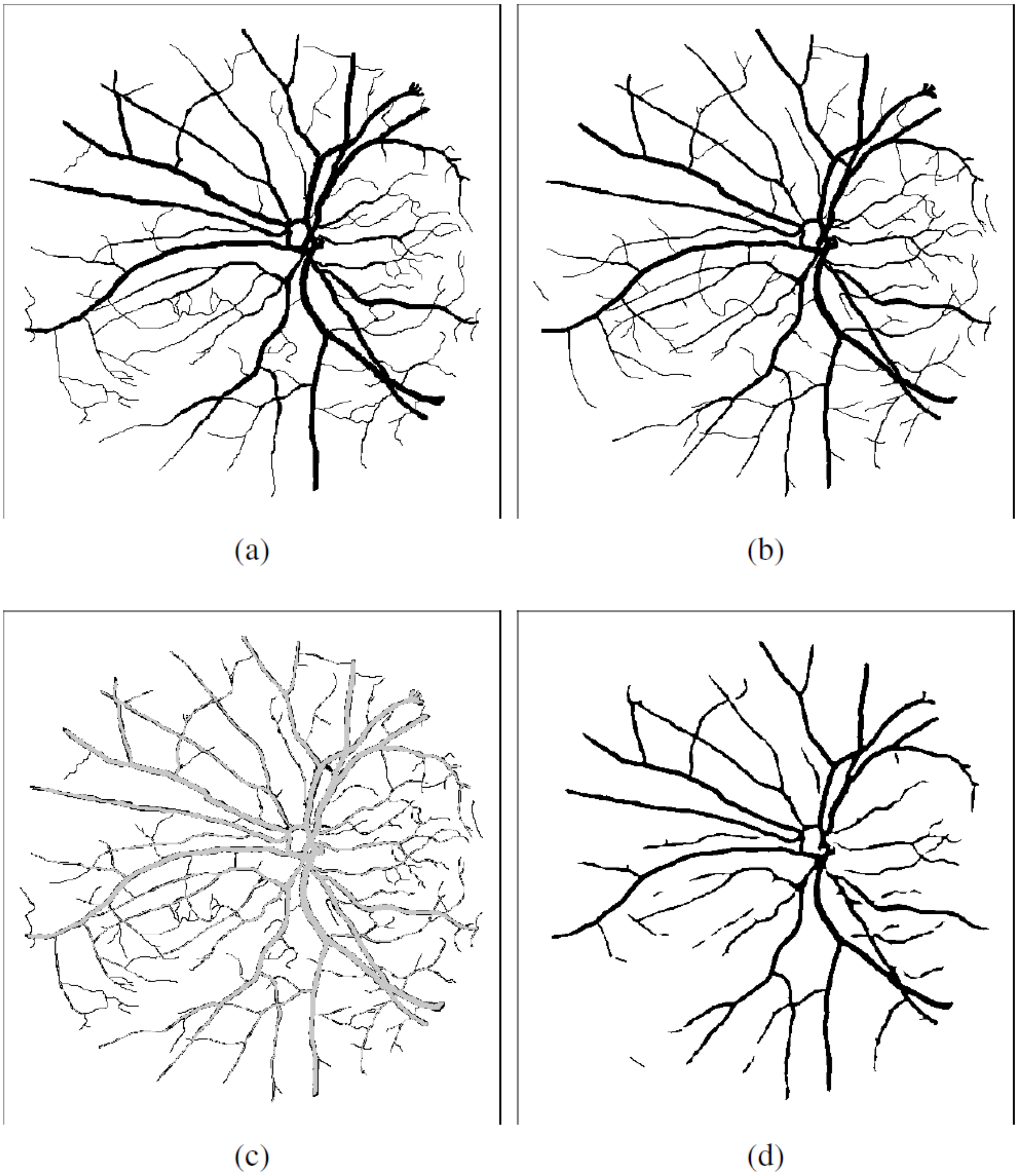


Fig. 3. Images used to show the tolerance of CAL to small tracing differences in expert labeled images: (a) and (b) Labeled images generated by two different human experts; (c) Composition of (a) and (b) showing coincidence (gray) and disagreement (black); (d) Segmentation produced by an automated method.

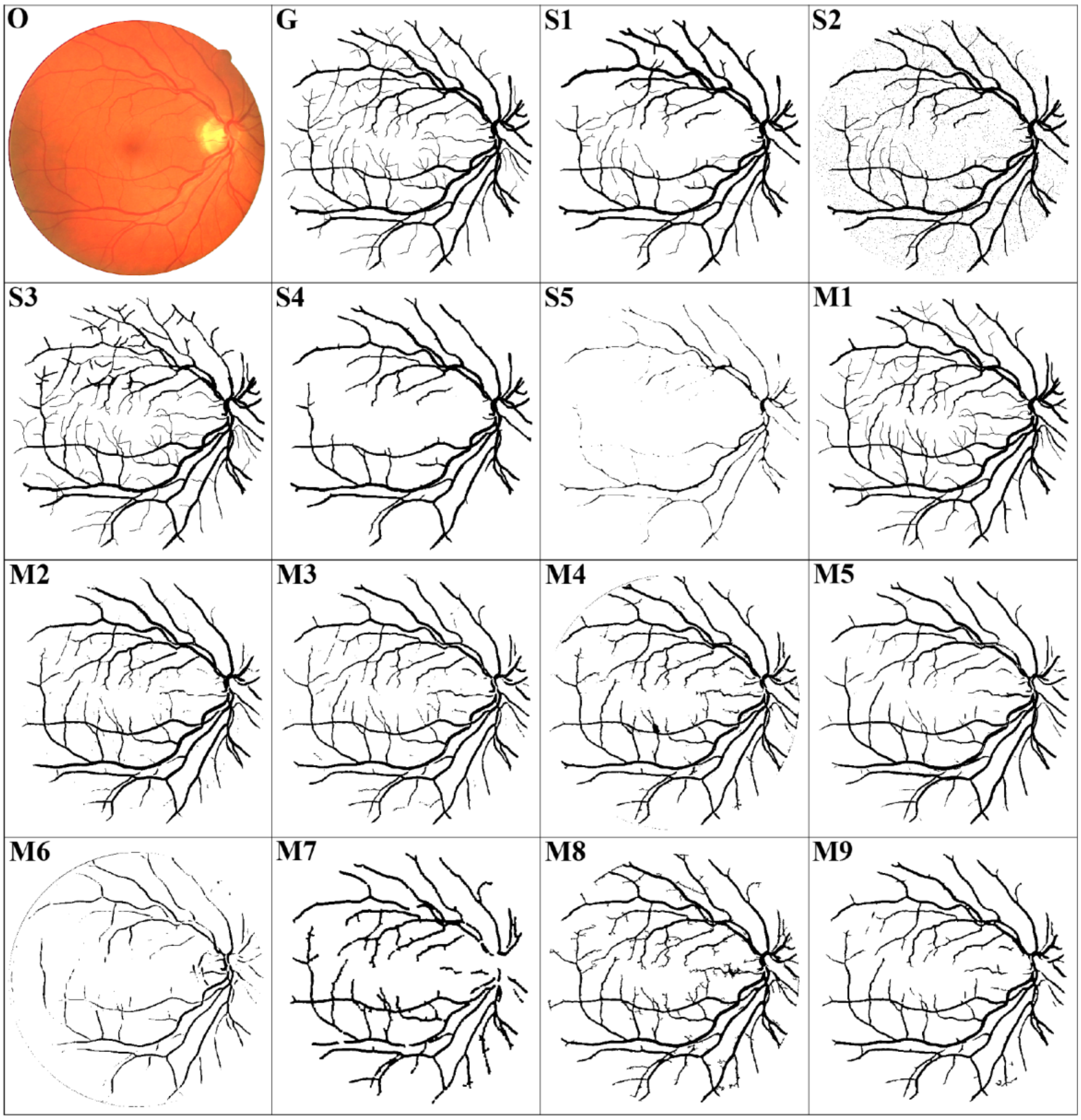


Fig. 4. Set of images corresponding to one of the five eye-fundus images used in the experimentation: O and G are the original and reference-standard images, respectively, while S1-S5 are vessel synthetic segmentations, and M1-M9 are segmentations provided by real algorithms.

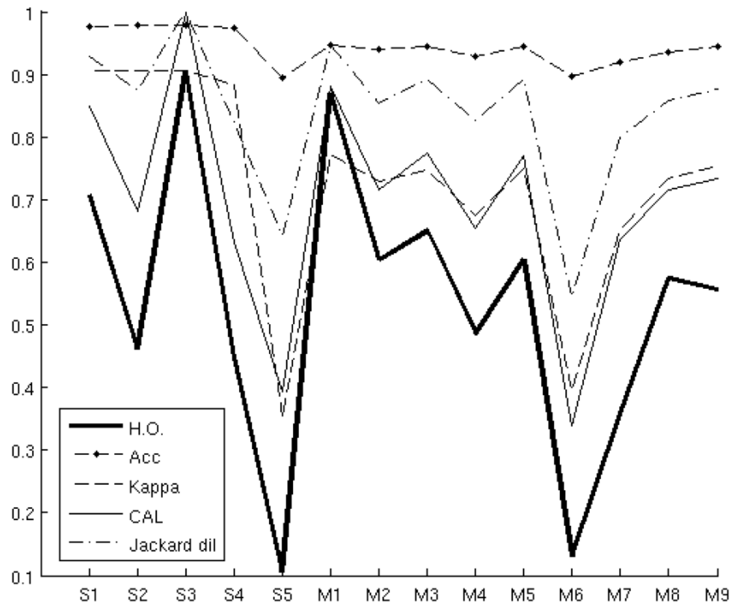


Fig. 5. Averages of human scores ($\overline{HO}$) and CAL- and other QEF-computed evaluations for the whole set of synthetic and real algorithm images

Tables

TABLE I

QUALITY VALUES FOR THE VESSEL SEGMENTATIONS SHOWN IN IMAGES (B), (C) AND (D) OF FIGURE 2 TAKING IMAGE (A) AS REFERENCE STANDARD.

Image	*Se*	*Sp*	*Acc*	*C*	*A*	*L*	*CAL*
(b)	0.9134	0.9758	0.9681	1.0	1.0	1.0	1.0
(c)	0.9134	0.9758	0.9681	1.0	0.8812	0.8372	0.7377
(d)	0.9134	0.9758	0.9681	0.8868	0.8181	0.7286	0.5286

TABLE II

"INDIVIDUAL MEASURES": QUALITY VALUES FOR THE VESSEL SEGMENTATION SHOWN IN IMAGE (D) OF FIGURE 3 TAKING IMAGES (A) AND (B) AS REFERENCE STANDARDS (DENOTED BY RS_1 AND RS_2). "AVERAGED MEASURES": AVERAGED QUALITY VALUES OF 20 AUTOMATED SEGMENTATIONS COMPUTED FOR THE TEST SET OF THE DRIVE DATABASE.

	Individual		Averaged Measures (std)	
	RS_1	RS_2	RS_1	RS_2
Se	0.7032	0.7386	0.7077	0.7399
Sp	0.9855	0.9857	0.9801	0.9809
Acc	0.9479	0.9543	0.9452	0.9510
C	0.9992	0.9993	0.9990	0.9990
A	0.8610	0.8548	0.8405	0.8506
L	0.7728	0.7836	0.7919	0.7996
CAL	0.6648	0.6694	0.6665	0.6824

TABLE III

AVERAGED DATASET OF QUALITY EVALUATION

FROM HUMAN PERCEPTION AND ALL

CONSIDERED QEFS.

	Set of Synthetic Images					Set of Real Algorithm Images								
	S1	S2	S3	S4	S5	M1	M2	M3	M4	M5	M6	M7	M8	M9
$\overline{HO}$	0.709	0.461	0.903	0.449	0.105	0.869	0.605	0.651	0.486	0.604	0.130	0.358	0.575	0.556
$HO\sigma$	0.097	0.112	0.067	0.121	0.071	0.060	0.140	0.143	0.191	0.150	0.084	0.165	0.141	0.160
$\overline{Se}$	0.918	0.918	0.919	0.813	0.242	0.778	0.685	0.716	0.663	0.715	0.291	0.666	0.765	0.729
$\overline{Sp}$	0.987	0.987	0.987	1.000	1.000	0.974	0.982	0.981	0.972	0.981	0.996	0.961	0.964	0.980
$\overline{Acc}$	0.977	0.977	0.977	0.974	0.894	0.947	0.941	0.944	0.930	0.944	0.897	0.920	0.936	0.945
$\overline{Kappa}$	0.906	0.906	0.906	0.882	0.352	0.772	0.729	0.748	0.675	0.749	0.396	0.652	0.733	0.755
$\overline{CAL}$	0.849	0.681	1.000	0.631	0.393	0.881	0.716	0.775	0.654	0.771	0.340	0.635	0.715	0.733
$\overline{J_0}$	0.849	0.849	0.849	0.814	0.241	0.670	0.615	0.637	0.554	0.639	0.274	0.536	0.625	0.646
$\overline{J_2}$	0.929	0.873	1.000	0.821	0.642	0.947	0.854	0.893	0.827	0.893	0.545	0.799	0.859	0.876
$\overline{D_0}$	0.919	0.919	0.919	0.897	0.386	0.802	0.761	0.777	0.708	0.779	0.424	0.697	0.769	0.785
$\overline{D_2}$	0.995	0.944	0.925	0.906	0.969	0.881	0.946	0.917	0.943	0.919	0.847	0.962	0.946	0.939
$\overline{N(ASD_0)}$	0.404	0.498	0.768	0.182	0.089	0.486	0.310	0.387	0.283	0.358	0.104	0.220	0.297	0.300
$\overline{N(ASD_2)}$	0.437	0.508	0.999	0.183	0.097	0.652	0.370	0.480	0.349	0.440	0.113	0.255	0.350	0.419
$\overline{N(RSMD_0)}$	0.160	0.240	0.643	0.079	0.053	0.276	0.155	0.205	0.152	0.182	0.061	0.113	0.146	0.147
$\overline{N(RSMD_2)}$	0.161	0.240	0.941	0.079	0.053	0.286	0.156	0.209	0.154	0.185	0.061	0.114	0.147	0.149
$\overline{N(MSD_0)}$	0.021	0.025	0.205	0.014	0.011	0.030	0.019	0.026	0.019	0.023	0.012	0.016	0.019	0.021
$\overline{N(MSD_2)}$	0.021	0.025	0.205	0.014	0.011	0.030	0.019	0.026	0.019	0.023	0.012	0.016	0.019	0.021

TABLE IV

RESULTS OF CONSISTENCE AND CORRELATION-BASED METHODOLOGIES FOR THE WHOLE SET OF ANALYZED QEFS.

	Se	*Sp*	*Acc*	*Kappa*	*CAL*	J_0	J_2	D_0	D_2	ASD_0	ASD_2	$RMSD_0$	$RMSD_2$	MSD_0	MSD_2
Δ_{R1}	1.000	1.623	0.979	1.023	0.363	0.984	0.695	1.048	1.322	0.747	0.752	1.180	1.367	1.615	1.615
Δ_{R2}	2.571	5.429	2.429	2.429	0.857	2.714	1.000	2.429	4.857	1.429	1.429	1.429	1.429	1.429	1.429
ρ	0.799	-0.336	0.711	0.776	0.977	0.178	0.937	0.964	0.752	0.937	0.784	0.039	0.848	0.870	0.726