This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2778748, IEEE Transactions on Medical Imaging

A Skeletal Similarity Metric for Quality Evaluation of Retinal Vessel Segmentation

Zengqiang Yan, Xin Yang, and Kwang-Ting (Tim) Cheng, *Fellow, IEEE*

***Abstract*—The most commonly used evaluation metrics for quality assessment of retinal vessel segmentation are sensitivity, specificity and accuracy, which are based on pixel-to-pixel matching. However, due to the inter-observer problem that vessels annotated by different observers vary in both thickness and location, pixel-to-pixel matching is too restrictive to fairly evaluate the results of vessel segmentation. In this paper, the proposed skeletal similarity metric is constructed by comparing the skeleton maps generated from the reference and the source vessel segmentation maps. To address the inter-observer problem, instead of using a pixel-to-pixel matching strategy, each skeleton segment in the reference skeleton map is adaptively assigned with a searching range whose radius is determined based on its vessel thickness. Pixels in the source skeleton map located within the searching range are then selected for similarity calculation. The skeletal similarity consists of a curve similarity which measures the structural similarity between the reference and the source skeleton maps and a thickness similarity which measures the thickness consistency between the reference and the source vessel segmentation maps. In contrast to other metrics that provide a global score for the overall performance, we modify the definitions of true positive, false negative, true negative and false positive based on the skeletal similarity, based on which sensitivity, specificity, accuracy and other objective measurements can be constructed. More importantly, the skeletal similarity metric has better potential to be used as a pixel-wise loss function for training deep learning models for retinal vessel segmentation. Through comparison of a set of examples, we demonstrate that the redefined metrics based on the skeletal similarity are more effective for quality evaluation, especially with greater tolerance to the inter-observer problem.**

***Index Terms*—Retinal vessel segmentation, quality evaluation, inter-observer problem, skeletal similarity.**

I. Introduction

RETINAL vessel segmentation [1], [2], [3], [4], is an essential step for the diagnosis of eye-related diseases, such as macular degeneration, diabetic retinopathy and glaucoma. Since manual annotation by a human observer is time consuming, automated retinal vessel segmentation has been widely studied over decades. However, how to effectively evaluate the quality of retinal vessel segmentation remains an open issue.

Similar to other types of digital images, objective image quality metrics could be used to assess the quality of vessel segmentation. Objective image quality evaluation metrics can be roughly classified based on whether a reference image is available or not. If a reference image is not available, usually a *no-reference* quality assessment approach would be adopted. Fortunately, for retinal vessel segmentation, in most cases a manually annotated image is available and used as the reference image. When a reference image is available, a *full-reference* approach for quality evaluation is used. For natural images, the most popular *full-reference* quality evaluation metrics are the mean squared error (MSE) that calculates the average squared intensity differences of pixels in the source image (*i.e.* image under assessment) and the reference image; the peak signal-to-noise ratio (PSNR) which is constructed based on the MSE; and the structural similarity (SSIM) index [5] which measures the similarity between two images. However, considering the sparsity and non-uniformity of vessels in fundus image, these metrics are not suitable for quality evaluation of vessel segmentation.

For quality evaluation of retinal vessel segmentation, Gegúndez-Arias *et al.* [6] proposed a function $f(C, A, L)$. Parameter C penalizes fragmented segmentations by comparing the number of connected segments in the source vessel segmentation and in the reference manual annotation. Parameter A measures the degree of overlapping areas between the source vessel segmentation and the reference manual annotation. Parameter L compares the lengths of the skeletons in the source vessel segmentation and in the reference manual annotation. Though skeleton maps are utilized for comparison, $f(C, A, L)$ does not measure their skeletal similarity. In addition, based on the global score of $f(C, A, L)$, it is difficult to examine the details of vessel segmentation or derive other evaluation metrics.

Currently, the most commonly used metrics for retinal vessel segmentation are sensitivity (Se), specificity (Sp) and accuracy (Acc) [7], [8], [9], [10], [11], where Se measures the ratio that vessel pixels are correctly segmented, Sp measures the ratio that non-vessel pixels are successfully identified and Acc evaluates the overall performance which can be seen as a weighted sum of Se and Sp. All these metrics are calculated based on a pixel-to-pixel matching strategy, namely by comparing each pair of pixels in the source vessel segmentation and in the reference manual annotation, with the basic assumption that the manual annotation is the ground truth and is absolutely correct. However, this assumption is often not valid for the vessel segmentation problem, due to the inter-observer problem where vessels annotated by different observers may vary in both thickness and location. More details of the inter-observer problem are discussed in the next

Z. Yan and K. -T. Cheng are with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clearwater Bay, Kowloon, Hong Kong (e-mail: z.yan@connect.ust.hk, timcheng@ust.hk).

X. Yang is with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China (e-mail: xinyang2014@hust.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2778748, IEEE Transactions on Medical Imaging

section.

In [21], Sofka *et al.* proposed to redefine true positives and false negatives in a given thinned vessel (centerline) segmentation, where a true positive is any detected point within the 2-pixel range of a point in the thinned manual segmentation and a false positive is any detected point located beyond the 2-pixel neighborhood of a centerline point in the thinned manual segmentation. Though the 2-pixel range somewhat alleviates the inter-observer problem compared to the pixel-to-pixel matching strategy, directly denoting all the points within the range as true positives fails to analyze the distortion of vessel centerlines.

In this paper, we propose a new metric for quality evaluation of retinal vessel segmentation, which is constructed by comparing the skeletons generated from the source and the reference vessel segmentations. To solve the inter-observer problem, we define a searching range for each pixel in the reference skeleton map to find the corresponding pixel(s) in the source skeleton map. In contrast to the method using a fixed 2-pixel range in [21], the searching range of each pixel is adaptively determined based on its vessel thickness so that thicker vessels would have a smaller searching range and tiny vessels would have a larger searching range. This adaptive strategy is based on the observation that the inter-observer problem is usually more serious in tiny vessels. Furthermore, instead of directly denoting all the pixels in the source vessel segmentation located within the searching range as true positives, we define *skeletal similarity* to measure the structural similarity and the thickness consistency between the reference and the source vessel segmentations. Based on the skeletal similarity metric, we then redefine Se, Sp and Acc and compare with other evaluation metrics. We demonstrate through experimental evaluation that the proposed metrics are more effective for quality evaluation of vessel segmentation especially taking into account the inter-observer problem.

II. Problem Analysis

In the calculation of the metrics Se, Sp and Acc, pixels in the source segmentation map are classified into four classes. Given a segmentation map, manually annotated vessel pixels that are correctly detected as vessel pixels are denoted as true positives (*TP*) and those wrongly classified as non-vessel pixels are counted as false negatives (*FN*). Meanwhile, manually annotated non-vessel pixels that are correctly classified are denoted as true negatives (*TN*) and those wrongly classified as vessel pixels are counted as false positives (*FP*). Then, the evaluation metrics Se, Sp and Acc are defined as follows:

$$Se = \frac{TP}{TP+FN}, Sp = \frac{TN}{TN+FP}, Acc = \frac{TP+TN}{N}, \quad (1)$$

where $N = TN+TP+FN+FP$. According to these definitions, the evaluation metrics Se, Sp and Acc are constructed based on pixel-to-pixel matching.

For retinal vessel segmentation in fundus image, we can further classify the vessels into the *basic vessel structure* and the *tiny vessels* as shown in Fig. 1. Visually, the vessels belong to the basic vessel structure are much thicker than those tiny vessels. If we define the vessels whose thickness is less than

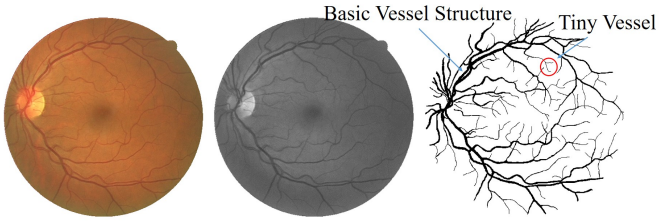

Fig. 1: Vessel segmentation in the fundus image of the DRIVE dataset [12]. From left to right: the fundus image, the green channel of the fundus image and the manual annotation.

3 pixels as the tiny vessels and those thicker vessels as parts of the basic vessel structure, nearly 77% of vessel pixels belong to the basic vessel structure and the tiny vessels only account for 23%. If all pixels are considered equally important according to pixel-to-pixel matching in the evaluation metrics Se, Sp and Acc, the basic vessel structure would be the dominating factor due to the imbalanced ratio (77% v.s. 23%). In other words, these evaluation metrics emphasize more on the matching of the basic vessel structure rather than the completeness of the whole vessel tree.

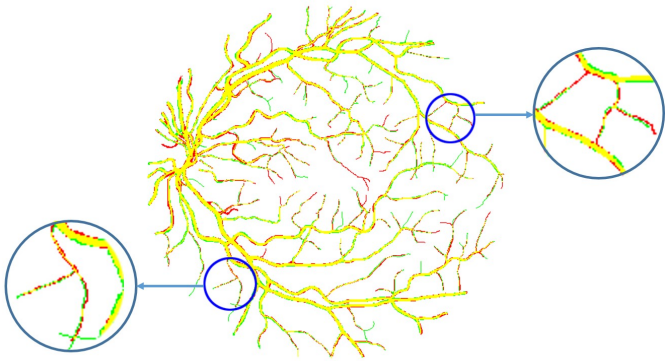

Fig. 2: The inter-observer problem in manual annotations: vessels annotated by the first observer (red pixels), vessels annotated by the second observer (green pixels) and common vessels annotated by the two observers (yellow pixels).

In addition, limited by the resolution of fundus image, manually annotated vessels in different annotations are often not perfectly matched as shown in Fig. 2. For the basic vessel structure (or thick vessels), locations annotated by different observers are quite similar, but vessel thickness could be slightly different. Though the thickness variation is limited, their influence can be disproportional due to the highly imbalanced ratio between the basic vessel structure and tiny vessels. In different manual annotations made to the same fundus image, vessel thickness of tiny vessels is usually similar (not necessarily exactly the same) but their locations can be quite different. In the enlarged region in Fig. 2, though the structural properties of these annotated tiny vessels are similar, it's possible that the annotated vessel pixels of the same segment in different annotations may not have any overlap. In other words, for tiny vessels, location variation in different annotations can be quite serious. Based on the analysis on the entire dataset,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2778748, IEEE Transactions on Medical Imaging

we found that the inter-observer problem is quite common and it is not unusual at all that non-vessel pixels in one manual annotation can be vessel pixels in another manual annotation. Therefore, the pixel-to-pixel matching strategy should not be directly used for evaluation, and a fair evaluation metric must tolerate such variations caused by the inter-observer problem.

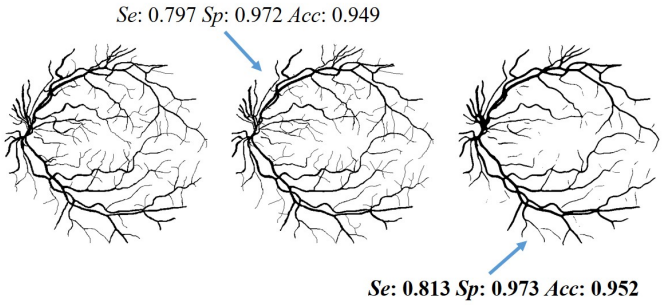

Fig. 3: Hard segmentation maps of the fundus image in Fig. 1. From left to right: the manual annotation made by the first observer, the manual annotation made by the second observer and the vessel segmentation map generated by N^4[19].

Due to the inter-observer problem in thickness variation and location variation, when we use the first observer's manual annotation as the reference, the scores of Se, Sp and Acc for the second observer's manual annotation will be 0.797, 0.972 and 0.949 respectively. Meanwhile the corresponding results for the third segmentation map (produced by the method of [19]) are 0.813, 0.973 and 0.952 respectively, all better than those for the second observer's. However, it is clear that the vessel tree in the second manual annotation is much better than that of the third segmentation map, indicating the ineffectiveness of these metrics.

Based on the above analysis, a good evaluation metric for retinal vessel segmentation should treat thick vessels and tiny vessels with equal importance. In addition, due to the inter-observer problem, the pixel-to-pixel matching strategy should not be used for comparison. Therefore, we propose to use the skeleton map for evaluation, which would make all vessels equally important (instead of making all pixels equally important as in traditional metrics). In contrast to pixel-to-pixel matching, each skeleton segment in the reference skeleton map is adaptively assigned with a searching range for similarity calculation, which helps achieve better tolerance to the inter-observer problem.

III. Skeletal Similarity Metric

The basic idea behind the proposed metric is evaluating the quality by analyzing the distortion of the segmented vessel tree. We design the skeletal similarity to address both location variation and thickness variation in different manual annotations which are two common issues caused by the inter-observer problem. Specifically, the skeletal similarity metric consists of a curve similarity to measure the structural similarity between the reference and the source skeleton maps and a thickness similarity to measure the thickness

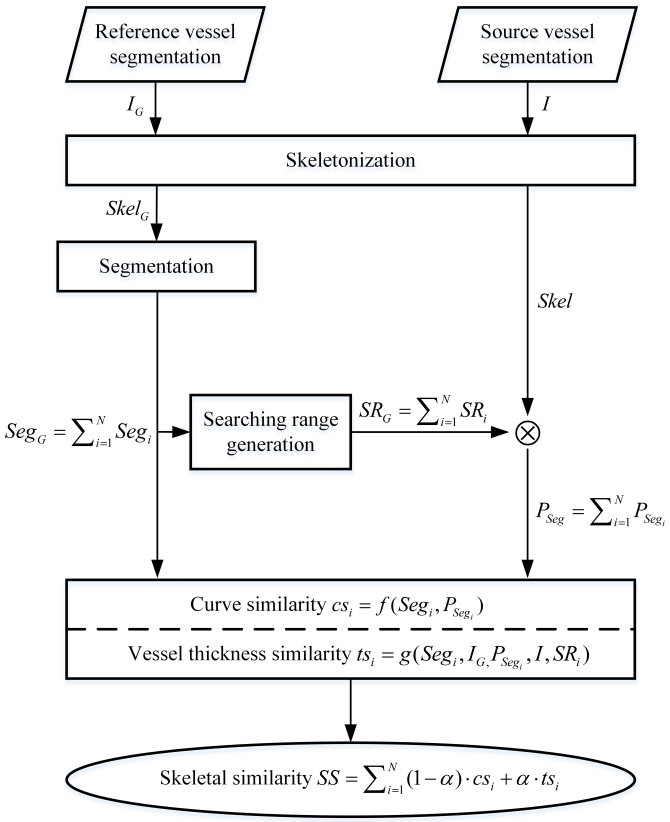

Fig. 4: Flowchart of the skeletal similarity metric evaluation process.

consistency between the reference and the source vessel segmentations. The flowchart of the skeletal similarity metric evaluation process is shown in Fig. 4. Given a reference vessel segmentation I_G and a source vessel segmentation I, skeletonization is applied to I_G and I to generate the corresponding skeleton maps $Skel_G$ and $Skel$. Then we segment $Skel_G$ to divide the entire skeleton into N segments, namely $Skel_G \rightarrow Seg_G = \sum_{i=1}^{N} Seg_i$. After that, we define a searching range for each segment in Seg_G based on its vessel thickness in I_G and the corresponding searching range of Seg_G derived as $SR_G = \sum_{i=1}^{N} SR_i$ where SR_i represents the searching range of the skeleton segment Seg_i. For each segment Seg_i, we find the corresponding pixels (denoted as P_{Seg_i}) in $Skel$ located within SR_i, namely $Skel \rightarrow P_{Seg} = \sum_{i=1}^{N} P_{Seg_i}$. Based on the segment correspondence, we calculate the curve similarity $cs_i = f(Seg_i, P_{Seg_i})$ and the thickness similarity $ts_i = g(Seg_i, I_G, P_{Seg_i}, I, SR_i)$. Accordingly, the final skeletal similarity is calculated as a summation of a weighted sum of cs_i and ts_i over all segments as: $SS = \sum_{i=1}^{N} (1-\alpha) \cdot cs_i + \alpha \cdot ts_i$. In following subsections, we describe each step in more details.

A. Skeletonization

Given a reference vessel segmentation I_G and a source vessel segmentation I, the first step is to extract the corresponding skeleton maps $Skel_G$ and $Skel$. Since the structure of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2778748, IEEE Transactions on Medical Imaging

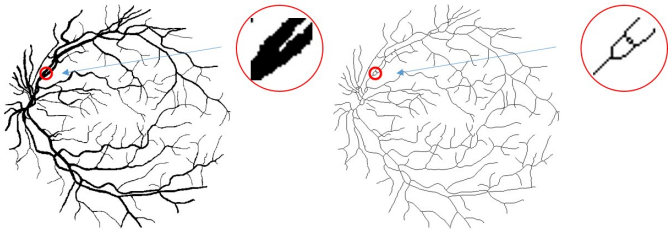

Fig. 5: Skeletonization of a given manual annotation. From left to right: the original manual annotation, the enlarged vessels, the corresponding skeleton map and the enlarged skeletons.

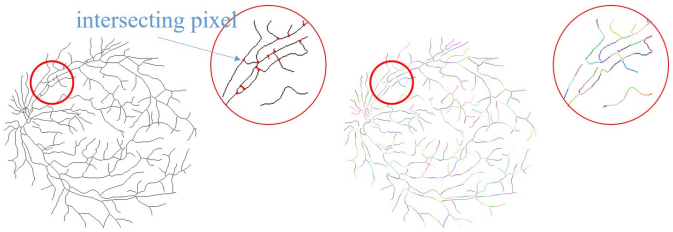

Fig. 6: Skeleton segmentation process. From left to right: the skeleton map, the enlarged skeletons, the skeleton segmentation and the enlarged skeleton segments. In the skeleton segmentation, different skeleton segments are assigned with different colors.

retinal vessel tree in fundus image is quite simple, we use the thinning method in [13] for skeletonization.

Illustrated by the skeleton map in Fig. 5, the thinning method can effectively extract skeletons. However, just like the enlarged region of the manually annotated vessel segmentation, vessels might be incorrectly connected by discrete pixels due to the limited resolution of the fundus image, which would be identified as skeletons in skeletonization, and these wrong skeletons should be removed for better evaluation. To achieve this goal, we divide the whole skeleton into segments. Details of the skeleton segmentation process are discussed in the following subsection *B*.

B. Skeleton Segmentation

Directly calculating the similarity between two skeleton maps is difficult. Instead, we divide the skeleton $Skel_G$ into small segments and perform a segment-level quality evaluation. Before segmenting the whole skeleton, we remove wrong skeletons generated by the skeletonizaiton process as discussed in subsection *A*. Since wrong skeletons are often caused by incorrectly connected pixels between two close vessels, these skeletons are usually quite short. After detecting all the intersecting pixels (namely pixels where different skeletons intersect as shown in Fig. 6) in $Skel_G$, we can divide all pixels in $Skel_G$ into different components. By removing the connected components which are less than a predefined minimum length $minLength$ (the default value is 4 pixels), we can effectively remove wrong skeletons. Meanwhile, for any component longer than the predefined maximum length $maxLength$ (the default value is 15 pixels), we further divide the component into smaller components, which would result in more effective skeletal similarity calculation. After the segmentation process, the reference skeleton map $Skel_G$ is represented by a set of skeleton segments (components) as

$$Seg_G = \sum_{i=1}^{N} Seg_i, \tag{2}$$

where N is the number of skeleton segments and the length of each segment is within the range of $[minLength, maxLength]$.

Illustrated by the skeleton segmentation in Fig. 6, all wrong skeletons in $Skel_G$ have been effectively removed. After the segmentation process, some true skeleton segments which are shorter than $minLength$ would also be removed. To roughly estimate how many true skeletons are included for the following skeletal similarity calculation, we define the $confidence$ as below:

$$confidence = \frac{\#\{Seg_G\}}{\#\{Skel_G\}}, \tag{3}$$

where $\#\{Seg_G\}$ is the number of pixels in Seg_G and $\#\{Skel_G\}$ is the number of pixels in $Skel_G$. Since $\#\{Skel_G\}$ includes the wrong skeletons generated by the skeletonization process, the real confidence should be slightly higher. According to the experimental results in this paper, the overall confidence is around 0.995. We believe that the overall confidence is enough for fair quality evaluation, which indicates the effectiveness of the skeleton segmentation process.

C. Searching Range Generation

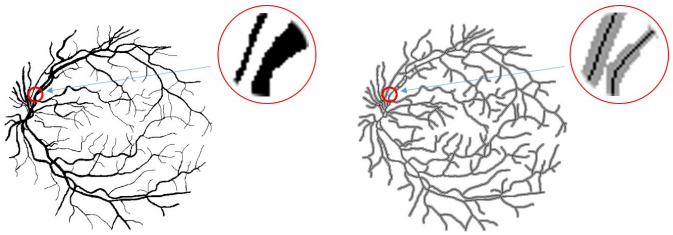

Fig. 7: Searching range generation. From left to right: the manual annotation, the enlarged vessels, the corresponding searching range and the enlarged parts of the searching range.

As discussed in Section II, the inter-observer variation problem incurs limited location variation yet noticeable thickness variation for thick vessels. On the contrary, for tiny vessels, location variation would be more serious than thickness variation. As a result, after skeletonization, skeleton locations of thick vessels are quite close in different manual annotations, while skeleton locations of tiny vessels can be quite different. To better tolerate these inter-observer variations, tiny vessels should be assigned with larger searching ranges than those of thick vessels. Therefore, the searching range of each pixel in Seg_G is adaptively determined according to its vessel thickness in I_G. For each pixel in Seg_G, we first calculate the minimum inscribed circle in I_G centered at the pixel which is completely covered by vessel pixels, and use the diameter

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2778748, IEEE Transactions on Medical Imaging

as its vessel thickness t. Then, for each pixel the searching radius r is

$$r = \begin{cases} R & \text{if } T_{\max} = T_{\min} \\ \left\lceil \frac{T_{\max} - t + \varepsilon}{\lfloor T_{\max} - T_{\min} \rfloor} \cdot R \right\rceil & \text{otherwise} \end{cases}, \quad (4)$$

where T_{max} and T_{min} represent the maximum and the minimum vessel thickness in I_G, R is the predefined maximum searching radius (the default value is 2 pixels), ε is a small positive value to guarantee that the minimum value of r is 1 pixel, $\lceil\rceil$ denotes the round-up operator and $\lfloor\rfloor$ is the round-down operator. Then, the whole searching range of Seg_G is defined as

$$SR_G = \sum_{i=1}^{N} SR_i, \quad (5)$$

where SR_i represents the searching range of the segment Seg_i. Fig. 7 shows exemplar results of the searching range generation process.

D. Skeletal Similarity Calculation

For each skeleton segment Seg_i in Seg_G, the pixels in $Skel$ located within the searching range SR_i are selected for the skeletal similarity calculation. Then the source skeleton map $Skel$ is converted to

$$P_{Seg} = \sum_{i=1}^{N} P_{Seg_i}, \quad (6)$$

where P_{Seg_i} represents the pixels in $Skel$ located within SR_i, and P_{Seg_i} can be $\emptyset$ for certain segments Seg_i.

Given the skeleton segment Seg_i and the corresponding pixels P_{Seg_i}, the skeletal similarity consists of the curve similarity cs_i and the thickness similarity ts_i. The curve similarity cs_i measures the structural consistency between Seg_i and P_{Seg_i}, regardless of the vessel thickness of Seg_i and P_{Seg_i}. Thus, it can better measure the completeness of the segmented vessel tree with respect to the reference which will be more suitable for clinical applications demanding accurate and complete tiny vessel segmentation, *e.g.* neovascularization detection in retinal images for diabetic retinopathy (DR) diagnosis [22], [23], [24] and registration of retinal images [25], [26], [27]. The thickness similarity ts_i measures the thickness variation between Seg_i and P_{Seg_i}. By defining the thickness similarity ts, we can measure the overall thickness consistency between the reference and the source vessel trees, which can be used as geometric features for several computer-aided diagnosis tasks [28], [29], [30].

1) Curve Similarity: To evaluate the structural similarity between Seg_i and P_{Seg_i}, the segment Seg_i is approximated by a fitted curve. Then, the same curve fitting method is applied to P_{Seg_i} to approximate its distribution in $Skel$. The similarity between Seg_i and P_{Seg_i} is measured based on the similarity between their approximated curve functions. Since the length of Seg_i is limited to $[minLength, maxLength]$ (namely [4, 15] pixels by default), a cubic function would be sufficient for approximation. Fig. 8 shows exemplar fitting results on Seg_i and P_{Seg_i}.

Letting $P_1(x) = a_1x^3 + b_1x^2 + c_1x + d_1$ represent the fitted curve of Seg_i and $P_2(x) = a_2x^3 + b_2x^2 + c_2x + d_2$ denote the fitted cubic function of P_{Seg_i}, we construct the

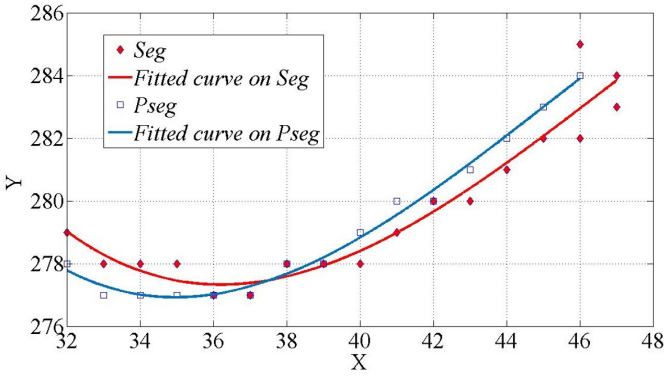

Fig. 8: Fitted curves on Seg_i and P_{Seg_i}. (x, y) in the figure represents coordinates of the pixel in fundus image.

corresponding coefficient vectors as $F_1 = [a_1, b_1, c_1]$ and $F_2 = [a_2, b_2, c_2]$. Here d_1 and d_2 are not used for the curve similarity calculation, as they would not contribute to the structural distortion analysis. The curve similarity cs_i between Seg_i and P_{Seg_i} is defined as

$$cs_i = \left| \frac{< F_1, F_2 >}{|F_1|\,|F_2|} \right| \in [0, 1], \quad (7)$$

where $<>$ is the dot product operation, and $||$ measures the length of the vector.

2) Thickness Similarity: To measure the thickness consistency between Seg_i and P_{Seg_i}, we first calculate the average vessel thickness $\bar{W}_{Seg_i}$ of Seg_i and the average vessel thickness $\bar{W}_{P_{Seg_i}}$ of P_{Seg_i}, based on the same method described in subsection C. As discussed in Section II, thickness variation of tiny vessels, commonly incurred in manual annotations from multiple observers, is more common than that of thick vessels. Therefore, the constraint on thickness consistency of thick vessels should be more restrictive than that of tiny vessels. Accordingly, the thickness similarity ts_i between Seg_i and P_{Seg_i} is defined as

$$ts_i = \begin{cases} 0 & if \frac{\left|\bar{W}_{Seg_i} - \bar{W}_{P_{Seg_i}}\right|}{\bar{W}_{SR_i}} \geq 1 \\ 1 - \frac{\left|\bar{W}_{Seg_i} - \bar{W}_{P_{Seg_i}}\right|}{\bar{W}_{SR_i}} & otherwise \end{cases}, \quad (8)$$

where $\bar{W}_{SR_i}$ is the average width of the searching range SR_i. As $\bar{W}_{SR_i}$ would be smaller for thicker vessels, the penalty on thickness variation of thick vessels is relatively greater.

As discussed before, the relative importance of structural similarity and thickness consistency is application dependent. To better meet the requirements of different applications, we adjust the relative impact of the curve similarity cs and the thickness similarity ts on the final skeletal similarity using a tunable parameter α as:

$$ss_i = (1 - \alpha) \cdot cs_i + \alpha \cdot ts_i. \quad (9)$$

In addition, if $|P_{Seg_i}| < 0.6 \cdot l_{Seg_i}$ is satisfied for a segment Seg_i, its skeletal similarity ss_i is automatically set to 0. After

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2778748, IEEE Transactions on Medical Imaging

calculating ss_i for each skeleton segment Seg_i, the overall skeletal similarity SS is defined as

$$SS=\frac{\sum_{Seg_i \in Seg_G} ss_i \times l_{Seg_i}}{\sum_{Seg_i \in Seg_G} l_{Seg_i}}, \tag{10}$$

where l_{Seg_i} is the length of the segment Seg_i.

E. Evaluation Metrics Definition

Fig. 9: Redefined vessels pixels (P_v) and non-vessel pixels (P_{nv}) in the reference image. P_v is $I_G \cup SR_G$.

Based on the overall skeletal similarity SS, we first redefine TP, FN, TN and FP. Since the skeletal similarity is not based on pixel-to-pixel matching, we need to redefine "vessel pixels" and "non-vessel pixels" in the redefined reference vessel segmentation I'_G. As all pixels located within the searching range SR_G contribute to the curve similarity calculation, pixels in I'_G located within the searching range are denoted as "vessel pixels". Meanwhile, in the thickness similarity calculation, since the vessel thickness of each vessel segment is calculated by referring to the reference vessel segmentation I_G, all annotated vessel pixels in I_G also are counted for "vessel pixels" in I'_G. As a result, in the redefined reference vessel segmentation I'_G, pixels are divided into P_v and P_{nv} as shown in Fig. 9, where P_v represents vessel pixels ($I_G \cup SR_G$) and P_{nv} represents non-vessel pixels in the FOV of I'_G.

Based on the division of P_v and P_{nv}, we redefine TP, FN, TN and FP as follows:

$$\begin{aligned} TP &= SS \times P_v \\ FN &= (1-SS) \times P_v \\ FP &= P_{nv}(1) \\ TN &= P_{nv}(0) \end{aligned}, \tag{11}$$

where $P_{nv}(1)$ represents the number of pixels that are wrongly classified as vessel pixels in P_{nv}, and $P_{nv}(0)$ is the number of pixels that are correctly classified as non-vessel pixels in P_{nv}. Then, we construct rSe, rSp and $rAcc$ as

$$\begin{aligned} rSe &= \frac{TP}{TP+FN} = SS \\ rSp &= \frac{TN}{TN+FP} = \frac{P_{nv}(0)}{P_{nv}} \\ rAcc &= \frac{TP+TN}{P_v+P_{nv}} = \frac{SS \times P_v + P_{nv}(0)}{P_v+P_{nv}} \end{aligned}. \tag{12}$$

We can further rewrite the definition of $rAcc$ as

$$rAcc = \frac{P_v}{P_v+P_{nv}} rSe + \frac{P_{nv}}{P_v+P_{nv}} rSp, \tag{13}$$

where $rAcc$ is a weighted sum of rSe and rSp. According to the division of P_v and P_{nv}, the size of P_v in I'_G is larger than the number of original vessel pixels in I_G. As a result, the weight $\frac{P_v}{P_v+P_{nv}}$ of rSe is greater compared to that of Se in the calculation of Acc. Thus, the weights of rSe and rSp in the calculation of $rAcc$ are better rebalanced, which makes $rAcc$ more reasonable for the overall performance evaluation.

IV. Evaluation

In this section, we manually implement the CAL [1] function according to the descriptions in [6], and compare the proposed skeletal similarity based metrics rSe, rSp and $rAcc$ with metrics Se, Sp, Acc and CAL on two public datasets DRIVE and STARE [14]. By adjusting the parameter α, we testify the performance of the proposed metrics for different purposes. In addition to quality evaluation of vessel segmentation, we further set $\alpha = 0$ to apply the proposed metric SS for quality evaluation of vessel centerline detection.

A. Metrics

In addition to the metrics Se, Sp and Acc defined in Section II, we also made comparison with the recently proposed metric CAL which was designed for quality evaluation of retinal vessel segmentation. Due to the lack of publicly available source code, we tried our best to implement the CAL function based the descriptions in [6]. Specifically, the CAL function consists of three factors C, A and L as follows.

$$C(I, I_G) = 1 - \min(1, \frac{|\#_C(I_G) - \#_C(I)|}{\#(I_G)}), \tag{14}$$

where $\#_C(I_G)$ and $\#_C(I)$ stand for the number of connected components in I_G and in I respectively, and $\#(I_G)$ is the cardinality of I_G. Factor C measures the fragmentation degree between I and I_G.

$$A(I, I_G) = \frac{\#(\delta_2(I) \cap I_G) \cup (I \cap \delta_2(I_G))}{\#(I \cup I_G)}, \tag{15}$$

where δ_2 is a morphological dilation using a disc of 2 pixels in radius. Factor A measures the degree of overlapping areas between I and I_G.

$$L(I, I_G) = \frac{\#((Skel \cap \delta_2(I_G)) \cup (\delta_2(I) \cap Skel_G))}{\#(Skel \cup Skel_G)}. \tag{16}$$

Factor L measures the degree of coincidence between I and I_G in terms of the total length. Finally, the CAL function is defined as

$$f(C, A, L) = C \cdot A \cdot L. \tag{17}$$

Here, all relevant parameters were set to the default values suggested in [6].

B. Comparison on DRIVE

For the DRIVE dataset, exemplar segmentation results of both manual annotations, HED [15], DeepVessel [16], CRF [17], DRIU [18] and N^4 [19] are shown in Fig. 10. The corresponding objective results of different metrics are shown in Table I. One interesting observation is that if we compare

[1] All experimental results and MATLAB source code of all metrics can be found at https://github.com/ZengqiangYan/SkeletalSimilarityMetric

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2778748, IEEE Transactions on Medical Imaging

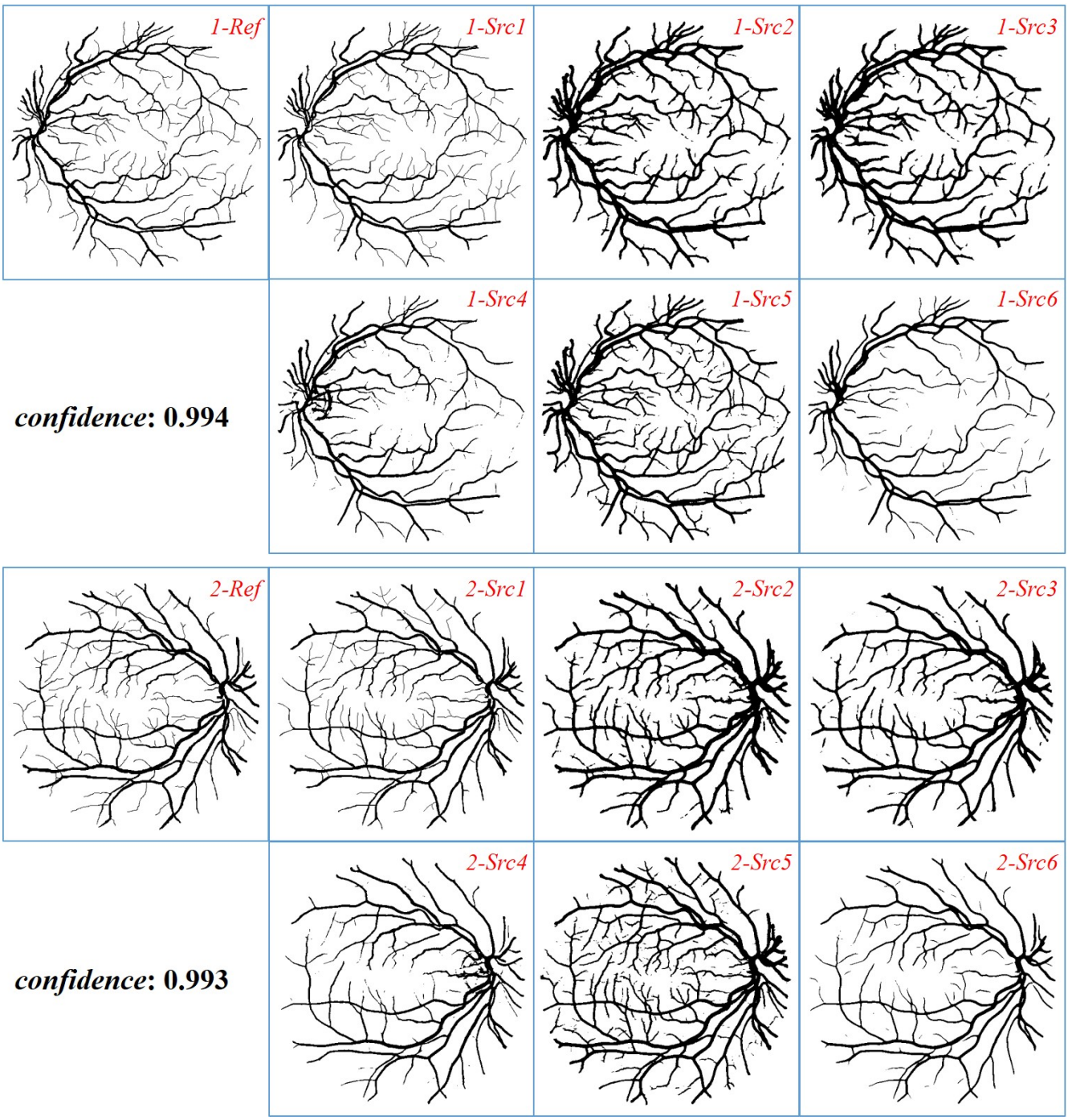

Fig. 10: Quality evaluation of retinal vessel segmentation for the DRIVE dataset. For each group, *Ref* is the reference image (*e.g.* the manual annotation generated by the first observer), and *Src1-Src6* are the vessel segmentations generated by the second human observer, HED [15], DeepVessel [16], CRF [17], DRIU [18] and N^4 [19] respectively.

the reference vessel segmentation with itself, the resulting scores of rSe and $rAcc$ might not be exactly 1.000, due to the imperfect skeleton segmentation process discussed in Section III.*B*. The overall scores of $confidence$ are 0.994 and 0.993 respectively, which we believe are high enough for fair evaluation. For the DRIVE dataset, two manual annotations are quite similar, namely two observers annotate both the basic vessel structure and tiny vessels consistently. Thus, in Table I, the scores of rSe, rSp and $rAcc$ for *1-Src1* and *2-Src1* are significantly improved compared to those of Se, Sp and Acc respectively. Obviously, with the increase of α, the results of rSe, rSp and $rAcc$ for *1-Src1* and *2-Src1* would decrease gradually, due to thickness variation caused by the inter-observer problem. Among different methods, if we set $\alpha = 0$, *1-Src5* and *2-Src5* have the best results in terms of rSe and $rAcc$, due to the completeness of the segmented vessel

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2778748, IEEE Transactions on Medical Imaging

TABLE I: Quantitative results of different metrics on DRIVE dataset

	Ref	Src						Ref	Src					
Metrics	1-Ref	1-Src1	1-Src2	1-Src3	1-Src4	1-Src5	1-Src6	2-Ref	2-Src1	2-Src2	2-Src3	2-Src4	2-Src5	2-Src6
Se	1.000	0.797	**0.968**	0.953	0.833	**0.979**	0.813	1.000	0.824	**0.965**	0.917	0.796	**0.974**	0.810
Sp	1.000	**0.972**	0.870	0.859	0.962	0.892	**0.973**	1.000	0.971	0.879	0.912	**0.974**	0.896	**0.980**
Acc	1.000	**0.949**	0.883	0.872	0.946	0.903	**0.952**	1.000	**0.949**	0.892	0.912	0.948	0.907	**0.955**
CAL	1.000	**0.901**	**0.883**	0.855	0.810	0.864	0.799	1.000	**0.890**	**0.862**	0.830	0.763	0.821	0.816
$\alpha = 0$														
rSe	0.990	**0.940**	0.861	0.865	0.781	**0.948**	0.720	0.980	**0.897**	0.846	0.754	0.718	**0.929**	0.743
rSp	1.000	**0.994**	0.956	0.951	0.989	0.966	**0.996**	1.000	**0.994**	0.948	0.968	0.992	0.959	**0.997**
rAcc	0.997	**0.980**	0.931	0.928	0.934	**0.962**	0.922	0.995	**0.968**	0.921	0.911	0.918	**0.951**	0.929
$\alpha = 1$														
rSe	0.999	**0.854**	0.309	0.250	**0.648**	0.512	0.603	0.998	**0.801**	0.326	0.363	0.615	0.547	**0.651**
rSp	1.000	**0.994**	0.956	0.951	0.989	0.966	**0.996**	1.000	**0.994**	0.948	0.968	0.992	0.959	**0.997**
rAcc	1.000	**0.957**	0.784	0.765	**0.899**	0.846	0.891	0.999	**0.942**	0.781	0.806	0.891	0.849	**0.904**

TABLE II: Quantitative results of different metrics on STARE dataset

	Ref	Src						Ref	Src					
Metrics	3-Ref	3-Src1	3-Src2	3-Src3	3-Src4	3-Src5	3-Src6	4-Ref	4-Src1	4-Src2	4-Src3	4-Src4	4-Src5	4-Src6
Se	1.000	**0.964**	**0.967**	0.957	0.742	0.721	0.951	1.000	0.978	**0.981**	**0.981**	0.760	0.779	0.976
Sp	1.000	0.931	0.936	0.897	**0.980**	**0.986**	0.884	1.000	0.938	0.938	0.909	**0.982**	**0.989**	0.890
Acc	1.000	0.934	0.939	0.903	**0.957**	**0.960**	0.891	1.000	0.942	0.943	0.917	**0.959**	**0.967**	0.899
CAL	1.000	0.649	**0.878**	**0.859**	0.700	0.628	0.834	1.000	0.771	**0.859**	**0.886**	0.789	0.744	0.844
$\alpha = 0$														
rSe	0.990	**0.987**	**0.925**	0.875	0.671	0.577	0.862	0.981	**0.975**	**0.959**	0.909	0.753	0.680	0.912
rSp	1.000	0.968	0.986	0.973	**0.998**	**0.999**	0.968	1.000	0.971	0.976	0.963	**0.995**	**0.997**	0.950
rAcc	0.998	**0.972**	**0.973**	0.951	0.925	0.904	0.945	0.997	**0.972**	**0.973**	0.953	0.952	0.940	0.943
$\alpha = 1$														
rSe	0.998	**0.802**	**0.616**	0.333	0.553	0.495	0.248	0.999	**0.674**	0.592	0.291	**0.596**	0.595	0.180
rSp	1.000	0.968	0.986	0.973	**0.998**	**0.999**	0.968	1.000	0.971	0.976	0.963	**0.995**	**0.997**	0.950
rAcc	1.000	**0.931**	**0.904**	0.830	0.898	0.886	0.807	1.000	0.917	0.907	0.842	**0.924**	**0.925**	0.811

trees. If α is set to 1, *1-Src4* and *2-Src6* would achieve the best results of rSe and $rAcc$ due to better thickness consistency. In terms of Se, *1-Src5* and *2-Src5* get better results than those of *1-Src1* and *2-Src1*, while *1-Src6* and *2-Src6* achieve better results of Sp and Acc than those for *1-Src1* and *2-Src1*. However, the annotated vessel trees in *1-Src1* and *2-Src1* are much better than those generated by DRIU [18] and N^4 [19], which indicates the ineffectiveness of Se for quality evaluation. In terms of CAL, *1-Src2* and *2-Src2* have the best results among different methods, and the results of different methods are quite similar and not distinguishable. Although *1-Src5* has better thickness consistency and vessel completeness than that of *1-Src2*, the CAL score of *1-Src2* is better than that of *1-Src5*. In addition, it is difficult to infer image content of the corresponding vessel segmentation map based on its CAL score.

C. Comparison on STARE

For the STARE dataset, exemplar segmentation results from both manual annotations, DRIU [18], HED [15], Wavelet [20] and DeepVessel [16] are shown in Fig. 11. The objective results of different evaluation metrics are provided in Table II. The overall scores of $confidence$ are 0.992 and 0.997 respectively. For the STARE dataset, the first observer mainly annotates the basic vessel structure, while the second observer annotates both the basic vessel structure and tiny vessels. Generally, vessels annotated by the second observer are thicker than those vessels annotated by the first observer. As a result, the results for *3-Src1* and *4-Src1* in terms of rSe, rSp and $rAcc$ decrease significantly with the increase of α. Among different methods, *3-Src2* achieves the best results of rSe and $rAcc$ under different settings of α, which is consistency with the subjective results in Fig. 11. Similarly, when α is set to 0, *4-Src2* achieves the best results of rSe and $rAcc$ compared to other vessel segmentations. When α is set to 1, *4-Src4* would have the best results of rSe as the skeletal similarity only measures the thickness consistency. Comparatively, although vessels in *4-Src3* are much thicker than those in *4-Src2*, *4-Src2* and *4-Src3* have the same results of Se. *3-Src5* and *4-Src5* achieve the best results of Sp and Acc, but their segmented vessel trees are quite incomplete compared to those of *3-Src2* and *4-Src2*. In terms of CAL, *3-Src2* and *4-Src3* achieve the best results respectively while the scores of *3-Src1* and *4-Src1* are much lower.

D. Quality evaluation on entire datasets

To evaluate the stability of the proposed evaluation metrics, we conduct experiments on the entire DRIVE and STARE

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2778748, IEEE Transactions on Medical Imaging

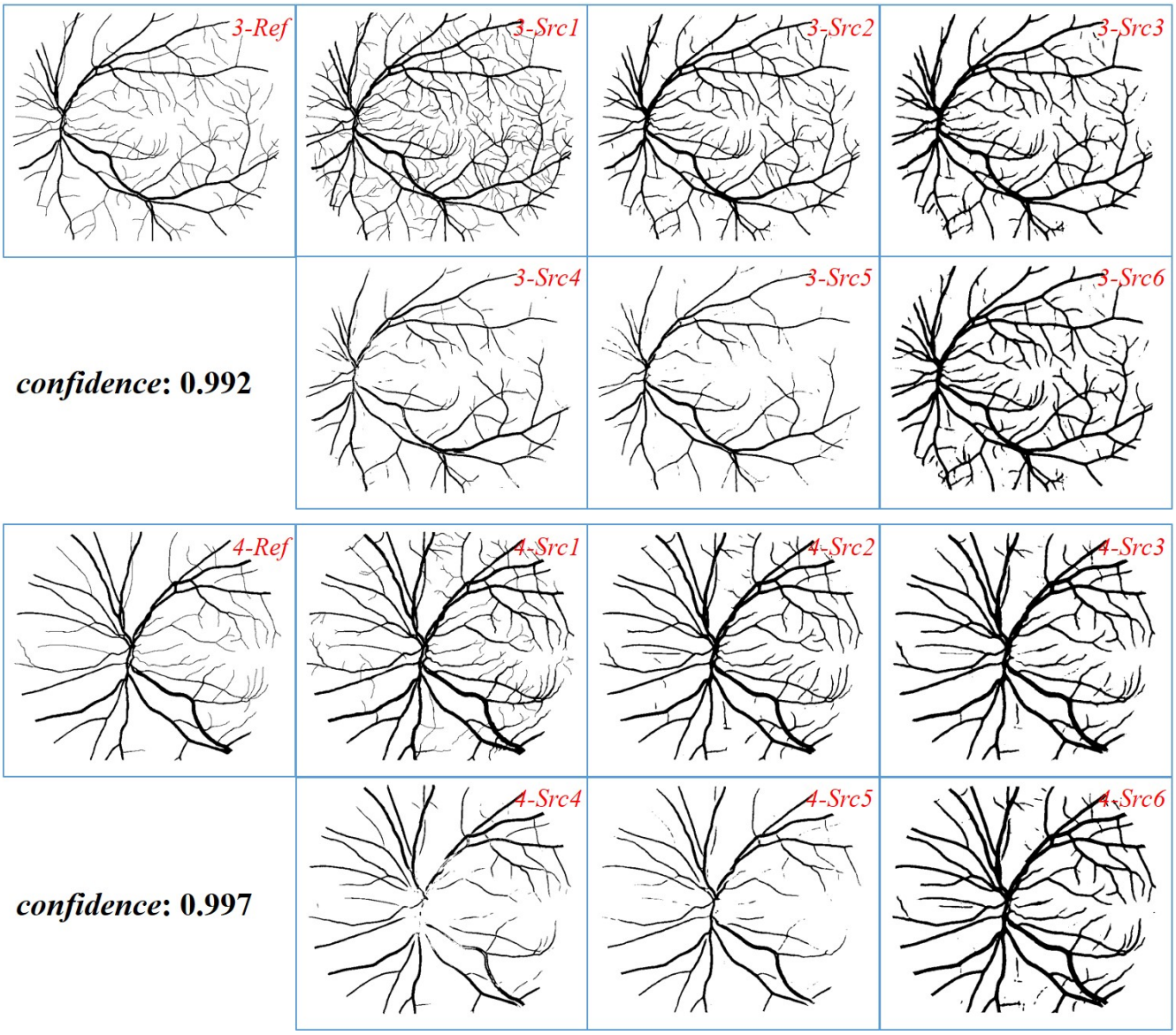

Fig. 11: Quality evaluation of retinal vessel segmentation for the STARE dataset. For each group, *Ref* is the reference image (*e.g.* the manual annotation generated by the first observer), and *Src1-Src6* are the vessel segmentations generated by the second human observer, DRIU [18], HED [15], Wavelet [20] and DeepVessel [16] (under two thresholds) respectively.

datasets by comparing the manual annotations generated by the second observer with those made by the first observer. Objective results of different evaluation metrics for the two datasets are shown in Fig. 12 and Fig. 13 respectively. For the DRIVE dataset, due to location variation in different annotations, the Se scores for the entire dataset are relatively low, even though the vessel structures annotated by different observers are quite similar. When parameter α is set as 0, the metric rSe (denoted as $csSe$) is able to solve the location variation problem which effectively improves the results, compared to those of Se. As described before, for the STARE dataset, the first observer mainly annotates thick vessels while the second observer annotates both thick vessels and thin vessels. In addition, since vessels annotated by the second observer generally are thicker than those annotated by the first observer, the results of $csSe$ and Se are similar except the first five annotations. Observing these five annotations, we found that vessels annotated by the first observer are thicker than those made by the second observer, which makes the improvements from Se to $csSe$ more obvious. In terms of the thickness similarity (denoted by $tsSe$), as the thickness inconsistency between annotations of the DRIVE dataset is less serious than that of the STARE dataset, the average results of $tsSe$ for the DRIVE dataset are better than those for the STARE dataset. By analyzing the performance of the proposed metric across different datasets, we demonstrate that the skeletal similarity metric is able to effectively handle the inter-observer problem with sufficient robustness.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2778748, IEEE Transactions on Medical Imaging

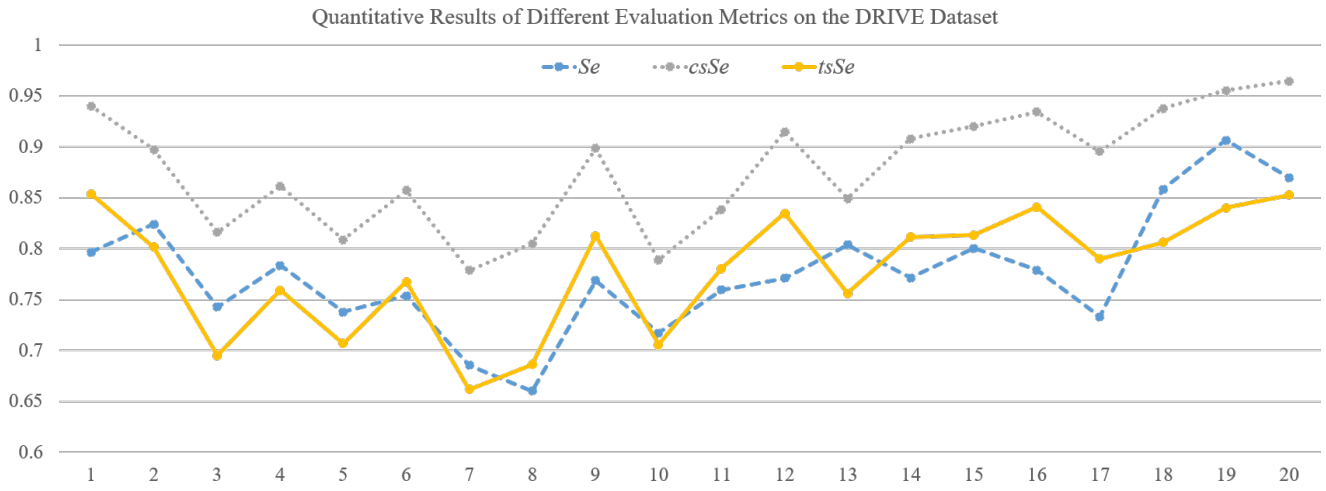

Fig. 12: Quantitative results of different evaluation metrics Se, $csSe$ and $tsSe$ for the DRIVE dataset by comparing the manual annotations generated by the second observer with those made by the first observer. Here, $csSe$ represents the metric rSe when α is set as 0 and $tsSe$ is the metric rSe when α is 1.

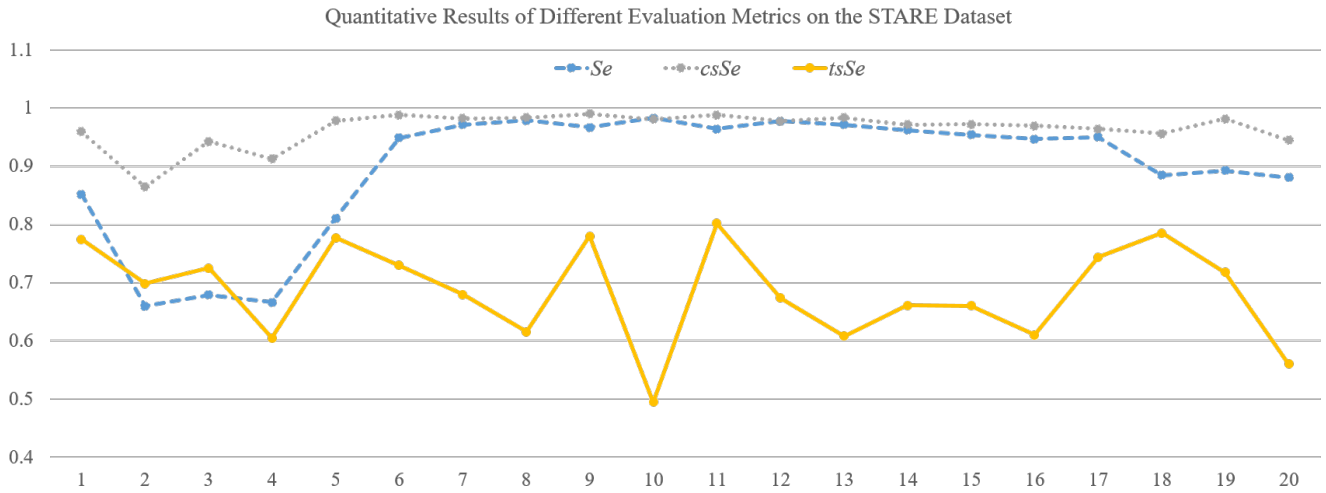

Fig. 13: Quantitative results of different evaluation metrics Se, $csSe$ and $tsSe$ for the STARE dataset by comparing the manual annotations generated by the second observer with those made by the first observer. Here, $csSe$ represents the metric rSe when α is set as 0 and $tsSe$ is the metric rSe when α is 1.

E. Quality evaluation of vessel centerline detection

When α is set to 0, the skeletal similarity is the same as the curve similarity which measures the structural similarity between two skeleton maps. Thus, the proposed skeletal similarity metric is applicable for quality evaluation of vessel centerline detection. For retinal vessel centerline detection, most recent algorithms [21], [31], [32] were developed based on the datasets of DRIVE and STARE, which provide no manually annotated vessel centerline maps. The ground true centerline maps are manually generated by applying a thinning method to the annotated vessel segmentation maps. In our experiment, the ground true vessel centerline maps are generated in the same way. Given the reference centerline map CL_{Ref}, we can generate the corresponding searching range map CL_{SR} according to the same method in Section III.C. Given a source centerline map CL_{Src}, the skeletal similarity SS is calculated by setting α to 0. Since the curve similarity is calculated based on the skeletons in CL_{Src} located within the searching range CL_{SR}, we further define R_{nc} to penalize the pixels in CL_{Src} located outside CL_{SR}. The definition of R_{nc} is as follows:

$$R_{nc} = \frac{\#(CL_{Src}) - \#(CL_{Src} \cap CL_{SR})}{\#(CL_{Ref})}, \qquad (18)$$

which measures the number of outliers in CL_{Src} with regard to the total number of pixels in CL_{Ref}.

Fig. 14 shows the generated vessel centerline maps by applying the thinning method [13] to the segmentation maps generated by different methods. In the reference vessel centerline map, according to the definition of the searching radius in (4), all pixels would have the same searching radius as the hyper-parameter R. By adjust the parameter R, the skeletal

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2778748, IEEE Transactions on Medical Imaging

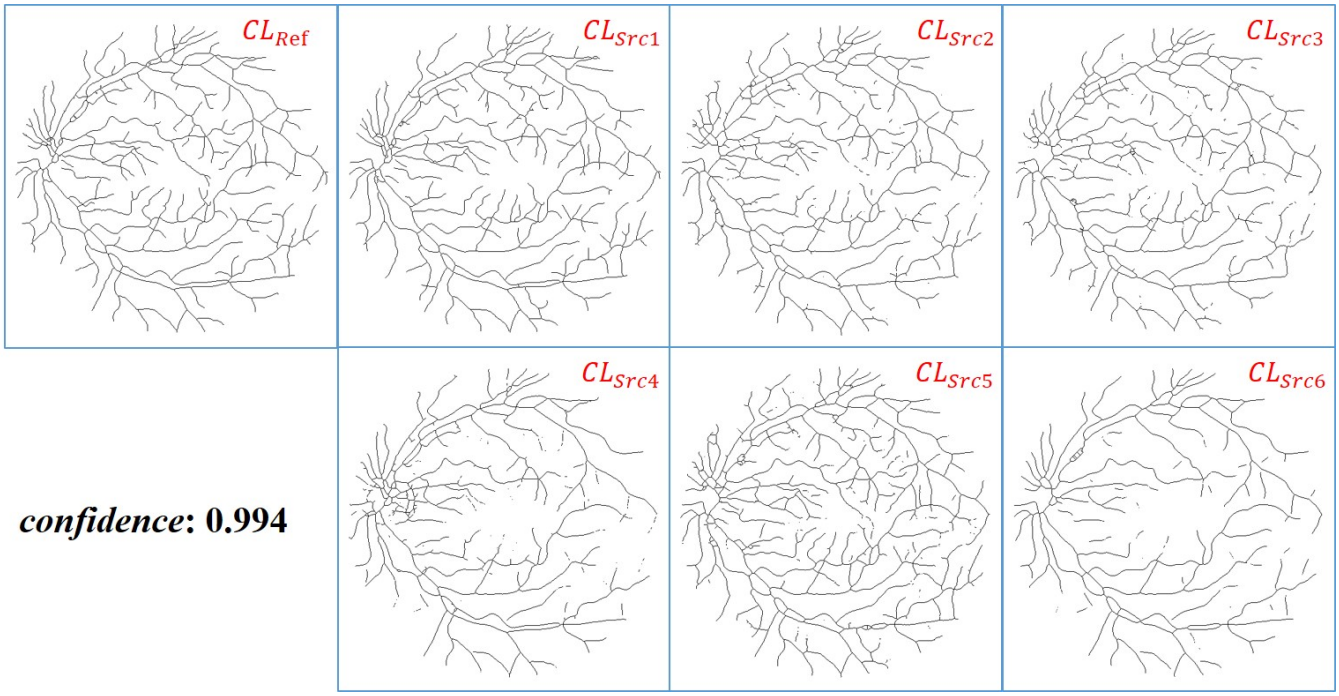

Fig. 14: Vessel centerline maps obtained by applying the thinning method [13] to the segmentation maps *1-Ref* and *1-Src1*∼*1-Src6* in Fig. 10.

TABLE III: Quantitative results of SS and R_{nc} for quality evaluation of vessel centerline detection

		Ref	Src					
R	Metric	CL_{Ref}	CL_{Src1}	CL_{Src2}	CL_{Src3}	CL_{Src4}	CL_{Src5}	CL_{Src6}
$minLength = 4, confidence = 0.994$								
1	SS	0.990	**0.933**	0.817	0.798	0.778	**0.923**	0.702
	R_{nc}	0.001	0.087	0.110	0.120	**0.044**	0.151	**0.024**
2	SS	0.990	**0.941**	0.864	0.867	0.781	**0.950**	0.721
	R_{nc}	0.001	0.055	0.059	0.059	**0.025**	0.110	**0.004**
3	SS	0.990	**0.941**	0.883	0.882	0.781	**0.956**	0.725
	R_{nc}	0.001	0.047	0.034	0.033	**0.020**	0.093	**0.001**
$minLength = 8, confidence = 0.979$								
1	SS	0.990	**0.936**	0.826	0.803	0.776	**0.926**	0.706
	R_{nc}	0.001	0.087	0.110	0.120	**0.044**	0.151	**0.024**
2	SS	0.990	**0.943**	0.870	0.871	0.781	**0.954**	0.724
	R_{nc}	0.001	0.055	0.059	0.059	**0.025**	0.110	**0.004**
3	SS	0.990	**0.943**	0.889	0.885	0.781	**0.961**	0.729
	R_{nc}	0.001	0.047	0.034	0.033	**0.020**	0.093	**0.001**

similarity can assess the quality of vessel centerline detection under different constraints. As shown in Table III, the results of SS for different centerline maps would increase with the increase of R while the corresponding results of R_{nc} would gradually decrease. Among centerline maps generated by different methods, CL_{Src5} achieves the best results of SS but has the worst results of R_{nc}. In terms of R_{nc}, CL_{Src6} achieves the best results under different settings of R, but the completeness of CL_{Src6} is quite low. Based on the results of both SS and R_{nc}, CL_{Src1} achieves the best overall accuracy.

To evaluate the stability of the proposed skeletal similarity metric, we further increase the value of $minLength$. As discussed in Section III.B, increasing the value of $minLength$ would remove more skeleton segments and decrease the overall confidence. However, from the results under different settings of $minLength$ in Table III, we find that results under different settings of $minLength$ are quite similar, which indicates that removing short skeletons in the skeleton segmentation process would not have non-trivial influence on the overall quality evaluation. Therefore, based on the requirements of different applications, we can choose different values of $minLength$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2778748, IEEE Transactions on Medical Imaging

V. Discussion

A. Parameter Selection of α

According to the definition of the skeletal similarity in (9), the parameter α is used to balance the relative importance between the curve similarity cs and the thickness similarity ts. The proposed metric would focus on the completeness of segmented vessel trees when α is set to 0. On the other hand, when α is set to 1, the metric measures the thickness consistency between the segmented vessel trees and the manually annotated vessel trees. The introduction of this tunable parameter α adjusts the proposed metric with different combinations of the two similarity measures for different target applications. However, given a specific application, it might be difficult to accurately determine the optimal value of the tunable parameter α. Furthermore, in practice, since the curve similarity cs and the thickness similarity ts evaluate the quality of vessel segmentations from different perspectives, it might make better sense for some applications to evaluate the segmentation results with respect to these two measures separately rather than jointly. Therefore, we first set $\alpha = 0$ to evaluate the vessel completeness of the segmented vessel trees. If the curve similarity is not sufficient to distinguish the vessel segmentation results under comparison, we then set $\alpha = 1$ to further assess the thickness consistency of the vessel segmentation results under comparison. On the other hand, for some other applications, if the curve similarity and the thickness similarity should be simultaneously considered for quality evaluation, the optimal value of α could be roughly estimated by selecting the value which .pngs the inter-observer variation between different manual annotations. Once the relative weight between the curve similarity and the thickness similarity is determined, the resulting combined skeletal similarity metric can then be applied for overall quality evaluation.

B. Thickness Similarity ts vs Se

As discussed in Section II, due to the pixel-to-pixel matching strategy and the highly imbalanced ratio between the basic vessel structure and tiny vessels, thickness variation of thick vessels dominates the calculation of Se. Comparatively, the thickness similarity ts measures thickness variation with respect to vessel segments instead of vessel pixels. Therefore, even for the case when one vessel segment contains several mismatched vessel pixels, it will not disproportionately influence the overall thickness similarity if the overall thickness variation is limited. As a result, using ts to calculate thickness similarity, the relative importance between thick vessels and tiny vessels is more balanced.

C. Limitation

The proposed skeletal similarity metric incurs a higher computational complexity. In our experiment, the average processing time for evaluating a testing image of the DRIVE and the STARE datasets is 24.79s and 37.28s respectively. The higher computational complexity is mainly due to the skeleton segmentation step and the curve fitting step.

D. Future Work

Since the proposed skeletal similarity metric can be regarded as a pixel-wise metric, the metric could also be used as a pixel-wise loss function to train deep learning models for retinal vessel segmentation or vessel centerline extraction. Following this direction, our future work includes exploring the performance of a deep learning network for vessel segmentation or vessel centerline extraction based on the skeletal similarity loss and reducing the time computational complexity of the skeletal similarity calculation for accelerating the training process.

VI. Conclusion

We have analyzed the most popular metrics sensitivity (Se), specificity (Sp) and accuracy (Acc) for quality evaluation of retinal vessel segmentation. Due to the inter-observer problem, the pixel-to-pixel matching strategy used for deriving these metrics is too restrictive to fairly assess the quality of vessel segmentation. In this paper, we propose to evaluate the quality of vessel segmentation by analyzing the distortion of the corresponding skeleton map. For each skeleton segment in the reference skeleton map, a searching range is adaptively generated, and pixels in the source skeleton map located within the searching range are used for quality evaluation. Rather than just counting the number of pixels within the searching range, we further define skeletal similarity to measure the structural similarity and the thickness consistency between the reference and the source vessel segmentations. Based on the skeletal similarity, we redefine the basic units TP, FN TN and FP and further redefine the evaluation metrics Se, Sp and Acc as rSe, rSp and $rAcc$ respectively according to these basic units. Experimental results have shown that rSe, rSp and $rAcc$ are more effective than the current metrics, especially when dealing with the inter-observer problem which is a common problem in the medical imaging field. Compared with other quality evaluation methods that provide a global score for overall evaluation, the proposed skeletal similarity metric has better extendibility to construct additional evaluation metrics, and has better potential to be used as a loss function for deep learning based pixel-wise retinal vessel segmentation.

References

[1] Y. Zhao, L. Rada, K. Chen, S. P. Harding, and Y. Zheng, "Automated vessel segmentation using infinite perimeter active contour model with hybrid region information with application to retinal images," *IEEE Trans. Med. Imag.*, vol. 34, no. 9, pp. 1797-1807, Mar. 2015.

[2] B. S. Y. Lam and Y. Hong, "A novel vessel segmentation algorithm for pathological retina images based on the divergence of vector fields," *IEEE Trans. Med. Imag*, vol. 27, no. 2, pp. 237-246, Feb. 2008.

[3] H. Narasimha-Iyer, J. M. Beach, B. Khoobehi, and B. Roysam, "Automatic identification of retinal arteries and veins from dual-wavelength images using structural and functional features," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 8, pp. 1427-1435, Aug. 2007.

[4] J. Orlando, E. Prokofyeva, and M. Blaschko, "A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 1, pp. 16-27, Jan. 2017.

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600-612, Apr. 2004.

[6] M. E. Gegúndez-Arias, A. Aquino, J. M. Bravo, and D. Marin, "A function for quality evaluation of retinal vessel segmentations," *IEEE Trans. Med. Imag.*, vol. 31, no. 2, pp. 231-239, Feb. 2012.

[7] G. Azzopardi, N. Strisciuglio, M. Vento, and N. Petkov, "Trainable COSFIRE filters for vessel delineation with application to retinal images," *Med. Image Anal.*, vol. 19, no. 1, pp. 46-57, 2015.

[8] B. Yin, H. Li, B. Sheng, X. Hou, Y. Chen, W. Wu, P. Li, R. Shen, Y. Bao, and W. Jia, "Vessel extraction from non-fluorescein fundus images using orientation-aware detector," *Med. Image Anal.*, vol. 26, no. 1, pp. 232-242, 2015.

[9] J. Zhang, B. Dashtbozorg, E. Bekkers, J. P. Pluim, R. Duits, and B. M. ter Haar Romeny, "Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores," *IEEE Trans. Med. Imag.*, vol. 35, no. 12, pp. 2631-2644, Aug. 2016.

[10] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, and T. Wang, "A cross-modality learning approach for vessel segmentation in retinal images," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 109-118, Jan. 2016.

[11] H. Fu, Y. Xu, D. W. K. Wong, and J. Liu, "Retinal vessel segmentation via deep learning network and fully-connected conditional random fields," in *Proc. ISBI*, 2016, pp. 698-701.

[12] J. Staal, M. D. Abramoff, 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, no. 4, pp. 501-509, Apr. 2004.

[13] L. Lam, S. -W. Lee, and C. Y. Suen, "Thinning methodologies-A comprehensive survey," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 14, no. 9, pp. 869-885, Sep. 1992.

[14] A. D. 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, no. 3, pp. 203-210, Mar. 2000.

[15] S. Xie, and Z. Tu, "Holistically-nested edge detection," in *Proc. ICCV*, 2015, pp. 1395-1403.

[16] H. Fu, Y. Xu, S. Lin, D. W. K. Wong, and J. Liu, "DeepVessel: Retinal vessel segmentation via deep learning and conditional random field," in *Proc. MICCAI*, 2016, PP. 132-139.

[17] J. I. Orlando, and M. Blaschko, "Learning fully-connected crfs for blood vessel segmentation in retinal images," in *Proc. MICCAI*, 2014, pp. 634-641.

[18] K. K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool, "Deep retinal image understanding," in *Proc. MICCAI*, 2016, pp. 140-148.

[19] Y. Ganin, and V. Lempitsky, "N^4-fields: Neural network nearest neighbor fields for image transforms," in *Proc. ACCV*, 2014, pp. 536-551.

[20] J. V. B. Soares, J. J. G. Leandro, R. M. Cesar, H. F. Jelinek, and M. J. Cree, "Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification," *IEEE Trans. Med. Imag.*, vol. 25, no. 9, pp. 1214-1222, Sep. 2006.

[21] M. Sofka, and C. V. Stewart, "Retinal vessel centerline extraction using multiscale matched filters, confidence and edge measures," *IEEE Trans. Med. Imag.*, vol. 25, no. 12, pp. 1531-1546, Dec. 2006.

[22] C. Heneghan, J. Flynn, M. O'Keefe, and M. Cahill, "Characterization of changes in blood vessel width and tortuosity in retinopathy of prematurity using image analysis," *Med. Image Anal.*, vol. 6, no. 4, pp. 407-429, 2002.

[23] A. Houben, M. Canoy, H. Paling, P. Derhaag, and P. de Leeuw, "Quantitative analysis of retinal vascular changes in essential and renovascular hypertension," *J. of Hypertension*, vol. 13, pp. 1729-1733, 1995.

[24] L. D. Hubbard, R. J. Brothers, W. N. King, L. X. Clegg, R. Klein, L. S. Cooper, A. R. Sharrett, M. D. Davis, and J. Cai, "Methods for evaluation of retinal microvascular abnormalities associated with hypertension/sclerosis in the atherosclerosis risk in communities study," *Ophthalmology*, vol. 106, pp. 2269-2280, Dec. 1999.

[25] 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, no. 5, pp. 419-428, 1999.

[26] 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, no. 12, pp. 1557-1563, 2004.

[27] C. Stewart, C. -L. Tsai, and B. Roysam, "The dual-bootstrap iterative closest point algorithm with application to retinal image registration," *IEEE Trans. Med. Imag.*, vol. 22, no. 11, pp. 1379-1394, 2003.

[28] A. Kifley, J. J. Wang, S. Cugati, T. Y. Wong, and P. Mitchell, "Retinal vascular caliber and the long-term risk of diabetes and impaired fasting glucose: the blue mountains eye study," *Microcirculation*, vol. 15, no. 5, pp. 373-377, 2008.

[29] M. K. Ikram, J. A. Janssen, A. M. Roos, I. Rietveld, J. C. Witteman, M. M. Breteler, A. Hofman, C. M. Van Duijn, and P. T. de Jong, "Retinal vessel diameters and risk of impaired fasting glucose or diabetes the rotterdam study," *Diabetes*, vol. 55, no. 2, pp. 506-510, 2006.

[30] M. L. Baker, P. J. Hand, J. J. Wang, and T. Y. Wong, "Retinal signs and stroke revisiting the link between the eye and brain," *Stroke*, vol. 39, no. 4, pp. 1371-1379, 2008.

[31] O. Wink, W. J. Niessen, and M. A. Viergever, "Multiscale vessel tracking," *IEEE Trans. Med. Imag.*, vol. 23, no. 1, pp. 130-133, 2004.

[32] M. M. Fraz, S. A. Barman, P. Remagnino, A. Hoppe, A. Basit, B. Uyyanonvara, A. R. Rudnicka, and C. G. Owen, "An approach to localize the retinal blood vessels using bit planes and centerline detection," *Comput. Methods Programs Biomed.*, vol. 108, no. 2, pp. 600-616, 2012.