Novel Evaluation Metrics for Vascular Structure Segmentation

Marcel Reimann[1], Weilin Fu[1], Andreas Maier[1]

[1]Pattern Recognition Lab, FAU Erlangen-Nürnberg
marcel.reimann@fau.de

Abstract. For the diagnosis of eye-related diseases segmentation of the retinal vessels and the analysis of the tortuousness, completeness, and thickness of these vessels are the fundamental steps. The assessment of the quality of the retinal vessel segmentation, therefore, plays a crucial role. Conventionally, different evaluation metrics for retinal vessel segmentation have been proposed. Most of them are based on pixel matching. Recently, a novel non-global measure has been introduced. It focuses on the skeletal similarity between vessel segments rather than the pixel-wise overlay and redefines the terms of the confusion matrix. In our work, we re-implement this evaluation algorithm and discover the design flaws in the algorithm. Therefore, we propose modifications to the metric. The basic structure of the algorithm, which combines the thickness and curve similarity is preserved. Meanwhile, the calculation of the curve similarity is modified and extended. Furthermore, our modifications enable us to apply the evaluation metric to three-dimensional data. We show that compared to the conventional pixel matching-based metrics our proposed metric is more representative for cases where vessels are missing, disoriented, or inconsistent in their thickness.

1 Introduction

The analysis of fundus images plays an important role in the diagnosis of many eye-related diseases, such as Diabetic Retinopathy (DR), Glaucoma and age-related macular degeneration, which are the leading causes of vision loss according to [1]. Segmentation of the retinal vessels is the fundamental step for fundus image analysis, and provides the physician with information on the thickness, tortuousness, and completeness of the retinal vessels [2]. During the past decades, many algorithms have been developed to generate accurate, fast, and robust retinal vessel segmentation. However, the evaluation metrics to assess the quality of these segmentation results are mainly based on pixel-wise overlapping and fail to capture the vascular structures of the vessels [2].

Recently, a new metric which utilizes the skeletal similarities between vessel segments to redefine each term in the confusion matrix has been proposed by Yan *et al.* [3]. The metric is a weighted combination of the thickness and curve similarity between the vessel segments of the ground truth and the test segmentation. Their metric overcomes the inter-observer problems of ground truths in evaluation data sets and ensures the completeness of the vessel tree in the test segmentation. These are significant improvements compared to traditional metrics. However, when re-implementing their algorithm, several design flaws are observed. In this work, we closely examine the algorithm by Yan *et al.*, point out their mathematical shortcomings, and propose solutions to the discovered problems. In this way, we can obtain a correct 2-D skeletal similarity-based evaluation metric and potentially lift the algorithm into 3-D. This enables the usage for more applications, such as the evaluation of airway segmentation.

2 Materials and Methods

2.1 Existing algorithm

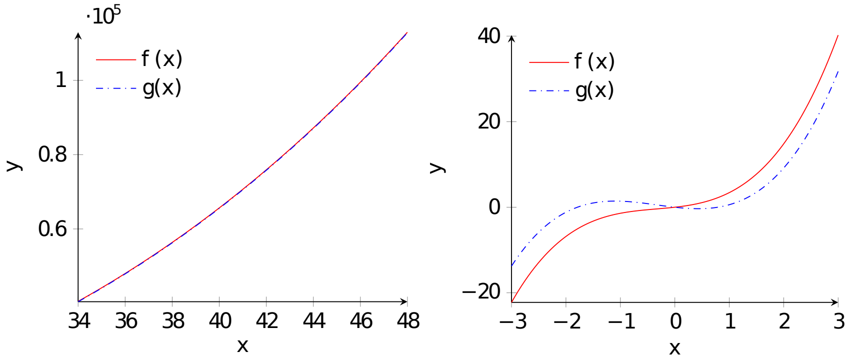
In the algorithm by Yan *et al.* the test and reference segmentations are skeletonized and the thickness of the vessel at each point of the skeleton is computed. Afterwards, the skeleton of the reference is cut into smaller components with a length in the range of $[4, 15]$ pixels. In the next step, each of the components gets assigned a searching range of $[1, 2]$ pixels and the average vessel thickness of the component is computed. Afterwards, a third order polynomial is fitted to the points of the reference skeleton and the points of the test skeleton within the searching range. Next, the first three coefficients of the cubic functions are compared using the cosine similarity which results in the curve similarity of that component. The thickness similarity can be evaluated by comparing the average thickness of the vessels in that part of the segmentation. The curve and thickness similarity are then balanced by the parameter $\alpha \in\ [0, 1]$ resulting in the skeletal similarity of the components mentioned above. The overall score is computed by summing up the individual scores weighed by the length of the corresponding reference component. Eventually, the values for true positives, false negatives, false positives, and true negatives are redefined using the searching range and the computed skeletal similarity.

While reviewing their algorithm, we discovered the computation of the curve similarity to be inaccurate. The coefficients of cubic functions may describe the curvature globally. However, looking at small intervals of the function, the coefficients alone cannot be used to calculate the curve similarity. As an example, we demonstrate the obvious similarity in curvature of $f(x) = x^3 + x^2 + \sqrt{2}x$ and $g(x) = x^3 + x^2 - \sqrt{2}x$ on the interval $[34, 48]$ compared to the dissimilarity for values close to the origin (Fig. 1).

However, if we compute the cosine similarity of the coefficients, the similarity results to 0.0. In conclusion, even though the algorithm by Yan *et al.* seems to provide accurate results, the metric is mathematically inaccurate.

2.2 Modification of the algorithm

Even though there is an error in the algorithm, we follow the idea behind it and try to modify the computation to yield correct and reasonable results.



(a) Interval $[34, 48]$ (b) Interval $[-3, 3]$

Fig. 1. Curvature of the functions $f(x)$ and $g(x)$ on different intervals.

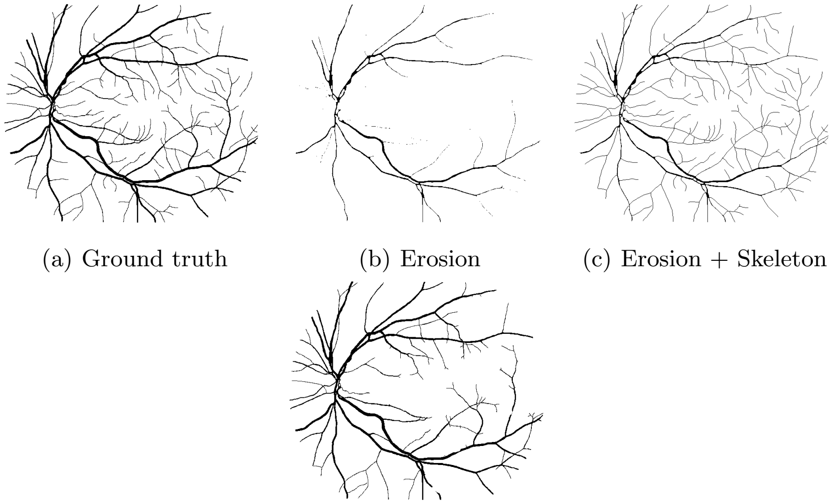
Piece-wise curvature of retinal vessel trees Firstly, we had a look at the curvature of all components of the skeleton, to check whether a third order polynomial as an approximation is necessary. We tried cubic, quadratic, and linear approximations to the components and evaluated the Mean Squared Error (MSE). As expected, the MSE decreased by an increasing order of the polynomial. However, a linear approximation yields a MSE of 3.389 pixels on average. We consider this a negligible error, which is underlined by Heneghan *et al.* [4], stating that retinal vessel trees are piece-wise linear. Therefore, we are able to proceed with linear instead of cubic approximations.

Modifications based on singular value decomposition To make our metric expandable to three-dimensional data, we choose to perform a singular value decomposition on the mean subtracted data and select the first eigenvector as the orientation vector of the linear approximation. The curve similarity is then calculated using the cosine similarity of the two corresponding orientation vectors.

Modifications based on directed Hausdorff distance Another option to replace the curve similarity cs is a point-wise comparison of the components A of the reference and B of the test segmentation. The maximum distance $\check{H}(A, B)$ is then weighed by the length l_{ref} of the reference component and subtracted from 1 in order to receive similarity scores in the interval $[0, 1]$

$$cs = 1 - \frac{\check{H}(A, B)}{l_{ref}} \tag{1}$$

As comparison metric we choose the directed Hausdorff distance, because it is commonly known, and used for the evaluation of image segmentation, even for three-dimensional data [5].



(a) Ground truth (b) Erosion (c) Erosion + Skeleton

(d) Deletions

Fig. 2. Evaluation data set based on STARE image no. 162.

2.3 Database

In this work, the STARE database [6] is utilized to evaluate the proposed metric. The STARE database contains 20 fundus images of shape 605 × 704 pixels. Each image is provided with two manually annotated label maps. The manual annotations from the first expert is utilized as the ground truth. Different modifications, such as morphological operations and small vessel removal are applied on the ground truth to generate test images. In this work we show exemplary results for erosion, erosion plus skeleton and deletions as shown in Figure 2.

3 Results

In the evaluation process we compare the scores of specificity, sensitivity, and accuracy of the traditional computation, the algorithm by Yan *et al.* and the two proposed modifications. To assess the performance of the curve and the thickness similarity, we set the parameter α to 0 and 1, respectively. In addition to the results (Tab. 1, 2, 3), we first verify the metric by comparing the ground truth image with itself, receiving higher scores than the existing algorithm. For Hausdorff we receive an average sensitivity of 99.98%, for SVD 99.99% and for Yan's algorithm 97.66% with α set to 0. When α is set to 1 the score of the modifications with 99.739% is almost indiscernible higher than of Yan's metric with an average sensitivity of 99.738%. The difference is caused by the restriction in Yan's metric that the test component of the skeleton must be at least 0.6 times the size of the reference, otherwise the similarity score is directly set to 0. In our metric we do not impose this restriction leading to higher scores and a better comparison of all the pixels in the segmentation.

Table 1. Erosion evaluation of STARE image No. 162.

	alpha = 0.0			alpha = 1.0		
	Se	Sp	Acc	Se	Sp	Acc
Traditional	29.529	100.000	94.979	29.529	100.000	94.979
Yan *et al.*	31.828	100.000	84.759	14.640	100.000	80.917
Hausdorff	35.393	100.000	85.556	26.934	100.000	83.665
SVD	43.851	100.000	87.447	26.934	100.000	83.665

Table 2. Erosion + Skeleton evaluation of STARE image No. 162.

	alpha = 0.0			alpha = 1.0		
	Se	Sp	Acc	Se	Sp	Acc
Traditional	52.706	100.000	96.630	52.706	100.000	96.630
Yan *et al.*	97.597	100.000	99.463	78.263	100.000	95.140
Hausdorff	99.617	100.000	99.914	78.263	100.000	95.140
SVD	99.945	100.000	99.988	78.263	100.000	95.140

Table 3. Deletion evaluation of STARE image No. 162.

	alpha = 0.0			alpha = 1.0		
	Se	Sp	Acc	Se	Sp	Acc
Traditional	88.746	100.000	99.198	88.746	100.000	99.198
Yan *et al.*	74.303	100.000	94.255	75.885	100.000	94.609
Hausdorff	76.716	100.000	94.794	80.266	100.000	95.588
SVD	80.473	100.000	95.634	80.266	100.000	95.588

Table 4. Noise evaluation of STARE image no. 162.

	alpha = 0.0			alpha = 1.0		
	Se	Sp	Acc	Se	Sp	Acc
Traditional	97.660	100.000	99.833	97.660	100.000	99.833
Yan *et al.*	98.026	100.000	99.559	89.875	100.000	97.736
Hausdorff	96.383	100.000	99.191	89.875	100.000	97.736
SVD	99.859	100.000	99.969	89.875	100.000	97.736

4 Discussion

In this work, we propose an effective evaluation metric for retinal vessel segmentation, which focuses on both the skeletal and thickness similarities of the vessel segments. The metric is applied on the STARE database, to compare the annotated images which are modified using morphological operations or vessel removal with the ground truth annotations. As shown in Table 2, for the case of eroded ground truth combined with the skeleton, the skeletal similarity is preserved, and our metric maintains high sensitivity and accuracy values. Meanwhile, our evaluation method is more sensitive than traditional metrics to cases where the completeness of the vessel tree is impaired, for instance when small vessels are removed as shown in Table 3. The properties of the Hausdorff distance make the respective modification sensitive to changes in tortuousness within the components and also their length. In comparison, the score of the SVD modification is affected only by changes regarding the orientation of the vessels' components and cannot reflect their individual length. Therefore, the scores shown in Table 1 are higher for the SVD modification. However, since the Hausdorff value is sensitive to outliers, it is observed that additive noise greatly affects the Hausdorff metric (Tab. 4). This is a disadvantage of that modification. In Table 1 we give an example for a combination of thickness and skeleton variation. In general, we recommend looking at different values for α, as low values introduce the curve similarity and high values the thickness similarity into the overall scores.

For future work, a user study could be carried out to confirm the effectiveness and superiority of the proposed metrics in a clinical environment. The proposed evaluation metrics could be directly applied on assessing the quality of the segmentation of other 2-D vascular structures. More experiments could be conducted to extend our method into 3-D and apply it on assessing the quality of more complicated tasks, such as airway segmentation.

References

1. Cheloni R, Gandolfi SA, Signorelli C, et al. Global prevalence of diabetic retinopathy: protocol for a systematic review and meta-analysis. BMJ Open. 2019;9(3):e022188.
2. Chetan L Srinidhi, P Aparna, Jeny Rajan. Recent advancements in retinal vessel segmentation. J Med Syst. 2017;41(4):1–22.
3. Yan Z, Yang X, Cheng KT. A skeletal similarity metric for quality evaluation of retinal vessel segmentation. IEEE Trans Med Imaging. 2018;37(4):1045–1057.
4. Heneghan C, Flynn J, O'Keefe M, et al. Characterization of changes in blood vessel width and tortuosity in retinopathy of prematurity using image analysis. Med Image Anal. 2002;6(4):407–429.
5. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015;15.
6. Hoover AD, Kouznetsova V, Goldbaum M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Trans Med Imaging. 2000;19(3):203–210.