International Journal of Computer Vision and Signal Processing, 1(1), 1-8(2012) ORIGINAL ARTICLE

A Quantitative Measure for Retinal Blood Vessel Segmentation Evaluation

Uyen T. V. Nguyen*, Kotagiri Ramamohanarao

Department of Computing and Information Systems
The University of Melbourne, Australia

Laurence A. F. Park

School of Computing, Engineering and Mathematics
University of Western Sydney, Australia

Liang Wang

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, China

Alauddin Bhuiyan

ICT Centre
Commonwealth Scientific and Industrial Research Organization, Australia

Abstract

Analysis of retinal blood vessels allows us to identify individuals with the onset of cardiovascular diseases, diabetes and hypertension. Unfortunately, this analysis requires a specialist to identify specific retinal features which is not always possible. Automation of this process will allow the analysis to be performed in regions where specialists are non-existent and also large scale analysis. Many algorithms have been designed to extract the retinal features from fundus images. However, to date, these algorithms have been evaluated using generic image similarity measures without any justification of the reliability of these measures. In this article, we study the applicability of different measures for retinal vessel segmentation evaluation task. In addition, we propose an evaluation measure, F_1, which is based on precision, recall and F-measure concept to deal with this evaluation task. An important property of F_1 is its tolerance of small localization errors which often appear in a segmented image, but do not affect the desired retinal features. The performances of different measures are tested on both real and synthetic datasets which take into account the important properties of retinal blood vessels. The results show that F_1 provides the greatest correlation to the desired evaluation measure in all experiments. Thus, it is the most suitable measure for retinal segmentation evaluation task.

Keywords: evaluation measure, retinal image, segmentation, F-measure

IJCVSP
International Journal of Computer Vision and Signal Processing

ISSN: 2186-1390 (Online)
http://www.ijcvsp.com

Article History:
Received: 9 August 2011
Revised: 14 March 2012
Accepted: 5 September 2012
Published Online: 8 September 2012

1. INTRODUCTION

Changes in retinal blood vessel features (e.g. vessel caliber, branching angle, etc.) are precursors of serious diseases such as cardiovascular diseases and stroke [1]. Therefore, an analysis of retinal vessel features can assist in detecting these changes and allow the patient to take action while the disease is still in its early stage. An automated retinal analysis system would allow us to perform reti-

*Corresponding author
Email addresses: thivun@student.unimelb.edu.au (Uyen T. V. Nguyen), kotagiri@unimelb.edu.au (Kotagiri Ramamohanarao), lapark@scm.uws.edu.au (Laurence A. F. Park), wangliang@nlpr.ia.ac.cn (Liang Wang), alauddin.bhuiyan@csiro.au (Alauddin Bhuiyan)

nal analysis in regions of the world where specialists are not available and reduce the cost associated with trained graders and remove the issue of inconsistency introduced by manual grading. Among different retinal analysis tasks, retinal blood vessel extraction plays an extremely important role as it provides the prerequisite result before any measurements may be made. Although many algorithms have been designed for retinal blood vessel segmentation [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17], there has been no investigation of how we should evaluate the effectiveness of these algorithms. So far, evaluation of these algorithms has been performed using generic binary image similarity measures such as the accuracy ([6, 10, 7, 11, 9, 5, 8, 4, 13, 12, 14]), true positive rate and false positive rate ([10, 7, 4, 12]). However, they are used without any justification or analysis of the reliability of these measures. Motivated by this, this paper studies the suitability of different measures to address this evaluation task. A modified version of F_1 score (called F_1) which allows small localization error which is often present in a segmented image is also proposed to deal with this task. A set of experiments tested on synthetic and real datasets show that the proposed measure provides the most suitable evaluation scores and thus is the best measure to be used for retinal image segmentation evaluation.

The rest of the paper is organized as follows. In section 2, a brief introduction of currently used measures for retinal vessel extraction evaluation is presented. A formal description of the proposed measure, F_1, is given in section 3. In section 4 and 5, we present the results of the experiments on synthetic and real datasets, respectively.

2. EXISTING MEASURES

Before introducing general measures used for comparing binary images, we first present the notations used in this paper based on the definitions in [18]. Suppose that I_A and I_B are two binary images representing the ground truth and the estimated segmentations of a given image, respectively. Each pixel of I_A and I_B takes one of two values: 0 (background pixels) and 1 (object pixels). We can describe I_A and I_B by their set of object pixel coordinates A and B, respectively. Let X denote the set of all possible pixel coordinates of the given image and $n(\cdot)$ be the set cardinality operator. For example, $n(X)$ is the number of elements of set X. Set operators such as minus ($\setminus$), union ($\cup$), intersection ($\cap$), and negation ($\neg$) are defined as in set context. For example, $A \setminus B$ is the set of all object pixels belonging to A but not B, $A \cup B$ is a set of all object pixels belonging to either A or B, $A \cap B$ is the set of all object pixels belonging to both A and B, $\neg A$ is the set of all pixels of the image except those pixels in A (or background pixels).

2.1. Existing measures for retinal blood vessel segmentation

The accuracy measure (ACC), true positive rate (TPR) and false positive rate (FPR) are prevalent in retinal vessel segmentation evaluation. They are defined as follows:

$$ACC(A,B) = \frac{n((A \cap B) \cup (\neg A \cap \neg B))}{n(X)} \quad (1)$$

$$TPR(A,B) = \frac{n(A \cap B)}{n(A)} \quad (2)$$

$$FPR(A,B) = \frac{n(B \setminus A)}{n(X \setminus A)} \quad (3)$$

They are called pixel based measures since the evaluation score is computed based on the number of correctly or incorrectly classified pixels. A drawback of these measures is their inability to deal with small localization errors. For example, if two objects detected by two different segmentations are identical but one of them is a shifted version (by a small distance) of the other, there are many pixels identified as incorrectly detected and the similarity score returned by these measures would be lower than expected. Such errors can easily appear in a segmentation result especially at the boundary of an object since the exact location of an object edge is difficult to establish. This quality leads us to hypothesize that pixel based measures are not suitable for retinal image vessel segmentation evaluation.

2.2. Distance based measures

In contrast to pixel based measures which only utilize the number of mis-segmented pixels, distance measures are based on the position of mis-segmented pixels. Typical measures are the mean squared error distance (MSE) [18], the Hausdorff distance (H) [18], the figure-of-merit (FOM) [19], and the p-order mean difference (Δ^p) [18].

Suppose that A is the set of object pixels of a binary image and $d(x, A)$ denotes the shortest distance from pixel x to A. $d(x, A)$ is defined as:

$$d(x,A) = \min\{\rho(x,y) : y \in A\} \quad (4)$$

where $\rho(\cdot,\cdot)$ is a metric and $\rho(x, y)$ is the distance between pixel x and y. The mean squared error distance (MSE) [18] is defined as the average squared shortest distance of all pixels in the estimated image B to A:

$$MSE(A,B) = \frac{1}{n(B)} \sum_{x \in B} d^2(x,A) \quad (5)$$

The Hausdorff distance (H) [18] measures the largest distance among all of the shortest distance connecting every pixel from A to B and from B to A:

$$HM(A,B) = \max\{\max_{x \in B} d(x,A), \max_{x \in A} d(x,B)\} \quad (6)$$

Figure of merit (FOM) [19] is a popular measure for evaluating edge detection algorithms:

$$FOM(A,B) = \frac{1}{\max\{n(A), n(B)\}} \sum_{x \in B} \frac{1}{1 + \alpha d^2(x, A)} \quad (7)$$

where α is a constant and is often set to 1/9 [18]. The p-order mean difference (Δ^p) [18] is a defined as:

$$\Delta^p(A,B) = \left[\frac{1}{n(X)} \sum_{x \in X} |w(d(x,A)) - w(d(x,B))|^p \right]^{1/p} \quad (8)$$

The function $w(\cdot)$ is called cutoff transformation and defined as:

$$w(t) = \min\{t, c\} \quad (9)$$

where the cutoff distance, c, is a constant positive value. It is set to 5 (as in [18]) in all of our experiments.

By taking into account the distance information, distance based measures can overcome the limitation of pixel based methods to some extent. Therefore, the performance and applicability of these methods will be validated through our experiments.

3. PROPOSED MEASURE

The proposed measure is a modifier of F_1 score which is widely used for information retrieval evaluation. The underlying difference between the proposed measure and the traditional F_1 score is the way precision and recall are computed to take into account small localization errors which are often present in a segmented image. Due to the similarity between these two measures, the proposed measure is also referred to as F_1 in the whole paper.

Given a ground truth segmentation A and an estimated segmentation B, our measure first identifies the number of pixels correctly detected in the segmentation A when compared to the other segmentation B and vice versa. We define:

$$BA = \{x | x \in B, d(x, A) \leq t\} \quad (10)$$

$$AB = \{x | x \in A, d(x, B) \leq t\} \quad (11)$$

where $d(\cdot, \cdot)$ is defined in Eq. (4), BA is the set of correct pixels of B when compared to A and vice versa for AB. A pixel of B is considered as correct if it is close to at least one pixel of A. The closeness is identified by the threshold value t. For example, if $t = 1$, a pixel of B is correct if it is less than or equals to one pixel distance to any pixel of A. The distance metric $\rho(\cdot, \cdot)$ used is the chessboard distance. In other words, if $t = 1$, a pixel of B is considered as correct if it belongs to the 8-neighborhood of at least 1 pixel of A. This allows us to identify the number of correct pixels in each binary image and allows small localization errors. The number of correct pixels is defined as the minimum of $n(BA)$ and $n(AB)$:

$$M = \min\{n(BA), n(AB)\} \quad (12)$$

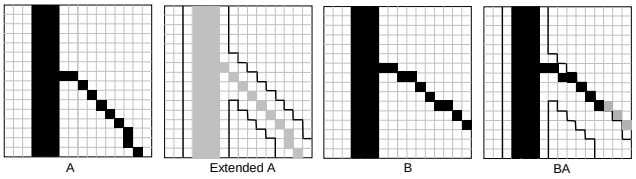

Figure 1: Computing BA when $t = 1$. We first extend the boundary of A by 1 pixel to obtain "Extended A". The set BA is the intersection of B and "Extended A" (shown as the black pixels in the image BA).

Precision and recall are then employed to partially evaluate the exactness and completeness of the result:

$$precision = \frac{M}{n(B)}; \quad recall = \frac{M}{n(A)} \quad (13)$$

Precision is defined as the fraction of the number of correct pixels over the total number of pixels detected in the estimated image. Thus, it measures the exactness or the accuracy of the result. Recall, on the other hand, measures the completeness of the result as it is defined as the fraction of correct pixels over the number of pixels of the ground truth segmentation. The F_1 measure is a combination of precision and recall:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \quad (14)$$

where F_1 ranges from 0 to 1. By taking the role of the ground truth and the estimated segmentation equally in the evaluation, F_1 is a symmetric measure. F_1 has one parameter, t, known as the threshold. The parameter t can take any positive integer (including zero); the greater the value of t, the more tolerant F_1 is towards error.

Implementation of our method is straight forward. Given two binary images, A and B, the first task is to identify the set of pixels in BA. We first extend the boundary of A by t pixels. This is demonstrated in Fig. 1 with $t = 1$. BA is then identified as the intersection of B and the extended version of A. This is done similarly for AB. Once BA and AB are computed, they are combined to produce the F_1 measure.

4. VALIDATION ON SYNTHETIC DATA

In this section, the performances of different measures are assessed based on their behavior on a set of synthetic data. The synthetic data set was designed to resemble the vessels at a fine resolution scale. In all experiments, the performances of the existing distance based measures (MSE, H, Δ^p with $p = 2$ and cutoff $c = 5$, and FOM with $\alpha = 1/9$) and the measures currently used for retinal segmentation evaluation (TPR, FPR, and ACC) are evaluated along with our proposed measure F_1. To implement distance based measures, the distance transformation proposed by [20] was used to compute the shortest distance $d(x, A)$ and $d(x, B)$.

4.1. Criteria of a good evaluation measure

This section aims at identifying the target criteria of a good evaluation measure. We first identify important properties of blood vessels that should be reflected in each segmented image. Then we determine possible small deformations that can happen between two segmentations and identify which changes should affect the score and which should not. This analysis helps us to design a set of synthetic data to test the performance of different measures.

Important features of retinal blood vessels that are used for diagnosis are vessel diameter, branching angle, bifurcation angle, and vessel gap. Vessel diameter is a measure of the width of a vessel. Branching angle is the angle between the main vessel and one of its branches. Bifurcation angle is the angle between two branch vessels at which the main vessel splits into two vessels. Vessel gap indicates the discontinuity of the vessel at a certain place. All of these features are demonstrated in Fig. 2. Many research studies indicate that the changes of these features are precursors of many diseases. For example, the decrease of vessel arterial diameter or bifurcation angle is found in hypertensive and diabetic patients [21]. The presence of the gap is an extreme case of vessel nicking, which is in turn a sign of stroke [22]. Therefore, a good segmentation method should retain these properties unchanged in the segmented image. Otherwise, the following diagnosis steps based on the segmentation result will produce unreliable results.

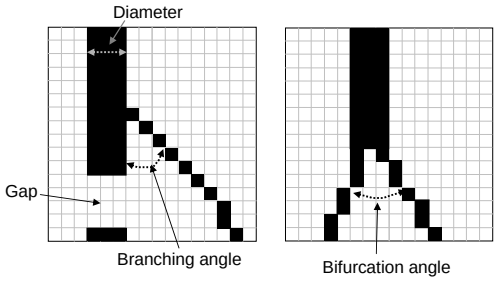

Figure 2: Important properties of retinal blood vessels: vessel diameter, branching angle, bifurcation angle and vessel gap (representing the disconnection of a vessel).

Small differences between segmentation results can be divided into two groups: those that do not affect the vessel properties and those that do. Shift and noise are two possible deformations that belong to the first group. Shift is where the whole or a part of a vessel in an image is shifted by some small distance compared to the corresponding vessel in the other image. This often happens in a segmentation result since the boundary of a vessel is often difficult to establish. Noise refers to the presence of additional pixels that do not affect the visual quality of the desired features. Such noise often happens in the segmentation results produced by computerized algorithms. Generally, these differences do not affect the vessel properties and hence should not be penalized by an evaluation measure.

The second group contains changes that affect important vessel properties. This can be the increasing or decreasing of the branching and bifurcation angle, or the expansion or contraction of the vessel which affect the vessel diameter. This can also be the presence of vessel gap in only one of the two considering segmentations, leading to a false detection of the vessel gap. All of these changes should be penalized by the evaluation measure.

From this point of view, a good evaluation measure for retinal blood vessel segmentation should be tolerant to small deformations that do not affect vessel properties but be able to recognize and penalize changes that affect important vessel properties. In other words, a good measure should provide lower penalty to changes that do not affect the important vessel properties, while providing a high penalty to changes that do affect important blood vessel properties.

4.2. Experiment 1: Comparing Types of Change

In the first experiment, we test the ability to distinguish the changes in two deformation groups as discussed in section 4.1. Changes in the first group are represented by $B1$ and $B2$ while changes in second group include $C1$, $C2$ and $C3$ as shown in Fig. 3. A good evaluation measure should provide higher similarity scores to $B1$ and $B2$ than $C1$, $C2$ and $C3$. Moreover, compared to the shift case ($B1$), the noise case ($B2$) is more likely to affect the whole image. So the expected ranking for this example is: [B1 B2 C1 C2 C3] = [1 2 3 3 3] (1 is the best rank).

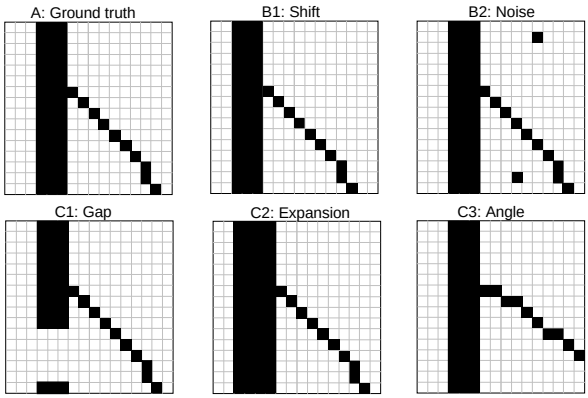

Figure 3: Possible small deformations when comparing two segmentations, used as the first synthetic test cases. Images B1 and B2 show changes that do not affect important vessel properties. Images C1, C2 and C3 show changes that do affect important vessel properties.

Table 1 shows the evaluation scores obtained by all measures on this example. These measures are divided into two groups: similarity measures (F_1, TPR, ACC, FOM) and dissimilarity measures (FPR, MSE, H, Δ^2). A similarity measure increases as the images increase in similarity, while a dissimilarity measure decreases as the images increase in similarity. The results show that only F_1 gives the expected result. TPR is in favor of expansion case ($C2$) while FPR is biased towards the gap case ($C1$). TPR and FPR show low similarity for the shift case. On

the other hand, distance measures MSE and H show low similarity for the noise case ($B2$). ACC gives the lowest similarity to the shift case ($B1$) while FOM provides the greatest similarity to the expansion case ($C2$). Δ^2 shows higher similarity for the gap and expansion cases ($C1$ and $C2$) compared to the noise and shift cases ($B1$ and $B2$).

To measure the correlation between the ranking obtained by each measure with the expected ranking, the correlation scores were computed using Kendall's tau coefficient and presented at the last row of the table. The correlation scores range from -1 to $+1$ with a value of $+1$ means that two rankings are identical. The correlation scores indicate that F_1 provides the most desirable ranking, followed by FOM.

4.3. Experiment 2: Comparing Degree of Change

In this experiment, we test the ability to produce decreasing similarity scores when the differences in the important vessel properties increase. In Fig. 4, from the top to the bottom row, changes in vessel gap, vessel diameter, branching angle and bifurcation angle are shown. The first column shows the ground truth images for each case while $B1$, $B2$ and $B3$ columns are images showing changes in an increasing order. For each case, the expected ranking of [B1 B2 B3] is [1 2 3].

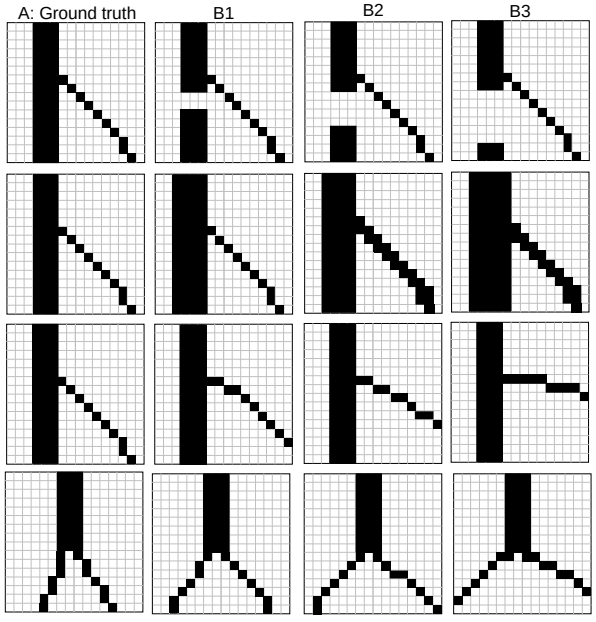

Figure 4: Increasing changes in important vessel properties, used as the second set of synthetic test cases. The important features varied are (from top to bottom row) gap, expansion, branching angle, bifurcation angle.

Table 2 shows the evaluation and correlation scores obtained (using Kendall's tau coefficient) by all measures through these four cases. The hyphen symbol (-) in this table means that the Kendall's tau value is undefined (corresponding to the case: [B1 B2 B3] = [1 1 1]). It is shown that only 3 measures (F_1, FOM and Δ^2) satisfy this test, meaning that they can recognize and evaluate these changes reasonably.

5. VALIDATION ON REAL DATA

In this study, the performances of all measures are assessed based on their evaluation of six retinal segmentation methods (Manual, Soares [11], Staal [13], Niemeijer [14], Perez [10] and Jiang [15]) on 20 test images from the DRIVE database [14]. Each image in the test set is manually segmented twice and this results in two manual segmentations for each image. In this experiment, the first manual segmentations were used as the ground truth while the second manual segmentations were used as the results of a manual method. The segmentation results of Soares method were obtained from their website[1]. The segmentations of all remaining methods were obtained from the Image Sciences Institute website[2] and thresholded at appropriate levels. These six segmentation methods are used for comparing the performance of different measures since the order of their segmentation quality can be clearly seen. The evaluation of these methods was done by a visual inspection on their segmentation results using many criteria (i.e. rewarding these methods that produce accurate vessel width and those that can detect small vessels while penalizing these methods the produce vessel disconnection (or vessel gaps)).

The Manual method is ranked highest as it produces high accurate vessel width and it does not introduce any vessel gap in their segmentation results. Soares and Staal are the next two methods that give good segmentation around the vessel boundary. However, compared to Soares method, Staal method often introduces vessel gaps in a segmentation, especially at bifurcation and branching regions. Niemeijer method produces similar segmentation as Staal method but it can not detect many small vessels as Staal method does. Perez method is ranked next as there is noise associated with small vessels (this affects to the vessel width). Jiang method is ranked last as only big vessels are detected and there are many vessel gaps present in their segmentation. These properties are observed consistently on the segmentation results of these methods on 20 DRIVE test images. Hence, the expected ranking of these methods (in a decreasing order) should be: Manual, Soare, Staal, Niemeijer, Perez, Jiang.

The evaluation results are presented in Table 3. To measure the correlation between the evaluation produced by each measure and the expected ranking, Kendall's tau coefficient was used. The correlation scores are presented at the last row of the table. It is shown that four measures, F_1, ACC, H and Δ^2, follow the expected ranking. However, compared to ACC, F_1 scores spread a wider range

[1] http://sourceforge.net/projects/retinal/files/mlvessel/
[2] http://www.isi.uu.nl/Research/Databases/DRIVE/

Table 1: Evaluation and correlation scores of all measures on the test cases from Fig. 3.

Case	Similarity measures				Dissimilarity measures			
	F_1	TPR	ACC	FOM	FPR	MSE	H	Δ^2
B1	**1**	.569	.805	.957	.126	.431	1	.810
B2	**.983**	1	.992	.976	.010	.867	6	.948
C1	**.852**	.741	.941	.741	0	0	3	.773
C2	**.879**	1	.938	.978	.081	.216	1	.433
C3	**.931**	.845	.930	.950	.045	.776	4	1.127
Corr	**0.84**	-0.25	-0.12	0.12	-0.36	-0.36	0	-0.12

Table 2: Evaluation and correlation scores of all measures on the test cases from Fig. 4.

Case		Similarity measures				Dissimilarity measures			
		F_1	TPR	ACC	FOM	FPR	MSE	HM	Δ^2
Gap	B1	.945	.897	.977	.897	0	0	1	.234
	B2	.885	.793	.953	.793	0	0	2	.512
	B3	.816	.690	.930	.690	0	0	3	.848
	Corr	***1***	***1***	***1***	***1***	-	-	***1***	***1***
Expansion	B1	.879	1	.937	.978	.081	.216	1	.433
	B2	.817	1	.898	.969	.131	.309	1	.590
	B3	.744	1	.844	.959	.202	.408	1	.734
	Corr	***1***	-	***1***	***1***	***1***	***1***	-	***1***
Branching angle	B1	.965	.845	.930	.959	.045	.517	3	.976
	B2	.897	.845	.930	.938	.045	1.19	5	1.32
	B3	.862	.845	.930	.907	.045	3.53	8	1.82
	Corr	***1***	-	-	***1***	-	***1***	***1***	***1***
Bifurcation angle	B1	.951	.732	.914	.953	0.051	.561	2	.803
	B2	.843	.732	.910	.899	0.056	1.59	3	1.19
	B3	.771	.707	.902	.866	0.060	2.81	5	1.59
	Corr	***1***	*0.82*	***1***	***1***	***1***	***1***	***1***	***1***

which means that it has higher discrimination capability than ACC.

To study the dependence of F_1 to the threshold t, an experiment was performed to examine F_1 behavior when t changes. Table 4 shows the evaluation of F_1 measure on six segmentation methods when t increases from 0 to 10. The last row of this table presents the correlation between the ranking obtained and the expected ranking. It is shown that F_1 measure gives expected evaluation with the first 5 values of t and the correlation scores decrease as t increases for the remaining values of t. On DRIVE dataset, the largest vessel width is about 8-9 pixels and a reasonable setting of t is from 0 to 4. So with a new image set, t should be set as smaller than half of the largest vessel width in order to obtain a good evaluation.

If a small amount of noise is added to a vessel segmentation image, it does not affect the visual quality of the desired retinal features. Therefore, it should have little effect on the similarity score. To test the ability to deal with noise of all measures, we add salt and pepper noise with increasing density to a segmented image and analyze the behavior of these measures. Fig. 5 shows the first manual segmentation of an image in the DRIVE database and its sub-image which is used as the ground truth for this experiment. A set of test images used for comparison are presented in Fig. 6. The image labeled Manual was extracted from the second manual segmentation of the same image. The three images named Noise 1, Noise 2 and Noise 3 are three variations of the Manual image by adding salt and pepper noise with the noise density of 0.01, 0.02, and 0.03, respectively. The last two images, Perez and Jiang, were extracted from the segmentations obtained by the Perez and Jiang methods correspondingly.

The images in both figures show that the Manual image is the most similar to the ground truth image. On the other hand, the quality of the Perez and Jiang segmentations are much worse than the Manual one. In addition, the salt and pepper noise does not affect the visual quality of the desired features in the Manual segmentation. Therefore, the three images Noise 1, Noise 2, Noise 3 should get higher similarity scores than those of Perez and Jiang. Thus, the expected ranking of [Manual, Noise1, Noise 2,

Table 3: Average evaluation scores and correlation scores of all measures when evaluating 6 retinal segmentation methods across 20 DRIVE test images.

Method	Similarity measures				Dissimilarity measures			
	F_1	TPR	ACC	FOM	FPR	MSE	H	Δ^2
Manual	**.918**	.776	**.947**	.889	.028	5.1	**41.6**	**.743**
Soares [11]	**.897**	.728	**.947**	.849	.021	2.3	**52.9**	**.873**
Staal [13]	**.883**	.735	**.944**	.848	.025	8.4	**53.5**	**.920**
Niemeijer [14]	**.849**	.673	**.942**	.770	.019	5.6	**61.6**	**1**
Perez [10]	**.838**	.744	**.932**	.852	.041	14.2	**70.6**	**1.096**
Jiang [15]	**.761**	.648	**.922**	.828	.037	3.3	**82.3**	**1.176**
Corr	**1**	0.47	**1**	0.47	0.2	0.2	**1**	**1**

Table 4: Evaluation and correlation scores of F_1 measure when evaluating 6 retinal segmentation methods with threshold t changes from 0 to 10.

t	0	1	2	3	4	5	6	7	8	9	10
Manual	.788	.918	.928	.932	.934	.937	.939	.940	.942	.943	.944
Soares [11]	.776	.897	.909	.915	.918	.920	.922	.923	.924	.926	.926
Staal [13]	.768	.883	.896	.903	.907	.910	.913	.915	.917	.918	.920
Niemeijer [14]	.743	.849	.861	.864	.867	.869	.871	.872	.874	.874	.875
Perez [10]	.697	.796	.808	.816	.823	.829	.835	.840	.845	.850	.854
Jiang [15]	.678	.761	.792	.815	.835	.849	.860	.868	.876	.882	.888
Corr	1	1	1	1	1	0.87	0.87	0.87	0.73	0.73	0.73

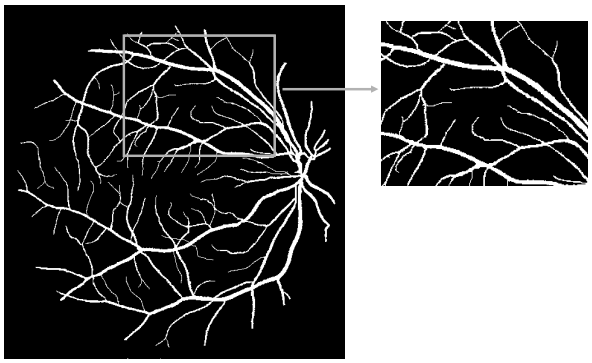

Figure 5: The first manual segmentation of test image 19 (from the DRIVE database) and its selected sub image used as the ground truth in the last experiment.

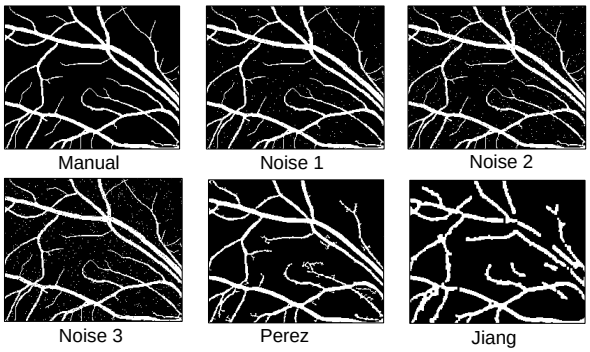

Figure 6: The sub image from the second manual segmentation (Manual) with varying levels of added noise (Noise 1, Noise 2, Noise 3), and the associated sub images from the Perez and Jiang segmentations.

Noise 3, Perez, Jiang] should be [1, 2, 3, 4, 5, 6].

The evaluation results of all measures are presented in Table 5. The last row of this table presents the correlation values obtained by each measure when compared to the expected ranking. The results show that distance based measures (FOM, MSE, H and Δ^2) give very low correlation scores. TPR and ACC produce better evaluation with higher correlation score (0.6 for TPR and 0.87 for ACC) but they put higher value to Perez and Jiang images than those noisy images. Only F_1 and FPR give the expected ranking with a perfect correlation score of 1. This indicates that F_1 and FPR is more robust to noise than the remaining measures. However, FPR measure does not give expected evaluation in previous experiments. Hence, F_1 is the most suitable measure for retinal image segmentation evaluation.

6. CONCLUSIONS

Although many algorithms have been designed to automate the retinal blood vessel extraction task, there has been no investigation of how we should evaluate the ef-

Table 5: Evaluation and correlation scores of all measures when evaluating sub-images presented in Fig. 6.

Method	Similarity measures				Dissimilarity measures			
	F_1	TPR	ACC	FOM	FPR	MSE	H	Δ^2
Manual	.918	.895	.952	.958	.038	4.41	34	0.81
Noise 1	.912	.890	.948	.941	.042	7.01	36	1.08
Noise 2	.901	.886	.943	.924	.048	8.65	34	1.31
Noise 3	.888	.885	.938	.909	.052	10.97	34	1.49
Perez	.842	.895	.940	.938	.053	2.52	34	1.10
Jiang	.746	.796	.901	.929	.080	1.14	37	1.24
Corr	1	0.6	0.87	-0.47	1	-0.2	0.26	0.47

fectiveness of these algorithms. So far, the evaluation of these algorithms has been performed using generic binary image similarity measures, which do not seem appropriate for the task. In this article, we have investigated the qualities of different evaluation measures and proposed a new measure, called F_1, that utilizes the precision, recall and F-measure concepts. A set of synthetic data was carefully designed to evaluate the performances of these measures. These measures were also validated on real data to examine the performances of different retinal image segmentation algorithms. From the results of all experiments, we conclude that our proposed measure, F_1, is the most suitable measure for retinal image segmentation. It provides the greatest correlation to the desired measure behavior and thus is the most suitable measure for studying the performance of different retinal segmentation algorithms. Since F_1 is a binary image similarity measure, it can be used as an evaluation tool for any image segmentation or even edge detector methods where small localization error is acceptable.

References

[1] N. Lai, Clinical ophthalmology: A systematic approach, Optometry and Vision Science 81 (5) (2004) 295.

[2] Y. Yin, M. Adel, S. Bourennane, Retinal vessel segmentation using a probabilistic tracking method, Pattern Recognition.

[3] X. You, Q. Peng, Y. Yuan, Y. Cheung, J. Lei, Segmentation of retinal blood vessels using the radial projection and semi-supervised approach, Pattern Recognition.

[4] D. Marn, A. Aquino, M. Gegndez-Arias, J. Bravo, A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features, Medical Imaging, IEEE Transactions on 30 (1) (2011) 146–158.

[5] B. Lam, Y. Gao, A. Liew, General retinal vessel segmentation using regularization-based multiconcavity modeling, Medical Imaging, IEEE Transactions on 29 (7) (2010) 1369–1381.

[6] C. Lupascu, D. Tegolo, E. Trucco, Fabc: retinal vessel segmentation using adaboost, Information Technology in Biomedicine, IEEE Transactions on 14 (5) (2010) 1267–1274.

[7] B. Zhang, L. Zhang, F. Karray, Retinal vessel extraction by matched filter with first-order derivative of gaussian, Computers in Biology and Medicine 40 (4) (2010) 438–445.

[8] B. Lam, H. Yan, A novel vessel segmentation algorithm for pathological retina images based on the divergence of vector fields, Medical Imaging, IEEE Transactions on 27 (2) (2008) 237–246.

[9] E. Ricci, R. Perfetti, Retinal blood vessel segmentation using line operators and support vector classification, Medical Imaging, IEEE Transactions on 26 (10) (2007) 1357–1365.

[10] M. Martinez-Perez, A. Hughes, S. Thom, A. Bharath, K. Parker, Segmentation of blood vessels from red-free and fluorescein retinal images, Medical Image Analysis 11 (1) (2007) 47–61.

[11] J. Soares, J. Leandro, R. C. Jr, H. Jelinek, M. Cree, Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification, IEEE Transactions on Medical Imaging 25 (9).

[12] A. Mendonca, A. Campilho, Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction, Medical Imaging, IEEE Transactions on 25 (9) (2006) 1200–1213.

[13] J. Staal, M. Abrmoff, M. Niemeijer, M. Viergever, B. van Ginneken, Ridge-based vessel segmentation in color images of the retina, IEEE Transactions on Medical Imaging 23 (4) (2004) 501–509.

[14] M. Niemeijer, J. Staal, B. van Ginneken, M. Loog, M. Abramoff, Comparative study of retinal vessel segmentation methods on a new publicly available database, in: Proceedings of SPIE, Vol. 5370, 2004, p. 648.

[15] X. Jiang, D. Motion, Adaptive local thresholding by verification-based multithreshold probing with application to vessel detection in retinal images, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (1) (2003) 131–137.

[16] A. Hoover, V. Kouznetsova, M. Goldbaum, Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response, Medical Imaging, IEEE Transactions on 19 (3) (2000) 203–210.

[17] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, M. Goldbaum, Detection of blood vessels in retinal images using two-dimensional matched filters, Medical Imaging, IEEE Transactions on 8 (3) (1989) 263–269.

[18] A. Baddeley, An error metric for binary images, Robust Computer Vision (1992) 5978.

[19] I. Abdou, W. Pratt, Quantitative design and evaluation of enhancement/thresholding edge detectors, in: IEEE, Proceedings, Vol. 67, 1979, pp. 753–763.

[20] A. Rosenfeld, J. Pfaltz, Sequential operations in digital picture processing, Journal of the ACM (JACM) 13 (4) (1966) 471–494.

[21] A. Stanton, B. Wasan, A. Cerutti, S. Ford, R. Marsh, P. Sever, S. Thom, A. Hughes, Vascular network changes in the retina with age and hypertension., Journal of hypertension 13 (12 Pt 2) (1995) 1724.

[22] Y. Wong, Is retinal photography useful in the measurement of stroke risk?, The Lancet Neurology 3 (3) (2004) 179–183.