Pattern Recognition Letters 33 (2012) 2048–2056

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Structural performance evaluation of curvilinear structure detection algorithms with application to retinal vessel segmentation

Xiaoyi Jiang [a,*], Martin Lambers [b], Horst Bunke [c]

[a] Department of Mathematics and Computer Science, University of Münster, Germany
[b] Computer Graphics Group, University of Siegen, Germany
[c] Institute of Computer Science and Applied Mathematics, University of Bern, Switzerland

ARTICLE INFO

Article history:
Available online 23 May 2012

Keywords:
Performance evaluation
Curvilinear structure
Vessel network
Airway tree
Graph matching

ABSTRACT

Curvilinear structures are useful features in a variety of applications, particularly in medical image analysis. Compared to other commonly used features such as edges and regions, there is relatively few work on performance evaluation methodologies for curvilinear structure detection algorithms. For instance, a pixel-wise comparison with ground truth has been used in all recent publications on vessel detection in retinal images. In this paper we propose a novel structure-based methodology for evaluating the performance of 2D and 3D curvilinear structure detection algorithms. We consider the two aspects of performance, namely detection rate and detection accuracy, separately, in contrast to their mixed handling in earlier approaches that typically produces biased impression of detection quality. By doing so, the proposed performance measures give us a more informative and precise performance characterization. Experiments on both synthetic and real examples will be given to demonstrate the advantages of our approach.

1. Introduction

Performance evaluation is an important issue in pattern recognition (Bharkad and Kokare, 2011; Cardoso and Sousa, 2011). Extensive early work exists for evaluating algorithms for edge detection and region-based segmentation. Already 1997, for instance, Heath et al. (1997) listed 12 edge detection evaluation methods. A discussion of related literature can be found in (Jiang, 2005; Jiang et al., 2006). In contrast there is only very little work on evaluating algorithms for curvilinear structure detection.

The term *curvilinear structure* denotes a line or a curve with some *width*. In contrast to edges, they have the same shape but non-negligible and varying width. Curvilinear structures are useful features for a variety of applications (finding roads or rivers in aerial images, detecting lanes for traffic tasks, etc.). Particularly in medical imaging they belong to the most widely observed and important features; examples are blood vessels, bones, airway trees, and other thin structures.

It is the purpose of this work to discuss the weaknesses of an approach widely used in medical image analysis literature and to propose an improved evaluation methodology. Throughout this paper our discussion will be exemplified by the task of detecting blood vessels in retinal images. However, it is important to point out that our approach is in no way bounded to retinal images only, but instead applicable in the general context of the evaluation of 2D and 3D curvilinear structure detection algorithms.

Reliable segmentation of the vasculature in retinal images is a nontrivial task for image analysis and has immense clinical relevance. Blood vessel appearance is an important indicator for diagnoses including diabetes, hypertension, and arteriosclerosis. For this purpose one needs a quantification of features of veins and arteries such as color, diameter, tortuosity, and opacity. Vessel detection provides the fundament for this kind of diagnosis making and indirectly also for other problems. Automatic detection algorithms for pathologies like microaneurysms may be improved if regions containing vasculature is excluded from the analysis (Frame et al., 1998). In addition, knowledge about the location of vessels can aid in registration of retinal images (Zana and Klein, 1999) and detection of other features like optic disc and fovea (Hoover and Goldbaum, 2003). The retinal vessels can be further separated into arteries and veins (Rothaus et al., 2009), which are fundamental to computing the important AV-ratio (ratio of artery to vein calibres). For all these reasons reliable vessel segmentation in retinal images and evaluation of vessel segmentation algorithms is of interest both from a theoretical and a clinical point of view. Similar conclusion applies to other curvilinear structure detection tasks in medical image analysis as well.

We motivate our work with a detailed discussion of the drawbacks of early approaches in Section 2. Then, we describe an improved evaluation methodology in Section 3. Experimental work demonstrating the usefulness of our approach follows in Section 4.

* Corresponding author. Tel.: +49 251 8333759; fax: +49 251 8333755.
E-mail address: xjiang@math.uni-muenster.de (X. Jiang).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.05.008

Finally, some discussions conclude the paper. This paper is an extended version of the conference paper (Ji ang et al., 2011) and contains more detailed literature review, additional technical details, and substantially extended experimental work.

2. Drawbacks of non-structural approaches

The efforts of performance evaluation in computer vision can generally be classified into four distinct categories: theory-based, human evaluation, ground truth (GT) based, and task-based. Our methodology falls into GT-based evaluation. The term ground truth is used to denote some reference result that represents the expected ideal segmentation. The basic idea of GT-based evaluation is then to compute some measure of differences between machine segmentation result and the ground truth.

Many algorithms have been proposed for vessel segmentation in retinal images (Chaudhuri et al., 1989; Fang et al., 2005; Hoover et al., 2000; Jiang and Mojon, 2003; Lam and Yan, 2008; Lam et al., 2010; Martinez-Perez et al., 1999; Staal et al., 2004; Zana and Klein, 2001). While visual inspection has been applied in early approaches for performance evaluation, recent works report on experimental results based on large datasets with manually specified ground truth. Hoover et al. (2000) have collected a dataset STARE (STructured Analysis of the REtina) of 20 retinal images which were manually segmented by two observers. The DRIVE (Digital Retinal Images for Vessel Extraction) dataset (Niemeijer et al., 2004; Staal et al., 2004) consists of 40 images with manual segmentation from three observers. Also the authors of Fang et al. (2005) reported on a dataset of 35 retinal images with ground truth. Two of the datasets (STARE and DRIVE) are publicly available.

In all those works using manually specified ground truth, a straightforward method is used for performance evaluation. Given a machine-segmented result image (MS) and its corresponding hand-labeled ground truth image (GT), any pixel which is marked as vessel in both MS and GT is counted as a true positive. Any pixel which is marked as vessel in MS but not in GT is counted as a false positive. The true positive rate (TPR) is established by dividing the number of true positives by the total number of vessel pixels in GT. The false positive rate (FPR) is computed by dividing the number of false positives by the total number of non-vessel pixels in GT. As an alternative, the FPR can also be based on the total number of non-vessels pixels within the circular field of view[1] (FOV) only (Niemeijer et al., 2004). This latter version seems to be more reasonable and thus will be consistently used in this work. If different pairs of sensitivity and specificity can be achieved, for instance by thresholding a soft classification or various parameter sets, the performance of a system can be investigated by receiver operating curves (ROC). The closer a ROC approaches the top left corner (TPR = 100%, FPR = 0%), the better the performance of the system. In (Lam and Yan, 2008) a variant of TPR is introduced to emphasize on pathological regions that are especially important but difficult to deal with.

Similar pixel-wise comparison has also been used for evaluating binarization methods (Lee et al., 1990) and building extraction from aerial imagery (Shufelt, 1999). While this approach is suitable there for comparing large regions, its application to curvilinear structures as elongated and thin regions is more questionable. This is illustrated in Fig. 1 with several modified versions of the GT. MS_{thin} results from thinning the GT at some places while in MS_{del} some vessel sections are deleted and others remain unchanged. For both MS_{thin} and MS_{del} we obtain TPR = 85.1% and FPR = 0.0%, indicating an equal rate of 85.1% correct detection and no spurious vessels. But in reality there are substantial differences between the two MS images. In MS_{thin} the entire vessel network is correctly detected, but some vessels have a smaller width than GT. In contrast MS_{del} perfectly equals GT except the deleted parts. A more objective performance measure would be TPR (MS_{thin}) = 1.00 and TPR (MS_{del}) < 1.00, indicating the percentage of the correctly detected part of the vessel network. The correctly detected parts of the vessel network can be further evaluated with respect to the detection accuracy, i.e., the width error. Then, we would expect a non-zero width error for MS_{thin} and zero width error for MS_{del}, respectively.

A second situation in Fig. 1(e) and (f) illustrates a related problem. Again, the two MS images MS_{exp} and MS_{ins} have equal performance measures TPR = 100% and FPR = 1.7%, implying a full detection of the vessel network and 1.7% spurious vessels in both cases. Here MS_{exp} emerges from GT by locally expanding GT while MS_{ins} equals GT plus eight spurious (diagonal) vessel parts. Different from MS_{ins}, the spurious vessel pixels in MS_{exp} cause vessel width errors, but do not change the vessel network structure in any way. Intuitively, a measure FPR (MS_{exp}) = 0 and FPR (MS_{ins}) > 0 thus makes more sense. Accordingly, MS_{exp} and MS_{ins} should be associated with non-zero and zero width error, respectively.

The examples above clearly show the drawbacks of the early approach to evaluating the performance of vessel segmentation algorithms. Due to the nature of curvilinear structures being thin and elongated regions, a pixel-wise comparison is not the most meaningful way of performance assessment. As a matter of fact, the overall performance measures TPR and FPR are both a mixture of two different aspects of performance, namely detection rate (how much of the vessel network structure is detected) and detection accuracy (what is the accuracy of the detected network structure). As shown in the examples above, such a mixture may result in a biased impression of the detection quality.

The non-structural nature of TPR and FPR has been implicitly acknowledged by other authors. Fang et al. (2005), for instance, state "Visual inspection is a way to examine extraction results for blood vessels in retinal images. Our method is able to recover an almost perfect morphological structure for high contrast images". Later, they use TPR and FPR to measure the detection quality without any further discussion about the accuracy of structure detection. This simply means that despite of the use of TPR and FPR, they were only able to obtain some qualitative impression of structure detection accuracy by visual inspection.

Another problem is discussed in (Niemeijer et al., 2004): "*A disadvantage is that in this way the wider vessels have a larger influence on the end result than the smaller vessels. [...] Many of the smallest vessels of the gold standard and the second manual segmentation are not or partly visible in the automatic segmentations. In an application where small vessel detection is critical a method with a lower overall accuracy could still be marked the better method if it would segment more of the smallest vessels*". With the current definition of TPR and FPR we do not distinguish between wider and smaller vessels in any way and thus a more sophisticated performance assessment as formulated by the authors of (Niemeijer et al., 2004) is not possible.

Based on the discussion above we believe that the key of a more meaningful way of performance assessment is to separate the two factors detection rate and detection accuracy. In our previous work (Jiang and Mojon, 2002) we represent the structure of a vessel network by its thinned version of midline points of one pixel width. The structures of the GT and the MS vessel network are compared by a point matching process. Then, the detection rate is measured by TPR and FPR defined by means of the matched midline points, reflecting the structural differences of the two networks. Afterwards, the computation of detection accuracy is based on the width information of matched midline points. While this approach is exactly what we need to alleviate the problems addressed above,

[1] The retinal images typically have a limited field of view, mainly due to the curvedness of human retina. If needed, multiple images can be fused using image registration techniques to form a montage with a larger field of view.

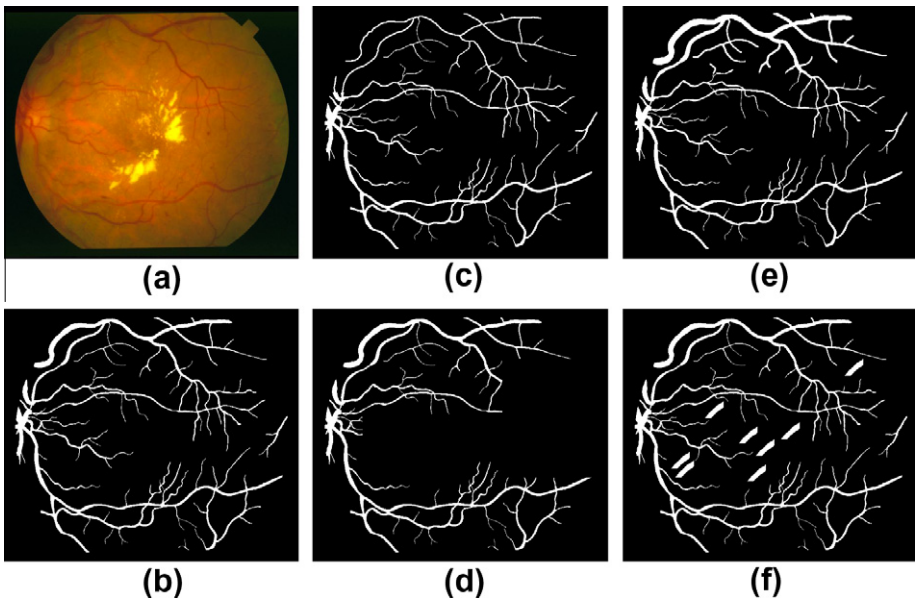

Fig. 1. (a) Retinal image; (b) GT; (c) MS_{thin}: partial thinning of GT; (d) MS_{del}: deletions in GT; (e) MS_{exp}: partial expanding of GT; (f) MS_{ins}: insertions in GT.

the point matching process is an ad hoc one and results in many non-optimal matchings. In this work we further develop the approach of (Jiang and Mojon, 2002) by introducing an optimal point matching process. In addition the experimental validation has been substantially extended to demonstrate the usefulness of our approach.

Similar philosophy has been applied in (Niemeijer et al., 2010) to evaluate the detection accuracy of microaneurysms in retinal images. Instead of counting pixels the performance is measured by comparing microaneurysms as a whole (by using a simple correspondence procedure).

3. Structural evaluation methodology

The basic assumption is that for each test dataset (image or volume), we have a corresponding GT dataset with the curvilinear structures manually specified. Our approach is thus a supervised one.

Given a binary dataset V, for example a blood vessel image, we define its structure as the set of midline points along with the width information. Such a representation fully characterizes the curvilinear network by two information sources, allowing us to investigate the detection rate and the detection accuracy separately. We extract this structure in the following way:

- Find the midline points by computing the skeleton of V. The skeleton is denoted by V_t. There are many thinning and skeletonization methods, each with different properties. We use the method from Cardoner and Thomas (1997) because it guarantees that (a) the skeleton is thin (single-pixel wide), (b) the skeleton is connected, and (c) the skeleton can be used to reconstruct the original image with a tolerance of one pixel. Furthermore, it can be generalized to higher dimensions (Romero et al., 2000). Note that for elongated shapes like curvilinear structures under consideration in this work the relatively simple thinning algorithm from Cardoner and Thomas (1997) suffices; for thinning general binary shapes more sophisticated approaches like (Tang et al., 2010) will be needed.
- Compute a distance map of V: Each structure point (i.e., point from the skeleton V_t) is assigned its Euclidean distance d to the background. Then, each structure point p in V_t receives a corresponding width value $w_p = 2d_p$. There are multiple methods to compute an exact squared Euclidean distance transform for this purpose in linear time. We chose the one presented in (Maurer et al., 2003).

Given a MS and GT, we propose to measure the detection rate by comparing MS_t and GT_t only, i.e., how much of the GT curvilinear network structure is detected in MS. In a second step the width of matched MS_t and GT_t structure points is compared to give a measure of detection accuracy.

The most crucial part of our approach is how to match GT_t and MS_t. We need to match as many structure points of GT_t as possible to structure points in MS_t and vice versa, in a way that ensures that two matched structure points are as similar as possible with respect to both their position and width. We formulate this problem as one of optimal graph matching.

3.1. Graph matching

A graph G is bipartite if its vertices form two disjoint subsets so that no edge exists between vertices in the same subset. The disjoint structure point sets GT_t and MS_t therefore form a bipartite graph G_{gm} if every edge connects a structure point $p \in GT_t$ with a structure point $q \in MS_t$. Such an edge represents a *candidate* for a match between the structure points p and q. Every edge is associated with cost that depends on the distance of p and q and on the difference of their width information.

A match between the two disjoint vertex sets of a bipartite graph is a set of edges so that each vertex is endpoint of at most one edge. Such a match will be called a *structural matching* in the following. In a structural matching, every structure point in GT_t is matched to at most one structure point in MS_t and vice versa.

The task can now be expressed as the problem of finding an optimal structural matching with minimum cost among all structural matchings with the maximum number of edges in G_{gm}. Such a matching is called a *maximum-cardinality minimum-cost matching*.

To build the graph G_{gm} we need to determine the set of match candidates and to specify their cost. Given the graph G_{gm}, we have to develop a procedure for finding its optimal structural matching $\mathcal{M}$. Based on this optimal structural matching we finally define a

new set of performance measures. The details of these steps are given in the following subsections.

3.2. Selecting match candidates

Not every pair (p,q) should be a match candidate: The Euclidean distance $d(p,q)$ should not be too high, and p, q should not represent structures of very different width.

To determine the match candidates and therefore the edges in G_{gm}, two thresholds $d_{\max}$ and $w_{\max}$ are necessary. A pair (p,q) is a match candidate if and only if

$$d(p,q) \leqslant d_{\max} \wedge |w_p - w_q| \leqslant w_{\max}$$

These thresholds are not independent of each other. In the case of thick structures, the allowed difference in position may be higher than in the case of thin structures, where it is more important to match the position exactly. To reflect this, $w_{\max}$ is determined from GT_t:

$$w_{\max} = c_w \cdot \max\{w_p | p \in \mathrm{GT}_t\}.$$

Then, $d_{\max}$ is determined from $w_{\max}$:

$$d_{\max} = c_d \cdot w_{\max}.$$

The factors c_w and c_d are parameters and have to be chosen in advance. Details of choosing parameter values will be discussed in Section 3.6.

3.3. Cost of match candidates

For each match candidate $(p,q) \in G_{gm}$, $p \in \mathrm{GT}_t$, $q \in \mathrm{MS}_t$, its cost $c(p,q)$ should be proportional both to the Euclidean distance $d(p,q)$ and to the difference $|w_p - w_q|$ of the structure widths. Additionally, the cost should be normalized to [0,1] to ease the task of defining quality measures later. Because $d(p,q)$ is bounded by $d_{\max}$ and $|w_p - w_q|$ is bounded by $w_{\max}$, the following definition fulfills these requirements:

$$c(p,q) = 1 - \left(1 - \frac{d(p,q)}{d_{\max}}\right) \cdot \left(1 - \frac{|w_p - w_q|}{w_{\max}}\right) \in [0,1]$$

A good match candidate (p,q) means small value of $d(p,q)$ and $|w_p - w_q|$, which results in small $c(p,q)$. The minimum of $c(p,q)$, 0, is reached only in case of perfect matches. Note that this cost depends on the ground truth segmentation GT since $w_{\max}$ and $d_{\max}$ are computed from GT_t.

Based on this cost for match candidates, the cost $C(M)$ of a structural matching M between GT_t and MS_t is defined as follows:

$$C(M) = \sum_{(p,q)\in M} c(p,q) \in [0, |M|]$$

summing up the costs of all individual matches in M.

3.4. Computing optimal structural matching

The problem of determining a maximum-cardinality minimum-cost matching on G_{gm} can be expressed as a special case of the assignment problem for which efficient algorithms are known.

For this purpose, the problem can first be reduced to the computation of a minimum-cost perfect match in an auxiliary graph G'_{gm}; see (Gabow and Tarjan, 1989), Section 202. (A perfect match in a bipartite graph $G = A \cup B$ is a match so that each vertex of A is matched to exactly one vertex of B and vice versa.) We form the auxiliary bipartite graph G'_{gm} as follows. It is created by putting G_{gm} and a copy of G_{gm} together. Then, we connect each vertex in G_{gm} with its copy and each such new edge is assigned the cost $N \cdot c_{\max}$, where N is the number of vertices in G_{gm} and $c_{\max}$ is the maximum cost assigned to an edge in G_{gm} (in our case $c_{\max} = 1$). G'_{gm} is again bipartite with the two disjoint vertex sets $\mathrm{GT}_t^1 \bigcup \mathrm{MS}_t^2$ and $\mathrm{MS}_t^1 \bigcup \mathrm{GT}_t^2$ (each represented by a dashed ellipse in Fig. 2), where GT_t^1 and MS_t^1 are the vertices of GT and MS part of G_{gm}, respectively, and GT_t^2 and MS_t^2 are the corresponding vertices from the copy of G_{gm}. It can be shown that G'_{gm} contains a minimum-cost perfect match, which corresponds to a maximum-cardinality minimum-cost match in G_{gm} when all edges that end in a vertex of the copy of G_{gm} are eliminated. More details can be found in (Gabow and Tarjan, 1989).

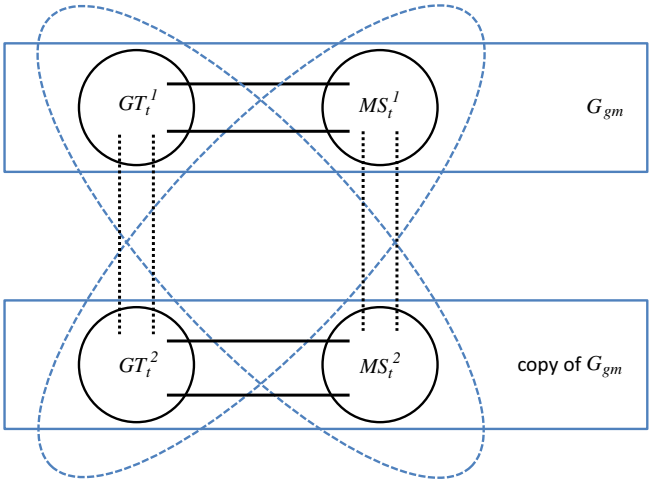

Fig. 2. Construction of auxiliary graph G'_{gm}: The dotted lines represent the edges connecting each vertex in G_{gm} with its copy.

The maximum-cardinality minimum-cost match in G_{gm} is the optimal structural matching $\mathcal{M}$ we are looking for. The remaining problem of finding a minimum-cost perfect match in G'_{gm} is a special case of the assignment problem and can be solved efficiently using the Cost Scale Assignment (CSA) algorithm from Goldberg and Kennedy (1995).

3.5. Quality measures

Based on the optimal structural matching $\mathcal{M}$ we define the following quality measures.

True positives: The successfully detected structure points of GT_t are those that have a corresponding structure point in MS_t according to the optimal structural matching $\mathcal{M}$. The portion of these successfully detected structure points in $\mathcal{M}$ to the total number of structure points in GT_t is the *true positives rate* (TPR):

$$\mathrm{TPR} = \frac{|\mathcal{M}|}{\#\text{ stucture points in } \mathrm{GT}_t}$$

The definition of TPR tells us how much of the GT curvilinear network structure is successfully detected in the machine segmentation. A measure of the matching quality of the true positives is the *detection error* (DE):

$$\mathrm{DE} = \frac{C(\mathcal{M})}{|\mathcal{M}|} \in [0,1]$$

Recall that $C(\mathcal{M})$ is the cost function bounded by $[0, |\mathcal{M}|]$ with zero indicating the best case. The detection error can also be split into two values to separately measure the *position error* (PE) and *width error* (WE) of the successfully detected curvilinear structure:

$$\mathrm{PE} = \frac{1}{|\mathcal{M}|} \sum_{(p,q)\in\mathcal{M}} d(p,q); \qquad \mathrm{WE} = \frac{1}{|\mathcal{M}|} \sum_{(p,q)\in\mathcal{M}} |w_p - w_q|$$

Lower values of DE, PE, and WE correspond to higher accuracy. Note that all three error measures are scaled by $1/|\mathcal{M}|$ because they are

intended to represent the error per pair of correctly detected structure point and its GT correspondence.

False positives: Those structure points of MS_t that have no match in GT_t according to $\mathcal{M}$ are false positives. The *false positives rate* (FPR) is defined as:

$$\mathrm{FPR} = \frac{\#\text{ structure points in } MS_t - |\mathcal{M}|}{\#\text{ non-structure points in FOV of } GT_t}$$

Note that the number of non-structure points in the denominator may alternatively be counted in GT_t. In this work we follow the more reasonable convention from Niemeijer et al. (2004) to restrict the consideration to FOV only.

In addition to compute the false positive rate (equivalently the number of false positives in MS_t) it is also interesting to ask about the characteristic, for instance the width, of these spurious structures. It is probably more problematic to erroneously detect thick structures than thin ones. To obtain this information we can establish a width histogram of the false positives.

False negatives: Those structure points of GT_t that have no match in MS_t according to $\mathcal{M}$ are false negatives. The *false negatives rate* (FNR) is defined as:

$$\mathrm{FNR} = \frac{\#\text{ structure points in } GT_t - |\mathcal{M}|}{\#\text{ structure points in } GT_t}$$

This measure indicates how much of the GT structure is missing in the machine segmentation. Likewise we can investigate the width characteristics of these false negatives by a width histogram.

The structure-based evaluation procedure is summarized as follows.

```
/* detection rate */
construct GT_t from GT and MS_t from MS;
find the optimal structural matching M between GT_t and MS_t;
NoOfTruePositives= |M|;
NoOfFalsePositives = (# structure points in MS_t)-|M|;
NoOfFalseNegatives = (# structure points in GT_t)-|M|;
TPR = NoOfTruePositives/ (# structure points in GT_t);
FPR = NoOfFalsePositives/ (# non-structure points in FOV of
  GT_t);
FNR = NoOfFalseNegatives/ (# structure points in GT_t);
  /* detection error */
Compute position error PE of true positive structures;
Compute width error WE of true positive structures;
```

3.6. Choosing parameter values

Two parameters c_d and c_w are used during the selection of match candidates. These parameters affect the number of match candidates as well as their cost, and therefore also the optimal match $\mathcal{M}$ and the induced quality measures. Fortunately, it turns out that the quality measures are fairly robust to parameter changes.

Since there is no stringent reason to treat distance in position differently from difference in width, $c_d = 1$ and therefore $d_{\max} = w_{\max}$ is a suitable choice. This leaves only c_w to be determined.

The influence on the number of match candidates is more critical than the influence on their cost. Since the optimal matching $\mathcal{M}$ is a maximum-cardinality match, too many match candidates inevitably lead to nonsense matches in $\mathcal{M}$. Therefore, c_w must not be chosen too large. On the other hand, c_w must not be chosen too small either, to avoid the exclusion of reasonable match candidates. Suppose the segmentation algorithm tends to mark structures wider than they really are. Then a small value of c_w quickly leads to the exclusion of reasonable match candidates.

It turns out that all reasonable match candidates are already included for $c_w \approx 0.5$: The true positives rate conforms to the expectations. The higher TPR observed for increasing values of c_w comes at the cost of match quality: The position error value PE and the width error value WE increase rapidly already for $c_w \approx 1$. A study of this parameter reported in Section 4.3 indicates quite stable results for $c_w \in [0.4,0.7]$. For this reason, a good parameter choice is $c_w = 0.5$, $c_d = 1$ and the experimental results reported in Section 4 are based on this parameter setting.

In principle, it would also be possible to choose $w_{\max}$ and $d_{\max}$ locally. For the structure point $p \in GT_t$ currently under examination, one could choose $w_{\max} = c_w \cdot w_p$ and $d_{\max} = c_d \cdot w_{\max}$. The expectation would be that the selection of match candidates adapts to the nature of the different structure regions and therefore further improves the quality measures. It turns out, however, that this method is very sensitive to the choice of parameters. Furthermore, the quality measures show no significant improvements even when suitable values are found. For this reason we will consistently apply the global version with $c_w = 0.5$, $c_d = 1$ in all experiments reported in the next section.

4. Experimental results

The motivation of our work is to develop a structure-based evaluation methodology so that we can overcome the bias problems discussed in Section 2. A series of experiments using both synthetic and real data have been conducted to demonstrate the effectiveness of our approach.

4.1. Synthetic data

First we show how our method evaluates the four images in Fig. 1(c)–(f), see Table 1. As wanted, MS_{thin} has TPR near 100%, implying a full detection of the vessel network structure. The fact that the detected vessels are thinner than GT is expressed by the width error 0.452 (pixel). The width error indirectly results in a position error 0.191. Note that TPR is not perfectly 100% because the thinned version MS_{thin} generally results in a skeleton which slightly differs from that of GT. In contrast MS_{del} leads to TPR = 77.6% only and accordingly 22.4% of the vessel network structure undetected. Since no error has been added to the correctly detected 77.6% of the vessel network in synthesizing MS_{del}, the error measures are all negligible in this case. The missing vessels in MS_{del} are expressed by the high number of false negatives 1729, meaning that 1729 of the structure points of GT_t cannot be matched to the segmentation result. In comparison MS_{thin} only has 32 missing structure points. Based on the histogram of false negatives we see further that the missing vessels are relatively thin; 62.2% of the missing parts have a width up to 3. The

Table 1
Evaluation results for MS_{thin}, MS_{del}, MS_{exp}, and MS_{ins}.

	MS_{thin}	MS_{del}
TPR	99.6%	77.6%
Detection error DE	0.060	0.000
Position error PE	0.191	0.002
Width error WE	0.452	0.002
False positives	0	0
False negatives	32	1729
FN histogram	–	1–2:50.3%, 2–3:11.9%, 3–4:25.0%, 4–5:9.4%
	MS_{exp}	MS_{ins}
TPR	100.0%	100.0%
Detection error DE	0.068	0.000
Position error PE	0.226	0.000
Width error WE	0.462	0.000
False positives	6	400
FP histogram	–	1–2:8.0%, 2–3:4.0%, 3–4:4.0%, ⩾4:84.0%
False negatives	0	0

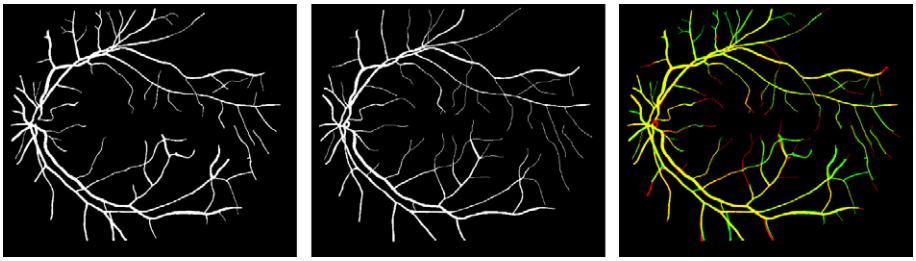

Fig. 3. Two hand-labelings of a retinal image from STARE database and their color overlay for making the tiny differences clearer. The first hand-labeling (left) is coded in green und the second (middle) in red. Common labeling points thus appear yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

interpretation of these evaluation measures is exactly what we postulated for more informative and precise performance evaluation in contrast to the pixel-wise evaluation method.

Similar improvement can also be observed for MS_{exp} and MS_{ins}. Although both MS_{exp} and MS_{ins} have 100% TPR, MS_{exp} is not a perfect detection. The partial expanding in MS_{exp} results in a width error 0.462, indicating some detection inaccuracy. Basically, no false positive is found in MS_{exp}, compared to 400 in MS_{ins}. Furthermore, the spurious vessels in MS_{ins} are mainly thick ones with 84.0% being at least 4 pixels wide (corresponding to the thick added diagonal vessels). The measures on this second image pair again confirm our impression of the segmentation results and demonstrate the more informative and precise nature of the proposed evaluation methodology.

4.2. STARE database

The STARE database (Hoover et al., 2000) contains 20 images of digitized slides (available at http://www.parl.clemson.edu/stare/probing/, 700 × 605 pixels, 8 bits per color channel). There are two hand-labelings made by two different persons (computer scientists with knowledge in ophthalmology), see Fig. 3 for an example. The first hand-labeling, which is usually used as ground truth in performance evaluation (Hoover et al., 2000; Jiang and Mojon, 2003; Staal et al., 2004), took a more conservative view of the vessel boundaries and in the identification of small vessels than the second hand-labeling.

4.2.1. Single image case (STARE)

We start with the single retinal image shown in Fig. 4, together with the corresponding ground truth, and vessel detection MS_1 and MS_2 from two different algorithms (Jiang and Mojon, 2003 and Martinez-Perez et al., 1999). Using the pixel-wise evaluation we obtain: MS_1: TPR = 91.9%, MS_2: TPR = 80.3%. There is a large difference (11.6%) in TPR. The evaluation measures based on our approach are: MS_1: TPR = 89.3%, MS_2: TPR = 87.2%. Actually, MS_1 only detects 2.1% more of the vessel network structure than MS_2. The much larger difference of 11.6% above is explained by the fact that MS_1 tends to be thicker than GT. Thus, it produces a better pixel-wise matching result. Measured by our method, this is expressed by a larger width error for MS_1 (1.129) than MS_2 (0.693). Here our performance measures clearly provide a more precise description of the differences between algorithmic results and ground truth.

4.2.2. Whole database (STARE)

We compare our evaluation approach with the early pixel-wise method in three different situations. The first hand-labeling is used as ground truth in all three of them. The results are summarized in Fig. 5.

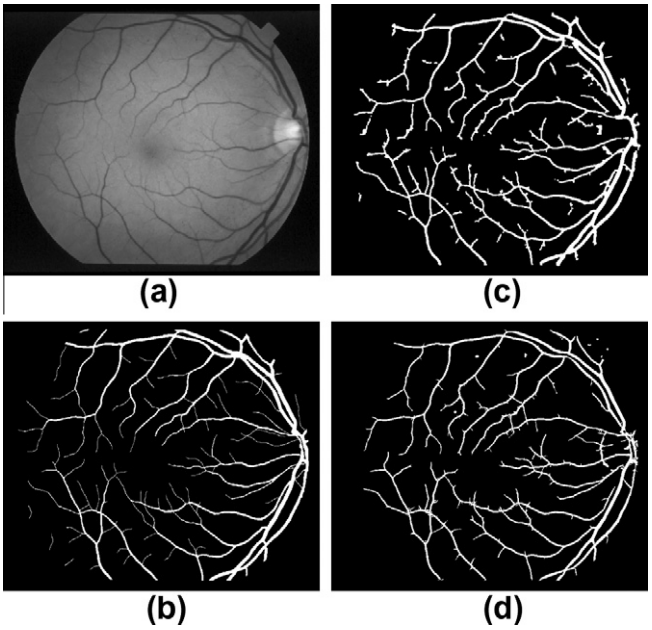

Fig. 4. (a) Retinal image; (b) GT; (c) first detection results MS_1; (d) second detection result MS_2.

*Verification-based adaptive local thresholding (*Jiang and Mojon, 2003*)*: This vessel detection method has been evaluated on the STARE database. Based on eight parameter sets the ROC is plotted in Fig. 5 ("multi-threshold probing"). Note that the 20 retinal images are divided into a subset of normal and a subset of abnormal cases. The performance study thus can be done for three test instances (all, normals, abnormals). In this case both evaluation methods have similar TPR values. The reason lies in the fact that the results from Jiang and Mojon (2003) tend to be thicker than the ground truth. Therefore, as soon as some part of the vessel network is detected, most of the vessel pixels of that part will be marked, leading to a local TPR value near 100% comparable to the local TPR from our evaluation approach. On the other hand, the FPR has much smaller values due to the use of spurious midline pixels only in our approach instead of all spurious vessel pixels.

*Piecewise threshold probing of matched filter response (*Hoover et al., 2000*)*: For this method only one result per image for a particular parameter set is available. Looking at Fig. 5 ("filter response analysis"), we see that our evaluation method rates the TPR considerably more positive (from lower than 70% to almost 80%). The low TPR value of pixel-wise evaluation is caused by the algorithm's tendency of not fully marking all local vessel pixels even if the middle part, thus the local network structure, is correctly found. Our structure-based evaluation approach considers the aspects of structure

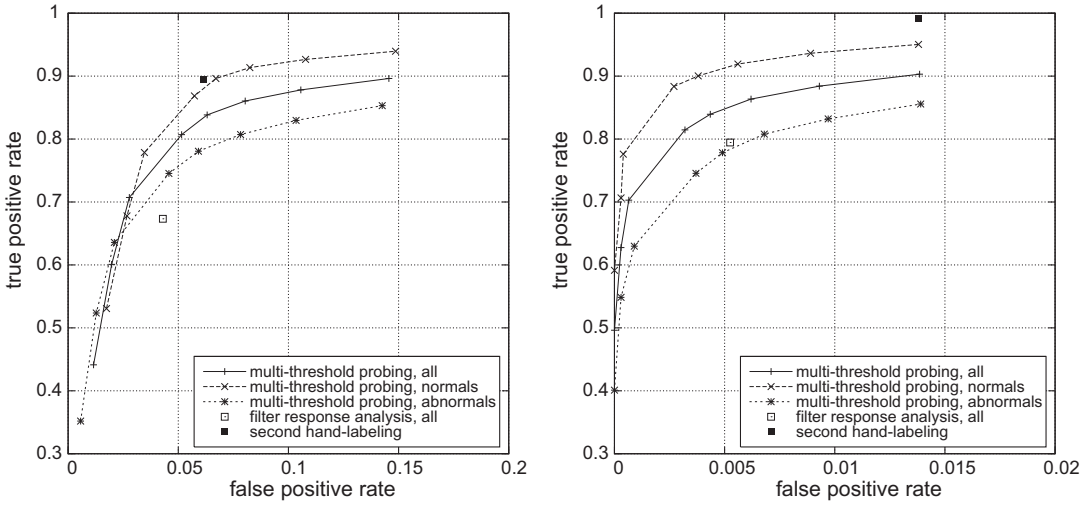

Fig. 5. Evaluation of multi-threshold probing, filter response analysis, and second hand-labeling on STARE database: pixel-wise evaluation (left) and our approach (right). In our case the FPR has much smaller values due to the use of spurious midline pixels only instead of all spurious vessel pixels.

Table 2
Evaluation results for comparing the second labeling in Fig. 3 against the first labeling.

TPR	92.2%
Detection error DE	0.375
Position error PE	1.573
Width error WE	0.634
False positives	955
FP histogram	1–2:90.1%, ⩾2:9.9%
False negatives	558
FN histogram	1–2:70.1%, 2–3:5.9%, 3–4:17.4%, ⩾4:6.6%

detection and local detection accuracy separately and is therefore able to characterize the behavior of an algorithm more precisely.

Second hand-labeling: In (Hoover et al., 2000; Jiang and Mojon, 2003) the second hand-labeling has been used as "machine-segmented result images" and compared to the first hand-labeling. The detection performance measures are then regarded as a target performance level. In Fig. 5 this level is indicated by an isolated mark in each plot ("second hand-labeling"). Although the second observer masked the vessels more completely, the pixel-wise TPR only amounts to about 90% because the second labeling is partly thinner than the first one. This assessment is obviously against our intuition and expectation. Using our approach the TPR increases to almost 100%.

Table 2 gives the details of this comparison for the retinal image shown in Fig. 3. The pixel-wise evaluation results in a TPR value of only 66.0% for this image. On the other hand, our approach indicates that a much higher rate of 92.2% of the vessel network structure has been correctly segmented by the second observer. The large divergence is caused by the differences in position and width of the marked vessels by the two observers, which is signified by quite large position and width errors in our case. The second observer labels more small vessels. This is documented by the number of false positives, namely 955. Among them 90.1% are midline pixels of thin vessels of one pixel width. This example makes once more our way of assessing the detection quality clear.

4.3. Study of parameter c_w

Our approach has one major parameter c_w. We have conducted a study using varying parameter values based on the multi-threshold probing method and the STARE database; see Fig. 6. The results indicate quite stable behavior for $c_w \in [0.4,0.7]$. For this

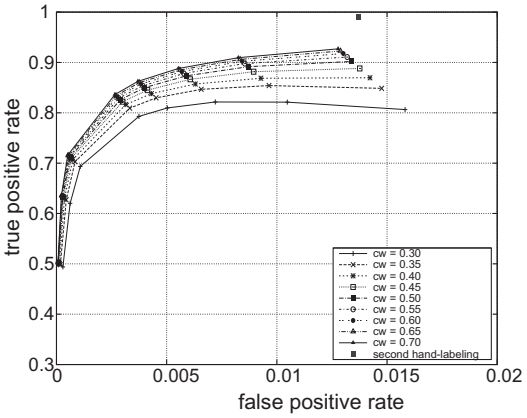

Fig. 6. Performance measure for varying values of parameter c_w (multi-threshold probing, STARE).

reason the parameter setting $c_w = 0.5$ was chosen and used for all experimental results reported in this paper.

4.4. Robustness issues of skeletonization

The basis for our proposed evaluation method is the skeletonization of detected vessel networks. Two robustness issues are considered here.

Image rotation: Ideally, the computed skeleton should be invariant to image translation and rotation. While translation invariance is mostly satisfied, the rotation invariance is more challenging. The relatively simple thinning algorithm from (Cardoner and Thomas, 1997), which is used in our current implementation, produces slightly different results for rotated binary images. To investigate the influence of rotation-dependent thinning results to the performance measures, we conducted the following experiment. The detection result in Fig. 4(c) was compared four times with the GT in Fig. 4(b): The original pair and counterclockwise rotated versions by 90°, 180°, and 270°. The resulting three main rates are listed in Table 3. It can be concluded that the influence of rotation-dependent thinning is not significant and thus using a thinning algorithm like (Cardoner and Thomas, 1997) should not have any impact on ranking vessel segmentation methods.

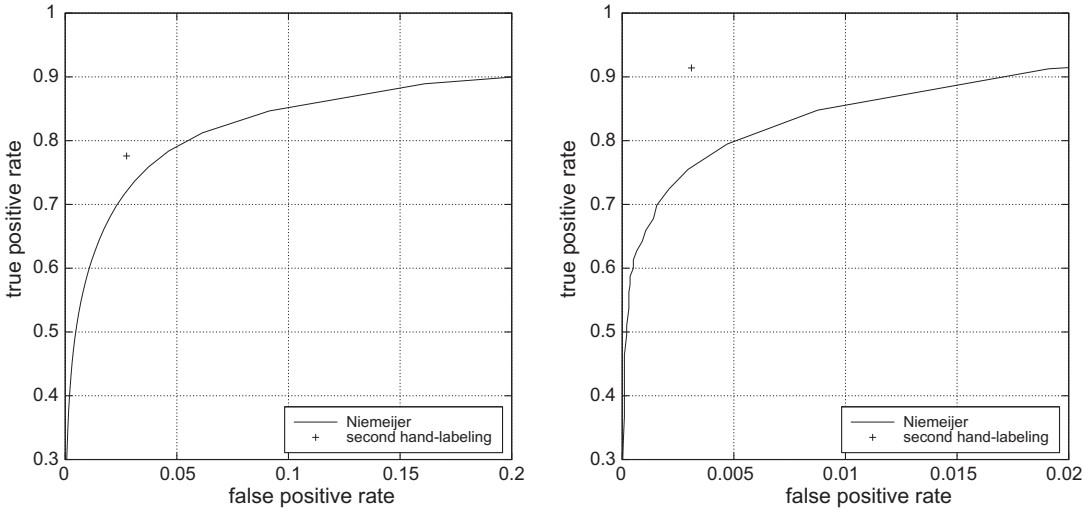

Fig. 7. Evaluation on DRIVE database: pixel-wise evaluation (left) and our approach (right). The term "Niemeijer" denotes the vessel detection method from Niemeijer et al. (2004). In our case the FPR has much smaller values due to the use of spurious midline pixels only instead of all spurious vessel pixels.

Spurious branches: Skeletonization potentially produces minor spurious branches. This phenomenon is not only observable in the relatively simple thinning algorithm (Cardoner and Thomas, 1997) used in our current implementation. Even more sophisticated thinning methods like (Tang et al., 2010) cannot fully avoid it. For the task under consideration, however, these spurious branches are of such small proportion that they should not significantly influence the performance measures. This expectation has been well confirmed by our experiments, in which such spurious branches were manually removed from both detection result and GT. The performance measures with and without the manual removal of spurious branches hardly differ.

4.5. DRIVE database

The DRIVE database consists of 40 images (available at http://www.isi.uu.nl/Research/Databases/DRIVE/, 768 × 584 pixels, 8 bits per color channel). The pixel classification approach to vessel detection from Niemeijer et al. (2004) has been evaluated using both methods, see Fig. 7. Similar to the evaluation of the verification-based adaptive local thresholding (Jiang and Mojon, 2003) on the STARE database, the performance measure TPR does not change too much in this case. Large changes occur, however, if we compare two different hand-labelings. Similar arguments apply here as well. Our structure-based evaluation indicates more precisely the structure detection rate.

4.6. Computational time

For comparing a pair of GT and MS image, the skeletonization takes about 2.5 s (STARE) and 1.6 s (DRIVE), respectively, on a standard notebook. In both cases the matching time is about 1.5 s. No code optimization has been tried so far. For the non-realtime task of performance evaluation the computation time is acceptable.

Table 3
Performance measures of rotated detection results.

	Original	90°	180°	270°
TPR	89.28%	89.22%	89.40%	89.14%
FPR	0.10%	0.10%	0.10%	0.09%
FNR	10.72%	10.78%	10.60%	10.86%

5. Discussions and conclusion

Compared to other commonly used features such as edges and regions, there is relatively little work on performance evaluation of algorithms for curvilinear structure detection. In this paper we have proposed a novel structure-based methodology for this purpose. We consider the two aspects of performance, namely detection rate and detection accuracy, separately, in contrast to their mixed handling in earlier approaches that typically produces biased impression of detection quality. By doing so, the proposed performance measures give us a more informative and precise performance characterization. The detailed information about width, for instance, helps the user select a suitable algorithm for a particular application as discussed in Section 2.

Both synthetic and real examples have been used to demonstrate the advantages of our approach. In fact, the use of our evaluation approach may change our thinking about the relative performance of algorithms. Concerned with the vessel detection method from Hoover et al. (2000), for instance, although the evaluation could only be done for one parameter setting, the results in Section 4.2 (i.e., the increased TPR value according to our structure-based evaluation) indicate that this algorithm has a higher "intrinsic" detection rate than assumed so far based on the pixel-wise comparison. The superior performance of the algorithm (Jiang and Mojon, 2003) is at least partly caused by its tendency of producing thicker vessels. More experiments (using vessel detection results not available to us yet) will be needed to fully clarify this point. But this incomplete comparison shows the potential of our approach to directing the evaluation to the intrinsic algorithmic performance.

It is not our intention in this work to conduct a rather complete performance evaluation for a large number of algorithms. Instead, the experiments reported in Section 4 are shown to demonstrate the principal usefulness of our structural performance evaluation. We will do the evaluation work in future by involving other researchers in a joint effort.

The description of the evaluation methodology and the experimental work have been embedded in the context of blood vessel detection in retinal images. It is important to point out that our approach is applicable in the general context of the evaluation of curvilinear structure detection algorithms. In particular, extraction of airway tree and other thin structures in volumetric data (Holtzman-Gazit et al., 2006; Tschirren et al., 2005) is a challenging task

and our evaluation technique will help assess the algorithm performance as well.

Acknowledgments

The authors want to thank A. Hoover, M. Niemeijer and his colleagues for making their retinal image databases and the ground truth data publicly available. They also provided us some of their experimental results. A. Goldberg and R. Kennedy kindly provided their CSA implementation for public use. Thanks also go to Peter Larysch for his support in experimental work.

References

Bharkad, S.D., Kokare, M., 2011. Performance evaluation of distance metrics: Application to fingerprint recognition. Int. J. Pattern Recog. Artif. Intell. 25 (6), 777–806.

Cardoner, R., Thomas, F., 1997. Residuals + directional gaps=skeletons. Pattern Recog. Lett. 18 (4), 343–353.

Cardoso, J.S., Sousa, R., 2011. Measuring the performance of ordinal classification. Int. J. Pattern Recog. Artif. Intell. 25 (8), 1173–1195.

Chaudhuri, S., Chatterjee, S., Katz, N., Nelson, M., Goldbaum, M., 1989. Detection of blood vessels in retinal images using two-dimensional matched filters. IEEE Trans. Med. Imag. 8 (3), 263–269.

Fang, B., You, X., Tang, Y.Y., Chen, W.S., 2005. Morphological structure reconstruction of retinal vessels in fundus images. Int. J. Pattern Recog. Artif. Intell. 19 (7), 937–948.

Frame, A., Undrill, P., Cree, M., Olson, J., McHardy, K., Sharp, P., Forrester, J., 1998. A comparison of computer based classification methods applied to the detection of microaneurysms in ophthalmic fluorescein agiograms. Comput. Biol. Med. 28, 225–238.

Gabow, H., Tarjan, R., 1989. Faster scaling algorithms for network problems. SIAM J. Comput. 18 (5), 1013–1036.

Goldberg, A., Kennedy, R., 1995. An efficient cost scaling Algorithm for the assignment problem. Math. Prog. 71, 153–178.

Heath, M.D., Sarkar, S., Sanocki, T., Bowyer, K.W., 1997. A robust visual method for assessing the relative performance of edge-detection algorithms. IEEE Trans. PAMI 19 (12), 1338–1359.

Holtzman-Gazit, M., Kimmel, R., Peled, N., Goldsher, D., 2006. Segmentation of thin structures in volumetric medical images. IEEE Trans. Image Process. 15 (2), 354–363.

Hoover, A., Goldbaum, M., 2003. Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels. IEEE Trans. Medical Imag. 22 (8), 951–958.

Hoover, A., Kouznetsova, V., Goldbaum, M., 2000. Locating blood vessels in retinal images by piece-wise threshold probing of a matched filter response. IEEE Trans. Med. Imag. 19 (3), 203–210.

Jiang, X., 2005. Performance evaluation of image segmentation algorithms. In: Chen, C.H., Wand, P.S.P. (Eds.), Handbook of Pattern Recognition and Computer Vision, 3rd ed. World Scientific, pp. 525–542.

Jiang, X, Mojon, D., 2002. Supervised evaluation methodology for curvilinear structure detection algorithms. In: Proc. 16th Int. Conf. on Pattern Recognition, vol. I, 2002, pp. 103–106.

Jiang, X., Mojon, D., 2003. Adaptive local thresholding by verification-based multithreshold probing with application to vessel detection in retinal images. IEEE Trans. PAMI 25 (1), 131–137.

Jiang, X., Marti, C., Irniger, C., Bunke, H., 2006. Distance measures for image segmentation evaluation. EURASIP J. Appl. Signal Processing, Special Issue on Performance Evaluation in Image Processing, 1–10.

Jiang, X., Lambers, M., Bunke, H., 2011. Structure-based evaluation methodology for curvilinear structure detection algorithms. In: Jiang, X., Ferrer, M., Torsello, A. (Eds.), Graph-Based Representations in Pattern Recognition, LNCS, vol. 6658. Springer, pp. 305–314.

Lam, B.S.Y., Yan, H., 2008. A novel vessel segmentation algorithm for pathological retina images based on the divergence of vector fields. IEEE Trans. Med. Imaging 27 (2), 237–246.

Lam, B.S.Y., Gao, Y., Liew, A.W.-C., 2010. General retinal vessel segmentation using regularization-based multiconcavity modeling. IEEE Trans. Med. Imaging 29 (7), 1369–1381.

Lee, S.U., Chung, S.Y., Park, R.H., 1990. A comparative performance study of several global thresholding techniques for segmentation. Computer Vision, Graphics, and Image Processing 52, 171–190.

Martinez-Perez, M., Hughes, A., Stanton, A., Thom, S., Bharath, A., Parker, K., 1999. Scale-space analysis for the characterization of retinal blood vessels. In: Proc. Medical Image Computing and Computed Assisted Intervention (MICCAI), 1999, pp. 90–97.

Maurer, C.R., Qi, R., Raghavan, V., 2003. A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. IEEE Trans. PAMI 25 (2), 265–270.

Niemeijer, M., Staal, J.J., van Ginneken, B., Loog, M., Abramoff, M.D., 2004. Comparative study of retinal vessel segmentation methods on a new publicly available database. In: Fitzpatrick, J., Sonka, M. (Eds.), SPIE Medical Imaging, vol. 5370, 2004, pp. 648–656.

Niemeijer, M., van Ginnecken, B., Cree, M.J., et al., 2010. Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs. IEEE Trans. Med. Imag. 29 (1), 185–195.

Romero, F., Ruos, L., Thomas, F., 2000. Fast skeletonization of spatially encoded objects. In: Proc. 15th Int. Conf. on Pattern Recognition, vol. 3, 2000, pp. 510–513.

Rothaus, K., Jiang, X., Rhiem, P., 2009. Separation of the retinal vascular graph in arteries and veins based upon structural knowledge. Image Vision Comput. 27 (7), 864–875.

Shufelt, J.A., 1999. Performance evaluation and analysis of monocular building extraction from aerial imagery. IEEE Trans. PAMI 21 (4), 311–326.

Staal, J., Abramoff, M.D., Niemeijer, M., Viergever, M.A., van Ginneken, B., 2004. Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imag. 23 (4), 501–509.

Tang, Y., Bai, X., Yang, X., Lin, L., Liu, S., Jan Latecki, L., 2010. Skeletonization with particle filters. Int. J. Pattern Recog. Artif. Intell. 24 (4), 619–634.

Tschirren, J., Hoffman, E.A., McLennan, G., Sonka, M., 2005. Intrathoracic airway trees: segmentation and airway morphology analysis from low-dose CT scans. IEEE Trans. Med. Imaging 24 (12), 1529–1539.

Zana, F., Klein, J., 1999. A multimodal registration algorithm of eye fundus images using vessels detection and Hough transform. IEEE Trans. Med. Imag. 18 (5), 419–428.

Zana, F., Klein, J., 2001. Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation. IEEE Trans. Med. Imag. 10 (7), 1010–1019.