# Dissimilarity criteria and their comparison for quantitative evaluation of image segmentation: application to human retina vessels

Yann Gavet, Mathieu Fernandes, Johan Debayle, Jean-Charles Pinoli

# Dissimilarity criteria and their comparison for quantitative evaluation of image segmentation. Application to human retina vessels.

Yann Gavet, Mathieu Fernandes, Johan Debayle and Jean-Charles Pinoli
École Nationale Supérieure des Mines de Saint-Etienne, France,
LGF UMR CNRS 5307,
158 Cours Fauriel, 42023 Saint-Etienne cedex, France
Tel.: +33 4 7742 0170; fax: +33 4 7749 9694.
E-mail address: gavet@emse.fr

**Abstract**

The quantitative evaluation of image segmentation is an important and difficult task that is required for making a decision on the choice of a segmentation method and for the optimal tuning of its parameter values. To perform this quantitative evaluation, dissimilarity criteria are relevant with respect to the human visual perception, contrary to metrics that have been shown to be visually not adapted. This article proposes to compare eleven dissimilarity criteria together. The field of retina vessels image segmentation is taken as an application issue to emphasize the comparison of five specific image segmentation methods, in regard to their degrees of consistency and discriminancy. The DRIVE and STARE databases of retina images are employed and the manual/visual segmentations are used as a reference and as a control method. The so-called $\epsilon$ criterion gives results in agreement with perceptually based criterions for achieving the quantitative comparison.

Please cite as follows:

```
@Article{Gavet2014,
  Title  = {Dissimilarity criteria and their comparison for quantitative evaluation of image segmentation:
         application to human retina vessels},
  Author = {Gavet, Yann and Fernandes, Mathieu and Debayle, Johan and Pinoli, Jean-Charles},
  Journal= {Machine Vision and Applications},
  Year   = {2014},
  Number = {8},
  Pages  = {1953-1966},
  Volume = {25},

  Doi    = {10.1007/s00138-014-0625-2},
  ISSN   = {0932-8092},
  Keywords = {Image segmentation; Quantitative evaluation;
          Segmentation evaluation; Dissimilarity criteria;
 Retinal blood vessels},
  Language = {English},
  Publisher= {Springer Berlin Heidelberg},
  Url      = {http://dx.doi.org/10.1007/s00138-014-0625-2}
}
```

## 1 Introduction

Image segmentation is one of the most important and hard-to-address steps in image analysis. An image segmentation process consists in partitioning the spatial support of an image into adjacent parts that are pairwise disjoint. When dealing with a 2-D image, that is to say with a 2-D support, such parts are either regions (two dimensional sets), contours (one dimensional sets) or isolated points (zero dimensional sets). Only Euclidean sets will thus be considered. Moreover, only binary images will be considered without the loss of generality for the purpose of this paper. The general segmentation case (segmentation into n different sets) will not be discussed as it implies different processes, like for example pairing the sets together (see also [22]).

The relevance of an image segmentation is firstly and mainly that of the resulting binary image. Comparing such resulting binary images coming from various segmentation methods (automatically or manually/visually performed by a computer and a human expert, respectively) is therefore required. Since these methods have been selected as performing in front of a particular

imaging application issue, they often output similar binary images, which makes their comparative study more tedious. The reader could refer to [18, 29, 30, 21] to have a view on some methods for segmentation evaluation.

The quantitative comparative evaluation either deals with the region-based paradigm or with the contour-based paradigm [17]. It can be performed by resorting to a reference method or not (see [3]). To perform this comparison, numerous criteria have been reported in the specialized literature. In this article, eleven numerical criteria are employed. They are supervised criteria, as they compare a binary image with a ground truth (usually performed manually/visually by an expert). In fact, the quantitative evaluation by means of comparative criteria is the so-called first-order comparative problem, since it is criterion-dependent. Consequently, this article also addresses the second-order comparative problem consisting in the quantitative comparison of the eleven criteria themselves. This article focuses on the so-called summary measures [22]. Although ROC curves are a lot richer, their comparison is not straightforward. It is not evident for example to decide whether it is preferable to have false negatives instead of false positives.

The field of ophthalmology with the segmentation of the blood vessels of retinal images is addressed here. Five image segmentation methods reported in the literature as particularly adapted to such an application issue have been arbitrarily selected. They are not state of the art methods, but they will provide different results for performing a comparison of the criteria. The article is organised in six sections. First, the image segmentation is described from a geometrical viewpoint. In section 2, the mathematical notions of metric and dissimilarity are presented and discussed, showing the relevance of the second one. Section 3 focuses on the eleven dissimilarity criteria used in this article for supervised segmentation evaluation. In section 4, five selected image segmentation methods are presented. Quantitative comparisons are detailed in section 5. Finally, a concluding discussion and perspectives end the article.

### 1.1 Image segmentation

An image segmentation process refers to the action of partitioning the spatial domain of an image into adjacent regions, each of them preserving a certain homogeneity with respect to a given criterion.

This article deals with the case where the image segmentation result is a binary image which partitions the spatial support $S$ (that will be considered as a bounded set in $\mathbb{R}^2$) such that $S = \bigcup_{i\in I} R_i$ where $I$ is finite and the regions $(R_i)_{i\in I}$ are pairwise disjoint (typically open) sets.

## 2 Metrics and dissimilarities

In many real applications, it is useful to compare segmentations together. This can be done for evaluating the efficiency of segmentation methods together or comparing them to a reference method (see [3]). Many approaches exist, mostly involving region criteria [30, 21]. Usually, these methods are mathematically based on distance functions, e.g., metrics.

Distance functions are adapted to perform comparisons of (mathematical) objects belonging to an "object collection" $\xi$ (for example, $\xi$ is the family of the segmentations on a given image, e.g., with a given definition domain $S$). A metric is a particular distance function $d$ that is defined on $\xi^2$ and valued in $\mathbb{R}_+$ (set of all positive real numbers) , satisfying the four following axioms (see [6]):

$$\begin{aligned}
&\text{(identity)} && \forall x \in \xi,\ d(x,x) = 0\\
&\text{(separation)} && \forall x, y \in \xi,\ d(x,y) = 0 \Rightarrow x = y\\
&\text{(symmetry)} && \forall x, y \in \xi,\ d(x,y) = d(y,x)\\
&\text{(triangle inequality)} && \forall x, y, z \in \xi,\ d(x,y) \leq d(x,z) + d(z,y)
\end{aligned}$$

Nevertheless, metrics are not necessarily the adapted notion. Indeed, it has been shown [26, 27, 10] that the metric notion is not consistent with the human visual perception. To be more precise, the three last axioms of separation, symmetry and triangle inequality do not hold for the human visual system:

- two objects of the same kind (for example, a computer-drawn square and a manually-drawn square) are considered as visually similar although different from a set-theoretic point of view.
- the order of observation of two different objects is important, thus, the symmetry is not verified by the human visual system.
- the triangle inequality is not verified. The Fig. 1 illustrates this assertion.

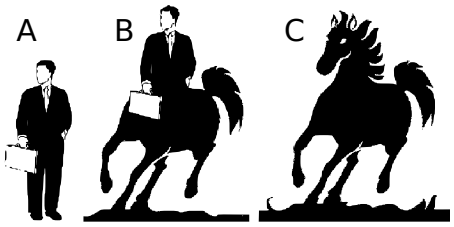



Figure 1: Illustration of the triangle inequality axiom from a visual point of view. The human visual system does not verify the triangle inequality, because A and B are partly close, B and C are also partly close, but A and C are really different Thus, if $d$ is the visual comparative criterion, yields $d(A,C) > d(A,B) + d(B,C)$.

Dissimilarity is the relevant notion, which defines a positive real valued function of two variables: a zero value means that two objects are similar, and a large value implies a high dissimilarity.

As the human expertise plays an important role in image segmentation evaluation, the reader should notice that there are two categories of segmentation evaluation: supervised and unsupervised [30]. The unsupervised evaluation involves an absolute criterion that characterizes the accuracy of the segmentation. The supervised evaluation is based on a distance between a reference segmentation result (called gold-standard in several applications fields; it can be the result given by an automatic reference method or by a human expert) and another segmentation result (generally computed).

# 3 Eleven dissimilarity criteria for supervised segmentation evaluation

The following notations are first introduced: $M$ (as Manual) is the binary result of a reference segmentation method and $X$ is the result of the evaluated segmentation method.

This section introduces some classical and recent dissimilarity criteria that will be considered for binary segmentation evaluation. This list is not exhaustive, interesting ideas and criteria can be found for example in [2, 15, 28], for general segmentation evaluation.

## 3.1 Figure of merit

One of the mostly used criterion for evaluating segmentation is the so-called "figure of merit" (fom) ([1, 25]) defined in Eq. 1. Be aware that the original definition does not involve a difference to 1. It has been introduced in this paper so that the fom criterion becomes a dissimilarity.

$$fom_M(X) = 1 - \frac{1}{\max\{\#(M), \#(X)\}} \sum_{p \in X} \frac{1}{1 + d^2(p, M)} \tag{1}$$

where $d(p, M)$ is the Euclidean distance between the pixel $p \in X$ to the closest pixel of $M$ and $\#$ designates the cardinality, namely the number of pixels of the considered sets $M$ or $X$.

## 3.2 Dice coefficient

Another widely used criterion is the Dice coefficient [7], related to the Jaccard index [14], and defined by Eq. 2:

$$dice_M(X) = 1 - \frac{2\#(M \cap X)}{\#(M) + \#(X)} \tag{2}$$

As for the figure of merit (fom), the original definition of the Dice coefficient does not involve the difference to 1, but this has been also introduced to become a dissimilarity.

## 3.3 Five criteria with pixel classification

Performance results can also be compared with the use of pixels classification (see Table 1) into four categories:

- True Positive ($TP$),
- False Positive ($FP$),

- True Negative ($TN$),
- False Negative ($FN$).

Positive or Negative refer to the detection, true or false refer to the reference from the ground truth. Then, several criteria are defined in Table 2, by evaluating the number of pixels in each category:

- sensitivity (measures the proportion of actual positives which are correctly identified as such, e.g. the percentage of pixels present in the ground truth which are correctly segmented , also known as recall rate ),
- specificity (measures the proportion of negatives which are correctly identified, e.g. the percentage of pixels absent of the ground truth which are present in the segmentation),
- positive predictive value (the proportion of detected pixels that are true positives, also known as precision ),
- negative predictive value (the proportion of negative pixels that will not be present in the segmentation result),
- accuracy (measures the closeness of measurements of a quantity to that quantity's actual (true) value).

Notice that the Dice coefficient can also be expressed in terms of pixel classification as $dice = \frac{2TP}{2TP+FN+FP}$.

| evaluated $X$ \ reference $M$ | Pixel value= 1 | Pixel value = 0 |
|---|---|---|
| Pixel value = 1 | True Positive ($TP$) | False Positive ($FP$) |
| Pixel value = 0 | False Negative ($FN$) | True Negative ($TN$) |

Table 1: Pixels classification. Each category set ($TP$, $FP$, $FN$ or $TN$) represents a number of pixels.

| | |
|---|---|
| Sensitivity | $Se = \frac{TP}{TP+FN}$ |
| Specificity | $Sp = \frac{TN}{TN+FP}$ |
| Positive predictive value | $Ppv = \frac{TP}{TP+FP}$ |
| Negative predictive value | $Npv = \frac{TN}{TN+FN}$ |
| Accuracy | $Acc = \frac{TP+TN}{TP+FN+TN+FP}$ |

Table 2: Some evaluation criteria based on pixel classification (Table 1 ).

### 3.4 The $\epsilon$ dissimilarity criterion

The $\epsilon$ dissimilarity criterion is based on the symmetric difference of sets, but involves a tolerance with the help of the Minkowski addition [20]. It is an extension of the Jaccard index [14] treated in [11], where a detailed study of its properties (tolerance to under- or over-segmentation, translation or distortion, smoothing and visual tolerance as well as mathematical properties) has been reported.

The $\epsilon$ dissimilarity criterion with the tolerance $\rho$ applied to segmented images is defined by the following equation (Eq. 3, see also [11]):

$$\epsilon_M^\rho(X) = \frac{\#\{(X \setminus M \oplus \rho N) \cup (M \setminus X \oplus \rho N)\}}{\#\{M \oplus \rho N\}} \tag{3}$$

with $N$ being the disk of radius 1 (as structuring element) and # designating the number of pixels within the set. The Minkowski addition symbol, denoted $\oplus$, is equivalent to the morphological dilation, $N$ being symmetric, where $N$ is called a structuring element (for example a disk). Practically, $\rho$ is the radius of the ball used to dilate the binary images.

### 3.5 Fuzzy Jaccard index

In the same objective of being perceptually relevant and visually tolerant, [18] proposes a so-called fuzzy Jaccard index. $X$ and $M$ are the segmented sets and the reference (manual) segmentation, respectively, $X^c$ and $M^c$ are the complements of these sets. Let $\mathcal{N}_x$ be the 8-neighborhood of a point $x \in \mathbb{Z}^2$.

$$B_X = \{x : x \in X \wedge \{\mathcal{N}_x \cap X^c \neq \emptyset\}\}$$
$$B_M = \{x : x \in M \wedge \{\mathcal{N}_x \cap M^c \neq \emptyset\}\}$$

A membership function, called fuzzy border pixel maps in [18], is defined as follows, for $K = X$ or $M$:

$$\tilde{B}_K(x) = exp\left(-\frac{\parallel x - \hat{x} \parallel^2}{2\sigma^2}\right)$$

$$\hat{x} = \arg\min_{y \in K} \parallel x - y \parallel$$

The parameter $\sigma$ is the "fuzzy" parameter, used to introduce a tolerance, and similar to $\rho$ for the $\epsilon$ parameter. Then, the fuzzy Jaccard index is defined as:

$$FJ = 1 - \frac{\sum_x \min\left(\tilde{B}_X(x), \tilde{B}_M(x)\right)}{\sum_x \max\left(\tilde{B}_X(x), \tilde{B}_M(x)\right)}$$

Notice that the values are between 0 and 1. This definition is slightly modified from the original definition from [18] to get 0 for an exact match.

### 3.6 Perceptually-weighted evaluation criterion

An evaluation criterion that tries to be perceptually significant, adapted to object present in videos, is proposed in [29]. For the purpose of binary image segmentation, we will restrict the definition to the static part of this criterion *qms*, which is defined for a fixed frame as follows:

$$\mathrm{QMs} = \mathrm{QMs}^+ + \mathrm{QMs}^-$$

where

$$\mathrm{QMs}^+ = \sum_{d=1}^{D^+_{\max}} w_+(d) \cdot \#\left[M_d^+ \cap X\right]$$

$$\mathrm{QMs}^- = \sum_{d=1}^{D^-_{\max}} w_+(d) \cdot \#\left[M_d^- \cap X^c\right]$$

and with the definitions (see [29]):

- $d$: distance to the reference mask border;
- $D^+_{\max}$ and $D^-_{\max}$: biggest distance $d$ for, respectively, false positives and false negatives;
- $w_+(d)$: weighting function for false positives according to distance $d$, expressed as $w_+(d) = B_1 + \frac{B_2}{d+B_3}$;
- $w_-(d)$: weighting function for false negatives, expressed as $w_-(d) = F_S \cdot d$, with $F_S$ a constant factor;

- $M_d^+$ and $M_d^-$: sets of pixels situated at distance $d$ from the reference mask border, outside or inside the mask, respectively ($dist$ is the distance from pixel $x$ to a set):

$$M_d^+ = \{x | x \in M^c, dist(x, M) = d\}$$
$$M_d^- = \{x | x \in M, dist(x, M^c) = d\}$$

- $X$ is the segmentation result.

Finally,

$$\text{qms} = \frac{QMS}{\#M}$$

. Practically, the histogram of the distances to the reference mask of all pixels is used to compute this evaluation criterion.

# 4 Presentation of the five selected human retina vessels image segmentation methods

Five image segmentation methods (see Fig. 2) that specifically address the human retina vessels extraction have been selected for the different tests on the dissimilarity criteria. This section presents them quickly.

## 4.1 Chaudhuri et al. method (Method 1)

This method [5] is based on the assumption that the retina vessels' intensity distributions may be approximated by the following equation (Eq. 4, where $f$ is the intensity profile in the image along a line perpendicular to a vessel):

$$f(x, y) = A\left(1 - ke^{-d^2/2\sigma^2}\right) \tag{4}$$

where $d$ is the perpendicular distance between the point $(x, y)$ and the straight line passing through the center of the blood vessel in the direction along its length, $\sigma$ denotes the thickness of the vessel, $L$ designates the length of the vessels and is used in Method 1, and $A$ and $k$ are real number constants reflecting the gray level intensity of the local background and the reflectance of the vessel. A matched filter (a correlation) is applied with these constraints ($\sigma = 2$ and $L = 10$ are the parameters of the algorithm) and the result is finally binarized.

The Chaudhuri et al method is summarized in Method 1.

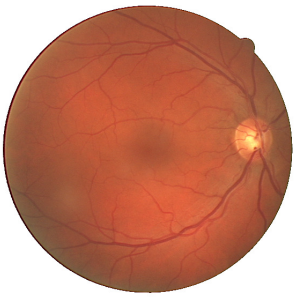
**Data** : *Input* ← Grayscale image of retina.
$\sigma$ ← Thickness of detected vessels (pixels).
$L$ ← Length of detected vessels (pixels).
*Direction*← Different orientations.
**Result** : $S$: Segmentation of the retina image.

```
begin
    foreach Direction do
        Matched filter: emphasize linear segments of length L and thickness σ;
    end
    Take the minimum result for all directions;
    Threshold and remove spurs to get a clean binary image;
    Apply a mask of the region of interest, if necessary;
end
```

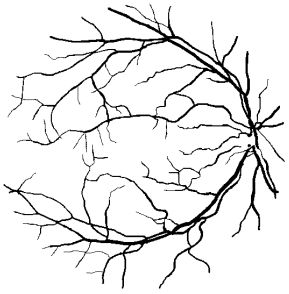
**Method 1:** Retina image segmentation from [5]. For the tests, the color images are converted into grayscale by only taking the Green channel. The values of $\sigma$ and $L$ are experimentally fixed at 2 and 10.

## 4.2 Chanwimaluang et al. method (Method 2)

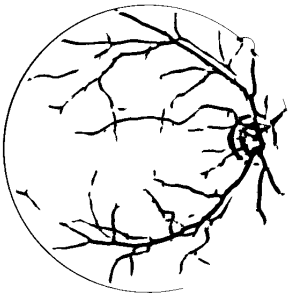
This second segmentation method ([4]) is based on the previous one, beginning by the same matched filter. The thresholding method changes (see Method 2): the connected components with a too small length are removed since it is based on a local entropy measure (see [4] for precisions).



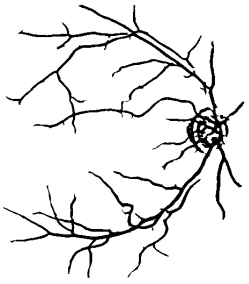
(a) Original image of retina.



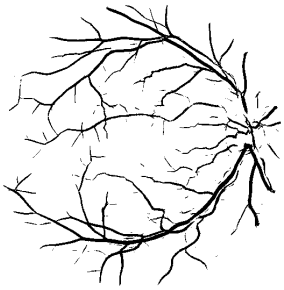
(b) Manual segmentation of (a).



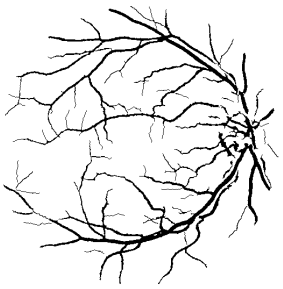
(c) Segmentation of (a) by the algorithm of Chaudhuri et al. [5] (Method 1).



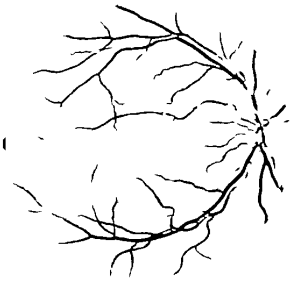
(d) Segmentation of (a) by the algorithm of Chanwimaluang et al. [4] (Method 2).



(e) Segmentation of (a) by the algorithm of Soares et al. [23] (Method 3).



(f) Segmentation of (a) by the algorithm of Mendonça et al. [19] (Method 4).



(g) Segmentation of (a) by the algorithm of Marin et al. [16] (Method 5).

Figure 2: Retina image (eye fundus) and its segmentations (from DRIVE database [24] and STARE database [12], by using the five presented methods of Chaudhuri et al. [5], Chanwimaluang et al. [4]), Soares et al. [23], Mendonça et al. [19] and Marin et al. [16].

**Data** : $Input \leftarrow$ Grayscale image of retina.
**Result** : $S$: Segmentation of the retina image.
**begin**
  Apply the matched filter as in Method 1. The parameters are fixed;
  Local entropy thresholding ;
  Length Filtering (considers small connected components as misclassified pixels);
**end**

**Method 2:** Retina image segmentation from [4].

### 4.3 Soares et al. method (Method 3)

Soares et al. [23] proposed a method applied on the (inverted) green channel and based on wavelet decomposition and pixel classification. The method is summarized in Method 3.

**Data** : $Input \leftarrow$ Inverted green channel of image of retina.
**Result** : $S$: Segmentation of the retina image.
**begin**
  Preprocessing: extending the Region of Interest (ROI);
  Wavelet decomposition (using 2D-Gabor wavelet);
  For each scale, keep the maximum value for all directions;
  Feature normalization (avoid dimensionality problems);
  Supervised classification (with manual segmentation on a training set);
**end**

**Method 3:** Retina image segmentation from [23].

### 4.4 Mendonça et al. method (Method 4)

A fourth segmentation method has been proposed by [19]. The main stages are presented in Method 4.

**Data** : $Input \leftarrow$ image of retina.
**Result** : $S$: Segmentation of the retina image.
**begin**
  Preprocessing: background normalization and thin vessel enhancement;
  Vessel centerline detection phase (Difference of Offset Gaussians filters);
  Vessel segmentation phase (based on mathematical morphology filters);
**end**

**Method 4:** Retina image segmentation from [19].

### 4.5 Marin et al. method (Method 5)

The fifth segmentation method presented in this article has appeared recently ([16]). The method is summarized in Method 5.

**Data** : $Input \leftarrow$ image of retina.
**Result** : $S$: Segmentation of the retina image.
**begin**
  Preprocessing: background homogenization and vessel enhancement;
  Feature extraction;
  Classification by Neural Networks;
  Postprocessing (filling pixel gaps);
**end**

**Method 5:** Retina image segmentation from [16].

## 5 Quantitative comparisons

The DRIVE database [24] is constituted of retina images and their segmentations. It contains two families of retina images (Fig. 2) : the **test** family and the **training** family. They contain images and their manual segmentations either performed by two experts (test set) or by only one (training set). The STARE database [12] also contains 20 images of segmented vessels of retina, with the manual reference given by two experts, denoted *ah* and *vk*.

This section presents the results of the evaluation of the different segmentations on both databases. The first subsection will begin a discussion on the tolerance parameter of the $\epsilon$ dissimilarity criterion. Then, all segmentation methods have been applied to the 40 images of the test set of the DRIVE database and of the STARE database. One manual segmentation is employed as

a reference (ground truth) and the second manual segmentation is presented as a control method. The results will be presented in different tables for the different dissimilarity criteria. The degrees of consistency and discriminancy (see Defs. 5.1 and 5.2 in Sect. 5.3) of the dissimilarity criteria together will be computed and presented to emphasize the good behaviour of the $\epsilon$ dissimilarity criterion.

## 5.1 Choice of the tolerance parameter of $\epsilon$

The $\epsilon$ dissimilarity criterion satisfies the identity and separation properties of a metric, but neither obeys the symmetry nor the triangle inequality axioms. For the symmetry property, it is justified by the fact that the reference segmentation already introduces a dissymmetry in the evaluation. Indeed, the $\epsilon$ dissimilarity criterion can almost be interpreted as a percentage of misdetected pixels. The triangle inequality property is more complicated to interpret in the case of segmentations. It has been deeply discussed in [11].

The monotonicity is the main property. Indeed, the $\epsilon$ dissimilarity criterion is decreasing in regard to the tolerance parameter $\rho$ (Eq. 5).

$$\forall(\rho_1, \rho_2) \in \mathbb{R}_+^2, \rho_1 > \rho_2 \Rightarrow \epsilon_M^{\rho_1}(X) \leq \epsilon_M^{\rho_2}(X) \tag{5}$$

As in every spatial analysis method where a scale factor is employed, the choice of the tolerance parameter is crucial and depends on the application issue. This section proposes a way to choose it. The **test** set of the **DRIVE** database is used to make the choice of the tolerance parameter value (two persons, trained by an expert and thus also considered as experts, have drawn the contours of the same retina image). The $\epsilon$ dissimilarity criterion has been used to compare every manual segmentation to the other. The mean value of the $\epsilon$ dissimilarity criterion is represented in the Fig. 3.

The reader can consider that two experts should always draw the contours at the same location, but within a certain spatial tolerance, depending on the image size and the precision of the drawing tool. The choice of the tolerance parameter then consists in determining the criterion value under which two segmentations are considered as identical.

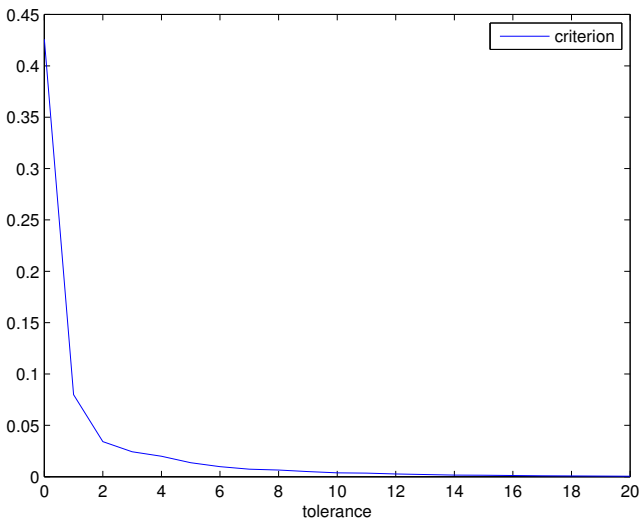



Figure 3: Method for fixing the tolerance parameter value. Two experts have manually segmented the 20 training images of the DRIVE database, and the $\epsilon$ dissimilarity criterion has been applied between the segmentation results. The average value for all the results is shown in the graph. In this example, $\rho$ is in pixels, and there is a strong gap between no tolerance ($\rho = 0$) and a tolerance of one pixel ($\rho = 1$).

For this (DRIVE) database and the considered size of images ($565 \times 583$ pixels), if we admit that an $\epsilon$ dissimilarity value such that $\epsilon < 0.02$ means "identical segmentations", we should then take a tolerance of $\rho = 4$ (see Fig. 3). The same value will be taken for image from the STARE database, as the scale of observation of the vessels is similar to those in the DRIVE database.

## 5.2 Comparison of the five segmentation methods by means of the eleven dissimilarity criteria

The 40 images from the databases have been segmented using the five methods, and their results have been compared to the reference (which is the first manual segmentation). The second manual segmentation is also included and compared to the

reference. As the two manual segmentations were performed in the same conditions, they should give similar results for their evaluations by the considered criteria.

The summary of the comparison is presented in the different subtables of Tables 3 and 2. The absolute values comparison between the criterions are not sufficient to evaluate a segmentation result. However, the relative comparison of the values obtained by the different segmentation methods is meaningful. The Figs. 4 and 5 illustrate in a visual way the results of Tables 3 and 2. The values are normalized for each criterion with the maximal value obtained among the different methods for the 40 images of the DRIVE and STARE database. This graph (Fig. 4) shows that $Sp$ and $Ppv$ criteria do not classify the manual segmentation as the best segmentation (i.e., with the lower dissimilarity value). Moreover, these two criteria do not sort the image segmentation methods in the same order as the others. Fig. 4 also shows that the $\epsilon$ dissimilarity criterion is shown to be more discriminating than the others. This is not true in the Fig. 5. It also highlights that it is very difficult to make a (visual) difference between the methods from Soares and Mendonça. The next subsection 5.3 will quantify these visual facts.

| | mean | std. dev. | median |
|---|---|---|---|
| **Chaudhuri et al** | 0.18 | 0.05 | 0.17 |
| **Chanwimaluang et al** | 0.12 | 0.07 | 0.09 |
| **Marin et al** | 0.08 | 0.03 | 0.08 |
| **Soares et al** | 0.06 | 0.05 | 0.05 |
| **Mendonca et al** | 0.07 | 0.07 | 0.05 |
| **Manual segmentation** | 0.04 | 0.02 | 0.03 |

(a) $\epsilon$ criterion, with $\rho = 4$

| | mean | std. dev. | median |
|---|---|---|---|
| **Chaudhuri et al** | 0.60 | 0.12 | 0.57 |
| **Chanwimaluang et al** | 0.32 | 0.10 | 0.29 |
| **Marin et al** | 0.35 | 0.11 | 0.35 |
| **Soares et al** | 0.31 | 0.13 | 0.28 |
| **Mendonca et al** | 0.32 | 0.15 | 0.28 |
| **Manual segmentation** | 0.25 | 0.10 | 0.22 |

(b) $fom$ criterion

| | mean | std. dev. | median |
|---|---|---|---|
| **Chaudhuri et al** | 0.50 | 0.12 | 0.48 |
| **Chanwimaluang et al** | 0.35 | 0.07 | 0.33 |
| **Marin et al** | 0.29 | 0.07 | 0.28 |
| **Soares et al** | 0.27 | 0.07 | 0.24 |
| **Mendonca et al** | 0.30 | 0.12 | 0.27 |
| **Manual segmentation** | 0.24 | 0.04 | 0.23 |

(c) $dice$ criterion

| | mean | std. dev. | median |
|---|---|---|---|
| **Chaudhuri et al** | 0.62 | 0.12 | 0.60 |
| **Chanwimaluang et al** | 0.35 | 0.09 | 0.35 |
| **Marin et al** | 0.39 | 0.10 | 0.39 |
| **Soares et al** | 0.35 | 0.10 | 0.33 |
| **Mendonca et al** | 0.37 | 0.14 | 0.32 |
| **Manual segmentation** | 0.29 | 0.09 | 0.28 |

(d) $Se$ criterion

| | mean | std. dev. | median |
|---|---|---|---|
| **Chaudhuri et al** | 0.01 | 0.01 | 0.01 |
| **Chanwimaluang et al** | 0.03 | 0.02 | 0.03 |
| **Marin et al** | 0.01 | 0.01 | 0.01 |
| **Soares et al** | 0.01 | 0.01 | 0.01 |
| **Mendonca et al** | 0.01 | 0.01 | 0.01 |
| **Manual segmentation** | 0.01 | 0.01 | 0.01 |

(e) $Sp$ criterion

| | mean | std. dev. | median |
|---|---|---|---|
| **Chaudhuri et al** | 0.18 | 0.13 | 0.14 |
| **Chanwimaluang et al** | 0.31 | 0.13 | 0.27 |
| **Marin et al** | 0.13 | 0.07 | 0.13 |
| **Soares et al** | 0.15 | 0.11 | 0.15 |
| **Mendonca et al** | 0.18 | 0.15 | 0.17 |
| **Manual segmentation** | 0.15 | 0.09 | 0.16 |

(f) $Ppv$ criterion

Table 3: Mean results of the eleven criteria for the five segmentation methods (and the second manual segmentation, which is a segmentation from another expert), over all images of DRIVE and STARE databases. Continues on Table 2 on page 11.

## 5.3 Comparison of the eleven dissimilarity criteria

The quantitative comparison of dissimilarity criteria is not an easy task. This article makes use the definition of consistency and discriminancy [8], more precisely the degree of consistency and the degree of discriminancy, as proposed in [13].

Intuitively, the degree of consistency evaluates the agreement between the different criteria, and the degree of discriminancy evaluates their capacity to distinguish the different methods. We have chosen not to used nonparametric tests (Friedman, Quade... see [9]) because the approach are similar (based on ranking methods together) and it would complicate the task of the reader.

**Definition 5.1** (Degree of Consistency)**.** *For two dissimilarity criteria $f$ and $g$ defined on the domain $\xi$ ($\xi$ is the family of the*

| | mean | std. dev. | median |
|---|---|---|---|
| **Chaudhuri et al** | 0.06 | 0.02 | 0.06 |
| **Chanwimaluang et al** | 0.04 | 0.01 | 0.04 |
| **Marin et al** | 0.04 | 0.02 | 0.03 |
| **Soares et al** | 0.04 | 0.02 | 0.03 |
| **Mendonca et al** | 0.04 | 0.02 | 0.03 |
| **Manual segmentation** | 0.03 | 0.02 | 0.03 |

(g) *Npv* criterion

| | mean | std. dev. | median |
|---|---|---|---|
| **Chaudhuri et al** | 0.07 | 0.02 | 0.07 |
| **Chanwimaluang et al** | 0.07 | 0.02 | 0.06 |
| **Marin et al** | 0.05 | 0.02 | 0.04 |
| **Soares et al** | 0.05 | 0.02 | 0.04 |
| **Mendonca et al** | 0.05 | 0.02 | 0.04 |
| **Manual segmentation** | 0.04 | 0.01 | 0.04 |

(h) *Acc* criterion

| | mean | std. dev. | median |
|---|---|---|---|
| **Chaudhuri et al** | 0.44 | 0.14 | 0.40 |
| **Chanwimaluang et al** | 0.21 | 0.10 | 0.19 |
| **Marin et al** | 0.20 | 0.09 | 0.19 |
| **Soares et al** | 0.18 | 0.10 | 0.15 |
| **Mendonca et al** | 0.19 | 0.13 | 0.13 |
| **Manual segmentation** | 0.13 | 0.08 | 0.12 |

(i) $F_1$ criterion

| | mean | std. dev. | median |
|---|---|---|---|
| **Chaudhuri et al** | 0.66 | 0.10 | 0.65 |
| **Chanwimaluang et al** | 0.46 | 0.06 | 0.47 |
| **Marin et al** | 0.44 | 0.10 | 0.45 |
| **Soares et al** | 0.36 | 0.10 | 0.36 |
| **Mendonca et al** | 0.35 | 0.11 | 0.34 |
| **Manual segmentation** | 0.29 | 0.09 | 0.28 |

(j) $FJ$ criterion

| | mean | std. dev. | median |
|---|---|---|---|
| **Chaudhuri et al** | 32.91 | 21.05 | 26.40 |
| **Chanwimaluang et al** | 11.64 | 6.22 | 9.98 |
| **Marin et al** | 8.78 | 4.77 | 7.43 |
| **Soares et al** | 5.26 | 2.87 | 4.34 |
| **Mendonca et al** | 5.48 | 3.91 | 4.22 |
| **Manual segmentation** | 3.58 | 2.28 | 2.70 |

(k) *qms* criterion

Table 2: Continued from Table 3. Mean results of the eleven criteria for the five segmentation methods (and the second manual segmentation, which is a segmentation from another expert), over all images of DRIVE and STARE databases. Notice that median and standard deviation values do not provide more information than the mean value does. The second manual segmentation is expected to be the best segmentation, i.e. to get the lowest value.

*segmentations on a given image, $X_1$ and $X_2$ are two segmentations belonging to $\xi$), let the sets $R$ and $S$ be defined by:*

$$R = \{(X_1, X_2) | X_1, X_2 \in \xi, \\ f(X_1) > f(X_2) \text{ and } g(X_1) > g(X_2)\}$$

$$S = \{(X_1, X_2) | X_1, X_2 \in \xi, \\ f(X_1) > f(X_2) \text{ and } g(X_1) < g(X_2)\}$$

*The degree of consistency of $f$ and $g$ is $d_C^\circ$ ($0 \leq d_C^\circ \leq 1$), where*

$$d_C^\circ = \frac{\#R}{\#R + \#S}$$

*(# stands for the number of elements of the sets). [13] precises that the degree of consistency $d_C^\circ$ is symmetric, i.e. $d_C^\circ(f, g) = d_C^\circ(g, f)$.*

**Definition 5.2** (Degree of Discriminancy)**.** *For two dissimilarity criteria $f$ and $g$ defined on the domain $\xi$, let*

$$P = \{(X_1, X_2) | X_1, X_2 \in \xi, \\ f(X_1) > f(X_2) \text{ and } g(X_1) = g(X_2)\}$$

$$Q = \{(X_1, X_2) | X_1, X_2 \in \psi, \\ g(X_1) > g(X_2) \text{ and } f(X_1) = f(X_2)\}$$

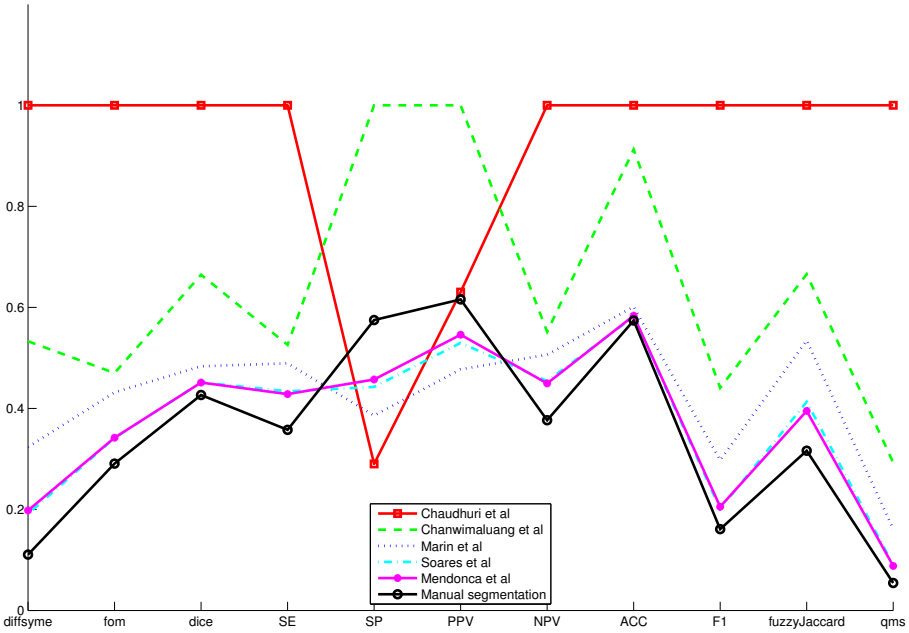



Figure 4: Quantitative comparison of the six image segmentation methods evaluated with the eleven dissimilarity criteria, for the images of the DRIVE database. The results have been normalized (i.e. the value 1 is given to the highest value for each dissimilarity criterion). The mean values (on all images) have been represented. The choice of representing the values with lines has been done because it emphasizes the fact that methods are consistent or not. In this graph, the comparison of numerical values has no sense. Instead, the rankings (see Def. 5.1) of the segmentation methods can be compared for the different criterions, as well as their discriminancy (see Def. 5.2). Continues with images of the STARE database on Fig. 5.

*The degree of discriminancy for $f$ over $g$ is*

$$d_D^\circ = \frac{\#P}{\#Q}$$

*Notice that the degree $d_D^\circ$ is not symmetric.*

It is stated in [13] that a criterion can be considered as better as another one if $d_C^\circ > 0.5$ (consistency) and $d_D^\circ > 1$ (discriminancy).

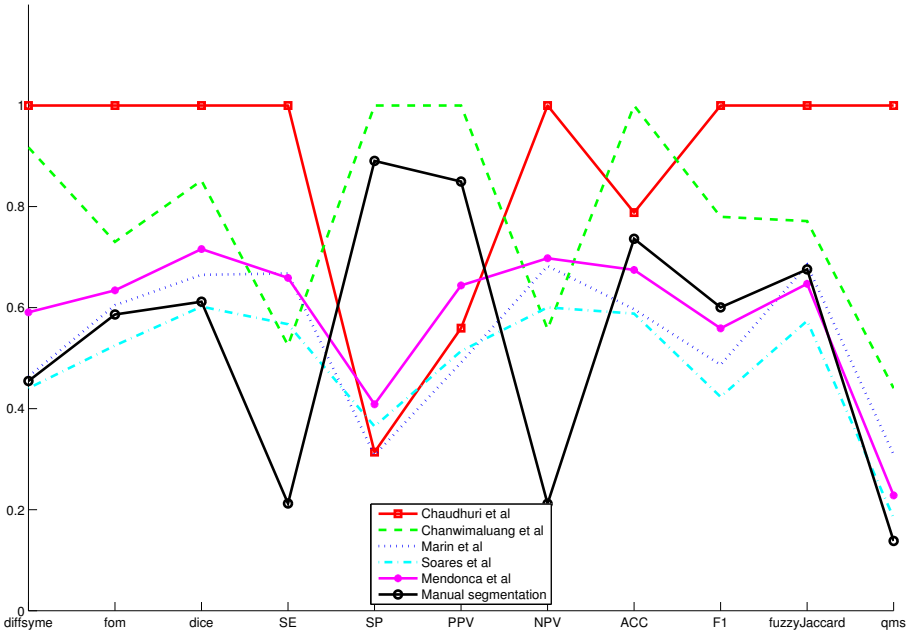
In the present paper, these two degrees are calculated on the 40 segmented images of DRIVE and STARE databases, evaluated for each dissimilarity criterion with the first manual segmentation taken as the ground truth.

The degrees of consistency are presented in Tables 3 and 4, the criteria can be separated into two groups (by taking values above 0.5): the first group is constituted of $\epsilon$, $fom$, $dice$, $Se$, $Npv$ and $Acc$, whereas the second group is composed of $Sp$ and $Ppv$.

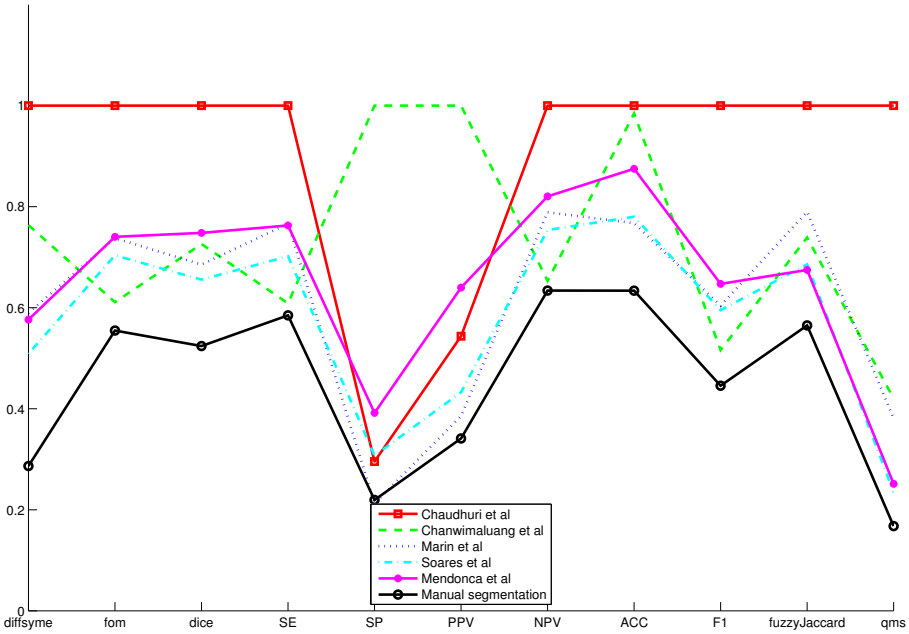
To compute the degree of discriminancy, as the values are float numerical values, the equality condition is subject to a small tolerance. The choice has been made to consider that two values are equal if their difference is lower than 5% (with the normalized values). The degrees of discriminancy are presented in Tables 5 and 6.

The Table 7 is the summary of this analysis. It shows that the $\epsilon$ dissimilarity criterion is a good choice. It has indeed a good consistency with $FJ$ and $qms$. One should notice that $Sp$ is considered as different, because not consistent with the other criteria. As expected, because they try to have a perceptual justification, $FJ$, $qms$, $fom$ and $\epsilon$ give good results. The $Acc$ criterion does not seem to be a good criterion.

As a discussion, the choice of one segmentation as a reference is arbitrary. It has no influence in the DRIVE database, because both experts propose almost similar segmentations, but references for the DRIVE database are really dissimilar ($vk$ segmentations present a lot more details than $ah$ segmentations). Thus, the global results show a different classification of the segmentation methods (see Figs. 4 and 5). The difference between the $FJ$ and the $\epsilon$ dissimilarity are less evident. The computed values of the degree of consistency are presented in Tables 3 and 4.




(a) The comparison is retricted to values of the STARE database, for expert 'ah'.




(b) The comparison is retricted to values of the STARE database, for expert 'vk'.

Figure 5: Continued from Fig. 4. The rankings are really difference, depending on the choosen expert. The manual segmentation is not always given the lowest value, because both references from experts 'ah' and 'vk' are really different (one contains detailed vessels, whereas the other is less detailed).

| | $\epsilon$ | **fom** | **dice** | **Se** | **Sp** | **Ppv** | **Npv** | **Acc** | **F1** | **FJ** | **qms** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\epsilon$ | 1.00 | 0.85 | 0.93 | 0.86 | 0.45 | 0.57 | 0.87 | 0.87 | 0.90 | 0.97 | 0.97 |
| **fom** | 0.85 | 1.00 | 0.86 | 0.86 | 0.34 | 0.46 | 0.86 | 0.79 | 0.92 | 0.86 | 0.84 |
| **dice** | 0.93 | 0.86 | 1.00 | 0.85 | 0.47 | 0.59 | 0.86 | 0.92 | 0.91 | 0.93 | 0.92 |
| **Se** | 0.86 | 0.86 | 0.85 | 1.00 | 0.33 | 0.44 | 0.99 | 0.78 | 0.87 | 0.86 | 0.87 |
| **Sp** | 0.45 | 0.34 | 0.47 | 0.33 | 1.00 | 0.89 | 0.34 | 0.55 | 0.41 | 0.45 | 0.45 |
| **Ppv** | 0.57 | 0.46 | 0.59 | 0.44 | 0.89 | 1.00 | 0.45 | 0.66 | 0.52 | 0.57 | 0.56 |
| **Npv** | 0.87 | 0.86 | 0.86 | 0.99 | 0.34 | 0.45 | 1.00 | 0.79 | 0.88 | 0.87 | 0.88 |
| **Acc** | 0.87 | 0.79 | 0.92 | 0.78 | 0.55 | 0.66 | 0.79 | 1.00 | 0.84 | 0.88 | 0.87 |
| **F1** | 0.90 | 0.92 | 0.91 | 0.87 | 0.41 | 0.52 | 0.88 | 0.84 | 1.00 | 0.91 | 0.90 |
| **FJ** | 0.97 | 0.86 | 0.93 | 0.86 | 0.45 | 0.57 | 0.87 | 0.88 | 0.91 | 1.00 | 0.98 |
| **qms** | 0.97 | 0.84 | 0.92 | 0.87 | 0.45 | 0.56 | 0.88 | 0.87 | 0.90 | 0.98 | 1.00 |

Table 3: Degree of consistency $d_C^\circ$ between the different dissimilarity criteria, for the DRIVE database. If $d_C^\circ > 0.5$, the two dissimilarity criteria are consistent. This evaluation is symmetric.

| | $\epsilon$ | **fom** | **dice** | **Se** | **Sp** | **Ppv** | **Npv** | **Acc** | **F1** | **FJ** | **qms** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\epsilon$ | 1.00 | 0.82 | 0.87 | 0.61 | 0.64 | 0.68 | 0.62 | 0.87 | 0.79 | 0.90 | 0.87 |
| **fom** | 0.82 | 1.00 | 0.84 | 0.68 | 0.55 | 0.59 | 0.68 | 0.76 | 0.86 | 0.85 | 0.81 |
| **dice** | 0.87 | 0.84 | 1.00 | 0.68 | 0.59 | 0.64 | 0.69 | 0.85 | 0.83 | 0.85 | 0.88 |
| **Se** | 0.61 | 0.68 | 0.68 | 1.00 | 0.27 | 0.32 | 0.99 | 0.53 | 0.58 | 0.63 | 0.73 |
| **Sp** | 0.64 | 0.55 | 0.59 | 0.27 | 1.00 | 0.95 | 0.28 | 0.74 | 0.65 | 0.63 | 0.54 |
| **Ppv** | 0.68 | 0.59 | 0.64 | 0.32 | 0.95 | 1.00 | 0.33 | 0.79 | 0.67 | 0.66 | 0.58 |
| **Npv** | 0.62 | 0.68 | 0.69 | 0.99 | 0.28 | 0.33 | 1.00 | 0.54 | 0.58 | 0.64 | 0.74 |
| **Acc** | 0.87 | 0.76 | 0.85 | 0.53 | 0.74 | 0.79 | 0.54 | 1.00 | 0.82 | 0.84 | 0.79 |
| **F1** | 0.79 | 0.86 | 0.83 | 0.58 | 0.65 | 0.67 | 0.58 | 0.82 | 1.00 | 0.82 | 0.76 |
| **FJ** | 0.90 | 0.85 | 0.85 | 0.63 | 0.63 | 0.66 | 0.64 | 0.84 | 0.82 | 1.00 | 0.87 |
| **qms** | 0.87 | 0.81 | 0.88 | 0.73 | 0.54 | 0.58 | 0.74 | 0.79 | 0.76 | 0.87 | 1.00 |

(a) Degree of consistency $d_C^\circ$ for the STARE database, with expert 'ah'.

| | $\epsilon$ | **fom** | **dice** | **Se** | **Sp** | **Ppv** | **Npv** | **Acc** | **F1** | **FJ** | **qms** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\epsilon$ | 1.00 | 0.77 | 0.91 | 0.76 | 0.61 | 0.68 | 0.78 | 0.87 | 0.76 | 0.92 | 0.93 |
| **fom** | 0.77 | 1.00 | 0.83 | 0.91 | 0.39 | 0.46 | 0.91 | 0.73 | 0.95 | 0.79 | 0.77 |
| **dice** | 0.91 | 0.83 | 1.00 | 0.83 | 0.55 | 0.62 | 0.85 | 0.89 | 0.82 | 0.88 | 0.87 |
| **Se** | 0.76 | 0.91 | 0.83 | 1.00 | 0.39 | 0.45 | 0.98 | 0.73 | 0.87 | 0.79 | 0.78 |
| **Sp** | 0.61 | 0.39 | 0.55 | 0.39 | 1.00 | 0.93 | 0.41 | 0.66 | 0.40 | 0.55 | 0.58 |
| **Ppv** | 0.68 | 0.46 | 0.62 | 0.45 | 0.93 | 1.00 | 0.47 | 0.73 | 0.46 | 0.62 | 0.64 |
| **Npv** | 0.78 | 0.91 | 0.85 | 0.98 | 0.41 | 0.47 | 1.00 | 0.75 | 0.86 | 0.81 | 0.79 |
| **Acc** | 0.87 | 0.73 | 0.89 | 0.73 | 0.66 | 0.73 | 0.75 | 1.00 | 0.72 | 0.80 | 0.83 |
| **F1** | 0.76 | 0.95 | 0.82 | 0.87 | 0.40 | 0.46 | 0.86 | 0.72 | 1.00 | 0.77 | 0.74 |
| **FJ** | 0.92 | 0.79 | 0.88 | 0.79 | 0.55 | 0.62 | 0.81 | 0.80 | 0.77 | 1.00 | 0.92 |
| **qms** | 0.93 | 0.77 | 0.87 | 0.78 | 0.58 | 0.64 | 0.79 | 0.83 | 0.74 | 0.92 | 1.00 |

(b) Degree of consistency $d_C^\circ$ for the STARE database, with expert 'vk'.

Table 4: Degree of consistency $d_C^\circ$ between the different dissimilarity criteria, for the STARE database. If $d_C^\circ > 0.5$, the two dissimilarity criteria are consistent. This evaluation is symmetric.

| | $\epsilon$ | **fom** | **dice** | **Se** | **Sp** | **Ppv** | **Npv** | **Acc** | **F1** | **FJ** | **qms** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\epsilon$ | NaN | 3.00 | 19.00 | 4.67 | 7.33 | 3.06 | 5.00 | 14.60 | 3.80 | 0.07 | Inf |
| **fom** | 0.33 | NaN | 2.26 | 2.08 | 2.92 | 1.16 | 2.38 | 2.65 | 0.90 | 0.17 | 2.50 |
| **dice** | 0.05 | 0.44 | NaN | 0.71 | 1.81 | 0.51 | 0.81 | 2.40 | 0.37 | 0.03 | 1.20 |
| **Se** | 0.21 | 0.48 | 1.40 | NaN | 2.07 | 0.76 | 5.00 | 1.83 | 0.50 | 0.10 | 1.50 |
| **Sp** | 0.14 | 0.34 | 0.55 | 0.48 | NaN | 0.23 | 0.53 | 0.84 | 0.32 | 0.06 | 0.64 |
| **Ppv** | 0.33 | 0.86 | 1.95 | 1.31 | 4.42 | NaN | 1.46 | 2.57 | 0.81 | 0.17 | 1.92 |
| **Npv** | 0.20 | 0.42 | 1.23 | 0.20 | 1.90 | 0.68 | NaN | 1.67 | 0.43 | 0.09 | 1.35 |
| **Acc** | 0.07 | 0.38 | 0.42 | 0.55 | 1.19 | 0.39 | 0.60 | NaN | 0.31 | 0.05 | 0.68 |
| **F1** | 0.26 | 1.11 | 2.73 | 2.00 | 3.09 | 1.24 | 2.33 | 3.22 | NaN | 0.13 | 3.23 |
| **FJ** | 14.00 | 5.78 | 34.50 | 10.50 | 15.83 | 5.80 | 11.17 | 21.25 | 7.83 | NaN | Inf |
| **qms** | 0.00 | 0.40 | 0.83 | 0.67 | 1.56 | 0.52 | 0.74 | 1.48 | 0.31 | 0.00 | NaN |

Table 5: Degree of discriminancy $d_D^\circ$ between the different dissimilarity criteria, for the DRIVE database. If $d_D^\circ > 1$, the criterion presented in column is more discriminant than the criterion presented in row (NaN stands for Not a Number. Inf stands for infinity; it is obtained when $Q = 0$, which means that there is no value where $g(X_1) > g(X_2)$ and $f(X_1) = f(X_2)$), for $f$ and $g$ two criterions, $X_1$ and $X_2$ two segmentation results.

| | $\epsilon$ | **fom** | **dice** | **Se** | **Sp** | **Ppv** | **Npv** | **Acc** | **F1** | **FJ** | **qms** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\epsilon$ | NaN | 0.63 | 0.54 | 0.13 | 4.95 | 0.40 | 8.60 | 15.86 | 0.50 | 0.26 | 0.00 |
| **fom** | 1.58 | NaN | 0.90 | 0.25 | 8.23 | 0.67 | 18.86 | 15.38 | 0.76 | 0.43 | 0.00 |
| **dice** | 1.87 | 1.11 | NaN | 0.28 | 8.38 | 0.74 | 16.88 | 24.40 | 0.83 | 0.44 | 0.00 |
| **Se** | 7.50 | 4.00 | 3.60 | NaN | 55.50 | 2.33 | Inf | 66.00 | 2.67 | 1.50 | 0.00 |
| **Sp** | 0.20 | 0.12 | 0.12 | 0.02 | NaN | 0.00 | 1.76 | 1.58 | 0.05 | 0.05 | 0.00 |
| **Ppv** | 2.50 | 1.50 | 1.36 | 0.43 | Inf | NaN | 17.50 | 41.67 | 1.17 | 0.64 | 0.00 |
| **Npv** | 0.12 | 0.05 | 0.06 | 0.00 | 0.57 | 0.06 | NaN | 0.85 | 0.04 | 0.02 | 0.00 |
| **Acc** | 0.06 | 0.07 | 0.04 | 0.02 | 0.63 | 0.02 | 1.17 | NaN | 0.01 | 0.02 | 0.00 |
| **F1** | 2.00 | 1.31 | 1.20 | 0.38 | 20.80 | 0.86 | 22.67 | 121.00 | NaN | 0.53 | 0.00 |
| **FJ** | 3.88 | 2.33 | 2.25 | 0.67 | 22.20 | 1.56 | 46.67 | 43.33 | 1.88 | NaN | 0.00 |
| **qms** | Inf | Inf | Inf | Inf | Inf | Inf | Inf | Inf | Inf | Inf | NaN |

(a) Degree of discriminancy for the STARE database, with expert 'ah'.

| | $\epsilon$ | **fom** | **dice** | **Se** | **Sp** | **Ppv** | **Npv** | **Acc** | **F1** | **FJ** | **qms** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\epsilon$ | NaN | 0.65 | 1.07 | 0.84 | 2.28 | 1.60 | 1.32 | 2.20 | 0.89 | 0.40 | 19.25 |
| **fom** | 1.55 | NaN | 2.05 | 2.11 | 3.96 | 2.54 | 7.20 | 3.21 | 3.17 | 0.93 | 6.69 |
| **dice** | 0.94 | 0.49 | NaN | 0.74 | 2.09 | 1.50 | 1.41 | 2.22 | 0.81 | 0.56 | 6.46 |
| **Se** | 1.20 | 0.47 | 1.36 | NaN | 2.66 | 1.85 | Inf | 2.14 | 1.14 | 0.72 | 6.06 |
| **Sp** | 0.44 | 0.25 | 0.48 | 0.38 | NaN | 0.19 | 0.58 | 0.66 | 0.38 | 0.28 | 1.41 |
| **Ppv** | 0.63 | 0.39 | 0.67 | 0.54 | 5.17 | NaN | 0.83 | 0.98 | 0.56 | 0.39 | 2.26 |
| **Npv** | 0.75 | 0.14 | 0.71 | 0.00 | 1.71 | 1.21 | NaN | 1.27 | 0.47 | 0.45 | 3.50 |
| **Acc** | 0.45 | 0.31 | 0.45 | 0.47 | 1.51 | 1.02 | 0.78 | NaN | 0.50 | 0.31 | 3.13 |
| **F1** | 1.12 | 0.32 | 1.23 | 0.88 | 2.67 | 1.79 | 2.12 | 2.00 | NaN | 0.67 | 5.59 |
| **FJ** | 2.50 | 1.08 | 1.79 | 1.38 | 3.54 | 2.59 | 2.21 | 3.25 | 1.48 | NaN | 19.80 |
| **qms** | 0.05 | 0.15 | 0.15 | 0.16 | 0.71 | 0.44 | 0.29 | 0.32 | 0.18 | 0.05 | NaN |

(b) Degree of discriminancy for the STARE database, with expert 'vk'.

Table 6: Degree of discriminancy $d_D^\circ$ between the different dissimilarity criteria, for the STARE database, using two different experts as the reference.

| | $\epsilon$ | **fom** | **dice** | **Se** | **Sp** | **Ppv** | **Npv** | **Acc** | **F1** | **FJ** | **qms** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\epsilon$ | | | * | * | * | * | * | * | * | | * |
| **fom** | * | | * | * | | | * | * | * | | * |
| **dice** | | | | | * | * | * | * | | | * |
| **Se** | | | * | | | | * | * | | | * |
| **Sp** | | | | | | | | | | | * |
| **Ppv** | | | | | * | | | * | | | * |
| **Npv** | | | | | | | | * | | | * |
| **Acc** | | | | | * | | | | | | * |
| **F1** | | | * | | | | * | * | | | * |
| **FJ** | * | * | * | * | * | * | * | * | * | | * |
| **qms** | | | | | | | | | | | |

(a) Results on images of both databases.

| | $\epsilon$ | **fom** | **dice** | **Se** | **Sp** | **Ppv** | **Npv** | **Acc** | **F1** | **FJ** | **qms** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\epsilon$ | | * | * | * | | * | * | * | * | | * |
| **fom** | | | * | * | | | * | * | | | * |
| **dice** | | | | | | | | * | | | * |
| **Se** | | | * | | | | * | * | | | * |
| **Sp** | | | | | | | | | | | |
| **Ppv** | | | * | | * | | | * | | | * |
| **Npv** | | | * | | | | | * | | | * |
| **Acc** | | | | | * | | | | | | |
| **F1** | | * | * | * | | * | * | * | | | * |
| **FJ** | * | * | * | * | | * | * | * | * | | * |
| **qms** | | | | | | | | * | | | |

(b) Results on images of the DRIVE database only.

Table 7: As stated in [13], the criterion $f$ is statistically consistent and more discriminating than $g$ if and only if $d_C^\circ > 0.5$ and $d_D^\circ > 1$. In this case, we assert that $f$ is a better dissimilarity criterion than $g$, which is represented with a star/green cell.

# 6 Concluding discussion and perspectives

## 6.1 Concluding discussion

The purpose of this article was to address the quantitative evaluation of image segmentation methods by means of comparative criteria (first-order comparative problem) and the quantitative comparison of the criteria themselves (second-order comparative problem). The application issue of retina vessels' image segmentation has been addressed with five specific segmentation methods and eleven dissimilarity criteria reported in the specialized literature.

This articles emphasizes the problem using expert's manual reference, in the case of the STARE database the two experts give really different segmentations. To enhance this study, it would be interesting to have a relative rank of all the methods given by one or several experts, so that the consistency and discriminancy values could reflect the experts analysis.

It has been shown that the $\epsilon$ dissimilarity criterion is consistent with other perceptually based criterions. The practical results on the DRIVE and STARE databases of retina images highlight its properties, particularly its tolerance to small variations in the binary images, and its robustness to various small perturbations in the initial images. The $\epsilon$ dissimilarity criterion ranks the manual segmentation with the lowest value, which means that it is in agreement with the visual perception.

Moreover, a pathway is now open to select the more appropriate image segmentation method for a particular application issue (e.g., the retina vessels extraction in the present article), in addition to choose the best parameter values for optimally tuning the algorithms of the segmentation methods.

## 6.2 Perspectives

In future studies, the $\epsilon$ dissimilarity criterion will be used to compare the results of some segmentation methods dedicated to corneal endothelium images, as well as the tuning of the parameter values.

# References


[1] Abdou, I., Pratt, W.: Qualitative design and evaluation of enhancement/thresholding edge detector. Proc. IEEE. **67**(5), 753–763 (1979)

[2] Cardoso, J., Corte-Real, L.: Toward a generic evaluation of image segmentation. Image Processing, IEEE Transactions on **14**(11), 1773–1782 (2005). DOI 10.1109/TIP.2005.854491

[3] Chalana, V., Kim, Y.: A methodology for evaluation of boundary detection algorithms on medical images. IEEE Trans. Med. Imaging **16**(5), 642–652 (1997)

[4] Chanwimaluang, T., Fan, G., Fransen, S.R.: Hybrid retinal image registration. IEEE Transactions on Information Technology in Biomedicine **10**(1), 129–142 (2006)

[5] Chaudhuri, S., Chatterjee, S., Katz, N., Nelson, M., Goldbaum, M.: Detection of blood vessels in retinal images using two-dimensional matched filters. Medical Imaging, IEEE Transactions on **8**(3), 263 –269 (1989). DOI 10.1109/42.34715

[6] Deza, M.M., Deza, E.: Dictionary of distances. Elsevier (2006)

[7] Dice, L.R.: Measures of the Amount of Ecologic Association Between Species. Ecology **26**(3), 297–302 (1945)

[8] Fix, E., Hodges, J.: Discriminatory analysis. nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique **57**(3), 238–247 (1989). URL http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA800276

[9] García, S., Fernández, A., Luengo, J., Herrera, F.: Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences **180**(10), 2044–2064 (2010)

[10] Gavet, Y.: Perception visuelle humaine, complétion des mosaïques et application à la reconstruction d'images de l'endothélium cornéen humain en microscopie optique spéculaire. Ph.D. thesis, École Nationale Supérieure des Mines de Saint-Etienne (2008)

[11] Gavet, Y., Pinoli, J.C.: A geometric dissimilarity criterion between jordan spatial mosaics. Theoretical aspects and application to segmentation evaluation. Journal of Mathematical Imaging and Vision **42**, 25–49 (2012). URL http://dx.doi.org/10.1007/s10851-011-0272-4. 10.1007/s10851-011-0272-4

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

[13] Huang, J., Ling, C.: Using auc and accuracy in evaluating learning algorithms. Knowledge and Data Engineering, IEEE Transactions on **17**(3), 299–310 (2005)

[14] Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin de la Société Vaudoise des Sciences Naturelles **37**, 547–579 (1901)

[15] Jiang, X., Marti, C., Irniger, C., Bunke, H.: Distance measures for image segmentation evaluation. EURASIP J. Appl. Signal Process. **2006**, 1–10 (2006). DOI http://dx.doi.org/10.1155/ASP/2006/35909

[16] Marin, D., Aquino, A., Gegundez-Arias, M., Bravo, J.: 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), 146–158 (2011)

[17] Marr, D.: Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Henry Holt & Company (1983)

[18] McGuinness, K., O'Connor, N.E.: A comparative evaluation of interactive segmentation algorithms. Pattern Recognition **43**(2), 434 – 444 (2010). DOI DOI:10.1016/j.patcog.2009.03.008. URL http://www.sciencedirect.com/science/article/B6V14-4VTVPT9-1/2/863e9be0e8f651f41146ef73f2898e0c. Interactive Imaging and Vision

[19] Mendonça, A., Campilho, A.: Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction. Medical Imaging, IEEE Transactions on **25**(9), 1200 –1213 (2006). DOI 10.1109/TMI.2006.879955

[20] Minkowski, H.: Volumen und Oberfläche. Mathematische Annalen **57**, 447–495 (1903)

[21] Philipp-Foliguet, S., Guigues, L.: Évaluation de la segmentation d'images: état de l'art, nouveaux indices et comparaison. Traitement du signal **23**(2), 109–124 (2006)

[22] Pont-Tuset, J., Marques, F.: Measures and meta-measures for the supervised evaluation of image segmentation. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pp. 2131–2138. IEEE (2013)

[23] Soares, J., Leandro, J., Cesar, R., Jelinek, H., Cree, M.: Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. Medical Imaging, IEEE Transactions on **25**(9), 1214 –1222 (2006). DOI 10.1109/TMI.2006.879967

[24] Staal, J., Abramoff, M., Niemeijer, M., Viergever, M., van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. Medical Imaging, IEEE Transactions on **23**(4), 501–509 (2004). DOI 10.1109/TMI.2004.825627

[25] Strasters, K.C., Gerbrands, J.J.: Three-dimensional image segmentation using a split, merge and group approach. Pattern Recognition Letters **12**(5), 307–325 (1991)

[26] Tversky, A.: Features of similarity. Psychological Review **84**(4), 327–352 (1977)

[27] Tversky, A., Gati, I.: Similarity, separability and the triangle inequality. Psychological Review **89**, 123–154 (1982)

[28] Unnikrishnan, R., Pantofaru, C., Hebert, M.: Toward objective evaluation of image segmentation algorithms. Pattern Analysis and Machine Intelligence, IEEE Transactions on **29**(6), 929–944 (2007)

[29] Villegas, P., Marichal, X.: Perceptually-weighted evaluation criteria for segmentation masks in video sequences. Image Processing, IEEE Transactions on **13**(8), 1092–1103 (2004). DOI 10.1109/TIP.2004.828433

[30] Zhang, Y.J.: A survey on evaluation methods for image segmentation. Pattern Recognition **29**(8), 1335–1346 (1996)