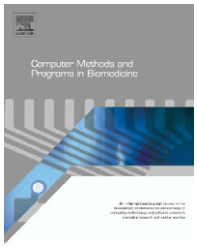
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 96 (2009) 108–124

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



A multidimensional segmentation evaluation for medical image data

Rubén Cárdenes[a,*], *Rodrigo de Luis-García*[a], *Meritxell Bach-Cuadra*[b]

[a] *Laboratory of Image Processing, University of Valladolid, 47011 Valladolid, Spain*
[b] *Signal Processing Laboratory (LTS5), Ecole Polytechnique Fédérale de Lausanne, EPFL, Switzerland*

ARTICLE INFO

Article history:
Received 21 August 2008
Received in revised form
13 April 2009
Accepted 15 April 2009

Keywords:
Segmentation evaluation
Principal Component Analysis
Multidimensional visualization
Image segmentation
MRI segmentation
Similarity measure

ABSTRACT

Evaluation of segmentation methods is a crucial aspect in image processing, especially in the medical imaging field, where small differences between segmented regions in the anatomy can be of paramount importance. Usually, segmentation evaluation is based on a measure that depends on the number of segmented voxels inside and outside of some reference regions that are called gold standards. Although some other measures have been also used, in this work we propose a set of new similarity measures, based on different features, such as the location and intensity values of the misclassified voxels, and the connectivity and the boundaries of the segmented data. Using the multidimensional information provided by these measures, we propose a new evaluation method whose results are visualized applying a Principal Component Analysis of the data, obtaining a simplified graphical method to compare different segmentation results. We have carried out an intensive study using several classic segmentation methods applied to a set of MRI simulated data of the brain with several noise and RF inhomogeneity levels, and also to real data, showing that the new measures proposed here and the results that we have obtained from the multidimensional evaluation, improve the robustness of the evaluation and provides better understanding about the difference between segmentation methods.

1. Introduction

Image segmentation is the task that consists in dividing an image into homogeneous regions for a set of relevant properties such as color, intensity, texture, etc. By homogeneous it is meant that all elements inside these regions share similar properties. The human brain carries out segmentation of common data with ease—a natural ability largely exploited in the medical environment. For instance, when radiologists look at magnetic resonance (MR) or computer tomography (CT) images to evaluate lesions or tumors, they first segment the suspicious growth mentally and then use their training and experience to assess its properties. However, several factors such as the ever-increasing amount of clinical data generated in clinical environments and their increasingly complex nature now make it necessary to rely more and more on computerized segmentation techniques. Their goal is to either assign a label to every voxel or estimate the relative amount of various classes within a voxel. Several technical factors make this goal hard if not impossible to achieve with the current technology. The first important factor is the quality of the collected data. The data acquisition process always introduces some noise as a result of signal attenuation, instrumentation noise, scattering, patient movement (for instance breathing) and many others. A related factor is the limited resolution of the data, which makes every voxel liable to correspond to more

* *Corresponding author*. Tel.: +34 983 423660.
E-mail addresses: ruben@lpi.tel.uva.es (R. Cárdenes), rodlui@tel.uva.es (R. de Luis-García), meritxell.bach@epfl.ch (M. Bach-Cuadra).
0169-2607/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2009.04.009

than one class in the patient anatomy. Such situation is known as partial volume effect. Another subtle factor, also related to the acquisition process, is the choice of imaging modality and its parameters. Poor choice will lead to poor contrast, poor resolution and generally speaking will prevent the full use of the anatomical knowledge that one may have.

A wide range of methods have been proposed to deal with these effects. Consequently the literature dedicated to the problem of medical data segmentation is large and varied. A thorough review of all these methods is beyond the scope of this paper. For interested readers, general reviews on image segmentation can be found in [1–3]. Other domain-specific studies on image segmentation and analysis are [4–6].

A crucial aspect of segmentation techniques (be they for medical data or not) is their reliance on contextual information for them to be effective. An important source of contextual information for medical data is the medical knowledge collected on the problem. Turning this medical knowledge into a set of criteria adapted to computer vision is one of the most difficult aspects of the development of computerized segmentation routines. It follows that segmentation techniques are best suited to specific applications and classes of data, for example to a type of data where an underlying assumption is true. No segmentation technique is better than the others for any purpose. Thus, for a particular problem we have to figure out what available method fits best into our needs in terms of a given criteria or a combination of them, as for instance accuracy, speed [7,8], reproducibility or user interaction [9].

This paper is an extended version of our previous work [10], and at the same time this is a continuation of that work: on the one hand we improve the results using a Principal Component Analysis (PCA) for dimensionality reduction in order to better discriminate graphically between the segmentation methods, and on the other hand, we validate our evaluation method using more segmentation methods, and a set of simulated data sets, a real data set, and a synthetic phantom. Here we first motivate the use of the new similarity measures[1] and we introduce a way to combine them for segmentation evaluation, in terms of accuracy, using a known ground truth. These new similarity measures are based on the location and the intensity values of the misclassified voxels and also based on the connectivity and the boundaries of the segmented data. We show how the combination of these measures can improve the quality of the evaluation. Furthermore, we show that using a single measure to evaluate segmentation is not enough, and therefore we propose to use a new global multidimensional measure between the segmented image and the gold standard. The study that we show here is carried out using several classic segmentation methods applied to a simulated MRI data set of the brain, and also applied to real data. We will show that our new measures and their combination improve the robustness of the evaluation and provides better understanding about the difference between segmentation methods. Moreover, we present in this paper a robust way to visualize all the proposed quality measures using PCA.

This paper is organized as follows. In the next section we will describe the state of the art on evaluation methods and the similarity measures for image segmentation proposed before in the literature. Then, in Section 3 we will provide an overview of the segmentation methods that we are going to employ in our evaluation method, describing the data sets used and the segmentation results. In Section 4 we will describe the evaluation method, describing the new similarity measures and a aggregated multidimensional similarity measure. Then, in Section 5 we will present some results to validate the evaluation methodology using simulated data sets and a real data set, as well as with a synthetic phantom, including the PCA for dimensionality reduction in the visualization of our results. Finally, the conclusions and future work will be presented in Section 6.

2. Background

2.1. State of the art on evaluation methods

Many works to evaluate segmentation methods have been reported in the last two decades. A good survey about segmentation evaluation can be found in [11]. This author distinguishes the evaluation methods between empirical (based on the study of the results) and analytical (based only on intrinsic features of the methods). The empirical methods are divided into discrepancy and goodness methods, where the former compare the results with a reference or ground truth, and the latter are based on the study of the results themselves. Among the discrepancy methods, there exist several features reported to measure the quality of the segmentation: number of misclassified voxels, position of misclassified voxels, number of objects in the image, feature values of segmented objects and other miscellaneous quantities.

Most of the methods in the literature for segmentation evaluation are based on classic discrepancy methods, limited to the computation of the number of voxels of the segmented classes in the results and in a gold standard. Other authors have introduced the location of the misclassified voxels as a feature to measure the discrepancy between segmented images, for example, Yasnoff et al. [12], Straters and Gerbrands [13] and later Pichon et al. [14] proposed to use an error distance from the misclassified voxels to the gold standard. Huttenlocher et al. [15] used the partial Hausdorff distance between set of voxels, and also [16] proposed an overlap distance using fuzzy set theory to take into account fractional labels coming from multiple test images. Other work proposed by Cardoso and Corte-Real [17] presented a general distance between segmentation partitions to measure the quality of a given segmentation.

One interesting work about segmentation evaluation is the one published by Udupa et al. [18]. In that work, the authors proposed a methodology that takes into account more aspects than just the accuracy of the segmentations. In addition, they present measures of the precision (reproducibility) and the efficiency (time taken), to finally conclude that the combina-

[1] The term *similarity measure* is more often used to guide image registration or segmentation, but this term is used instead of error measure because the measure values used here increase as the similarity between images increase.

tion of these factors are essential in the assessment of the performance of any segmentation method.

Some other methods have been proposed to perform segmentation evaluation without a ground truth, see for instance [19–21].

2.2. Similarity measures

This section describes the similarity measures used commonly in the literature for segmentation evaluation, namely the confusion tables, other classic measures based on voxel overlap, and some distance based similarity measures.

2.2.1. Classic similarity measures

Segmentation is often evaluated using similarity measures based on region overlap. One of the most common ways to show segmentation evaluation results are the confusion tables. Every value in a confusion table represents the number of overlapping voxels between two classes, divided by the number of voxels of the class in the gold standard

$$M_{ij} = \frac{|X_i \cap Y_j|}{|Y_j|}, \tag{1}$$

where the sub-indexes represent the classes. Other classic measures commonly used are the Jaccard (*JC*), Dice Similarity (*DS*), Tanimoto (*TN*), and Volume Similarity (*VS*) coefficients. All of them take values between 0 and 1. If X is the set of voxels segmented as class C in one volume, Y is the set of voxels of the same class in the other volume, a is the number of voxels in their intersection, b is the number of voxels in X not belonging to Y, c is the number of voxels of Y not belonging to X, and d is the number of voxels outside X and Y, we can define these measures with the following expressions,

$$JC := \frac{|X \cap Y|}{|X \cup Y|} = \frac{a}{a+b+c} \tag{2}$$

$$DS := \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2a}{2a+b+c} \tag{3}$$

These two coefficients are equal to one if X and Y are the same region, and zero if they are disjoint regions. In fact, they are related by $DS = 2JC/(JC+1)$, so they give equivalent values.

$$TN := \frac{|X \cap Y| + |\overline{X \cup Y}|}{|X \cup Y| + |\overline{X \cap Y}|} = \frac{a+d}{a+2b+2c+d} \tag{4}$$

TN is one if X is equal to Y, and zero if they are disjoint regions and they occupy all the image.

$$VS := 1 - \frac{||X| - |Y||}{|X| + |Y|} = 1 - \frac{|b-c|}{2a+b+c} \tag{5}$$

VS is one if the number of elements of X is equal to the number of elements of Y, and zero if one of them is empty.

2.2.2. Distance based similarity measures

The similarity measures described above are based only on the number of voxels of the classes in the segmented image and in the gold standard, their union, and their intersection. As Pichon et al. [14], and also Crum et al. [16] recently proposed, it is important to use the distances from the misclassified voxels to the ground truth in order to improve the similarity measures. Given two point sets X and Y, it is possible to define the distances from the points in one set to the other, as proposed in [14]:

$$d(r) := \begin{cases} 0, & r \in X \cap Y, \\ \min_{x \in X} ||x - r||, & r \in Y \setminus X, \\ \min_{y \in Y} ||y - r||, & r \in X \setminus Y. \end{cases} \tag{6}$$

There exist measures using this kind of distance definitions between point sets, such as the Yasnoff discrepancy measure [12], defined as

$$YM := \frac{1}{N} \sum_{i=1}^{N} d(r_i)^2 \tag{7}$$

and the Factor of Merit proposed by Straters [13]

$$FOM := \frac{1}{N} \sum_{i=1}^{N} \frac{1}{1 + d(r_i)^2} \tag{8}$$

where r_i are the misclassified voxels, N is the number of misclassified voxels, and $d(r_i)$ is the distance from r_i to the ground truth. Another popular distance between two point sets X and Y, is the Hausdorff distance, defined as

$$H(X,Y) := \max\{\max_{x \in X} \min_{y \in Y} ||x - y||, \max_{y \in Y} \min_{x \in X} ||x - y||\} \tag{9}$$

which is the maximum distance one set has to move its boundaries so that it would enclose the other set.

Some of these measures have been extended to surface to surface error measures as reported in the literature in [22], and for instance in [23] where the Hausdorff distance between surfaces is used, and also this kind of surface measures is especially useful to drive or evaluate deformable model algorithms, such is the case in [24].

3. Design considerations

In order to illustrate this new multidimensional evaluation methodology four segmentation techniques for brain tissue segmentation are assessed. We choose this application since brain tissue segmentation is a key issue in many applications of medical image analysis for quantitative studies [25–31], particularly in the study of many brain disorders, or as a preliminary step of image processing algorithms such as image registration. However, there is a problem inherent to this evaluation, because it is quite difficult to obtain a reliable reference segmentation data set. The most used approach is to use a manual segmentation, or a combination of several manual segmentations, from several experts if possible. There is though, the possibility to validate brain tissue segmentation methods on a brain *simulated* data set as the one proposed by the *Brain Web* MR simulator [32]. Their data is very well suited for this purpose since a ground-truth classification is known

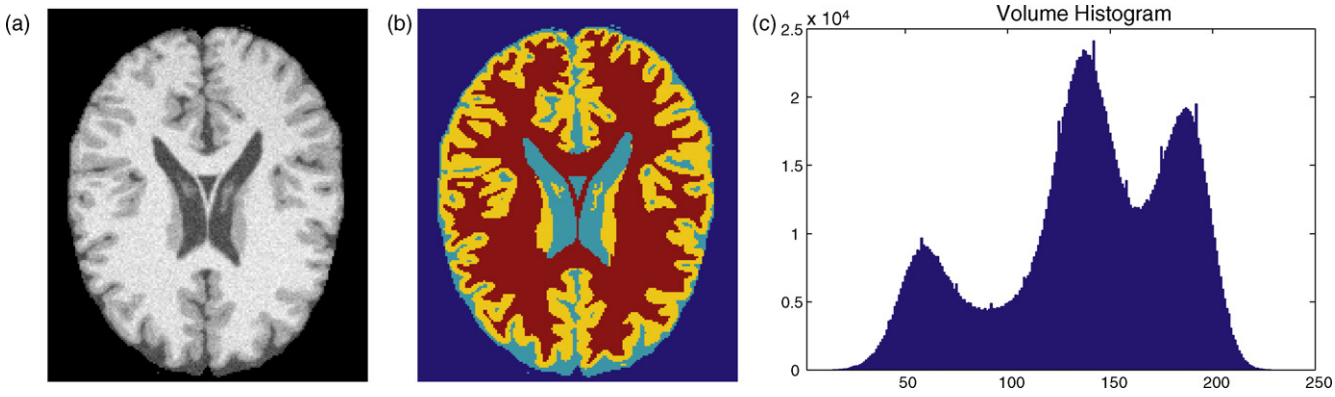


Fig. 1 – Axial slice of the brainweb simulated MRI (a), gold standard: in red is WM, in light blue is CSF, in yellow is GM, and in dark blue is the background (b) and volume histogram (c). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

while different types of MR modalities and image resolution and artifacts can be reproduced.

3.1. *Image segmentation techniques*

A large number of approaches have been proposed to deal with the MR brain image classification problem. In this work, we have selected four well known and widely used segmentation techniques[2]:

- A K-Means clustering method [33].
- A Gaussian mixture model using the Expectation Maximization (EM) algorithm [34].
- A statistical classification using Gaussian Hidden Markov Random Field Model (GHMRF) [35].
- A supervised method based on the *k* Nearest Neighbor (kNN) rule [36].

The first three methods are non-supervised parametric methods, and the last method is a supervised non-parametric method. All of them are intensity based and only the third one is context dependent method. We have also used a non-linear filter (NLF) approach, based on the work by [37], previous to the execution of the K-Means and the kNN segmentation in order to reduce noise without blurring the edges of the image (since these two approaches do not use context information). NLF is not applied to EM neither GHMRF since the underlying assumption of image intensity Gaussian distribution does not hold after the non-linear filtering. Thus, six different segmentations are finally evaluated: K-Means, EM, GHMRF, kNN, NLF + K-Means and NLF + kNN.

3.2. *Simulated data*

As we said above, the main image used in this study comes from the digital brain phantom from McConnell Brain Imaging Center [32]. They proposed a realistic anatomical brain model and a MRI simulator to generate images with different radio-frequency (RF) non-uniformity, also called bias field. In this work, all the methods have been applied to images with RF non-uniformity of 0%, 20%, and 40% and noise levels of 3%, 5%, 7%, and 9%, on the T1-weighted modality. The volume is $217 \times 181 \times 217$ voxels with isotropic 1 mm voxel size, and non-brain tissues have been removed previously. An axial slice of this volume is shown in Fig. 1(a), with the gold standard segmentation corresponding to that slice (Fig. 1(b)). We will consider only the three main classes in the brain: white matter (WM), gray matter (GM) and cerebro-spinal fluid (CSF).

3.3. *Real data*

While simulated data provides an excellent tool to validate and compare method performance in presence of a variety of artifacts, assessment on real data is ultimately needed since the final purpose of these methods is to classify a real MRI of the human brain for a specific application. In this work, we consider a single real MR image of a normal brain (female adult, no pathology): a 3D T1-weighted magnetization-prepared rapid acquisition gradient echo (MPRAGE) sequence (Siemens Vision®, 1.5T, Erlangen, Germany) TR 9.7 ms, TE 4 ms, FOV 280×280, matrix 256×256, 146 slices, $0.98\,\text{mm} \times 0.98\,\text{mm} \times 1.25\,\text{mm}$. We show an axial slice and the gold standard segmentation in Fig. 2.

Manual segmentations were performed by five experts and a ground truth for CSF, GM and WM was then estimated using STAPLE algorithm [19].

3.4. *Segmentation results*

We show in Fig. 3 an axial slice of the results obtained from different segmentation methods (GHMRF (a), kNN (b), K-Means (c), EM (d), NLF + kNN (e) and NLF + K-Means (f)), applied to the MRI data shown in Fig. 1(a).

For the GHMRF, the value of β is fixed empirically to 1.2, $U(x, \beta)$ follows the Potts model, and instead of computing Z, the conditional probabilities at a given point $P(x_i|x_{N_i})$ are forced to sum up one among all possible labels. For more details please refer to [35]. The kNN segmentation has been carried out using a training set of 194 points, and using $K = 9$. The non-linear

[2] Please remind that we are not focusing on the methods evaluated but on the validation process itself.

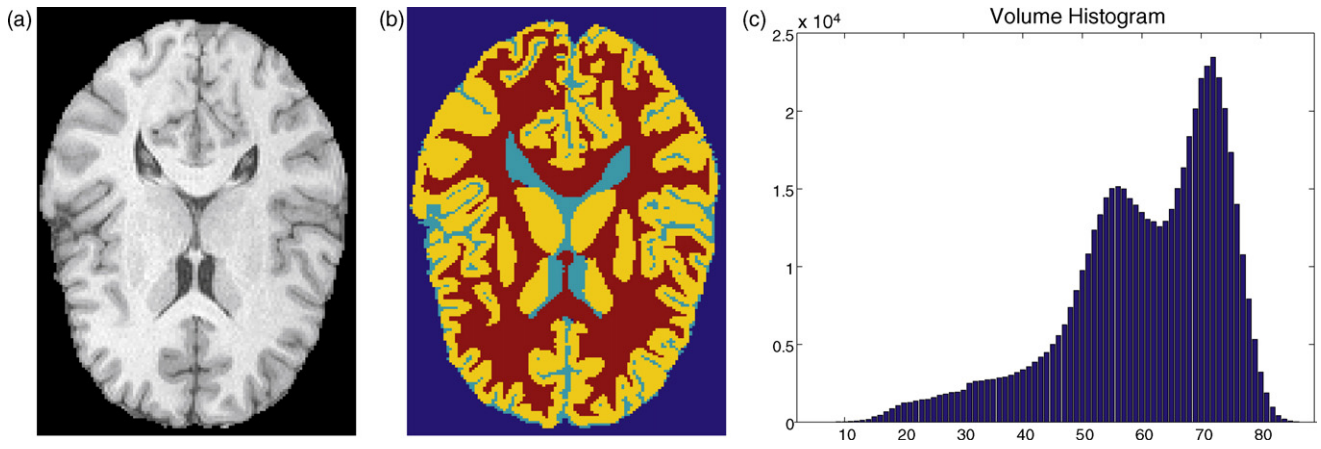


Fig. 2 – Axial slice of the real MRI (a), gold standard (b) and volume histogram (c).

filtering [37], has been done choosing temporal step $\tau = 100$, a size parameter $\sigma = 2$, and 5 iterations.

We show in Fig. 4 only the pixels that overlap with other classes in the same axial slice, in order to better show the misclassified voxels in that particular slice.

3.5. *Preliminary results using classic similarity measures*

In Fig. 5 we show the values of the similarity measures described in Section 2.2.1 computed for all methods, for the case with bias 20% and noise level 5%.

Notice that values in the TN coefficients differ from the values obtained by the other coefficients (for instance, the classes are ordered different than with the other three coefficients) and give values that hardly can differentiate the methods. This is because it depends on the number of voxels outside X and Y, that can be very large in our case, therefore leading to values near one, even if there is not too much overlapping. The *VS* coefficients present non-realistic results (notice an almost perfect classification of CSF in kNN, K-Means and NLF + K-Means methods), that is because *VS* depends only on the number of voxels of X and Y, and it can be one even if there exists no overlapping at all. Finally the *JC* and *DS* coefficients provide equivalent values as expected, that are also reasonable for this application. For these reasons, we will use the *JC* coefficient for our evaluation methodology.

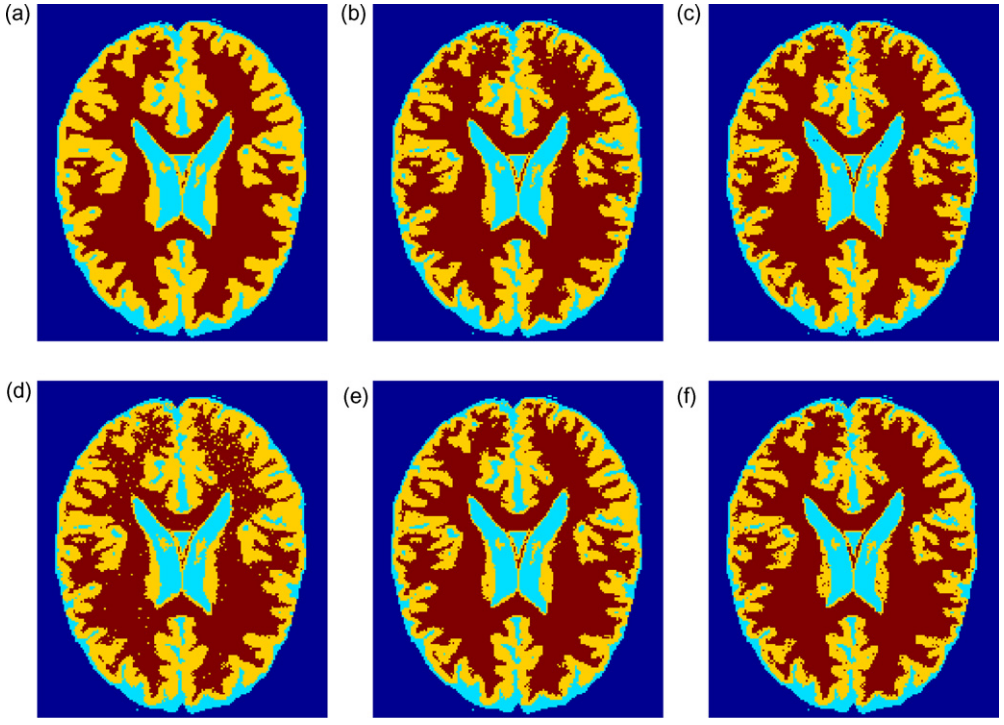


Fig. 3 – Axial slice of the segmentation results using GHMRF (a), kNN (b), K-Means (c), EM (d), NLF + kNN (e) and NLF + K-Means (f).

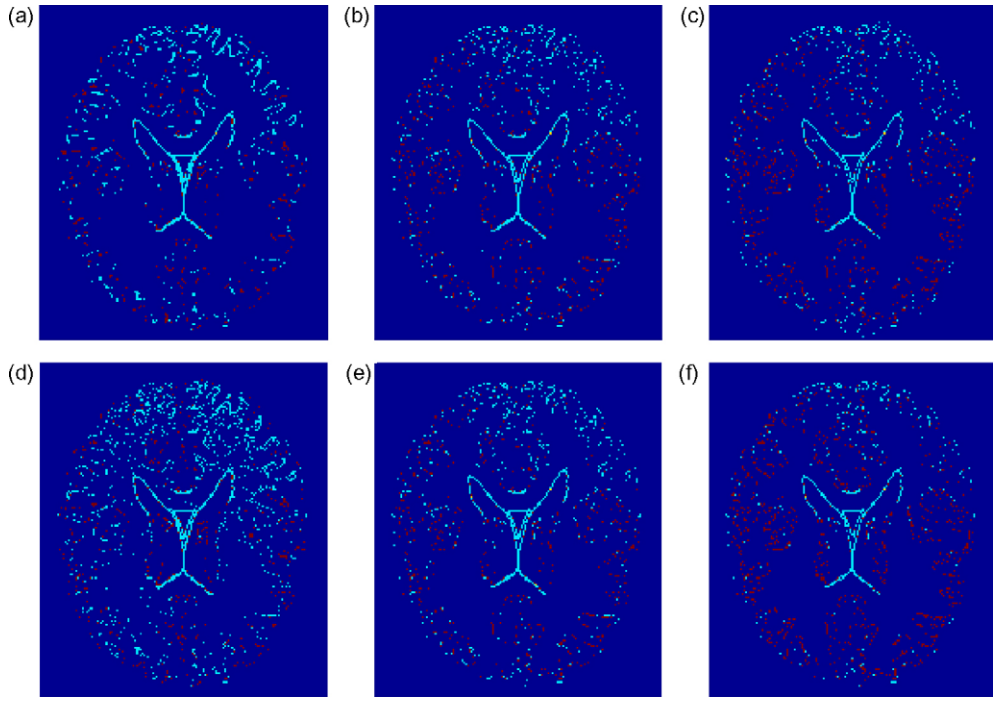


Fig. 4 – Error images from the segmentation of the axial slice of Fig. 1, using GHMRF (a), kNN (b), K-Means (c), EM (d), NLF + kNN (e) and NLF + K-Means (f).

4. Evaluation methodology

In the following we present our evaluation methodology, describing first some new similarity measures that take into account the position and the intensity from the misclassified voxels with respect to the ground truth, and other similarity measures based on connectivity and boundaries of the segmented images. All the measures proposed range between 0 and 1, with 0 the minimum similarity and 1 the maximum similarity between the segmentation and the gold standard. Finally we present the combination of the individual measures to obtain an aggregated one.

4.1. *Distance based similarity measures*

We propose to use the distance defined in Eq. (6), to define new similarity measures. The idea is to penalize more those voxels that are more distant from their corresponding class in the gold standard, i.e. to weight every misclassified voxel by its smallest Euclidean distance to the correct class it belongs to. To compute those Euclidean distances, it is enough to simply

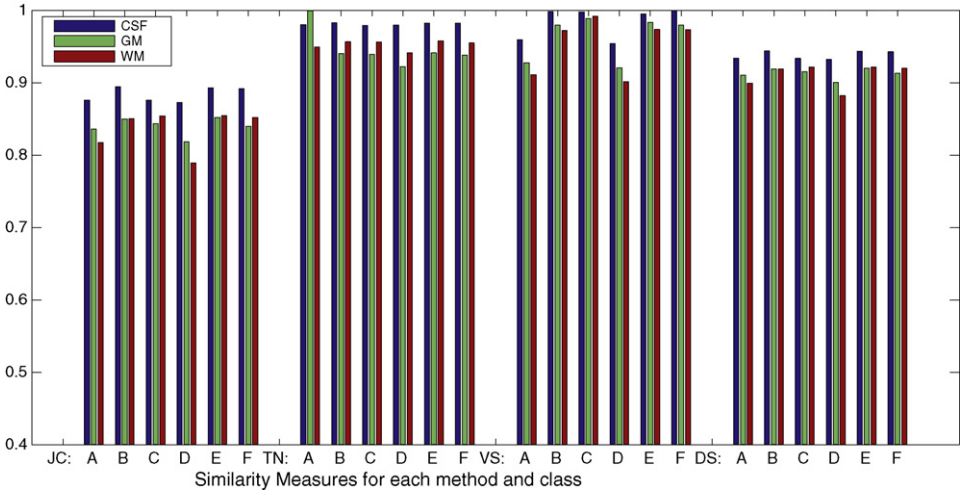


Fig. 5 – Classic similarity measures (JC, TN, VS and DS) computed for all methods for the simulated data set with bias 20% and noise level 5%. (A) GHMRF, (B) kNN, (C) K-Means, (D) EM, (E) NLF + kNN and (F) NLF + K-Means.

compute the distance transformation (DT) from a given class in the gold standard to the rest of the image, and look at the voxels of the DT at the positions of the misclassified voxels. We will use the squares of the distances in order to increase the penalization effect in the misclassified voxels.

We propose a new measure that we will call *JCd*, defined by substituting the values b and c from Eq. (2), by $\sum_i d^2(x_i)$ and $\sum_i d^2(y_i)$ respectively, where x_i are misclassified voxels of X that should be classified as Y, y_i are voxels of Y that should be classified as X, and $d()$ is the distance defined in Eq. (6):

$$JCd := \frac{a}{a + \sum_i d^2(x_i) + \sum_i d^2(y_i)} \tag{10}$$

The mean and standard deviation of the distances for every segmented class and for every method is shown in Fig. 6(a). In Fig. 6(b), we show the values of the new similarity measure defined, *JCd*.

All the values obtained now are lower than the measures that only count voxels, because we are penalizing each misclassified voxel by its square distance to the nearest classified voxel in the gold standard. These measures show the same performance as the classic ones but increasing their differences. This can be clearly seen in the example here after.

4.2. Similarity measures including intensity values

In this section we introduce another set of similarity measures, but this time, instead of using Euclidean distances in the image, we are going to use the intensity space. The idea is to penalize the misclassified voxels that should belong to a given class C, when those voxels are close to the theoretic mean of that class C, and far from the other classes. The reason is because voxels near the theoretic mean of a class and far from the others is easier to classify, than voxels that are, for instance, equidistant from two classes. The distance in the intensity space has to be measured taking into account the nature of the data. In the particular case of brain tissue segmentation we can use the mean and the variance of each class, in order to use the Mahalanobis distance from a given point in the intensity domain, x to a given class C:

$$d_{mh}(x, C) = \sqrt{(x - \mu_C)^T \Sigma_C^{-1} (x - \mu_C)}, \tag{11}$$

where μ_C and Σ_C are the mean and the covariance matrix of the class C. Using these distances, the new similarity measure can be defined as:

$$JCi = \frac{a}{a + \sum_{i=1}^{N} ((1/N_c) \sum_{j=1}^{N_c} d_{mh}(x_i, C_j) / d_{mh}(x_i, C))}, \quad C_j \neq C, \tag{12}$$

where this time x_i are misclassified voxels, N is the total amount of them, C is the class x_i should belong to, C_j are the rest of classes, and N_c is the number of classes C_j. In the previous equation each misclassified voxel contributes with its Mahalanobis distance to C, increasing the penalization if that distance is small, and decreasing that penalization effect when the average distance from x_i to the rest of classes C_j, is large. This means that if a misclassified voxel x_i is near the mean of the correct class C, it was likely to be classified as C, and thus we have to penalize it, and we also penalize x_i, when its intensity value is far from the mean of the rest of classes.

The results using this new measure are shown in Fig. 7. Again, we obtain lower values than with the classic measures due to the penalization on each misclassified voxel. Using these measures we observe a different situation than with *JCd*, because this measure will favor methods based on the histogram such as K-Means or K-NN, and decrease the performance of the methods that use neighborhood information such as GHMRF and those with the non-linear filtering.

4.3. Connectivity similarity measure

Connectivity is a property that can be used to measure the quality of a segmentation, and it is associated to the amount of islands or the granularity of a labeled image. If a segmented

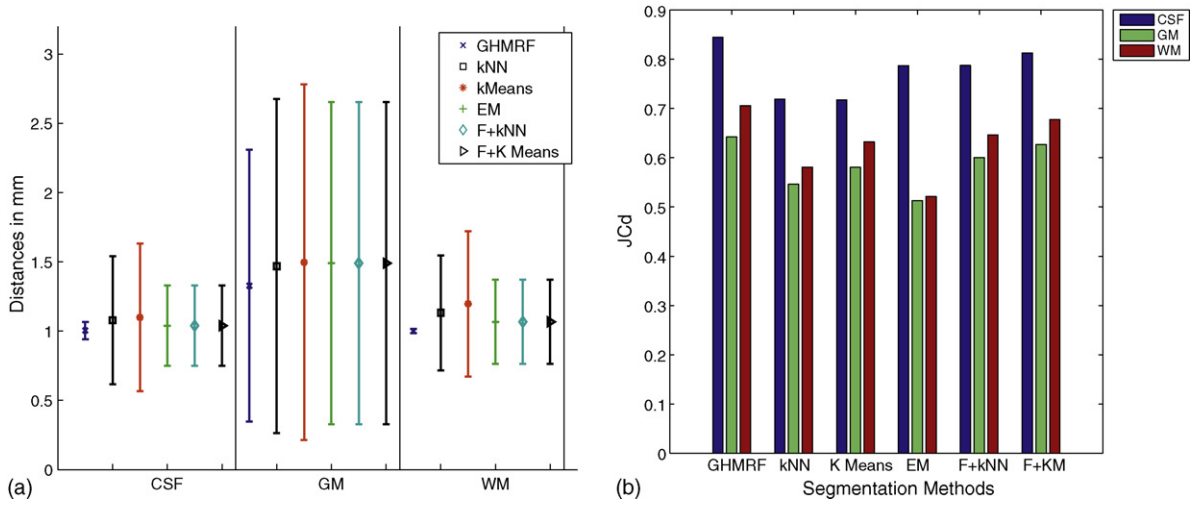


Fig. 6 – Average distances and standard deviation for the misclassified voxels (a) and distance based similarity measures, *JCd* (b), computed for all methods and for all the simulated data sets averaged (bias 0%, 20% and 40%, and noise levels 3%, 5%, 7% and 9%).

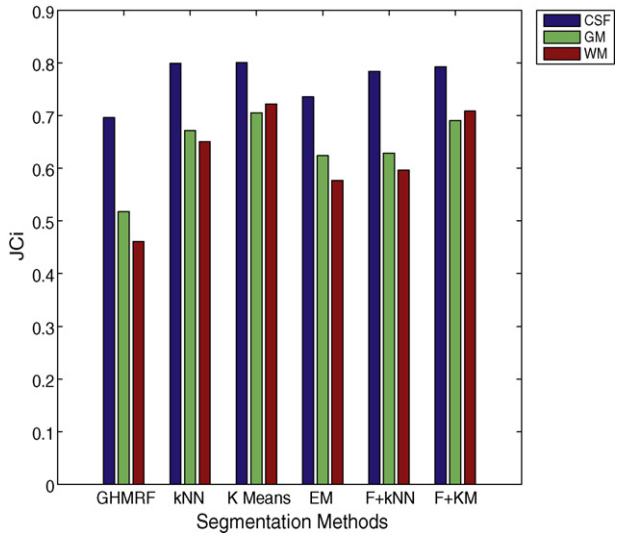


Fig. 7 – Similarity measures including intensity information computed for all methods and for all the simulated data sets averaged (bias 0%, 20% and 40%, and noise levels 3%, 5%, 7% and 9%).

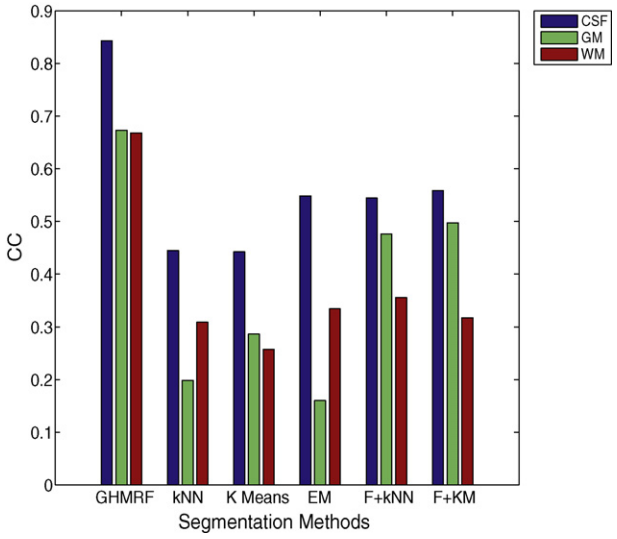


Fig. 8 – Connectivity similarity measures computed for all methods and for all the simulated data sets averaged (bias 0%, 20% and 40%, and noise levels 3%, 5%, 7% and 9%).

image has many disconnected regions, typically isolated voxels, compared to the gold standard, it has a quite negative visual impact that usually is not measured by classic similarity measures. With the definition of this measure in the evaluation method we include the granularity as an indicator of the quality of a segmentation. The connectivity of a region X, in a 3D regular grid is defined using a morphological dilation operator $\mathcal{D}_s$, with s a structuring element. We say that X is connected with other region Y, if

$$\mathcal{D}_s(X) \cap Y \neq \emptyset \tag{13}$$

We use a $3 \times 3 \times 3$ structuring element, thus defining a connectivity taking the 26 closest neighbors of each voxel. The number of connected components for each class N_{X_c} in the segmented volume can be compared with the number of connected components for the same class in the gold standard N_{Y_c}. The definition of a connectivity coefficient CC that takes values between 0 and 1 can be expressed as

$$CC_c := 2\frac{\min\{N_{X_c}, N_{Y_c}\}}{N_{X_c} + N_{Y_c}} \tag{14}$$

The results using this new measure are shown in Fig. 8, where this time more increased differences between methods can be observed, making this measure useful to discriminate between different methods.

4.4. *Similarity measures on the boundaries*

Another important feature to study are object boundaries, 3D surfaces for a volume and 2D contours for images. This is particularly important for some applications where surface or contour accuracy is more important than classic measures such as volume overlapping. In a general way, we can define a new measure, given the boundaries of one segmented class C, ∂X_c, and the boundary of that class in the ground truth, ∂Y_c, that we will call boundary JC:

$$BJC_c := \frac{|\partial X_c \cap \partial Y_c|}{|\partial X_c \cup \partial Y_c|} \tag{15}$$

As we said in Section 4.3, the segmented images may contain some small sets of voxels spread over the image, presenting granularity. It is important to highlight that these voxels will affect our boundary measure even if the main boundary of the ground truth really fits with the boundary of the segmented image. In order to prevent this effect, we will use only the main boundaries, thus removing the effect of granularity because it will provide a better measure of the contours of surfaces of our data and because the granularity is already measured with the connectivity measure proposed before. Therefore, we will use a modified boundary for every class in the segmented image $\partial X'_c$. We can express ∂X_c as the union of non-connected sets

$$\partial X_c = \bigcup_i \partial X_c^i \quad \forall i \text{ such as } \mathcal{D}_s(\partial X_c^i) \cap \partial X_c^j = \emptyset \quad \text{and} \quad i \neq j, \tag{16}$$

and $\mathcal{D}_s$ is the morphological dilation operator, as defined before. The new boundary $\partial X'(c)$ is then defined as

$$\partial X'_c = \bigcup_i \partial X_c^i \quad \forall i \text{ such as } \partial X_c^i \cap \partial Y_c \neq \emptyset \tag{17}$$

and ∂X_c^i are disjoint sets as before. With this new boundaries the modified measure is:

$$BJC'_c = \frac{|\partial X'_c \cap \partial Y_c|}{|\partial X'_c \cup \partial Y_c|} \tag{18}$$

The measure results using this new measure can be seen in Fig. 9. We also show in Fig. 10(a) a slice with the misclassified surface voxels for the GM in green, and in Fig. 10(b) the

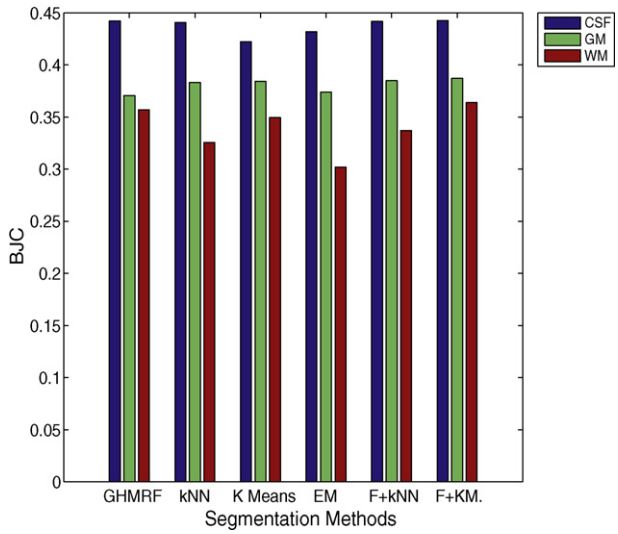


Fig. 9 – Boundary JC similarity measures computed for all methods and for all the simulated data sets averaged (bias 0%, 20% and 40%, and noise levels 3%, 5%, 7% and 9%).

misclassified surface is shown in red for a surface rendering of the GM. In this latter case only exterior surface errors can be seen.

4.5. *Aggregated multimodal similarity measure*

We propose to use the above definitions to combine different features to obtain more objective and reliable assessments. In this work we state that, as well as in the human vision, an intelligent system should employ several features to decide between different segmentation results. For that reason, a better similarity measure would emerge from the combination of the measures proposed before, a multidimensional similarity measure, that can be applied to any evaluation study. In order to define an aggregated similarity measure, let's construct a vector of similarity measures for a given class C:

$$\mathbf{v_c} = [JC_c, JCd_c, JCi_c, BJC'_c, CC_c] \tag{19}$$

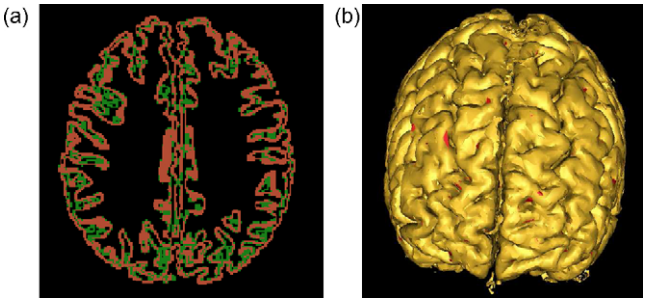


Fig. 10 – Boundary error voxels of the GM for a 2D slice, shown in green (a), and surface rendering of the GM with the exterior boundary errors in red (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

which is a vector composed of the classic Jaccard coefficient JC, the distance based Jaccard coefficient (JCd), the intensity Jaccard coefficient (JCi), the modified boundary Jaccard coefficient (BJC'), and the connectivity measure (CC). The aggregated similarity measure for a given class C, will be defined as

$$G_c := [\mathbf{v_c K v_c^T}]^{1/2} \tag{20}$$

where $\mathbf{K}$ is a matrix whose elements K_{ij} weights the different measures between them. For simplicity, we will use the identity matrix for $\mathbf{K}$, but depending on the application, it could be useful to increase or decrease some of the matrix coefficients, to give more importance to some of the measures with respect to others. To obtain a final value for the entire segmentation, we propose to combine the values obtained for each class, using the number of voxels of each class at the gold standard, $|Y(c)|$, as the weights:

$$G := \frac{\sum_c G_c |Y_c|}{\sum_c |Y_c|} \tag{21}$$

4.6. *Graphical data representation: Principal Component Analysis*

The aggregated measure presented in the previous section reduces the information contained in the similarity measure vectors to just one scalar value, which constitutes a simplification of the overall method. To overcome this situation, the multidimensional information of these vectors can be exploited in order to represent graphically the information contained for each segmentation method. In order to obtain effective graphical representations, that show clearly the behavior of the segmented data, we propose to use 2D plots of the obtained data. Two solutions are followed in this work to obtain these plots. The first one consists simply in the representation of pairs of measures, removing the information of the rest. An example of these plots are presented in the Section 5.3. However, using two similarity measures constitutes a limitation in the data representation, and therefore, we propose to use all the information included in the vectors in a second two dimensional plot of the data. This consists in a dimensionality reduction on the vectors, using a well-established tool such as PCA. In this case, let us consider the $N \times d$ matrix $\mathbf{X}$ containing the d different similarity measures obtained for N segmentation experiments. Using PCA, the linear transformation is given by

$$\mathbf{Y} = \mathbf{XH} \tag{22}$$

where $\mathbf{Y}$ is the obtained $N \times d'$ matrix of results ($d' \leq d$), and $\mathbf{H}$ is the transformation matrix whose columns are the eigenvectors corresponding to the d' largest eigenvalues of the $d \times d$ covariance matrix of the input data, $\mathbf{X}$. Using this transformation, we can reduce the dimensionality of the matrix of results so as to visualize it while retaining as much information as possible. In Section 5.4, the graphical representations results using this technique are presented.

Table 1 – Confusion tables for the simulated data with 20% bias and 5% noise level. (A) GHMRF, (B) kNN, (C) K-Means, (D) EM, (E) NLF + kNN and (F) NLF + K-Means. Values are in %.

	Reference			FP		Reference			FP
	CSF	GM	WM			CSF	GM	WM	
A					B				
CSF	89.77	1.03	0.01	1.04	CSF	94.29	2.27	0.02	2.29
GM	10.23	98.16	17.39	27.61	GM	5.71	91.13	7.07	12.79
WM	0.00	0.81	82.60	0.81	WM	0.00	6.60	92.91	6.60
FN	10.23	1.84	17.40		FN	5.71	8.87	7.09	
C					D				
CSF	93.58	2.89	0.02	2.91	CSF	89.12	0.88	0.00	0.89
GM	4.60	90.50	7.07	11.67	GM	10.88	97.79	19.70	30.58
WM	0.00	6.60	92.91	6.60	WM	0.00	1.33	80.30	1.33
FN	4.60	9.49	7.09		FN	10.88	2.21	19.70	
E					F				
CSF	94.84	2.60	0.02	2.62	CSF	94.27	2.29	0.02	2.31
GM	5.16	93.58	10.17	15.34	GM	5.72	89.47	5.45	11.17
WM	0.00	3.82	89.81	3.82	WM	0.00	8.24	94.53	8.24
FN	5.16	6.42	10.19		FN	5.72	10.53	5.47	

5. Experiments and results

In this section some experiments and results are shown. First, the results using the confusion tables obtained for one simulated data set are presented, in order to show the classic way to compare segmentation methods, then some validation experiments are described. The first one is an experiment showing the improvement in the similarity measures when distance information is included, using a synthetic data set. After that, two validation studies are shown, one for the simulated data sets shown in Section 3.2 and other for the real data set of Section 3.3.

5.1. *Results with confusion tables*

In Table 1, we show the confusion tables for the GHMRF, kNN, K-Means, EM, NLF + kNN and NLF + K-Means methods respectively, using the simulated data set with 3% noise and 0% bias. We also show the false positive and false negatives values obtained in each case.

5.2. *Validation experiments for distance based measures*

Similarly to the new *JCd* coefficient proposed in Section 4.1, we can define the *D* Sd, *TNd*, and *VSd* coefficients, also replacing b and c in Eqs. (3)–(5) by $\sum_i d^2(x_i)$ and $\sum_i d^2(y_i)$ respectively, as done before. In order to show the effect of the distance in these new coefficients, we have carried out an experiment using the synthetic data shown in Fig. 11. A squared shape embedded in a 256 × 256 2D image is considered as the ground truth, and a similar square rotated from 2° to 30° every 2° clockwise are considered as the segmentations. In Fig. 12 we show the values of the similarity measures for the rotated images using the *JC*, *TN*, *VS* and *DS* coefficients, and also for the *JCd*, *TNd*, *VSd* and *DSd* coefficients, in order to compare the behavior of the new measures with the distances incorporated. Compared to the classic similarity measures, it is clear that the sensitivity of the new distance based measures is higher, so they can be used to evaluate more accurately similar segmentations. Notice that in this case, the *VS* coefficient does not change because it only decreases if $|X| \neq |Y|$, and in this case both values remain always the same. This is the main reason that *VS* is not suitable to measure dissimilarity for label maps. When the distance is incorporated to *VS*, it becomes more useful, but still the other coefficients are preferable.

The distance values, coded in grayscale color map, from the misclassified voxels to their corresponding nearest classes are shown in Fig. 13, for the case where the square is rotated 30° from its original position. This image is used to compute the new similarity measures that we have proposed here.

5.3. *Experiment on a simulated data set*

We show in Fig. 14, the values for the aggregated similarity measures per class and for the whole segmentation, in the six methods studied, and in order to compare, we show in

Fig. 11 – Synthetic data. In white the gold standard square and in gray a square rotated from 0° to 30° every 6°.

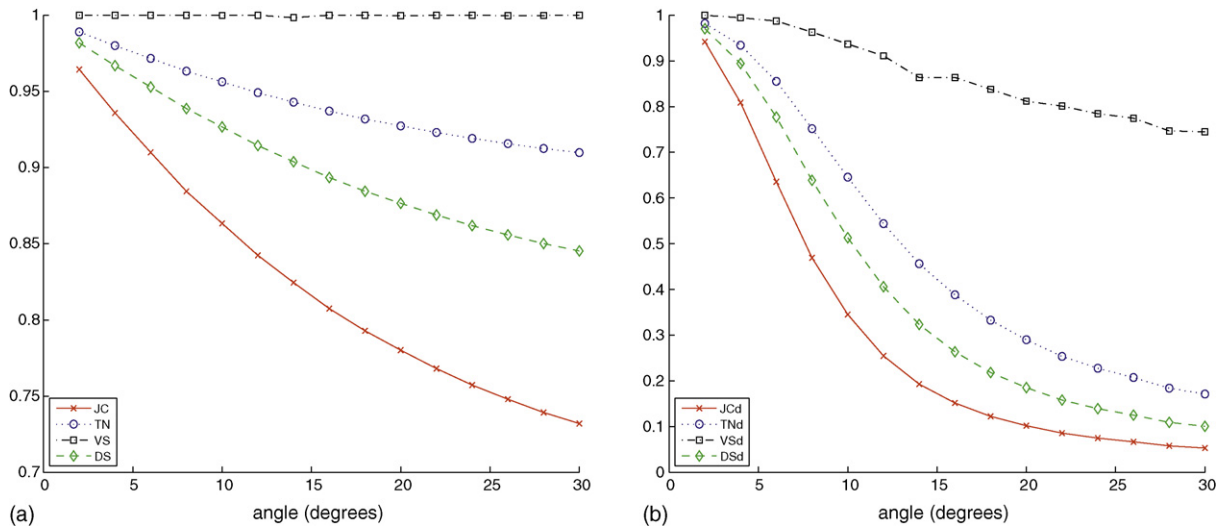


Fig. 12 – Similarity measures computed for the synthetic data of Fig. 11 for several angle values, without distances (a) and using distances (b).

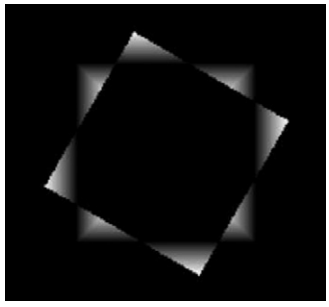


Fig. 13 – Distances from the misclassified voxels to the nearest class they should belong to, in the square rotated by 30°.

Fig. 15 the *JC* averaged values, scaled to a range of the same size (0.2). In both figures the ordering of the methods are: GHMRF, NLF + K-Means, NLF + kNN, K-Means, kNN, and EM, being the differences using the aggregated measures more noticeable than only using *JC*.

A quick look at all the similarity measures can be done also using a boxplot, see Fig. 16. Note that there is substantially more variability in the ratings of CC than in the rest of measures.

As mentioned in Section 4.6, in order to better represent the similarity measures, we have drawn a series of 2D plots of the similarity measure values, choosing pairs of them: one in the *x*-axis and the other in the *y*-axis. In these plots each point corresponds to the mean value across the three classes (CSF, GM and WM) of each segmentation. For each method an ellipse is also plotted representing the covariance of the data group, with the center representing the mean of the group.

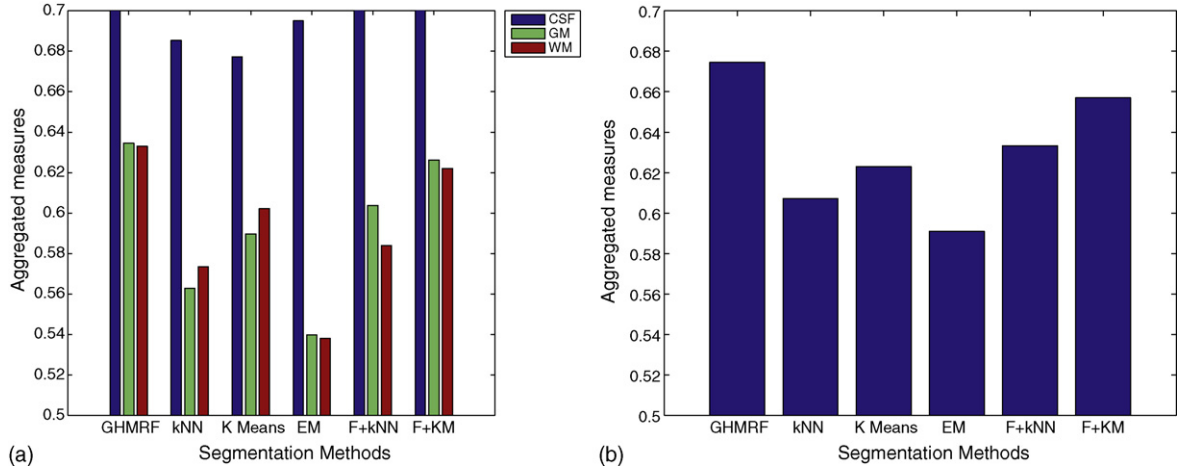


Fig. 14 – Aggregated similarity measures averaged for all the simulated data sets (bias: 0%, 20%, 40% and noise levels: 3%, 5%, 7%, 9%), separated by classes (a) and averaged (b).

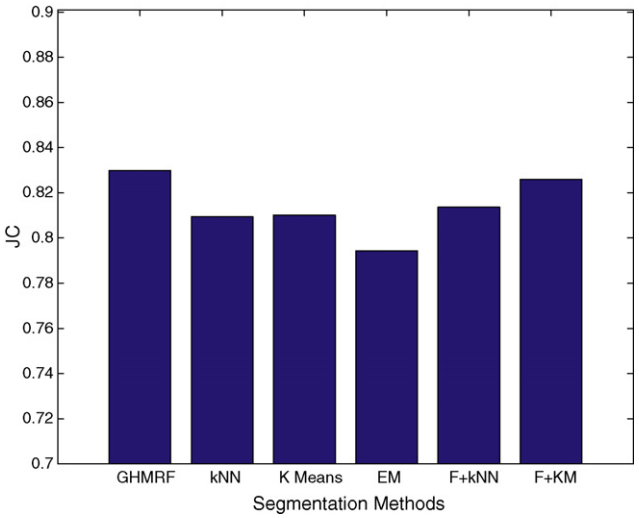


Fig. 15 – JC similarity measures averaged for all the simulated data sets (bias: 0%, 20%, 40% and noise levels: 3%, 5%, 7%, 9%).

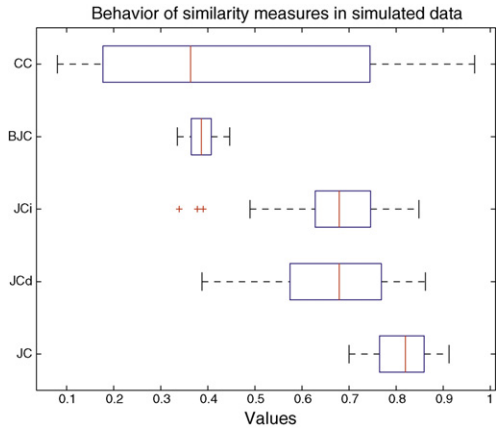


Fig. 16 – Boxplot representation of similarity measures, for the simulated data set (bias 0%, 20% and 40%, and noise levels 3%, 5%, 7% and 9%) and all methods.

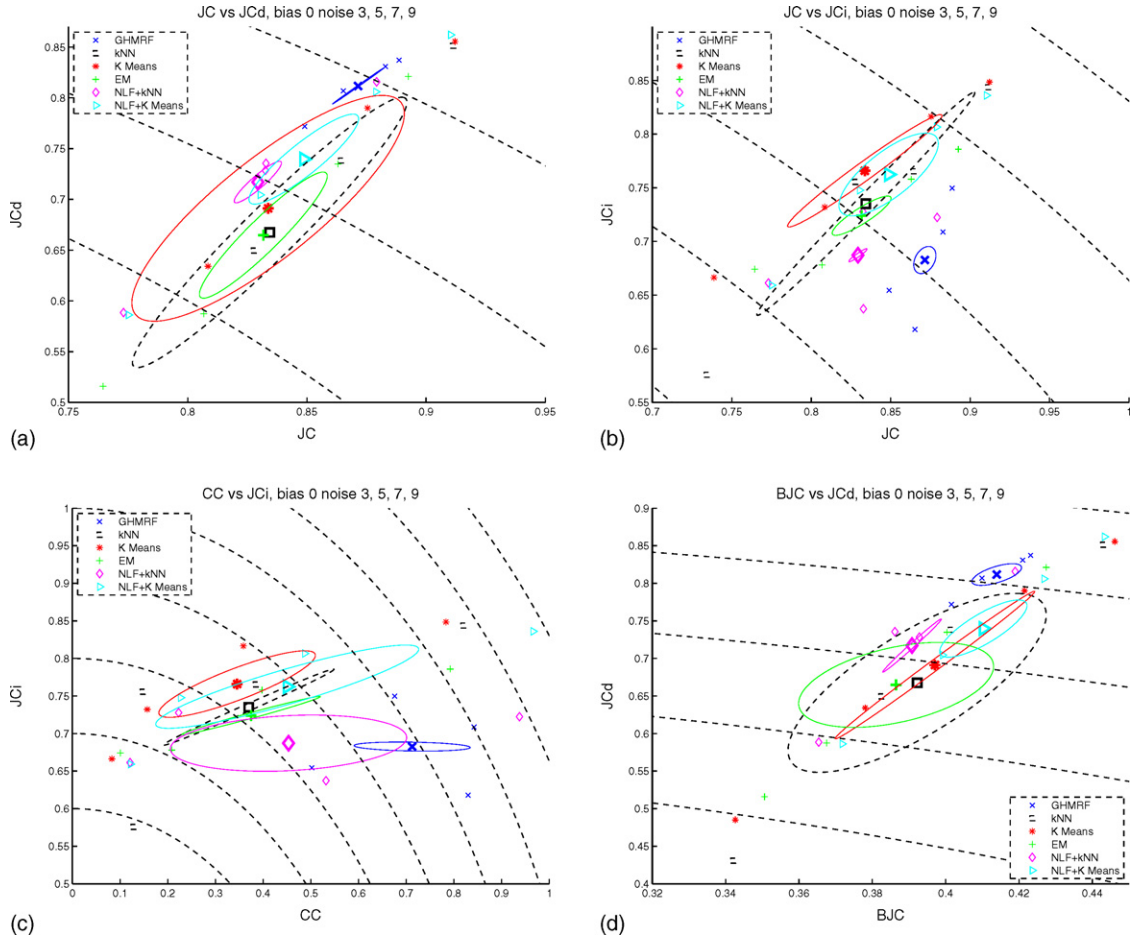


Fig. 17 – Joint similarity measures for the simulated data set, with bias 0% and noise levels 3%, 5%, 7% and 9%. JC vs JCd (a), JC vs JCi (b), CC vs JCi (c) and BJC vs JCd (d).

The sizes of the ellipses are proportional to the variance of the data.

Using this representation we can see more clearly the differences between several methods than in one-dimensional plots. Fig. 17 illustrates some of these plots for the simulated data, for bias 0%, and each group of points corresponds to different noise levels: 3%, 5%, 7%, and 9%. The center of each group has a marker of higher size, to distinguish it from the other markers. Notice that a perfect situation will be a circle of radius zero placed at the (1,1) position, therefore in order to see which method performs better than others we also show dotted circles centered at (0,0) with increasing radius values. It can be seen in Fig. 17, that GHMRF is the best method in most cases, the second one is NLF + K-Means, and the order of the others changes depending on the figure.

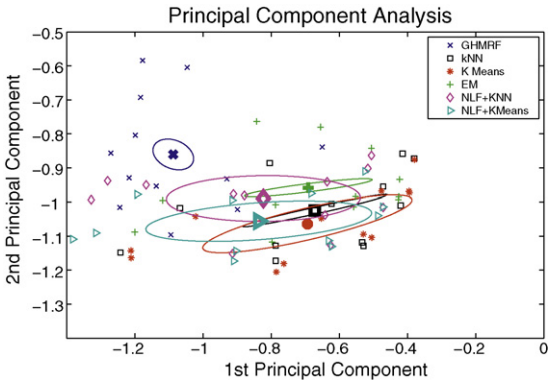


Fig. 18 – 2D plot representation using PCA, for the simulated data set (bias 0%, 20% and 40%, and noise levels 3%, 5%, 7% and 9%) and all methods.

5.4. *Principal Component Analysis*

Representing two similarity measures in a 2D plot provides a useful insight into the differences between several segmentation methods, but the similarity measures vectors can be used more efficiently using the PCA representation detailed in Section 4.6. In the PCA study shown here, the two first principal components (eigenvectors) represent the 83% and the 14.5% of the variance of the data. Thus, a representation using these two components, explain most of the variance in our data. This is shown in Fig. 18, where two first principal component coefficients are represented together for all methods and experiments. To understand such plot we have to be aware of the coefficients of the linear combinations of the original variables that generate the principal components. The first two

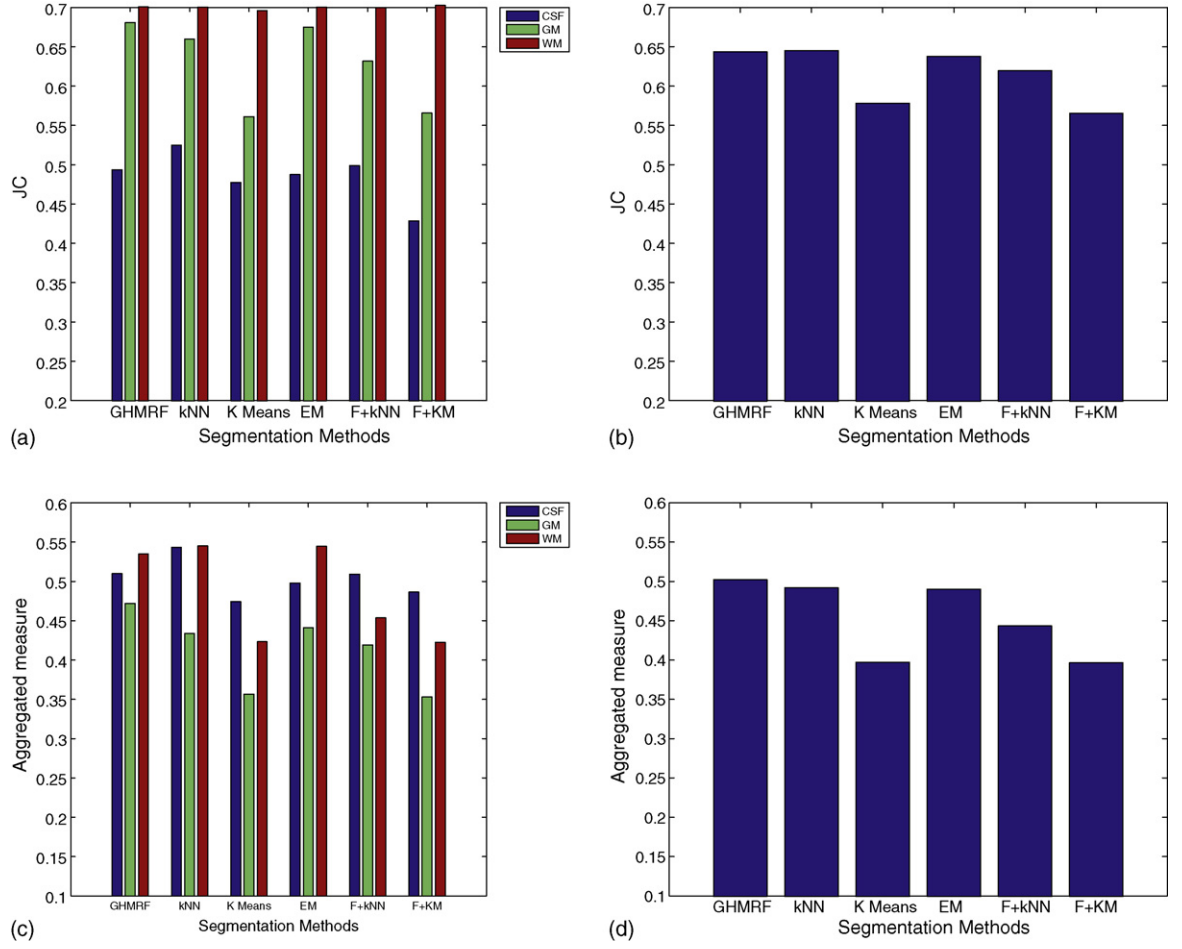


Fig. 19 – JC (a) and (b) and aggregated similarity measures (c) and (d) for the real data, by classes and averaged.

principal component coefficient vectors are:

1st principal component :

[−0.1384, −0.3346, −0.0683, −0.0711, −0.9269],

2nd principal component :

[−0.2460, −0.4598, −0.7966, −0.1395, 0.2721],

where the corresponding similarity measures for these coefficients are as in Eq. (19). All the coefficients of the first principal component have the same sign, making it a weighted average of all the original variables.

In the new PCA space, three groups of similar performance can be distinguished. Similarly to previous 2D plots, we bold the values that represent the mean of each method (among all the simulated images) and this is the center of a circle of radii variance of the data. The best performance is achieved clearly by GHMRF method (on the left part), followed by the methods with non-linear filtering: NLF + K-Means, NLF + kNN (on the middle), and finally with similar values: K-Means, EM and kNN (on the right part of the plot). Note however that the most negative the first component value is in the PCA, the best the method performs. This is because the linear transform coefficients are all negative. Actually, if we focus on the projection of such values to the first component only (83% of the variance explained), we conclude the same behavior than in Fig. 14 but in a more discriminant manner, except for the EM and kNN methods that present quite similar values.

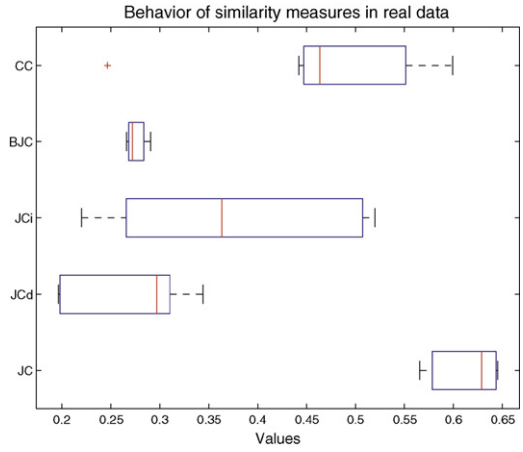


Fig. 20 – Boxplot of similarity measures for real data set.

5.5. *Experiments on real data*

We have performed an evaluation using the real MR data set shown in Fig. 2. In Fig. 19(a) and (b), the *JC* obtained is shown into separated classes and averaging among them, whereas in Fig. 19(c) and (d), we show the aggregated similarity measure obtained, all of them scaled to a range of the same size (0.5). Notice that the order of performance

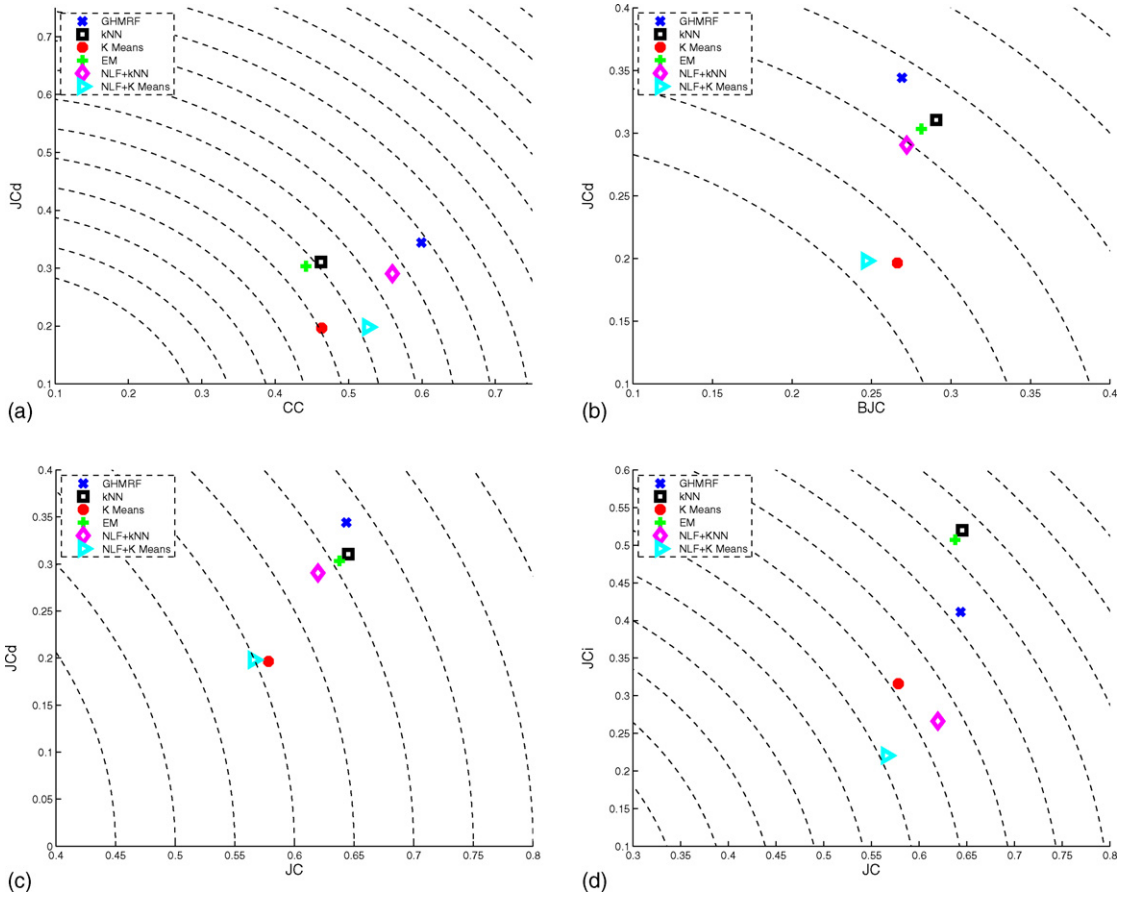


Fig. 21 – Joint similarity measures for the real data, CC vs JCd (a), BJC vs JCd (b), JC vs JCd (c) and JC vs JCi (d).

of the methods is almost the same in both cases. Looking at the aggregated similarity measures, the best method is GHMRF, as we expected, whereas kNN performs better under the *JC* measure. The rest of methods are ordered in the same way. In this real case, as opposite as in the simulated data, kNN performs better than EM, and EM performs better than K-Means, which is a more realistic order. We can also conclude that methods using filtering (NLF + kNN and NLF + K-Means) are not better than the methods without filtering (kNN and K-Means), which means that filtering affects the performance, probably because a excess of filtering have been applied. However, the most important result derived from the values obtained here is that the aggregated measures proposed are more discriminant than standard *J C* measures.

As we did in the previous section, we show in Fig. 20 the behavior of each similarity measure. In this case, *JCi* is the measure that show most variance of values across all methods. We also present 2D plots of pair of similarity measures in Fig. 21. GHMRF is the best method in most of the cases, and the rest of them are ordered different in each plot. However, we can see that EM and kNN methods are close to each other in all the figures, as well as K-Means and NLF + K-Means, showing similar behaviors in each pair of methods.

Here also a PCA analysis is done (Fig. 22). In this case the first two components represent well all the variance of the data (78% and 20%). Two first component representation is shown in Fig. 22. However, in this case we have only one point per method, so we cannot draw estimate the variance of the data. Please note that the linear transformation coefficients are not all of the same sign and thus, no conclusion can be drawn on which method performs better depending on their position in the PCA plane:

1st principal component :

[−0.1567, −0.2084, −0.9245, −0.0819, 0.2657]

2nd principal component :

[0.3305, 0.7058, −0.0406, 0.0518, 0.6231]

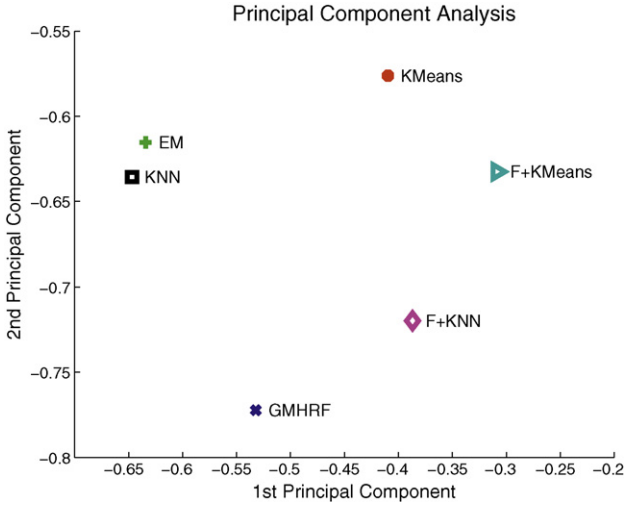


Fig. 22 – 2D plot representation using PCA for the real data and all methods.

We can see that the six methods are clustered in a different way than in the simulated data analysis. In this case, K-Means algorithm distinguishes from kNN and EM, the filtered methods are more separated between each other, and GHMRF remains differentiated as in the simulated data.

6. Conclusions and future works

We have described in this paper several similarity measures for the evaluation of segmented images, given a ground truth, and we have applied it to the segmentation of the brain into its three main tissues, CSF, GM and WM. Using the segmentation of a simulated and a real MRI data set, we have presented a new evaluation framework.

We have shown that classic similarity measures can provide non-realistic results, as illustrated graphically in Fig. 5. That is the case of *TN* and *VS* coefficients that give values that could arise in erroneous decisions, as discussed in Section 3.5. Other classic measures, such as *JC* and *DS* coefficients, give reasonable values and therefore *JC* has been used in this evaluation study. We have proposed a set of new similarity measures, adding complementary criteria to the sizes of volume overlapping. We propose new measures based on the position and intensity values of the misclassified voxels, as well as measures related to the granularity and the boundaries of the segmented images. As a result we have proposed a new aggregated multidimensional similarity measure that combines the similarity measures proposed to obtain better reliability in the evaluation of several segmentation methods, which is the main contribution of this work. This method provides more detailed and objective information about the quality of each method and presents some ideas of how much better one method can be with respect to others, and also how much a class can be better classified than the others within a given method. This is especially useful for 3D data sets, where the segmentation results are often difficult to assess visually, and in general for the evaluation of any segmentation result, disregarding the application, when a gold standard is available. As far as we know, the new measures described (*JCd*, *JCi*, *CC* and *BJC*) are completely new, and this is the first time that multiple similarity measures are combined in this way for segmentation evaluation.

We have also presented 2D plots of pairs of similarity measures that show how the combination of several measures improves the visual representation of the difference between several methods, and motivate the validity of the multidimensional aggregated measure proposed here. We have also developed a successful 2D representation using PCA in order to show in one single 2D plot the relation between all methods, and if the principal component coefficients are all the same sign it is possible to compare the performance of each method in such dimension. To overcome this limitation, and to obtain 2D plots with comparable values, we have to look for projection directions with all the coefficients of the same sign. Unfortunately this cannot be done with a standard PCA decomposition, and therefore, other methods for dimensionality reduction should be employed instead.

Taking into account the amount of information used, we can say that our methodology improves the segmentation

evaluation (visually and numerically) as compared to classic measures. We have therefore changed the classic evaluation philosophy where the objective was to look at how many voxels are correctly or wrongly segmented, with a new philosophy where the objective is to look at how many voxels are well segmented, how distant are the wrong voxels from their correct position (in the image and in the intensity domain), how good are the boundaries, and how good is the segmentation granularity, taken this all together.

The correspondence between visual inspection (by looking at one slice, see Fig. 4), and the numeric values of our aggregated measure fits quite well, resulting in a classification in order of decreasing quality: GHMRF, NLF + K-Means, NLF + kNN, K-Means, kNN and EM in the experiment with simulated data. This result is in part as expected, because GHMRF method is designed specifically for this particular application so it should be the one with better results, and the methods with a non-linear filtering give better results than without them. On the other hand, it is not expected that one simple clustering method such as K-Means performs in general better than EM and kNN, which only means that the simulated data is well suited for such clustering method. This is not the situation with real data, where the order is GHMRF, kNN, EM, NLF + kNN, K-Means, and NLF + K-Means, which is a more natural result.

The evaluation study done here is not intensive, and it should be considered as a good example of how our proposed evaluation method can be applied. This methodology can be also used to compare the results of a given method using different parameters in order to select the correct set of parameters. Notice also that new measures not related to accuracy, for instance measures based on reproducibility, efficiency and user interaction, can be included in our model, as proposed by Udupa et al. [18]. As a final remark, the coefficients of the matrix **K** should be selected appropriately for every application in order to provide an objective aggregated similarity measure.

Conflict of interest

None declared.

Acknowledgments

The authors would like to thank Prof. Thiran and Prof. Vandergheynst for their discussion and advice. This work has been funded by the National Spanish Grant TEC-2007-67073/TCM, the European network of excellence Similar, FP6-507609, by the Center for Biomedical Imaging (CIBM) of the Geneva - Lausanne Universities, the EPFL, as well as the foundations Leenaards and Louis-Jeantet.

REFERENCES

[1] N. Pal, S. Pal, A review on image segmentation techniques, Pattern Recognition 26 (1993) 1277–1294.
[2] L. Clarke, R. Velthuizen, M. Camacho, J. Heine, M. Vaidyanathan, L. Hall, R. Thatcher, M. Silbinger, MRI segmentation: methods and applications, Magnetic Resonance Imaging 13 (3) (1995) 343–368.
[3] D.L. Pham, C. Xu, J.L. Prince, Current methods in medical image segmentation, Annual Review of Biomedical Engineering 2 (2000) 315–338.
[4] J. Duncan, N. Ayache, Medical image analysis: progress over two decades and the challenges ahead, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (1) (2000) 85–106.
[5] J. Noble, D. Boukerroui, Ultrasound image segmentation: a survey, IEEE Transactions on Medical Imaging 25 (8) (2006) 987–1010.
[6] A. Frangi, W. Niessen, M. Viergever, Three-dimensional modeling for functional analysis of cardiac images: a review, IEEE Transactions on Medical Imaging 20 (1) (2001) 2–25.
[7] R. Cardenes, S. Warfield, E. Macías, J. Ruiz-Alzola, High performance supervised and unsupervised MRI brain segmentation, in: Proc. of the NeuroImaging Workshop, Eurocast, 2003, pp. 57–60.
[8] R. Cardenes, S. Warfield, E. Macías, J. Santana, J. Ruiz-Alzola, An efficient algorithm for multiple esclerosis lesion segmentation from brain MRI, Computer Aided System Theory-EUROCAST 2809 (2003) 542–551.
[9] S. Olabarriaga, A. Smeulders, Interaction in the segmentation of medical images: a survey, Medical Image Analysis 5 (2001) 127–142.
[10] R. Cárdenes, M. Bach, Y. Chi, I. Marras, R. de Luis, M. Anderson, P. Cashman, M. Bultelle, Multimodal evaluation for medical image segmentation, in: Proc. of the 12th International Conference on Computer Analysis of Images and Patterns, CAIP, Vienna, Austria, 2007, pp. 229–236.
[11] Y. Zhang, A review of recent evaluation methods for image segmentation, in: Proc. of the International Symposium on Signal Processing and its Applications, ISSPA, 2001, pp. 148–151.
[12] W. Yasnoff, J. Miu, J. Bacus, Error measures for scene segmentation, Pattern Recognition 9 (1977) 217–231.
[13] K. Straters, J. Gerbrands, Three-dimensional segmentation using a split, merge and group approach, Pattern Recognition Letters 12 (1991) 307–325.
[14] E. Pichon, A. Tannenbaum, R. Kikinis, A statistically based flow for image segmentation, Medical Image Analysis 8 (2004) 267–274.
[15] D. Huttenlocher, G. Klanderman, W. Rucklidge, Comparing images using the hausdorff distance, PAMI 15 (9) (1993) 850–863.
[16] W. Crum, O. Camara, D. Hill, Generalized overlap measures for evaluation and validation in medical image analysis, IEEE Transactions on Medical Imaging 25 (11) (2006) 1451–1461.
[17] J. Cardoso, L. Corte-Real, Toward a generic evaluation of image segmentation, IEEE Transactions on Image Processing 14 (11) (2005) 1773–1782.
[18] J. Udupa, V. LeBlanc, Y. Zhuge, H. Schmidt, L. Currie, B. Hirsch, J. Woodburn, A framework for evaluating image segmentation algorithms, Computerized Medical Imaging and Graphics 30 (2) (2006) 75–87.
[19] S. Warfield, K. Zou, W. Wells, Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation, IEEE Transactions on Medical Imaging 23 (2004) 903–921.
[20] F. Bello, A. Colchester, Measuring global and local spatial correspondence using information theory, in: Proc. of the 1st International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI, 1998, pp. 964–973.
[21] M. Martin-Fernandez, S. Bouix, L. Ungar, R.W. McCarley, M.E. Shenton, Two methods for validating brain tissue classifiers, in: Proc. of the 8th International Conference on Medical

Image Computing and Computer-Assisted Intervention, MICCAI, Palm Springs, CA, USA, 2005, pp. 515–522.
[22] H. Lockett, M. Guenova, Similarity measures for mid-surface quality evaluation, Computer Aided Design 40 (3) (2008) 368–380.
[23] N. Aspert, D. Santa-cruz, T. Ebrahimi, Mesh: measuring errors between surfaces using the hausdorff distance, in: IEEE International Conference in Multimedia and Expo, 2002, pp. 705–708.
[24] L. Hea, Z. Pengb, B. Everdingb, X. Wangb, C. Hanb, K. Weissc, W. Weeb, A comparative study of deformable contour methods on medical image segmentation, Image and Vision Computing 26 (2) (2008) 141–163.
[25] M. Brummer, R. Mersereau, R. Eisner, R. Lewine, Automatic detection of brain contours in MRI data sets, IEEE transactions on Medical Imaging 12 (2) (1993) 153–166.
[26] M. Atkins, B. Mackiewich, K. Whittall, Fully automatic segmentation of the brain in MRI, IEEE Transactions on Medical Imaging 17 (1998) 98–107.
[27] K. Held, E. Kops, B. Krause, W. Wells, R. Kikinis, et al., Markov random field segmentation of brain MR images, IEEE Transactions on Medical Imaging 16 (6) (1997) 878–887.
[28] T. Kapur, E. Grimson, W. Wells, R. Kikinis, Segmentation of brain tissue from magnetic resonances images, Medical Image Analysis 1 (1996) 109–127.
[29] C. Li, D. Godlgolf, L. Hall, Knowledge-based classification and tissue labeling of MR images of human brain, IEEE transactions on Medical Imaging 12 (4) (1993) 740–750.
[30] K. Lim, A. Pfefferbaum, Segmentation of MR brain images into cerebrospinal fluid spaces, white and gray matter, Journal of Computer Assisted Tomography 13 (4) (1989) 588–593.
[31] S. Warfield, M. Kaus, F.A. Jolesz, R. Kikinis, Adaptive, template moderated, spatially varying statistical classification, Medical Image Analysis 4 (1) (2000) 43–55.
[32] D. Collins, A. Zijdenbos, V. Kollokian, J. Sled, N. Kabani, C. Holmes, A. Evans, Design and construction of a realistic digital brain phantom, IEEE Transactions on Medical Imaging 17 (3) (1998) 463–468.
[33] J. Hartigan, M. Wong, A K-Means clustering algorithm, Journal of Royal Statistical Society Series C, Applied Statistics 28 (1979) 100–108.
[34] A. Dempster, N. Laird, D. Rubin, Maximum-likelihood from the incomplete data via the EM algorithm, Journal of Royal Statistical Society Series B 39.
[35] M. Bach-Cuadra, L. Cammoun, T. Butz, O. Cuisenaire, J. Thiran, Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images, IEEE Transactions on Medical Imaging 24 (12) (2005) 1548–1565.
[36] S. Warfield, Fast K-NN classification for multichannel image data, Pattern Recognition Letters 17 (7) (1996) 713–721.
[37] J. Weickert, B. ter Haar Romery, M. Viergever, Efficient and reliable schemes for nonlinear diffusion filtering, IEEE Transactions on Image Processing 7 (3) (1998) 398–410.