Eurographics Workshop on Visual Computing for Biology and Medicine (2014)
I. Viola, K. Bühler, and T. Ropinski (Editors)

Uncertainty-aware Ensemble of Classifiers for Segmenting Brain MRI Data

Ahmed Al-Taie[1,3], Horst K. Hahn[1,2], and Lars Linsen[1]

[1]Jacobs University, Bremen, Germany
[2]Fraunhofer MEVIS, Bremen, Germany
[3] Computer Science Department, College of Science for women, Baghdad University, Baghdad, Iraq

Abstract

Estimating and visualizing uncertainty in medical image segmentation has become an active research area due to the necessity of making medical experts aware of possibly wrong segmentation decisions. Still, to our knowledge all these methods are based on a single choice of the underlying segmentation approach. Segmentation using an ensemble of classifiers (or committee machine) use multiple classifiers to increase the performance when compared to applying a single classifier. In this paper, we propose methods to estimate uncertainties in segmentations produced by ensembles of classifiers. We investigate and compare the different combining strategies of the segmentation results of the ensemble members from an uncertainty point of view. We discuss why some combining strategies tend to perform better than others. Also, we visualize the estimated uncertainties using a color mapping in image space and propose a post-segmentations correction step to reclassify the noisy pixels in the final result based on the statistical uncertainty.

1. Introduction

In pattern recognition and machine learning, several studies confirmed the concept that combining the results of multiple classifiers can yields more reliable and accurate results when compared to the results of individual classifiers [KHDM98, Sha99, Die00, Kun04, FJ05]. This concept is known as committee machine, mixture of experts, or ensemble of classifiers [Mig10]. An important aspect of such an ensemble of classifiers is the diversity, i.e., that the complementary information of the individual classifiers can improve the final result when combining them. Hence, the individual classifiers are allowed to produce errors, but in order to be able to correct them, different classifiers shall not produce the same errors. Diversity can be achieved in multiple ways: Several instances of the same algorithm can be applied on different subsets of the input data or on the same data but initialized using different parameter values. Diversity can also be achieved using different data representations (e.g., the same input image is represented in different color spaces) or using different algorithms with diverse behaviors on the input data [FJ05, Mig10]. In the context of combining the members of the ensemble of classifiers, there are several combining rules proposed in the literature. Examples of these rules are majority votes, weighted majority votes, or probability rules such as product, sum, maximum, minimum, median, etc. There is no agreement on which rule would be the best for all cases, neither a clear theoretical explanation why certain rules are better than others for certain applications. In general, the concept of ensemble of classifiers was mostly used in machine learning applications for supervised classification.

In medical image analysis, the main focus of applying the concept of ensembles of classifiers was to solve the medical image segmentation problem using collection of atlas-based or human-rater segmentations. The ensembles of classifiers has been applied to estimate the performance level of individual segmentation algorithms [WZW04, LvdHK*10] or to achieve more accurate segmentation results [RM05, AMBdS09, APNY13].

Recently, many approaches have been presented to tackle uncertainty estimation and visualization including a few techniques in the context of medical image segmentation, but they typically address the uncertainty associated with a single segmentation approach [ATHL14, RPHL14, PGA13, PRH10, SMH10]. These studies show the importance of uncertainty-aware medical visualization in supporting the analysis and decision-making process. To our knowledge,

DOI: 10.2312/vcbm.20141182

there is no known method to estimate and visualize the uncertainty associated with ensemble-based image segmentation, although applying several approaches to solve the segmentation problem together is an important uncertainty source. In the presence of new measures to estimate the uncertainty associated with probabilistic image segmentations [ATHL14], we developed in this paper a new method to estimate the uncertainties associated with the results of segmentations from ensemble of classifiers. We compare different combining rules from an uncertainty point of view to explain why some combining rules tend to perform better than others. Also, we visualize the estimated uncertainties using a color mapping in image space and propose a post-segmentation correction step to re-classify erroneous pixels in the final result.

The main contributions of this paper can be summarized as: (1) Uncertainty estimation for segmentations from ensembles of classifiers and their numerical and visual assessment. (2) Combining segmentation results from ensembles of classifiers without given ground truth in form of atlases or manual segmentations. (3) Segmentation correction based on uncertainty estimation in segmentations from ensembles of classifiers.

2. Related Work

In recent years, combining ensembles of classifiers in order to improve their performance have witnessed a great attention by researchers across different fields to solve different classification problems. Kittler et al. have reviewed the combining rules and introduced a common theoretical framework of these rules [KHDM98]. Dietterich has reviewed the ensemble methods algorithms and explained from a statistical, computational, and representational point of views why ensembles can often performs better than any individual classifier [Die00]. Mignotte introduced the probabilistic Rand index (PRI) as combining strategy in a label field fusion Baysian model for image segmentation [Mig10]. Fred et al. have explored the idea of evidence accumulation for combining the results of multiple clusterings using different ways of producing data partitions in order to achieve the diversity for more improvement [FJ05]. Recently, Paci et al. proposed an ensemble-based texture classification system [PNS13].

As image segmentation plays an essential role in any medical visualization system, medical image segmentation is the most addressed problem to be solved using the ensembles of classifiers concept in the biomedical field. Several researchers exploited the concept of ensemble methods to tackle the drawbacks of the individual segmentation approaches [RM05, AMBdS09] or to estimate the accuracy of individual approaches [WZW04,LvdHK*10]. Rohlfing et al. proposed a multi-classifier framework for atlas-based image segmentation. Images from several subjects have been segmented using multiple individual atlases, or using one atlas registered with different parameter settings for different subjects. Then, the combining rules are used to produce the final segmentation [RM05]. Warfield et al. presented the STAPLE algorithm for the validation of image segmentation using a collection of segmentations produced by human raters or by automated segmentation algorithms. The algorithm uses an expectation-maximization approach in an iterative way to estimate a probabilistic ground truth. The estimated ground truth is then used for performance assessment of an automated image segmentation algorithm or for performance comparison of human raters and the automated algorithms [WZW04]. Langerak et al. have proposed the SIMPLE algorithm as improvement to the STAPLE algorithm by removing the segmentations with low accuracies from the ensemble in each iteration [LvdHK*10]. Wang et al. proposed a classifier ensemble based on the performance level estimation of the individual classifiers [WZH*09]. Artaechevarria et al. [AMBdS09] followed up on the idea by Rohlfing et al. [RM05] in combining multi-atlas-based image segmentation. They demonstrated that no fusion method outperforms others for all the regions and the performance of each method depends on the gray-level contrast characteristics of the segmented region. While combining rules that use local weights outperform global methods in segmenting high-contrast structures, the global techniques are less sensitive to noise in regions with low contrast between structures. Although these methods succeeded in improving the performance of the individual classifiers or in building probabilistic ground truth for accuracy level estimation, they suffer from several drawbacks. As pointed out by Rohlfing et al. [RM05] producing multiple atlases (also human rater segmentations for STAPLE or SIMPLE) is time consuming and tedious, such atlases are, in practice, not always available. Langerak et al. [LvdHK*10] referred to the shortcoming of atlas-based segmentation as being equivalent to the segmentation by human expert. They also discussed two important drawbacks of using multiple atlases: the large computational costs of the registration process and the shape variance in the atlas ensemble that is not always similar to that of the population from which the input image is drawn. These drawbacks may lead to the fact that the ensemble methods using atlas-based segmentations become impractical. Although Langerak et al. [LvdHK*10] tried to reduce the effects of these drawbacks by reducing the number of atlases through atlas selection procedure, the problem could not be solved completely.

In this paper, we combine the result of several unsupervised classification-based segmentations of the same input image using different segmentation approaches with acceptable accuracies. We achieve the required diversity and remove the above-mentioned drawbacks, i.e., the requirement for producing atlas-based or human-rater segmentations and for establishing the registration process.

Recently, several approaches presented robust methods to estimate and visualize the uncertainties associated with

probabilistic segmentations. These studies show how the methods can be useful for post-segmentation visual analysis and for decision-making support [ATHL14, RPHL14, PGA13, PRH10, SMH10]. Saad et al. [SMH10] introduce two-way and three-way interactive tools, which measure the difference between the first and second largest and between the second and third largest probabilities,respectively. These tools are used to highlight the uncertainty regions in the segmentation results. Praßni et al. [PRH10] use the probabilistic segmentation result of a random walker algorithm. After classifying the pixels into being certain or uncertain based on some selected probability thresholds, they use the gradient of the maximum probability of the uncertainty information to estimate the uncertain area at the boundary of segments. The approaches by Potter et al. [PGA13] and Al--Taie et al. [ATHL14] use concepts from information theory to estimate and visualize the uncertainty of a probabilistic segmentation result. Ristovski et al. [RPHL14] present a taxonomy to a wide range of uncertainty sources that encountered in the medical visualization pipeline. Yet, there is no method to estimate the uncertainties associated with ensemble of classifiers methods, although using several segmentation approaches is considered an important uncertainty source. In this paper, we exploit the recently developed uncertainty measure by Al-Taie et al. [ATHL14] to estimate the uncertainty associated with ensemble-based segmentations suitable for several combining rules. Furthermore, the proposed method does not rely on ground truth.

3. Combining segmentation ensembles

In the context of probabilistic segmentation, the output associated with each voxel x is the probability vector $P(x)$ where the i^{th} entry $P_i(x)$ of the vector denotes the probability that voxel x belongs to the segment (or class) i out of C segments (classes) such that $\sum_{i=1}^{C} P_i(x) = 1$ (i.e., $P_i(x)$ is the a posteriori probability for class i). Traditionally, the maximum a posteriori (MAP) Baysian principle is applied to obtain a hard (crisp) classification from this "soft" output.

In the framework of combining the results of L classifiers, some combining rules depend on the soft output (the a posteriori probabilities) of the individual classifiers such as the product, sum, max, min, and median rules, while other rules depend on the label field (i.e., on the hard classification output) such as the majority voting or the weighted majority voting (see [KHDM98]).

For quick reference, we rewrite these rules as defined in [KHDM98] here. To each pixel x, we assign the class that maximizes the value of the argument of the corresponding rule. Hence, we assign the following classes:

- Product Rule:

$$\arg\max_{k=1}^{C} P(k)^{-(L-1)} \prod_{j=1}^{L} P_{kj}(x),$$

where $P(k)$ is the a priori probability for class k, and P_{kj} is the a posteriori probability for class k obtained by classifier j. An issue with the product rule is that we lose the information in the product, if any of the probabilities has the value zero. Because of this issue, the product rule is not suitable for our purposes and we do not consider it further.

- Sum Rule:

$$\arg\max_{k=1}^{C} \left[(1-L)P(k) + \sum_{j=1}^{L} P_{kj}(x)\right]$$

Under the assumption of equal priors, the sum rule can be viewed as computing the average a posteriori probabilities for each class over all the classifier outputs [KHDM98] as follows:

$$\arg\max_{k=1}^{C} \frac{1}{L} \sum_{j=1}^{L} P_{kj}(x) \tag{1}$$

Using the same assumption, Kittler et al. derived the following max, min, median, and majority vote rules [KHDM98].

- Max Rule:

$$\arg\max_{k=1}^{C} \max_{j=1}^{L} P_{kj}(x) \tag{2}$$

- Min and Median Rule: for min and median rules, the $\max_{j=1}^{L}$ operator in Equation 2 is replaced with the $\min_{j=1}^{L}$ or the $\mathrm{med}_{j=1}^{L}$ operator, respectively.
- Majority Vote Rule: Applying the MAP Bayesian principle to the a posteriori probabilities P_{kj} produces a binary-valued function Δ_{kj} as

$$\Delta_{kj} = \begin{cases} 1 & \text{if } P_{kj}(x) = \max_{i=1}^{C} P_{ij}(x) \\ 0 & \text{otherwise.} \end{cases}$$

Then, under the assumption of equal priors, the majority vote rule simply counts the votes received for each class from the individual classifiers and selects as final decision the class with the largest number of votes:

$$\arg\max_{k=1}^{C} \sum_{j=1}^{L} \Delta_{kj}$$

- Weighted Majority Vote Rule: Based on some assumptions, the individual classifiers are assigned different weights (e.g., the accuracy level of the individual classifiers). In this case, the majority vote rule becomes a weighted majority vote rule

$$\arg\max_{k=1}^{C} \sum_{j=1}^{L} \omega_j \Delta_{kj},$$

where ω_j is the weight assigned to classifier j.

4. Uncertainty estimation for single classifier segmentation

Recently, Al-Taie et al. [ATHL14] proposed several forms of the normalized Kullback-Leibler divergence in addition to the normalized total-variation divergence as measures to estimate the uncertainty associated with the segmentation result of the probabilistic segmentation. Yet (to our knowledge), no known method is available to estimate the uncertainties associated with the segmentation obtained by combining ensemble of classifiers methods.

In this paper, we develop ways to estimate the uncertainties associated with the segmentation result of ensemble methods suitable for each of the above-mentioned combining rules. Our approach is based on producing probabilistic ensemble segmentations for each combining rule and we exploiting the uncertainty measure by Al-Taie et al. [ATHL14] to estimate the uncertainty associated with ensemble segmentation methods. Throughout this paper, we used only the second form of normalized Kullback-Leibler divergence uncertainty measure as it has been reported to be the measure with the best behavior among others in modeling the uncertainty. For quick reference, we rewrite the second form here: The uncertainty for voxel v using the second form of Al-Taie et al. [ATHL14] is defined by

$$U_{\overline{KLI}}(v) = 1 - \frac{D_{KL}(\mathbf{P}_v||\mathbf{P}_{max})}{D_{KL}(\mathbf{P}_{min}||\mathbf{P}_{max})},$$

where D_{KL} is the Kullback-Leibler divergence (for two probability distribution P and Q the Kullback-Leibler divergence defined as: $D_{KL}(\mathbf{P}||\mathbf{Q}) = \sum_i P_i log_2(P_i/Q_i)$). $\mathbf{P}_{min}$ represents the minimum (i.e., no) uncertainty which is obtained when one entry of the probability vector is 1 and all the others 0 (e.g., $\mathbf{P}_{min} = (1,0,\ldots,0)$), and $\mathbf{P}_{max}$ represents the maximum uncertainty which is obtained when a pixel is equally likely to belong to all segments, i.e., $\mathbf{P}_{max} = (1/c,\ldots,1/c)$. The $\mathbf{P}_v$ vector represents the segmentation probability vector for voxel v. The normalization term $D_{KL}(\mathbf{P}_{min}||\mathbf{P}_{max}))$ represents the maximum amount of randomness, which amounts to $log_2(c)$ for c segments. In case of no uncertainty, i.e., $\mathbf{P}_v = \mathbf{P}_{min}$, we obtain that $U(v) = 0$. Likewise, in case of maximum uncertainty, i.e., $\mathbf{P}_v = \mathbf{P}_{max}$, we obtain that $U(v) = 1$.

We also make use of the aggregated uncertainties proposed by Al-Taie et al. [ATHL14]. The aggregated uncertainty information can be computed based on the uncertainty measure $U(v)$ for all voxels v, where aggregation can be performed over the entire image, for a certain segment (or region), for a certain level of uncertainty, and - in case of known ground truth - for the misclassified area. The uncertain voxels can be distinguished after defining an uncertainty threshold θ, beyond which voxels are considered as uncertain. Here, we use $\theta = 0.2$, i.e., $U > 0.2$, similar to Al-Taie et al. [ATHL14] and Praßni et al. [PRH10].

The first aggregated measure we use here is the uncertainty area $UArea(s)$ of segment s, which is the number of uncertain voxels in segment s (also applicable to any region or the entire image).

The uncertainty ratio $URatio(s)$ is, then, the uncertainty area divided by the total number of voxels in segment s denoted by $|s|$:

$$URatio(s) = \frac{UArea(s)}{|s|}$$

The uncertainty mass $UMass(s)$ is defined as the sum of the uncertainties of all uncertain voxels:

$$UMass(s) = \sum_{v \in s, U(v) > \theta} U(v)$$

Finally, the uncertainty density $UDensity(s)$ is the uncertainty mass divided by the total number of voxels of segment s:

$$UDensity(s) = \frac{UMass(s)}{|s|}$$

5. Uncertainty estimation for ensemble segmentation

For ensemble segmentation, instead of using the probabilities obtained by a single classifier, the probability values (or votes for majority rule) of all classifiers compete to determine the winner as the final ensemble decision. To do so, the probability values are combined in a normalized fashion represent the probabilities assigned to the classes. The resulting probability vector that consists of these probabilities is the probabilistic ensemble segmentation. For each combining rule, the probabilistic ensemble segmentation is achieved by applying a respective combination step. The probabilistic ensemble segmentation can be fed to applying the MAP Bayesian principle to obtain a hard ensemble segmentation.

The probabilistic ensemble segmentation can be computed by using the respective combining rule without the application of the final maximum operator and normalizing the vectors. We produce probability vectors of the probabilistic version of the combining rule as follows:

$$\forall\, i,\ i \in [1 \ldots C]\ assign\ x \rightarrow P_i(x)$$

where $P_i(x)$ in the probabilistic combining rules represents the probability that voxel x belongs to class i according to the corresponding combining rule as detailed below. I.e., $P_i(x)$ represents the i^{th} entry of the probability vector of the probabilistic ensemble segmentation result at each voxel x. The values can be computed as follows:

- Probabilistic Sum Rule:

$$P_i(x) = \frac{\frac{1}{L}\sum_{j=1}^{L} P_{ij}(x)}{\sum_{k=1}^{C}\frac{1}{L}\sum_{j=1}^{L} P_{kj}(x)}$$

  The sum rule used here is the version with equal priors assumption in Equation (1) above.

- Probabilistic Max Rule:

$$P_i(x) = \frac{\max_{j=1}^{L} P_{ij}(x)}{\sum_{k=1}^{C} \max_{j=1}^{L} P_{kj}(x)}$$

- Probabilistic Min Rule:

$$P_i(x) = \frac{\min_{j=1}^{L} P_{ij}(x)}{\sum_{k=1}^{C} \min_{j=1}^{L} P_{kj}(x)}$$

- Probabilistic Median Rule:

$$P_i(x) = \frac{\mathrm{med}_{j=1}^{L} P_{ij}(x)}{\sum_{k=1}^{C} \mathrm{med}_{j=1}^{L} P_{kj}(x)}$$

- Probabilistic Majority Vote Rule:

$$P_i(x) = \frac{\sum_{j=1}^{L} \Delta_{ij}}{\sum_{k=1}^{C} \sum_{j=1}^{L} \Delta_{kj}}$$

- Probabilistic Weighted Majority Vote Rule:

$$P_i(x) = \frac{\sum_{j=1}^{L} \omega_j \Delta_{ij}}{\sum_{k=1}^{C} \sum_{j=1}^{L} \omega_j \Delta_{kj}}.$$

Since we have the probabilistic ensemble segmentation result now, it is straightforward to estimate the associated uncertainty at each voxel using any of Al-Taie et al. uncertainty measures [ATHL14] presented in Section 4. Consequently, we can also use the visualization and the aggregated uncertainties methods of Al-Taie et al. directly [ATHL14]. This can be useful for comparing the performance of different combining rules (in the sense that better rules have smaller uncertainty area and lower uncertainty density) and for comparing their behaviors from an uncertainty point of view (in term of which combining rule models better the misclassified area as uncertain region and the correctly classified area as certain region).

6. Uncertainty-based correction step

Recently, Saad et al. [SMH10] and Praßni et al. [PRH10] have suggested uncertainty-driven interaction tools as post-segmentation tools for further expert segmentation editing that are required to recover the imperfections in the segmentation result. The uncertainty information provided by these tools simplify the editing process and replace the usual manual editing which is relying largely on visual assessment. However, editing large amount of pixels or groups of pixels scattered over the image using such tools is still time consuming and impractical for large volumes. For an ensemble-based segmentation framework, the existence of large amount of information (such as the first and second most probable classes, the local statistics, and the uncertainty information) about each pixel/voxel that is derived from several classifiers in addition to the ensemble uncertainty information (available using the methods proposed above) enables us to develop an automated post-segmentation correction step (PSCS). Based on this information, the correction step reclassifies the suspicious pixels (i.e., pixels that have high probability to be misclassified) with the estimated correct class. The tool recognizes the suspicious pixels based on how high is their ensemble uncertainty level according to a certain threshold (e.g., the level of uncertainty density (U_Density) per uncertainty area (U_Area) in the image). The proposed step uses a simple set of fuzzy-logic rules that exploit the classes probabilities in the first and second most probably classes, local class statistics, and their corresponding uncertainties to estimate the correct class for suspicious pixels. The results show that this optional tool can improve the segmentation result significantly such that the performance of combining rules with very bad results is improved to compete or sometimes even outperform the best one (without PSCS). The logic used for this step is as follows:

The Algorithm

Let F, S, and N represent the sets of classes with (first) highest probability, (second) highest probability, and the highest probability observed among all (neighboring) pixels in the result of each individual classifier respectively. Hence, each of the three sets contains L members, where L is the number of classifiers in the ensemble. Let $U(d)$ represents the uncertainty associated with the possible decision d. The uncertainty associated with each member in F and S is simply the uncertainty of the corresponding classifier decision. The uncertainty for each member in the neighbor set N is computed based on neighbor classes probabilities in the corresponding individual classifier (the uncertainty for the three sets are estimated using Al-Taie et al.'s [ATHL14] methods). The logic used here in estimating the most possible class as the correct class for each pixel/voxel is the class that has the the maximum number of appearance in the three sets or the combination of the sets and if that does not lead to a clear winner, then the one that additionally has smallest uncertainty. For further explanation, we give an example for one pixel here: Let $L = 5$; $F = \{3,3,3,3,1\}$, $S = \{0,0,0,0,3\}$, and $N = \{0,0,0,0,0\}$, with the average uncertainty for the most probable class in F (i.e., 3) is $U(d = 3) = 0.6907$ and in N (i.e., 0) is $U(d = 0) = 0.4825$. Now to estimate the correct class using our rules, the competition is limited to three decisions: these are the first and second most probable classes in F (i.e., 3 and 1 in this example, respectively) and the first probable class in N (i.e., 0 in our example). The winner is the decision that appears as the first or as the second most probable class in both the FSN and FN combination sets, and additionally has smallest uncertainty. In this example, the competition is between the most probable class in F (3) and the most probable class in N (0), and the winner decision is 0 as it has the smallest uncertainty. Applying the PSCS in this example leads to correct the false decision (=3) to be equal the ground truth (=0).

7. Experimented set-up

The problem of medical image segmentation has been addressed in the framework of ensemble of classifiers methods using several atlas-based segmentations or several human rater segmentations. To avoid the drawbacks of using atlas-based segmentations mentioned above, we propose to combine the results of several automated segmentations with acceptable accuracies instead of combining the results of several atlases registered to the target image. The diversity is achieved through several unsupervised segmentations that use different approaches in the hope that the probability that the different approaches (with sufficiently high accuracy e.g.$>$ 0.80%) agree on error is very low. The individual segmentations used here are probabilistic or can be interpreted as being probabilistic such as Fuzzy c-means (FCM), since the probability-based combining rules require that the individual segmentation results to be probabilistic. Even for hard segmentation approaches such as K-means, we can produce probabilistic results easily by applying one FCM iteration initialized with the information derived from the hard segmentation result. In this paper, we use several variants of the fuzzy c-means (FCM) algorithm introduced by Bezdek [Bez81]. Fuzzy c-means is one of the most used algorithms for image segmentation [MAF99, CZ04, ZC04, CTC*06, CCZ07, AYM*02, YWC*05]. Its main advantages include a straightforward implementation, the applicability to multichannel data, its robustness in the absence of prior knowledge about cluster centers, robustness when the number of classes is known (as in the problem of human brain segmentation), and the ability to model uncertainty within the data. In addition to the above reasons, the existence of many FCM variants in the literatures that use different approaches to overcome the sensitivity of the standard FCM to noise and the ability of interpreting their results as probabilistic segmentation result encouraged us to choose a subset of these variants in our ensemble design. The FCM variants used in this paper are: (1) the modified fuzzy c-means (mFCM) [MAF99], (2) the Bias-corrected FCM (BCFCM) [AYM*02] and its improvement, (3) the Bias-corrected FCM with weighted α (BCFCM_WA) [YWC*05], (4) the spatial fuzzy c-means (sFCM) [CTC*06], (5) the kernelized fuzzy c-means (KFCM) and its spatial version, (6) the spatial kernelized fuzzy c-means (SKFCM) [CZ04, ZC04], (7) a variant of simplified fuzzy c-means methods (FCMS1) using mean filter, another such variant (8) (FCMS2) using median filter [CZ04], (9) the CLIC algorithm [LXAG09], (10) the enhanced fuzzy c-means algorithm (EnFCM) [SBSA03], (11) its variant of the fast generalized fuzzy c-means (FGFCM) [CCZ07], and finally (12) the fuzzy rule based system (FRBS) [TP98].

We present and compare the experimental results of applying the proposed methods on the synthetic image in Figure 1(a) after corrupting it with noise (Figure 1(b)) and simulated MRI brain images from [MNI97] (Figures 1(c) T2-weighted and (d) PD-weighted). The reasons for using digitally simulated images are the prior knowledge of the true tissue types (used for evaluating the results) and the control over image parameters such as mean intensity values and noise. For the synthetic image, we tried to mimic the main brain tissues of MR T1 and T2 images in a synthetic image (i.e. the background Bg, the white matter WM, the gray matter GM, and the cerebrospinal fluid CSF). We generate an example of four respective classes with complex structures as shown in Figure 1(a). We believe that our examples mimics the structures in an MR brain image better than the two-class synthetic images of [AYM*02] or [CZ04] and the four-class synthetic image of [CTC*06]. We corrupted our synthetic image with different types of noise that are common in medical data such as Gaussian, salt-and-pepper, or sinusoidal noise. In our experiments we use the synthetic image corrupted with a mixture of three types of noise as shown in Figure 1(b). Before describing the experiments, we show in Table 1 the segmentation accuracy for each of the above FCM variants (individual classifiers), which we will use later for comparison and analysis purposes. Al-Taie et al.'s [ATHL14] visualization methods will be used later to visualize the ensemble uncertainty estimated using our proposed methods for each combining rule. The visualization method is simply color-coding the uncertainty level at each pixel using a color map that assigns to value 0 (minimum uncertainty) the darkest color (dark purple) and gradually changes to brighter colors with increasing uncertainty level until value 1 (maximum uncertainty) is reached using the brightest color (yellow). The color map is shown to the right of each image. The segmentation accuracy that is used

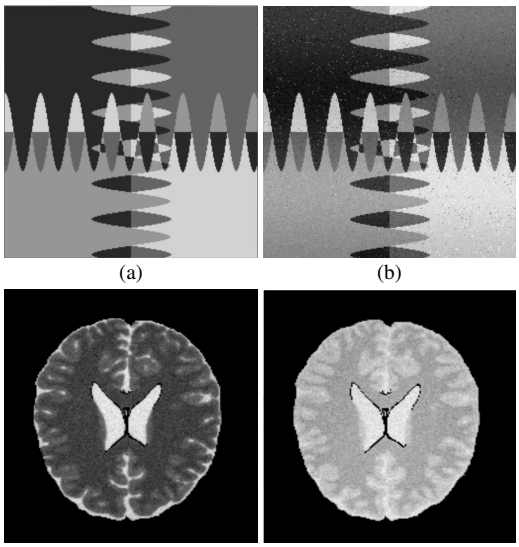


Figure 1: (a) Ground truth of synthetic image with four classes; (b) synthetic image with mixed noise; (c) simulated T2-weighted MR brain image; (d) simulated PD-weighted MR brain image.

Table 1: Segmentation accuracy (*SA*) in percentage of the synthetic image in Figure 1(b) for the modified FCM variants that are used as individual classifiers of the ensemble.

Method	SA% (low)	Method	SA% (middle)	Method	SA% (high)
BCFCM	85.33	BCFCM_WA	94.30	FCMS1	98.5565
CLIC	88.8062	EnFCM	94.5251	FRBS	98.5657
FGFCM	93.3868	FCMS2	96.5958	mFCM	98.7747
SKFCM	94.0079	KFCM	97.0963	sFCM	99.388

throughout this paper is computed by

$$SA = \frac{Number\ of\ correctly\ classified\ pixels}{Total\ number\ of\ pixels} \times 100\%.$$

8. Results and Discussion

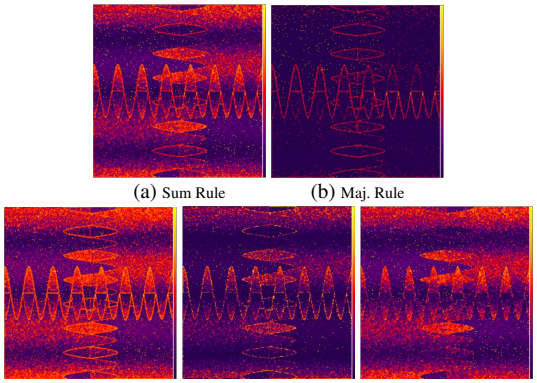
In the first group of experiments, we apply the proposed methods for estimating the uncertainty associated with the ensemble segmentation using the existing combining rules. We implement several experiments for ensembles with different sizes (3, 4, 5, and 6). For each ensemble size, different subsets from above FCM variants (the individual classifiers) are used. The strategy used in selecting the individual classifiers for each ensemble is based on grouping the classifiers according to their levels of accuracy (low, middle, and high). Then, the ensemble consists of members selected from each group separately or from a mixture of different groups. The aim of the experiments were: first, to present the uncertainty estimated using our proposed methods for each of the existing combining rules and, second, to compare the behaviors of the different combining rules from an uncertainty point of view using both the segmentation accuracy (SA) and the aggregated uncertainty measures proposed by Al-Taie et al. [ATHL14].

Figure 2 shows an example of the uncertainty estimated using the proposed methods for the existing combining rules when applied on the synthetic image with mixed noise shown in Figure 1(b). The ensemble used here is of size 4 with mFCM, FGFCM, KFCM, and FCMs1 algorithms as the individual classifiers. The visualization color-codes uncertainty levels (also for certain pixels). As the results show that the majority voting rule outperforms the weighted majority voting rule in all examples, we omit the results of the weighted majority in our experimental results.

Figure 3 shows four examples of segmentation accuracy comparison for ensembles of size four using the existing combining rules (note that the bars range over the interval [0.78,1] to better show the differences). In addition to the individual classifiers accuracies (SA of IC), each example includes comparisons of the segmentation accuracies for all rules (1) without applying the post-segmentation correction step (SA) and (2) after applying the post-segmentation correction step (SA with PSCS). While the example in Figure 3(a) uses a combination of classifiers with low accuracies, the examples in Figures 3(b), and 3(c) use classifiers with mixed levels of accuracies (low and high) and (middle and high), respectively. The last example in Figures 4(d) use classifiers with high accuracies only. Several conclusions can be drawn from this experimental results. (1) We can observe from the direct implementation of rules (i.e., first comparison) that not all combining rules outperform the individual segmentation results, but three rules most of the time are the best among all. Those three rules produce comparable results with a preference for the majority voting rule (the subsequent experiment will confirm this result). Hence, the winners are the majority voting rule (MajR- the best most of the time), the Median Rule (MedR), and the Sum Rule (SumR). If no improvement occurred (most likely because of a too low ensemble diversity) for some of three rules, then the accuracy is, at least, equal to the best individual classifier accuracy or very close to it. Figure 3(d) shows such an example where the best individual classifier has very high accuracy [99.3%], such that further improvement is difficult to achieve. (2) For the SA without vs. with correction step comparison, it can be observed that the correction step improves the performance of all the combination rules when compared to the first comparison results.

For further investigate the significance analysis of the correction step(PSCS) improvement, we implement the T-Test analysis on the set of the pixels that are distinguished as suspicious pixels and selected to apply PSCS on them. The goal was to test how significant the difference toward classification improvement for these group of pixels before and after applying the PSCS (i.e., to test whether the same group has different mean (average) scores on different -before and after- binary variables - correct/false decision-). The test is applied for all the experiments in Figure 3 using all the combining rules. As we got similar result for all experiments, we report here the result for the experiment in Figure 3(b). For large degrees of freedom (df>1000) as in our examples, with p=0.00001 on one-tailed T-Test, the calculated t-score must equal or exceed 4.271 to indicate statistical significance. The results we get from this experiment for all the combining rules is as follows: SumR(t=9.9005; df=61964; p=0.00001), MajR(t=4.51; df=23350; p=0.00001), MaxR(t=75.26; df=94300; p=0.00001), MinR(t=104.24; df=14070; p=0.00001), and MedR(t=4.38; df=39346; p=0.00001). We can observe that all results indicate the statistical significance of applying the correction step (PSCS) for improvement with the given p-value(we have only one exception from the four experiments with p=0.0015 in MajR(t=2.99; df=36028; p=0.0015)).

In the next experiment, we calculate the mean and the standard deviation of the segmentation accuracies that resulted from applying all the five combining rules for all ensembles with sizes (3, 4, and 6). The goal was to compare the performance of all rules according to their accuracies, which is similar to the first experiment but based on more comprehensive view. Figure 4 shows the comparison result



(a) Sum Rule (b) Maj. Rule (c) Max Rule (d) Min Rule (e) Med Rule

Figure 2: Uncertainty visualization of an ensemble segmentation result of the synthetic image in Figure 1(b) using the proposed uncertainty methods for each combining rule.

for the five combining rules. The result of the comparison confirms the conclusions drawn from the first experiment above. Again, the the majority rule MajR performed best, but the Median Rule MedR, and the Sum Rule SumR produced comparable results. Now, we compare the behaviors of these rules from an uncertainty point of view in sense that the winner is the rule, which best models the uncertainty. Hence, the winner shall have the smallest uncertainty ratio, shall recognize the misclassified area as uncertain and the correctly classified area as certain, and shall concentrate the high uncertainty density inside the misclassified area while keeping it low outside. Figure 5(a) shows that most rules concentrate the high uncertainty density inside the misclassified area (M_Area), while Figures 5(b) shows that the majority voting rule (MajR) is the best among the competing rules in modelling the misclassified area as uncertain while modeling the correctly classified area as certain (with low uncertainty density, as well).

In summary, the majority voting is the best in achieving the high segmentation accuracies and the minimum false positive and minimum false negative ratios in modeling the uncertainties. It is clear that better uncertainty modeling is important for post segmentation analysis, as it helps the user to focus on more accurate uncertain areas or on correcting fewer problematic pixels.

In addition to the synthetic image, we conduct similar experiments on the simulated T2- and PD-weighted MR brain images corrupted with 5% Gaussian noise and 20% intensity inhomogeneity shown in Figures 1(c) and (d), respectively. Figure 6 shows the visualization of ensemble segmentation uncertainty estimated for all rules on the T2-weighted image using the proposed uncertainty estimation methods. The ensemble consists of four members (CLIC, BCFCM_WA, FCMs2, and EnFCM) as individual classifiers. Again, we can observe that the majority voting has the smallest uncertainty area and the lowest uncertainty density

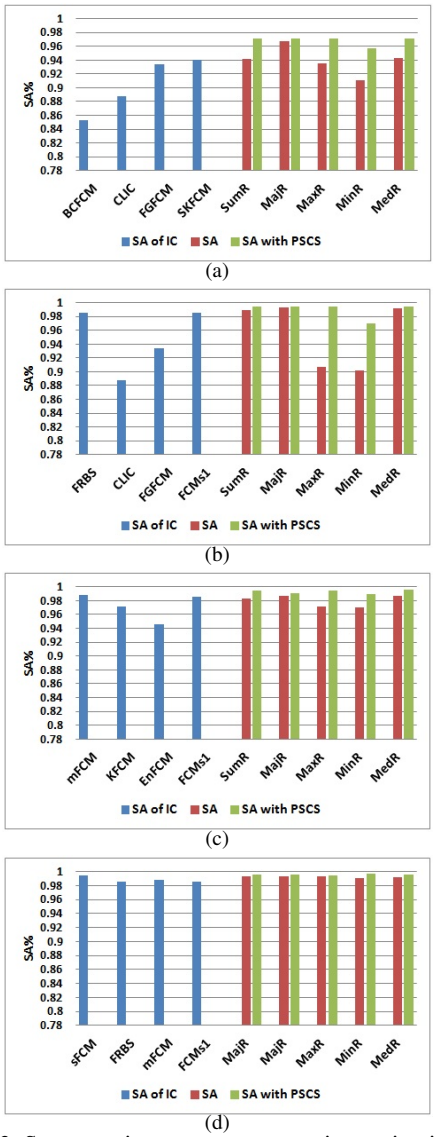


Figure 3: Segmentation accuracy comparison using individual classifiers (blue) and ensembles with the different combining rules without (red) and with post-segmentation correction step (green) on the synthetic image in Figure 1(b).

among the best 3 rules in terms of segmentation accuracy, see Figure 7(a). Figure 7(a) for the T2-weighted and 7(b) for the PD-weighted image shows the segmentation accuracy comparison for all rules before and after applying the post-segmentation correction step (PSCS). In Figure 7, we can observe that the PSCS improves the results in general.

9. Conclusions

In recent years, the concept of combining several classifiers to produce classification accuracy that outperforms the ac-

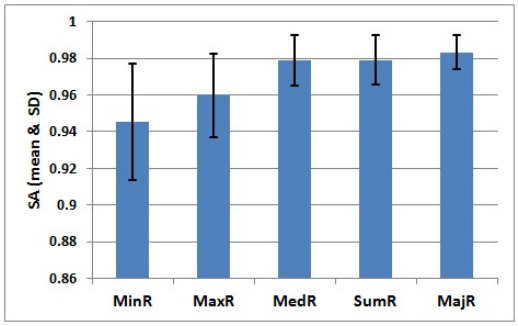
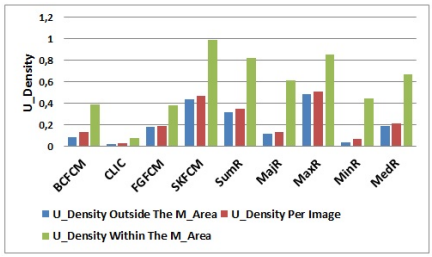
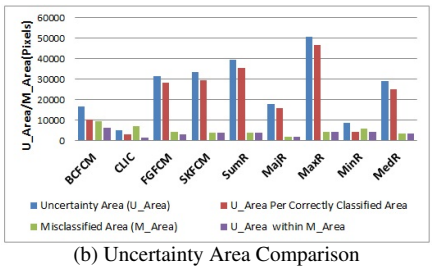


Figure 4: Mean and standard deviation of the segmentation accuracy comparison for the different combining rules on the synthetic image with mixed noise (Figure 1(b)).

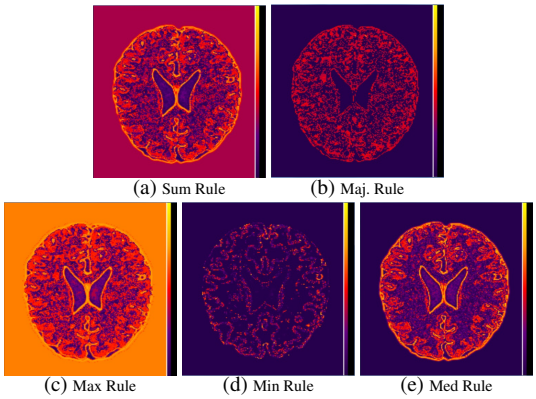


(a) Uncertainty Density Comparison



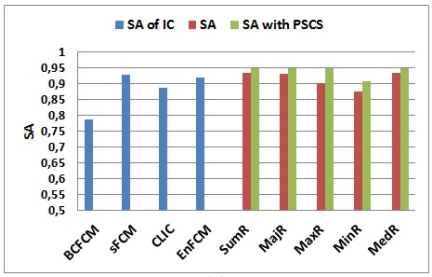
(b) Uncertainty Area Comparison

Figure 5: Uncertainty density (U_Density) and uncertainty area (U_Area) comparison using the different combining rules on the synthetic image with mixed noise in Figure 1(b).

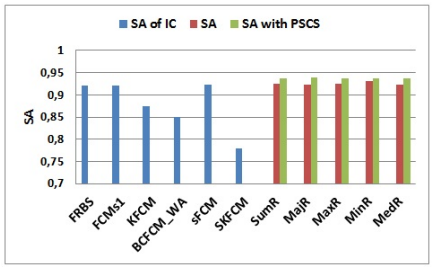


(a) Sum Rule (b) Maj. Rule (c) Max Rule (d) Min Rule (e) Med Rule

Figure 6: Uncertainty visualization of an ensemble segmentation result of the simulated T2-weighted MR image in Figure 1(c) using the proposed uncertainty methods for each of the combining rules.



(a)



(b)

Figure 7: Segmentation accuracy comparison using the different combining rules on the simulated MR images in Figure 1; (a) for T2-weighted and (b) for PD-weighted.

curacy of individual classifiers attracted the attention of researchers in the biomedical field to improve the segmentation accuracy or to evaluate the performance level of the individual segmentations. Most of these attempts are based on combining several atlas-based segmentations. On the other hand, several approaches have been developed to estimate the uncertainty associated with individual probabilistic segmentation results. We presented the first approach that is able to estimate and visualize the uncertainty associated with ensemble of classifiers segmentation. In this paper, first, we replaced atlas-based segmentations by unsupervised automated segmentations in the ensemble design to avoid the drawbacks associated with atlas generation and the required registration process. Second, we developed ways to estimate the uncertainties associated with ensemble segmentations for each of the existing combining rules. The estimated uncertainty is suitable to be visualized or to be used as a basis for the aggregated uncertainty measures using the recently proposed uncertainty visualization and aggregated uncertainty methods for single segmentation. These methods have been proven to be useful for further numerical and visual analysis [SMH10, ATHL14]. In addition, we show in this paper that the uncertainty information is not only important for segmentations performance comparison and post-segmentation analysis but can also be helpful for automatic segmentation correction within the ensemble segmentation environment. This is achieved by the post-segmentation correction step which is an uncertainty- and statistics-based step for an automatic correction of pixels that have been identified as suspicious. Finally, we compared the existing ensemble combiner using both the segmentation accuracy and the

aggregated uncertainty measure. In this sense, we show that the majority voting is the best combining rule, as it achieved high segmentation accuracies and minimum false positive and false negative ratios in modeling the uncertainties.

References

[AMBdS09] ARTAECHEVARRIA X., MUÑOZ-BARRUTIA A., DE SOLORZANO C. O.: Combination strategies in multi-atlas image segmentation: Application to brain mr data. *IEEE Transactions Medical Imaging 28*, 8 (2009), 1266–1277. 1, 2

[APNY13] AZMI R., PISHGOO B., NOROZI N., YEGANEH S.: Ensemble semi-supervised frame-work for brain magnetic resonance imaging tissue segmentation. *Journal of Medical Signals and Sensors 3*, 2 (Apr. 2013), 94–106. 1

[ATHL14] AL-TAIE A., HAHN H. K., LINSEN L.: Uncertainty estimation and visualization in probabilistic segmentation. *Computers & Graphics 39*, 0 (2014), 48 – 59. 1, 2, 3, 4, 5, 6, 7, 9

[AYM*02] AHMED M. N., YAMANY S. M., MOHAMED N., FARAG A. A., MORIARTY T.: A modified fuzzy c-means algorithm for bias field estimation and segmentation of mri data. *IEEE Transactions on Medical Imaging 21*, 3 (March 2002), 193–199. 6

[Bez81] BEZDEK J.: Pattern recognition with fuzzy objective function algorithms. *Plenum, NY.* (1981). 6

[CCZ07] CAI W., CHEN S., ZHANG D.: Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. *Pattern Recognition 40*, 3 (2007), 825–838. 6

[CTC*06] CHUANG K.-S., TZENG H.-L., CHEN S., WU J., CHEN T.-J.: Fuzzy c-means clustering with spatial information for image segmentation. *Computerized Medical Imaging and Graphics 30*, 1 (2006), 9 – 15. 6

[CZ04] CHEN S., ZHANG D.: Robust image segmentation using fcm with spatial constraints based on new kernel-induced distance metric. *IEEE Trans. on System, Man and Cybernetics-Part B 34*, 4 (2004), 1907–1916. 6

[Die00] DIETTERICH T. G.: Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems* (London, UK, UK, 2000), Springer-Verlag, pp. 1–15. 1, 2

[FJ05] FRED A., JAIN A.: Combining multiple clusterings using evidence accumulation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 27*, 6 (Jun 2005), 835–850. 1, 2

[KHDM98] KITTLER J., HATEF M., DUIN R. P. W., MATAS J.: On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 20*, 3 (Mar 1998), 226–239. 1, 2, 3

[Kun04] KUNCHEVA L. I.: *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004. 1

[LvdHK*10] LANGERAK R., VAN DER HEIDE U. A., KOTTE A. N. T. J., VIERGEVER M. A., VAN VULPEN M., PLUIM J. P. W.: Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (simple). *IEEE Transactions Medical Imaging 29*, 12 (2010), 2000–2008. 1, 2

[LXAG09] LI C., XU C., ANDERSON A., GORE J.: Mri tissue classification and bias field estimation based on coherent local intensity clustering: A unified energy minimization framework. In *Information Processing in Medical Imaging*, vol. 5636 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2009, pp. 288–299. 6

[MAF99] MOHAMED N., AHMED M., FARAG A.: Modified fuzzy c-mean in medical image segmentation. In *Proceedings IEEE International Conference on Acoustics, Speeeh, and Signal Processing, 1999, Piscataway, NI USA* (1999), vol. 6, pp. 3429–3432 vol.6. 6

[Mig10] MIGNOTTE M.: A label field fusion bayesian model and its penalized maximum rand estimator for image segmentation. *IEEE Transactions on Image Processing 19*, 6 (2010), 1610–1624. 1, 2

[MNI97] MNI: Brainweb, simulated brain database. Available at http://www.bic.mni.mcgill.ca/brainweb/, access time: on November 2012, 1997. 6

[PGA13] POTTER K. C., GERBER S., ANDERSON E. W.: Visualization of uncertainty without a mean. *IEEE Computer Graphics and Applications 33*, 1 (2013), 75–79. 1, 3

[PNS13] PACI M., NANNI L., SEVERI S.: An ensemble of classifiers based on different texture descriptors for texture classification. *Journal of King Saud University - Science 25*, 3 (2013), 235 – 244. 2

[PRH10] PRASSNI J., ROPINSKI T., HINRICHS K.: Uncertainty-aware guided volume segmentation. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1358–1365. 1, 3, 4, 5

[RM05] ROHLFING T., MAURER C. R. J.: Multi-classifier framework for atlas-based image segmentation. *Pattern Recognition Letters 26*, 13 (2005), 2070 – 2079. 1, 2

[RPHL14] RISTOVSKI G., PREUSSER T., HAHN H. K., LINSEN L.: Uncertainty in medical visualization: Towards a taxonomy. *Computers & Graphics 39*, 0 (2014), 60 – 73. 1, 3

[SBSA03] SZILÁGYI L., BENYO Z., SZILÁGYI S., ADAM H.: Mr brain image segmentation using an enhanced fuzzy c-means algorithm. In *Proceedings of the 25th Annual International Conference of the IEEE* (17-21 Sept. 2003), vol. 1, Engineering in Medicine and Biology Society, pp. 724 – 726. 6

[Sha99] SHARKEY A. J. C.: *Combining artificial neural nets: ensemble and modular multi-net systems*. Springer-Verlag, New York, 1999. 1

[SMH10] SAAD A., MÖLLER T., HAMARNEH G.: Probexplorer: Uncertainty-guided exploration and editing of probabilistic medical image segmentation. *Computer Graphics Forum 29*, 3 (2010), 1113–1122. 1, 3, 5, 9

[TP98] TOLIAS Y., PANAS S.: On applying spatial constraints in fuzzy image clustering using a fuzzy rule-based system. *Signal Processing Letters, IEEE 5*, 10 (1998), 245–247. 6

[WZH*09] WANG W., ZHU Y., HUANG X., LOPRESTI D. P., XUE Z., LONG L. R., ANTANI S., THOMA G. R.: A classifier ensemble based on performance level estimation. In *ISBI* (2009), IEEE, pp. 342–345. 2

[WZW04] WARFIELD S. K., ZOU K. H., WELLS W. M.: Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. *IEEE Transactions Medical Imaging 23* (2004), 903–921. 1, 2

[YWC*05] YUAN K., WU L., CHENG Q., BAO S., CHEN C., ZHANG H.: A novel fuzzy c-means algorithm and its application. *International Journal of Pattern Recognition and Artificial Intelligence 19*, 8 (2005), 1059–1066. 6

[ZC04] ZHANG D., CHEN S.: A novel kernelised fuzzy c-means algorithm with application in medical image segmentation. *Artificial Intelligence in Medicine 32*, 1 (2004), 37–50. 6