Atlas-based segmentation consists of a powerful baseline for brain tissue segmentation when compared to an machine learning-based approach

Cyril Albrecht Valerio Mollet Quentin Savary

***Abstract*—Medical image analysis usually uses atlas-based methods to segment brain tissues. Machine learning based methods could replace those older ones in the near future. Both types of segmentation have been compared, as well as the influence of affine or non-rigid registration before-hand. Only five tissues of the brain have been been segmented with 20 training images and 10 test images. The results are compared with the Dice score and Hausdorff distance.**

Atlas-based segmentations give better results with smaller brain parts (amygdala, hippocampus, and thalamus). Geometrically complex parts (grey and white matter) have better results when segmented with a machine learning method. The best overall solution is obtained with a locally weighted atlas-based segmentation.

***Index Terms*—Random forest segmentation; Atlas-based segmentation; Multi-atlas fusion; Non-rigid registration; Affine registration.**

Introduction

Medical images are widely used for all kinds of diagnoses such as tumour detection. Manual tissue segmentation is tedious and repetitive tasks for the medical personnel. A medical image analysis (MIA) pipeline is a tool that can reduce the time needed to perform said tasks. The current trend of machine learning offers new possibilities for this tool, especially in the classification step. This work aims to test the following hypothesis: *Atlas based segmentation consists of a powerful baseline for brain tissue segmentation when compared to an ML based approach*. To accept or reject this hypothesis, different atlas-based segmentations were implemented and compared with machine learning. In this project, four variations of atlas-based segmentation and one machine learning algorithm are compared. Those are majority voting, global and local weighted voting and shape based averaging for atlas-based segmentation versus random forest for the machine learning segmentation. The comparison is done with a Dice score and a Hausdorff distance, two widely used measurement tools. The same set of anonymized images have been used for the training and testing of the different methods. The registration quality have an obvious impact on an atlas-based segmentation performance. Intuitively, if the target image and the atlas image are the same, the only source of error is the segmentation. An non-rigid registration allows to reduce the dissimilarity between the target image and the atlas. Thus two kinds of registration have been implemented and compared: affine and non-rigid.

Related work

Organ segmentation is a big concern in medical image processing, and it exists numerous works on the subject. It has been shown that multi-atlas segmentation performs better than single-atlas for brain tissue segmentation [1]. Thus, no single-atlas method was used in this work. Multi-atlas fusion, including majority voting, local and global weighting strategies, and the combination of those methods are studied in [2]. Shape-based averaging method for multi-atlas fusion has been introduced by [3]. Each of the atlas-based methods implemented in this work for segmentation are described in detail in those papers.

Method

In order to compare the performance of an altas-based segmentation versus a machine learning approach, both methods have been applied to the same set of images. Five brain tissues have been segmented: grey and white matter, amygdala, hippocampus, and thalamus. The pipeline used as a basis for the algorithms implementation was intended for a machine learning segmentation approach. For the atlas-based segmentation implementation, the training data from the provided pipeline have been used to generate different atlas according to the methods described below. Since the training data are used to be compiled multi-atlas fusion, it will be refer to the atlas as the result of the fusion from the training images. This work includes the comparison of four different multi-atlas segmentation methods versus the provided machine learning approach described in the previous chapter. The data are composed of an atlas (N = 1), a training subset (N = 20) and a testing subset (N = 10). Each of those data contain a T1-weighted image (T1w), a T2-weighted image (T2w), a ground truth image, and a brain mask used for skull stripping.

MIA pipeline

The medical image analysis pipeline consists of five distinct steps: registration, pre-processing, feature extraction, classification and post-processing. It is illustrated in the fig. 1.

Registration matches an acquired image to a reference one, usually provided from an atlas. This image transformation can be affine as well as non-rigid, which allows local deformations. The pre-processing is used to improve the quality of an image. It uses different kind of filters and masks to get a higher quality image or remove unnecessary data. Intensity normalization is

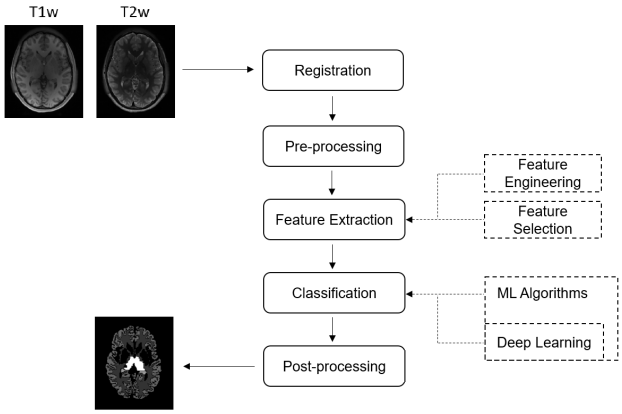

Fig. 1. Schematic of the medical image analysis pipeline.

also used, especially if machine learning is involved. Feature extraction tries to find position of anatomical landmarks. Contours and corners can be found with rather simple algorithms. Classification means to decide of the anatomical type for each voxel. This can be done in various ways, decisions trees (also called random forests) in machine learning or by growing regions algorithms. Finally, post-processing gives a cleaner final segmentation. It will remove small groups of voxels that should not correspond to any real anatomical part. As an example, a human body can only have one liver or two kidneys of similar size.

Registration

For the registration step, two different solutions have been compared. The first is affine transformation, which which allows translation, rotation and scaling. The second one is non-rigid, which transforms the volumes locally. This can lead to a better matching between the input image and the atlas. For the affine registration of two volumes, the volume to be registered is moved relative to the fixed volume by rotation, translation or scaling. After each displacement a similarity of the volumes is measured. This parameter is minimized by moving the volume further until a minimum is reached. The same is true for a non rigid registration, but there is an additional tuning factor which allows local displacements, rotations and scaling. We used the *Simple Elastix*[1] library to compute the non-rigid transformation. This library uses a B-spline registration to register the two volumes non rigidly.

Segmentation

For the segmentation step, a total of five methods have been compared. The first one is machine learning based, essentially a random forest with 10 estimators of maximum depth of 40. These parameters were found step by step with a manual optimization. The other four methods were atlas-based, where mainly the weight function has been modified. The simplest one is when all images have the same weight. To reduce potentially worse ground truth inputs, we used global and local weights. In order to make the segmentation smoother, an additional method called Shape based averaging was implemented.

Machine Learning: The machine learning approach was used in the previously described medical image analysis pipeline. The images are preprocessed with skullstripping, normalization and registration. Intensity and gradient intensity features of the training dataset are used for the training of the random forest. For a matter of time, the parameters optimization has been performed using a brute force approach. First the optimal depth is determined by keeping the number of estimator constant, varying the depth of the random forest and measuring the mean Hausdorff distance and Dice coefficient over the testing dataset. Once the optimal depth is known, the number of estimator is varied, and the same measurement are performed.

Majority Voting: Majority voting is the simplest method for multi-atlas based segmentation. The volumes of all atlas are registered to the registration atlas. The target being segmented is also registered to this atlas. Next, for each voxel of the target, the same voxel of each atlas is used to vote which label should apply. The label with the most votes wins and is segmented accordingly. In fig. 2. this method is illustrated with 3 atlases, for our method we used 20 atlases. This method will obviously struggle with variable brain parts. Since very different atlases compared to the target will have the same weights as very similar atlases. This problem can be reduced with a weighting of the atlas. This is the most straight forward method to perform a multi-atlas fusion. Since the atlas images are registered (indirectly) to the "target" image, each voxel v correspond to the same voxel in all other images. The intuitive way to understand this method is that each voxel's label is assign according to maximum number of atlas vote for this label. Explicitly, for $l(v^i)$ being the label of a given voxel v^i this can be expressed as $L = l_1, ..., l_D$:

$$Y^i = max[\sum]$$

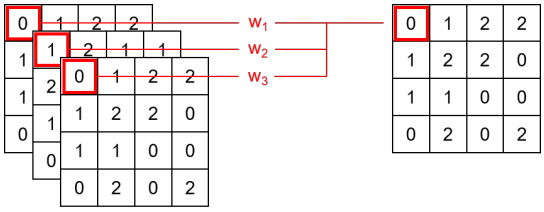

Fig. 2. Schematic representation of majority voting.

Global Weighted Voting: Global weighted voting is an atlas-based segmentation method that takes into account individual variations of the target being segmented with the available atlases. Voting of atlases with high similarity to the target being segmented are weighted higher. In our case, the T1w and T2w volumes of 20 atlases were compared with the T1w and T2w volumes of the target after a registration to an atlas volume. By measuring the mean square differences (MSD) averaged over both T1w and T2w volumes, each

[1]https://github.com/SuperElastix/SimpleElastix

atlas was given a corresponding weight. The weights were distributed by a soft maximization function so that the sum of all weights would equal one. This simplifies the calculation of the probability of the prediction. To get the segmentation, a majority voting is performed with the globally calculated weights.

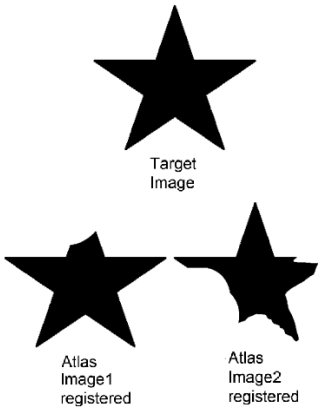

Fig. 3. Example for showing the problem of global weighted voting [2].

The problem of this method is that an atlas is very similar in most parts but local is very different compared to the target. The weighting is still high due to the overall similarity. The opposite can also be the case, so that the overall similarity is very low but the atlas still has a very high correlation with the target in local parts. Thus, this similarity is not accounted for by the global weighting. To give this local similarities weighting, a local weighted voting can be performed. This problem is well illustrated in fig. 3.

Local Weighted Voting: The segmentation method local weighted voting takes into account local similarities of atlases with the target being segmented. Thereby, like in global weighted voting, the T1w and T2w volumes of the target and the atlases are registered to a registration atlas. Now each voxel of the atlases gets its own weighting based on the local similarity with the target. The similarities can be obtained by calculating the cross correlation, squared differences or mutual information. (see fig. 4 for better overview) Of course, there are other similarity measures, but they have not been considered. Because the squared difference measurements give the best results for segmentation, this method was chosen [**?**]. Another factor is the size of the kernel which is compared locally with the target. This is not quite trivial, if the kernel is too large, a similar problem arises as with global weighted voting (see fig. 3). If the kernel is too small, structures can be compared less effectively. To simplify the problem, a kernel of size 1 was chosen. Now each atlas has a whole matrix of weights instead of one similarity score, as in the global weighted voting.

Shape-based Averaging: Shape-based averaging is a method introduced by [3]. This algorithm aims to reduce the number of unconnected region in the segmentation. Unlike the previous algorithm, this fusion method is not based on votes. The signed

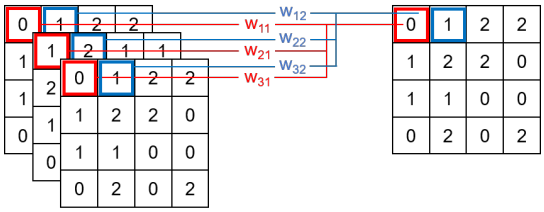

Fig. 4. Schematic representation of local weighted voting.

Euclidian distance map of each individual labels is summed up for each atlas. For each voxel, the label minimizing this sum is attributed. Formally, for D_i being the signed Euclidian distance of a specific voxel v for a label l in the atlas i.

$$l(v) = \arg\min_{l} \sum_{i=1}^{N} D_i(v, l)$$

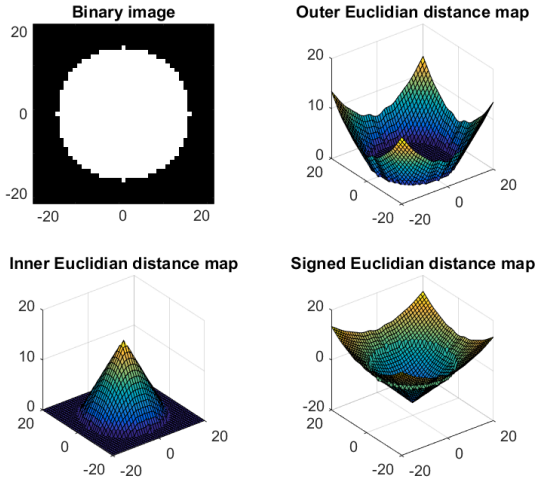

Fig. 5. Illustration of the computation steps of the signed Euclidian distance map from a simple binary image (adapted from [3])

Performance assessment: This work focused on the precision aspect of the algorithm and not on the time consumption or complexity. To measure the overall performance, several evaluation metrics can be used. The Dice coefficient tells how good the computed result and the ground truth do overlap. It goes from 0 if there is no overlap at all to 1 being exactly same segmentation. The Hausdorff distance is another metric. It measures the maximum distance from the computed segmentation contour to the ground truth one. There are many more metrics that can be used to test specific characteristics of a computed segmentation.

Results

The results are separated for registration and segmentation approaches. All were obtained with 20 training images and 10 test images. The results for each case are measured with the Dice score and the Hausdorff distance. They represented with boxplots to show the mean as well as the variability.

Random forest parameters optimization

For a fixed number of estimator of 10, the random forest approach almost reaches its optimal performance around 20. As the time consumption of the algorithm has not been considered in this work, a margin has been tolerated and the selected number of estimators was 40.

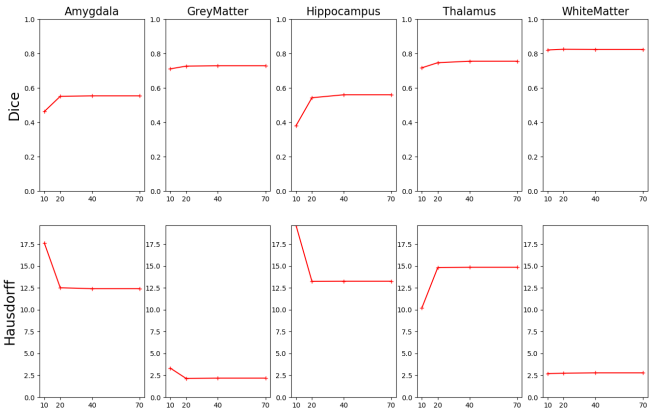

Fig. 6. Random forest mean Hausdorff distance and dice coefficient for a constant number of estimator of 10.

With a fixed number of estimator, except for the thalamus and hippocampus regions, the random forest shows the best results for a depth of 10.

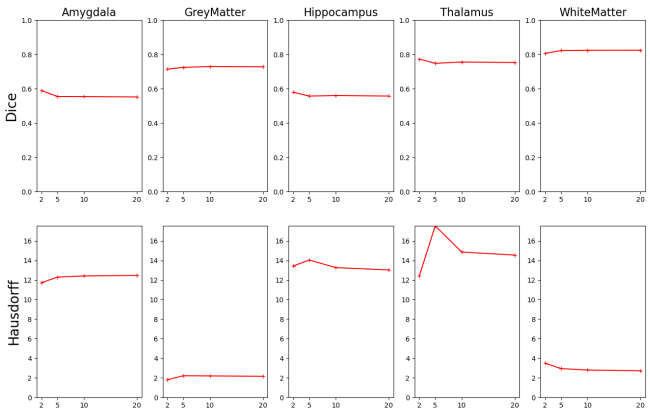

Fig. 7. Random forest mean Hausdorff distance and dice coefficient for a constant depth of 40.

Registration

Two registration methods have been implemented: affine (Aff) and non-rigid (NR). The first is easier to implement and faster to execute, the second one should match the atlas image better as more deformations are possible.

As can be seen in figure **??**. non-rigid registration in the gray matter region is better than affine registration because it can compensate for local changes. What is additionally important is that the non-rigid registration must be reversed after segmentation. Since it does not correspond to the truth. To compare the effect of the different registration methods,

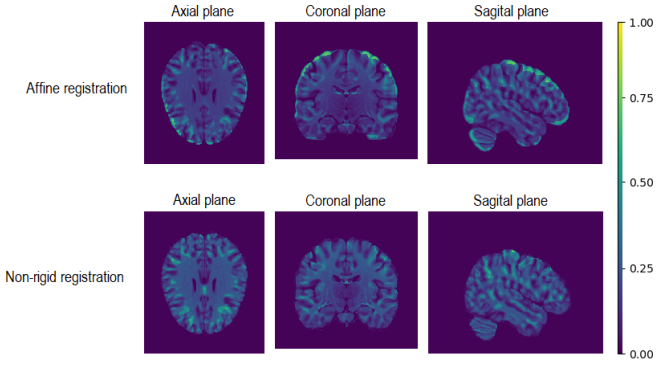

Fig. 8. Comparison between affine and non-rigid registration by similarity measurement of squared differences of the registered volume to the fixed volume, high values mean larger differences.

the pipeline was run for both methods with the segmentation methods majority voting (MV) and machine learning (ML). The results of this comparison are represented in figure 9.

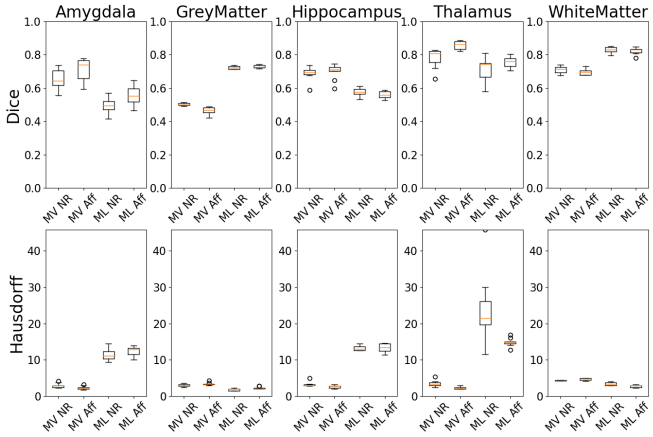

Fig. 9. Dice and Hausdorff distance result for majority voting (MV) and machine learning (ML) segmentation preceded by non-rigid (NR) or affine (Aff) registration.

From the graphic, we can see no clear advantage of the non-rigid registration over the affine one. The difference is too small and has too much variability to choose one registration over the other one.

Segmentation

Five segmentation methods have been implemented: majority voting (MV), global weighted (GW), local weighted (LW), shape based averaging (SBA) and machine learning (ML). All except the last one are atlas-based. The results with affine registration are represented in the figure 10.

For the smaller parts of the brain (Amygdala, Hippocampus and Thalamus), all atlas-based segmentation methods have similar results. Machine learning has more difficulties with those brain parts. On the other hand, it gives better results with grey and white matter. Local weighted atlas significantly performs better than any other atlas-based segmentation meth-

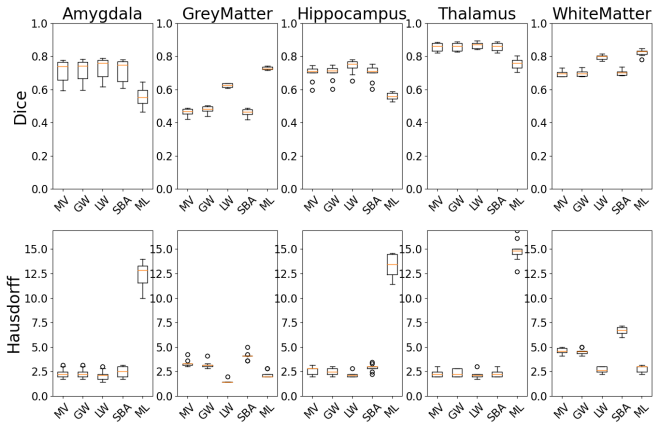

Fig. 10. Dice and Hausdorff distance result for majority voting (MV), global weighted (GW), local weighted (LW), shape based averaging (SBA) and machine learning (ML) segmentations preceded by affine registration.

ods for those brain parts. It is the best overall segmentation methods.

The same segmentation methods with non-rigid registration are represented in figure 11.

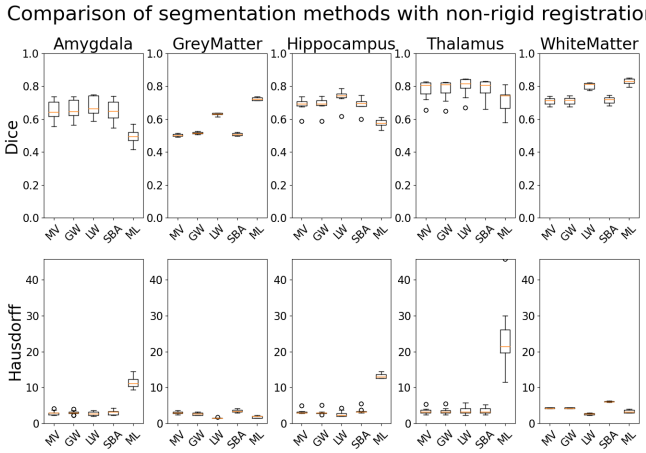

Fig. 11. Dice and Hausdorff distance result for majority voting (MV), global weighted (GW), local weighted (LW), shape based averaging (SBA) and machine learning (ML) segmentations preceded by non-rigid registration.

In general, the results are very similar to the one with affine registration. Machine learning seems to struggle even more with the Thalamus, as the Hausdorff distance suggests.

Discussion

The random forest parameters optimization could have been performed using a design of experiment instead. But this work focused more on the implementation of atlas-based segmentation methods.

The four atlas-based segmentation perform better with small brain parts, such as the amygdala, hippocampus and thalamus. This can be explained due to their relative static position and low distortion.

For machine learning segmentation, grey and white matter have better results as atlas-based ones. Machine learning is able to interpret the subtle geometrical changes of those. This is clearly visible in the figure 12.

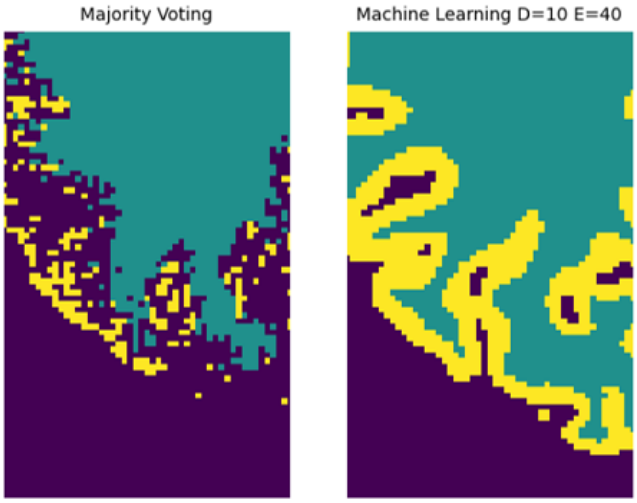

Fig. 12. Comparison between majority voting and machine learning segmentation for grey (yellow) and white matter (cyan).

When comparing all four atlas-based segmentation algorithms, there is no major difference between majority voting, global weighted voting and shape based averaging. Local weighted voting significantly improves with grey and white matter, as seen in the results. The fig. 13 shows this case precisely. The edges of white matter and the grey matter are much more realistic and complete for the global weighted voting segmentation. The other three segmentations are very sparse at the frontier between grey and white matter.

Overall, the best performing algorithm of all those tested is the local weighted voting segmentation. The amygdala, hippocampus and thalamus have similar score as the other three atlas-based methods. The grey and white matter have a score close to the one with machine learning segmentation.

Conclusion

At the end of the project it can be said that there are many different atlas based segmentation methods. Majority voting, global weighted voting, local weighted voting and shape based averaging were implemented and analyzed in detail. All atlas based methods perform better than machine learning for the smaller brain regions. However, for brain regions that differ from person to person, such as gray matter, atlas-based methods perform worse. With the locally weighted method, the dice can be increased a little. But it cannot reach the performance of machine learning. Besides, it shows that the registration methods do not have a big impact on the final result. Non-rigid registration does not perform better than affine registration. The hypothesis that Atlas based segmentation consists of a powerful baseline for brain tissue segmentation when compared to an ML based approach can be rejected according to the obtained findings. It can be used as a support to improve machine learning. Perhaps by optimizing only the gray and white matter with machine learning.

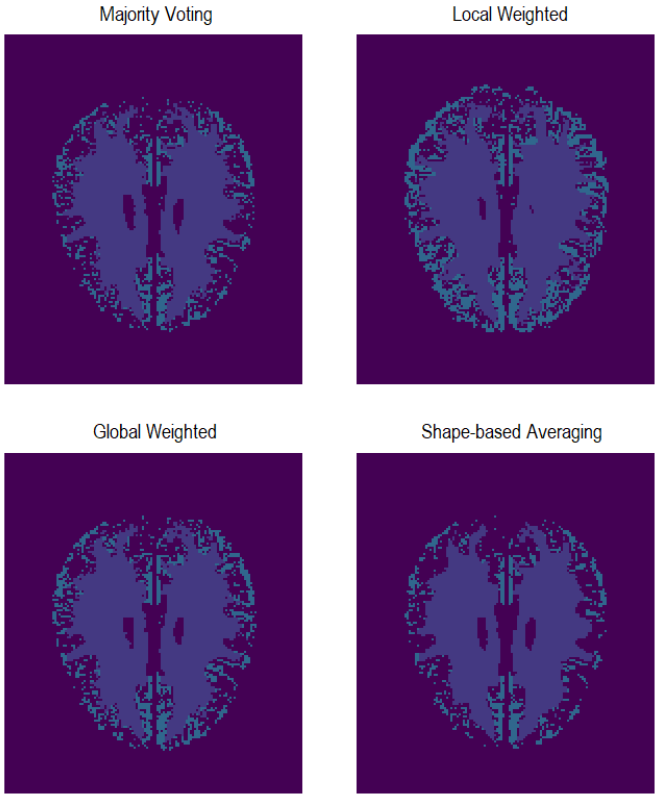

Fig. 13. Comparison between all 4 atlas-based segmentations for grey (cyan) and white matter (blue).

REFERENCES

[1] A. Klein, B. Mensh, S. Ghosh, J. Tourville, and J. Hirsch, "Mindboggle: Automated brain labeling with multiple atlases," *BMC Medical Imaging*, vol. 5, 10 2005.

[2] X. Artaechevarria, A. Muñoz-Barrutia, and C. O. de Solórzano, "Combination strategies in multi-atlas image segmentation: Application to brain mr data," *IEEE Transactions on Medical Imaging*, vol. 28, pp. 1266–1277, 2009.

[3] T. Rohlfing and C. R. Maurer, "Shape-based averaging," vol. 16, pp. 153–161, 2007.