# MOrphologically-aware Jaccard-based ITerative Optimization (MOJITO) for Consensus Segmentation

Dimitri Hamzaoui, Sarah Montagne, Raphaele Renard-Penna, Nicholas Ayache, Hervé Delingette

# MOrphologically-aware Jaccard-based ITerative Optimization (MOJITO) for Consensus Segmentation

Dimitri Hamzaoui[1], Sarah Montagne[2], Raphaële Renard-Penna[2], Nicholas Ayache[1], and Hervé Delingette[1]

[1] Université Côte d'Azur, Inria, Epione Project-team, Sophia Antipolis, France
dimitri.hamzaoui@inria.fr
[2] Radiology Department, CHU La Pitié Salpétrière/Tenon, Sorbonne Université, Paris, France

**Abstract.** The extraction of consensus segmentations from several binary or probabilistic masks is important to solve various tasks such as the analysis of inter-rater variability or the fusion of several neural network outputs. One of the most widely used method to obtain such a consensus segmentation is the STAPLE algorithm. In this paper, we first demonstrate that the output of that algorithm is heavily impacted by the background size of images and the choice of the prior. We then propose a new method to construct a binary or a probabilistic consensus segmentation based on the Fréchet means of Jaccard distances which make it totally independent of the image background size. We provide a heuristic approach to optimize this criteria such that a voxel's class is fully determined by its morphological distance, the connected component it belongs to and the group of raters who segmented it. We compared extensively our method on three datasets with the STAPLE method and the naive segmentation averaging method, showing that it leads to consensus masks of intermediate size between Majority Voting and STAPLE and to different posterior probabilities than those methods.

**Keywords:** Image segmentation, consensus, distance

## 1 Introduction

The fusion of several segmentations into a single consensus segmentation is a classical problem in the field of medical image analysis related to the need to merge multiple segmentations provided by several clinicians into a single "consensus" segmentation. This problem has been recently revived by the development of deep learning and the multiplication of ensemble methods based on neural networks [10]. One of the most well-known methods to obtain a consensus segmentation is the STAPLE algorithm [22], where an Expectation-Maximization algorithm is used to jointly construct a segmentation consensus and to estimate the raters' performances posed in terms of sensitivities and specificities. The seminal STAPLE method [8] creating a probabilistic consensus from a set

of binary segmentations was followed by several follow-up works. For instance, Asman *et al.* [2] replaced global indices of performance by spatially dependent performance fields and Commowick *et al.* [6] combined STAPLE with a sliding window approach, in order to allow spatial variations of rater performances. Another improvement consisted in introducing the original image intensity information [3]. Many other fusion methods were proposed based on generative models [4, 21], label fusion with intensity images [17] or simple majority voting (MV) [16, 1]. The STAPLE method is based on simple probabilistic models, is widely applicable [22, 8] but it suffers from several limitations, some of them already addressed in the literature [2, 6, 3] and some, to the best of our knowledge, never raised before.

In this article, we first analytically characterize the dependence of the STAPLE algorithm on the size of the background image and the choice of prior consensus probability. We then introduce an alternative consensus segmentation method, coined MOJITO, which is based on the minimization of the squared distance between each binary segmentation and the consensus. By adopting the Jaccard distance between binary or probabilistic shapes, the consensus is thus posed as the estimation of a Fréchet mean which is independent from the size of the background image. We show that the adoption of specific heuristics based on morphological distances during the optimization allows to provide a novel binary or probabilistic globally consistent consensus method which creates masks of intermediate size between Majority Voting and the STAPLE methods.

## 2 STAPLE dependence on background size and prior

### 2.1 STAPLE dependence on background size

The STAPLE consensus method [22] takes as input a set of $K$ binary segmentations $\mathcal{S} = \{S^1, ..., S^K\}$ of size $N$ and produces a single probabilistic consensus $T_n \in [0, 1]$,$1 \leq n \leq N$. The consensus prior probability $P(T_n) = w_n$ is an important parameter of the algorithm giving the prior probability that voxel $n$ belongs to the consensus. Each rater's performance is characterized by a sensitivity ($p_k$) and a specificity ($q_k$) parameter, that are estimated throughout the algorithm. From Bayes law, the consensus posterior probability $u_n = P(T_n|\mathcal{S})$ is :

$$u_n = \frac{w_n \prod_k p_k^{S_n^k}(1-p_k)^{1-S_n^k}}{w_n \prod_k p_k^{S_n^k}(1-p_k)^{1-S_n^k} + (1-w_n)\prod_k q_k^{1-S_n^k}(1-q_k)^{S_n^k}} \tag{1}$$

where the parameters $p_k$ and $q_k$ are updated as follows:

$$p_k = \frac{\sum_{n,S_n^k=1} u_n}{\sum_n u_n} = \frac{TP_k}{FN_k + TP_k} \quad q_k = \frac{\sum_{n,S_n^k=0}(1-u_n)}{\sum_n (1-u_n)} = \frac{TN_k}{TN_k + FP_k}$$

where $TP_k$, $TN_k$, $FP_k$, $FN_k$ are respectively the continuous extension of the number of true positive, true negative, false positive and false negative voxels

from rater $k$. It is easy to show that $\text{logit}(u_n) = \ln\left(\frac{u_n}{1-u_n}\right)$ can be expressed as

$$\text{logit}(u_n) = \text{logit}(w_n) + \sum_{k,S_n^k=1} \big(\ln(p_k) - \ln(1-q_k)\big) + \sum_{k,S_n^k=0} \big(\ln(1-p_k) - \ln(q_k)\big).$$

Thus, when the background size increases, $TN_k$ also increases whereas $FP_k$ is only marginally impacted after a critical size. So, $q_k \to 1$ when $N \to \infty$ (supposing a constant foreground size) and we can write $\text{logit}(u_n) \sim \text{logit}(w_n) + \sum_{k,S_n^k=1}(\ln(N - B_k) + \ln\left(\frac{p_k}{FP_k}\right)) + \sum_{k,S_n^k=0} \ln(1-p_k)$ with $B_k = TP_k + FN_k$ constants for N large enough.

### 2.2 Impact of the consensus prior $w_n$

In [22], Warfield *et al.* proposed to set $w_n$ as a spatially uniform value $w_n = w$ where $w$ is either a constant (typically $w = 0.5$) or defined as the average occurrence ratio ($w = \frac{1}{NK}\sum_{n,k} S_n^k$). We further consider more general priors of the form $w = \frac{A}{N^\alpha}$, with A a constant independent of the image size $N$, thus having $\text{logit}(w_n) = -\ln\left(\frac{N^\alpha - A}{A}\right)$. We show in the supplementary material that, for large values of $N$, $\text{logit}(u_n)$ takes the asymptotic value of $\big(\sum_{k,S_n^k=1} \ln(N) - \alpha \ln(N)\big) + \ln(A) + \sum_{k,S_n^k=1} \ln\left(\frac{p_k}{FP_k}\right) + \sum_{k,S_n^k=0} \ln(1-p_k)$

Therefore, the consensus posterior $u_n$ is largely influenced by the image (background) size $N$ and the exponent $\alpha$ of image size $N$ in the consensus prior. More precisely, the asymptotic value of $u_n$ when $N \to \infty$ depends only on the sign of $\sum_k S_n^k - \alpha$ in two cases: if $\sum_k S_n^k > \alpha$, then $u_n \to 1$ whereas if $\sum_k S_n^k < \alpha$, $u_n \to 0$. If $\sum_k S_n^k = \alpha$, this limit value depends on $A$, $p_k$ and $FP_k$.

An example of the impact of the background image size on the STAPLE algorithm is provided in Fig. 1. This dependence of the STAPLE consensus can be explained by the fact that it is a generative model that should explain both the foreground and the background voxels.

The use of local sliding windows in STAPLE as in [6] can somewhat mitigate the background size effect, but smallest structures in the image can still be impacted and the window size remains a hyperparameter which is difficult to set.

## 3 MOJITO algorithm

We propose an alternative framework to the STAPLE algorithm that is solely based on the distance between foreground masks which makes the estimation of the consensus independent from the background image size. Furthermore, unlike the MV consensus, it takes into account the shape of each binary connected component surrounding each voxel to decide whether or not a voxel should be in the consensus, instead of just looking at the voxel value.

More formally, we propose to set the probability of generating a rater mask $S^k$ from a consensus $T$ to be of the form : $p(S^k|T) \propto \exp(-\lambda d(T, S^k)^2)$ where $d(T, S^k)$ is a distance or a metric between the two masks $S^k$ and $T$. With this

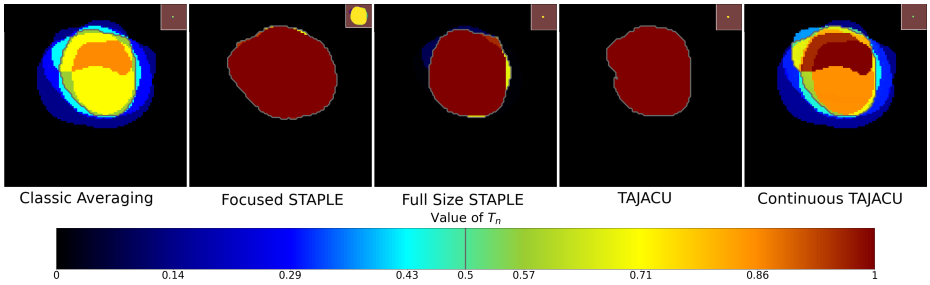



Fig. 1: Impact of the background size on STAPLE results between 7 segmentations (with an empty one) with $w = (\sum_{n,k} S_n^k)/NK$, computed with sizes respectively of size 67×61 (focused STAPLE) and 640×640 (full size STAPLE); and comparisons with other methods. The relative size of the structure used for computation can be seen at the top right corner. Contours after 0.5 thresholding are indicated in grey.

formulation, finding the consensus as the maximum likelihood is equivalent to minimizing the Fréchet variance $\mu_d = \text{argmin}_M \sum_k d(M, S^k)^2$. It is easy to show that when setting the distance as the square root of the symmetric difference between the 2 masks, $d(T, S^k) = \sqrt{|T \Delta S^k|} = \sqrt{|(T \cup S^k) \setminus (T \cap S^k)|}$, the Fréchet mean becomes the majority voting (MV) consensus (see the supplementary material). Yet, in the MV consensus, each voxel is processed independently of its neighbors which may lead to isolated voxels and non symmetric binary results.

### 3.1 MOJITO Binary algorithm

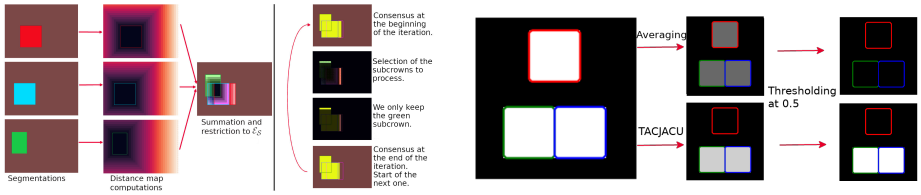
**Jaccard distance** Instead of the square root symmetric distance, we propose to use the Jaccard distance between binary masks $A$ and $B \in \{0,1\}^N$ defined as the complementary to the Jaccard index: $\mathrm{d}_J(A,B) = \frac{|A \Delta B|}{|A \cup B|} = 1 - \frac{|A \cap B|}{|A \cup B|}$. One can show [11] that this is a distance following the triangular inequality unlike the complementary to the Dice coefficient. By construction, the Jaccard consensus $\mu_J$ minimizes the mean squared Jaccard distance to $\mathcal{S}$ and only depends on the foreground binary masks and is independent of the background size :$\mu_J(\mathcal{S}) = \arg\min_{M \in \{0,1\}^N} \sum_{k=1}^K \text{dist}_J(M, S_k)^2$. Its computation can be restricted to the union of all rater masks : $\mathcal{E}_\mathcal{S} = \{n | \sum_{k=1}^K S_n^k > 0\}$. In addition, we consider that to decide if a voxel belongs to the consensus, one should only take into account the context associated with the connected components surrounding that voxel, since far away components are considered irrelevant. Therefore, we choose to minimize separately the Jaccard distance for each connected component $St$ of the masks union $\mathcal{E}_\mathcal{S}$ (i.e. each structure). This is equivalent to minimize the local mean squared Jaccard distance : $\text{lMSJD}(S, M) = \sum_{St} \mathrm{d}_J(S_{\|St}, M_{\|St})$. To lighten notations, we consider in the remainder only a single structure in $\mathcal{E}_\mathcal{S}$.

**Heuristic computation based on morphological distance and crowns** The minimization of the Fréchet variance is a combinatorial problem with a complexity of $2^{|\mathcal{E}_\mathcal{S}|}$ for the naive approach. Furthermore, it may lead to several global minima $\mu_J(S)$ when the number of raters $K$ is small. This is why we propose instead to seek a local minimum of the Fréchet variance by introducing some heuristics in the optimization. The resulting local minimum has a lower complexity to compute and is by construction maximally connected to avoid isolated voxels. More precisely, we take into account the global morphological relationships between each rater mask by decomposing the set $\mathcal{E}_\mathcal{S}$ into a set of *sub-crowns*. The algorithm then proceeds in a greedy fashion by iteratively removing or adding sub-crown to the current estimate of the consensus in order to minimize the mean square Jaccard distance. More precisely, we define $Dm_\mathcal{N}(S)$ as the distance map to the binary mask $S$ on $\mathcal{E}_\mathcal{S}$ according to the considered neighborhood $\mathcal{N}$, which can be either the 4 or 8 (resp. 6 or 26) connexity in 2D (resp. 3D) [9]. The distance is null for voxels inside each structure. The global morphological distance map is the sum of those maps $D_\mathcal{S}^\mathcal{N} = \sum_{S^k \in \mathcal{S}} Dm_\mathcal{N}(S^k)$ on $\mathcal{E}_\mathcal{S}$. A crown $C_d^\mathcal{N}$ is then defined as the set of voxels at a distance $d$ in the global map $D_\mathcal{S}^\mathcal{N}$. It can be shown that crowns realize a partition of $\mathcal{E}_\mathcal{S}$ ($\mathcal{E}_\mathcal{S} = \coprod_d C_d^\mathcal{N}$), and that the 0-crown corresponds to the intersection of all masks in $\mathcal{S}$. We propose to further partition each crown as a set of sub-crowns:

$$C_d^\mathcal{N} = \coprod_{g \in \mathcal{P}([\![1,K]\!])} (C_d^\mathcal{N})^g, \text{with } (C_d^\mathcal{N})^g = \{n | n \in C_d^\mathcal{N} \ \& \ \forall k \ S_n^k = (k \in g)\}$$

where $\mathcal{P}([\![1,K]\!])$ is the power set (i.e. the set of all subsets) of the first K integers. In other words, a sub-crown corresponds to a group of voxels located at the same morphological distance from the intersection (or union) of all rater masks and which have been segmented by exactly the same group of raters, as seen in Fig. 2a. Thus, our method leads to a consistent grouping since all voxels belonging to the same connected component, having the same morphological distance, and being generated by the same group of raters will end up in the same class.

**MOJITO algorithm** We proceed in a greedy approach by adding or removing sub-crowns until the lMSJD criteria stops decreasing. We use two concurrent strategies: either we start from the union of all masks (as seen in Fig. 2a) and then remove sub-crowns with decreasing distances or we start with the crown with the minimum distance and then add sub-crowns of increasing distances. Both growing and shrinking strategies are applied in order to mitigate the risk of falling into a local minimum and the consensus associated with the minimum lMSJD of either strategy is kept. Because it proceeds by adding or removing sub-crowns of increasing or decreasing distance, this algorithm enforces the compactness of the consensus with a low risk of having isolated voxels. Furthermore, adding or removing entire crowns would lead to suboptimal results because they can be fairly large. Therefore, we found this MOrphologically-aware Jaccard-based ITerative Optimization (MOJITO) approach to be a good compromise in terms




(a) (b)

Fig. 2: (a) Left: Preprocessing step of the MOJITO algorithm, with the construction of the crowns. Right: An iteration of the shrinking approach with selection of sub-crowns and the evaluation of their contribution to the lMSJD. (b) Application of averaging and MOCJITO on a toy example with three segmentations (red, green and blue contours). After thresholding, averaging gives an empty segmentation whereas the MOCJITO method is more inclusive and outputs one connected component.

of compactness, consistency and efficiency as seen in Fig. 1, with a number of iterations exponentially depending on $K$ but lower than the naive $2^{|\mathcal{E}_S|}$.

### 3.2 Continuous algorithm

Instead of seeking a binary consensus $T_n \in \{0,1\}$ between $K$ raters, we may be interested to get a soft consensus like a probability map $T_n \in [0,1]$ as in the STAPLE algorithm. The trivial consensus solution is the mean consensus $T_n = \frac{1}{K} S_n^k$ which corresponds to choosing $d_{\text{Mean}}(x,y) = \|x-y\|$. As for the MV, the mean consensus considers each voxel independent from its neighbors. Below, we introduce an extension of the MOJITO approach to the continuous case, called MOCJITO.

**Extension of Jaccard distance** Several extensions of the Jaccard distance to the continuous case have been proposed. Among them, the Soergel metric [18] $\mathrm{d}_{\text{Soergel}}(x,y) = \frac{\sum_i \max(x_i,y_i) - \min(x_i,y_i)}{\sum_i \max(x_i,y_i)}$ has the advantage of following the triangle inequality but is not differentiable. Instead, we consider the widely-used Tanimoto distance [23, 9, 13] $\mathrm{d}_{\text{Tan}}(x,y) = 1 - \frac{\sum_i x_i y_i}{\sum_i x_i^2 + y_i^2 - x_i y_i} = \frac{||x-y||^2}{||x-y||^2 + <x,y>}$.

**Continuous MOJITO algorithm** The brute force optimization of the sum of squared Tanimoto distances leads to optimize a sum of $K$ rational polynomials over a set of $|\mathcal{E}_S|$ scalars. So, we proceed in a greedy manner, separately on each connected component, by starting with the mean consensus and optimizing successively sub-crowns of increasing distances. All sub-crowns of increasing distances are iteratively considered until $\text{lMSJD}(T,\mathcal{S})$, similarly extended on the continuous domain with the Tanimoto distance, stops decreasing. For each sub-crown $r = (C_d^{\mathcal{N}})^g$, we optimize its scalar value $n_r \in [0,1]$ such that it minimizes

:

$$n_r = \underset{x\in[0,1]}{\operatorname{argmin}}(\mathrm{d}_{\mathrm{Tan}}(T^{r,x}, \mathcal{S})), \text{ with } T^{r,x} = \begin{cases} T^{r,x}_n = x \text{ if } n \in r \\ T^{r,x}_n = T_n \text{ otherwise} \end{cases};$$

For the optimization process we used the SLSQP algorithm [12] implemented in Scipy v1.7.3 [20]. Outputs of this method can be seen in Figs. 1 and 4, and the difference of behaviour with the classic averaging can be seen in Fig. 2b.

## 4 Results

### 4.1 Datasets and metrics

We applied our method on 2 datasets: a private database of transition zones of prostate MR images and the publicly available MICCAI MSSEG 2016 dataset of Multiple Sclerosis lesions segmentations [7]. The two datasets include 5 (resp. 7) raters' binary delineations for 40 (resp. 15) subjects. Images from the private dataset (resp. MSSEG dataset) have a size of [80-288]×[320-640]×[320-640] voxels (resp. [144-261]×[224-512]×[224-512] voxels). It was possible to extract from the private dataset bounding boxes of size [58-227]×[53-184]×[62-180] voxels. From the 3D private dataset, we created a 2D subset by extracting a single slice for each patient located at the basis of the prostate since this region is subject to a high inter-rater variability [5, 15].
We compared our algorithm to the STAPLE method implemented in SimpleITK v2.0.2 [14] and the naive segmentation averaging using the lMSJD (minimized by our method) and the mean squared error (MSE, minimized by the classic averaging). The statistical significance was evaluated with the Wilcoxon signed-rank test corrected with the Bonferroni-Holm method implemented in Pingouin v0.5.0 [19]. We used in all cases the neighborhood linked to the 8 or 26-connexity.

### 4.2 Evaluation of the different methods

We show in Fig. 1 comparison of our methods to the STAPLE method applied on the whole image and on a cropped image centered on the segmentation and to the MV. In Fig. 3 are represented the results for all considered methods on the three datasets, exact numerical results being available in supplementary material. Largest differences have been observed on the MSSEG dataset - an example being shown in Fig. 4. On the 3D private dataset, applying STAPLE on the whole image took $9.5 \pm 6.8s$ by image, against $14.7 \pm 15.9s$ for MOJITO and $30.7 \pm 33.2s$ for MOCJITO. On the MSSEG dataset the processing time was $45.5 \pm 56.6s$ for the binary version and $75.5 \pm 82.0s$ for the continuous one, against $20.5 \pm 20.4s$ for STAPLE.

**Discussion** In all datasets, we see the significant impact of the background size on the STAPLE result with a p-value$<$ .001 between the full size STAPLE, which produces very large consensuses, and other methods. MOJITO algorithms

also output consensuses that are significantly different from the ones produced by averaging especially on datasets where the inter-rater variability is higher. The MOJITO consensuses often include voxels segmented by less than half of the raters, (as seen in Fig. 4), and more rarely exclude voxels segmented by a majority of raters (as seen in the supplementary material). In general, our method appears to produce consensus segmentations of larger size than MV but smaller than those produced by full size STAPLE. This is particularly true on cases with a high inter-rater variability, as in the MSSEG dataset with 14 out of 15 cases with a MOJITO consensus strictly larger than the MV one (as shown in the supplementary material). Besides, posterior probabilities for a voxel to belong to the MOCJITO consensus differ from the ones obtained by averaging, as seen in Fig. 1. This is why significant statistical differences can be observed between averaging and MOCJITO in Fig. 3.

## 5 Conclusion

We have shown that the STAPLE method is impacted by the image background size and the choice of prior law. This dependence was also verified experimentally on two datasets. We have also introduced a new background-size independent method to generate a consensus based on the Jaccard distance. Our approach generalizes the Majority Voting and mean consensus, by taking into account local morphological configurations between rater masks and the proposed MOJITO and MOCJITO algorithms lead to consistent masks. Therefore, we believe that the MOJITO (resp. MOCJITO) algorithm is a good alternative to MV (resp. STAPLE) method to define segmentation consensus.

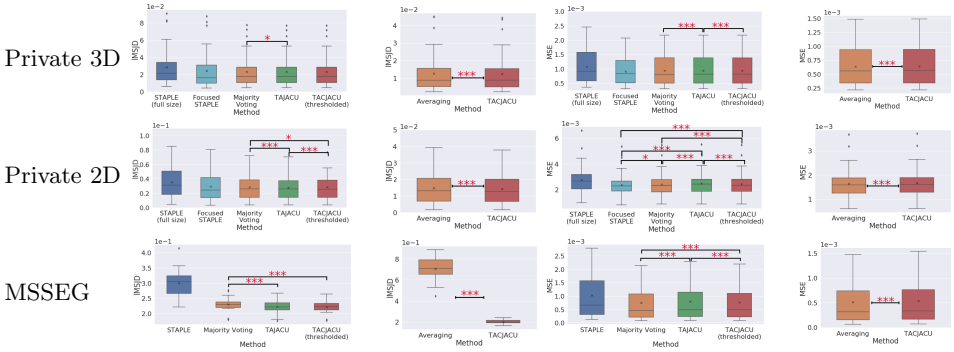



Fig. 3: Results on the private 3D (top) and 2D (middle) datasets and on the MSSEG dataset (bottom) with regards to the lMSJD (first two columns) and to the MSE (last two columns). Means are indicated by a cross. ***: p-val<.001; *: p-val<.05. All statistically significant differences are indicated except for full size STAPLE which differs for all other methods for all metrics on all datasets with a p-value<.001, so we did not represent its differences for clarity reasons.

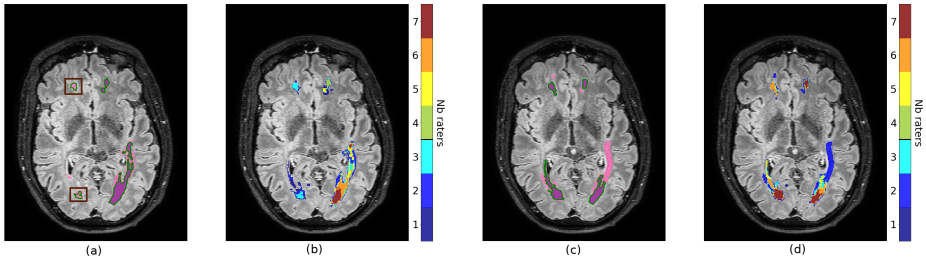



Fig. 4: Two consecutive slices of a MSSEG sample on which we applied STAPLE (pink), Majority Voting (purple) and our method (green contour) (a, c), and for each voxel of those slices the number of raters who segmented them (b, d). Differences between MV and MOCJITO are highlighted by brown squares

## References

1. Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., Rueckert, D.: Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. NeuroImage **46**(3), 726–738 (2009). https://doi.org/10.1016/j.neuroimage.2009.02.018
2. Asman, A., Landman, B.: Formulating Spatially Varying Performance in the Statistical Fusion Framework. Medical Imaging, IEEE Transactions on **31**, 1326–1336 (06 2012). https://doi.org/10.1109/TMI.2012.2190992
3. Asman, A.J., Landman, B.A.: Non-local statistical label fusion for multi-atlas segmentation. Medical Image Analysis **17**(2), 194–208 (2013). https://doi.org/10.1016/j.media.2012.10.002
4. Audelan, B., Hamzaoui, D., Montagne, S., Renard-Penna, R., Delingette, H.: Robust Fusion of Probability Maps. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2020. pp. 259–268. Springer International Publishing, Cham (2020)
5. Becker, A.S., Chaitanya, K., Schawkat, K., Muehlematter, U.J., Hötker, A.M., Konukoglu, E., Donati, O.F.: Variability of manual segmentation of the prostate in axial t2-weighted mri: A multi-reader study. European Journal of Radiology **121**, 108716 (2019). https://doi.org/https://doi.org/10.1016/j.ejrad.2019.108716, https://www.sciencedirect.com/science/article/pii/S0720048X19303663
6. Commowick, O., Akhondi-Asl, A., Warfield, S.K.: Estimating A Reference Standard Segmentation with Spatially Varying Performance Parameters: Local MAP STAPLE. IEEE Transactions on Medical Imaging **31**(8), 1593–1606 (Aug 2012). https://doi.org/10.1109/TMI.2012.2197406
7. Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S.C., Girard, P., Ameli, R., Ferré, J.C., Kerbrat, A., Tourdias, T., Cervenansky, F., Glatard, T., Beaumont, J., Doyle, S., Forbes, F., Knight, J., Khademi, A., Mahbod, A., Wang, C., Mckinley, R., Wagner, F., Muschelli, J., Sweeney, E., Roura, E., Llado, X., Santos, M., Santos, W.P., Silva-Filho, A.G., Tomas-Fernandez, X., Urien, H., Bloch, I., Valverde, S., Cabezas, M., Vera-Olmos, F.J., Malpica, N., Guttmann, C.R.G., Vukusic, S., Edan, G., Dojat, M., Styner, M., Warfield, S.K., Cotton, F., Barillot, C.: Objective Evaluation of Multiple Sclerosis Lesion Segmen-

tation using a Data Management and Processing Infrastructure. Scientific Reports **8**(1), 13650 (Dec 2018). https://doi.org/10.1038/s41598-018-31911-7
8. Dewalle-Vignion, A.S., Betrouni, N., Baillet, C., Vermandel, M.: Is STAPLE algorithm confident to assess segmentation methods in PET imaging? Physics in Medicine and Biology (11 2015). https://doi.org/10.1088/0031-9155/60/24/9473
9. Deza, M.M., Deza, E.: Distances and Similarities in Data Analysis. In: Encyclopedia of Distances. pp. 327–345. Springer Berlin Heidelberg, Berlin, Heidelberg (2016). https://doi.org/10.1007/978-3-662-52844-0˙17
10. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**, 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
11. Kosub, S.: A note on the triangle inequality for the Jaccard distance. Pattern Recognition Letters **120**, 36–38 (2019)
12. Kraft, D.: A software package for sequential quadratic programming. Tech. Rep. DFVLR-FB 88-28, DLR German Aerospace Center – Institute for Flight Mechanics, Koln, Germany (1988)
13. Leach, A.R., Gillet, V.J.: Similarity Methods. In: An Introduction To Chemoinformatics. pp. 99–117. Springer Netherlands, Dordrecht (2007). https://doi.org/10.1007/978-1-4020-6291-9˙5
14. Lowekamp, B., Chen, D., Ibanez, L., Blezek, D.: The Design of SimpleITK. Frontiers in Neuroinformatics **7** (2013). https://doi.org/10.3389/fninf.2013.00045
15. Montagne, S., Hamzaoui, D., Allera, A., Ezziane, M., Luzurier, A., Quint, R., Kalai, M., Ayache, N., Delingette, H., Renard Penna, R.: Challenge of prostate MRI segmentation on T2-weighted images: inter-observer variability and impact of prostate morphology. Insights into Imaging **12**(1) (Jun 2021). https://doi.org/10.1186/s13244-021-01010-9, https://hal.archives-ouvertes.fr/hal-03221227
16. Rohlfing, T., Maurer, C.R.: Shape-Based Averaging. IEEE Transactions on Image Processing **16**(1), 153–161 (2007). https://doi.org/10.1109/TIP.2006.884936
17. Sabuncu, M.R., Yeo, B.T.T., Van Leemput, K., Fischl, B., Golland, P.: A generative model for image segmentation based on label fusion. IEEE Transactions on Medical Imaging **29**(10), 1714–1729 (2010). https://doi.org/10.1109/TMI.2010.2050897
18. Späth, H.: The Minisum Location Problem for the Jaccard Metric. Operations-Research-Spektrum **3**, 91–94 (1981)
19. Vallat, R.: Pingouin: statistics in Python. Journal of Open Source Software **3**(31), 1026 (2018). https://doi.org/10.21105/joss.01026
20. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods **17**, 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2
21. Wang, Z., Demarcy, T., Vandersteen, C., Gnansia, D., Raffaelli, C., Guevara, N., Delingette, H.: Bayesian logistic shape model inference: Application to cochlear image segmentation. Medical Image Analysis **75**, 102268 (2022). https://doi.org/10.1016/j.media.2021.102268
22. Warfield, S., Zou, K., Wells, W.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image seg-

mentation. IEEE Transactions on Medical Imaging **23**(7), 903–921 (2004). https://doi.org/10.1109/TMI.2004.828354
23. Willett, P., Barnard, J.M., Downs, G.M.: Chemical Similarity Searching. Journal of Chemical Information and Computer Sciences **38**(6), 983–996 (1998). https://doi.org/10.1021/ci9800211