Robust Fusion of Probability Maps

Benoît Audelan, Dimitri Hamzaoui, Sarah Montagne, Raphaële Renard-Penna, Hervé Delingette

Robust Fusion of Probability Maps

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

[1] Université Côte d'Azur, Inria, Epione project-team, Sophia Antipolis, France
[2] Département de radiologie, CHU La Pitié Salpétrière/Tenon, Sorbonne Université, Paris, France

Abstract. The fusion of probability maps is required when trying to analyse a collection of image labels or probability maps produced by several segmentation algorithms or human raters. The challenge is to weight properly the combination of maps in order to reflect the agreement among raters, the presence of outliers and the spatial uncertainty in the consensus. In this paper, we address several shortcomings of prior work in continuous label fusion. We introduce a novel approach to jointly estimate a reliable consensus map and assess the production of outliers and the confidence in each rater. Our probabilistic model is based on Student's t-distributions allowing local estimates of raters' performances. The introduction of bias and spatial priors leads to proper rater bias estimates and a control over the smoothness of the consensus map. Image intensity information is incorporated by geodesic distance transform for binary masks. Finally, we propose an approach to cluster raters based on variational boosting thus producing possibly several alternative consensus maps. Our approach was successfully tested on the MICCAI 2016 MS lesions dataset, on MR prostate delineations and on deep learning based segmentation predictions of lung nodules from the LIDC dataset.

Keywords: Image segmentation · data fusion · consensus · mixture.

1 Introduction

The fusion of probability maps is required to solve at least two important problems related to image segmentation. The former is to establish the underlying ground truth segmentation given several binary or multi-class segmentations provided by human raters or segmentation algorithms (e.g. in the framework of multi-atlas segmentations). This is especially important because estimating a consensus segmentation and the inter-rater variability is the gold standard in assessing the performance of a segmentation algorithm in the absence of physical or virtual phantoms. The second related problem is the fusion of probability maps that are outputted by several segmentation algorithms such as neural networks. Indeed, it has been shown experimentally that combining the outputs of several segmentation algorithms often leads to improved performances [11].

Prior work has mainly focused on the fusion of binary masks, one of the most well known method being the STAPLE algorithm [14]. In this case, the

raters' binary segmentations are explained by Bernoulli distributions from the consensus segmentation and an Expectation-Maximization (EM) scheme allows to jointly build a consensus and estimate the raters' performances. Among known shortcomings of STAPLE, there is the constraint of having only global performance estimations of raters thus ignoring local variations [5,3]. One proposed solution [5] is to perform a STAPLE in a sliding window fashion or to extend the performance parameters to the pixel level [3]. Another limitation is that STAPLE only considers binary masks as input thus being agnostic to the image content and especially to the presence of large image gradients [4,10]. In [10], it was proposed to include in the STAPLE approach simple appearance models such as Gaussian distributions for the background and foreground, but this approach is only applicable to simple salient structures.

The extension of the STAPLE algorithm for continuous labels was proposed in [15] where raters' performances are captured by a set of biases and variances while assuming a Gaussian distribution for the raters' continuous labels. In [16], it was observed that to properly estimate raters' biases, the introduction of a prior was required. Furthermore, no spatial prior is used to regularize the consensus estimate and raters' performances are assumed to be global to the whole image.

In this paper, we introduce a comprehensive probabilistic framework that addresses many shortcomings of prior work on the fusion of continuous or categorical labels. First, we allow for a spatial assessment of raters' performances by replacing Gaussian with Student's t-likelihoods. Thus, image regions that largely differ from the consensus segmentation will be considered as outliers. Second, we introduce a bias prior and a label smoothness prior defined as a generalized linear model of spatially smooth kernels. Third, the proposed framework is posed within a proper metric, the Hellinger distance, in the space of probability maps through the introduction of a square root link function. In addition, probability maps are created from segmentation binary masks by using geodesic distance instead of Euclidean distance in order to take into account the image content. Finally, we address the unexplored issue of dissensus rather than consensus among raters. Indeed, fusing several probability maps into a single consensus map may not be meaningful when consistent patterns appear among raters. In [9], the worse performing raters' masks were removed from the consensus estimation process at each iteration. In [6], a comparison framework for the raters' maps based on the continuous STAPLE parameters was developed. In our approach, several consensus are iteratively estimated through a technique similar to variational boosting [12] and clusters of raters are identified.

We use variational Bayes (VB) inference to estimate the latent posterior distributions of variables and the unknown hyperparameters. The method has been applied on two databases of human expert segmentations of prostate and multiple sclerosis (MS) lesions and on the fusion of deep learning probability maps to segment lung nodules. We show that local variations of raters' performances were successfully identified and that improved segmentation performances were obtained after fusing probability maps.

2 Robust estimate of consensus probability map

2.1 Probabilistic framework

We are given as input a set of P probability maps $\mathbf{D}_n^p$, each map consisting of N categorical probability values in K classes, i.e. $\mathbf{D}_n^p \in S^{K-1} \in \mathbb{R}^K$ where S^{K-1} is the K unit simplex space such that $\sum_{k=1}^K \mathbf{D}_{nk}^p = 1$. Our objective is to estimate a consensus probability map $\mathbf{T}_n \in [0,1]^K$, $\sum_{k=1}^K \mathbf{T}_{nk} = 1$ over the input maps.

Each probability map is supposed to be derived from a consensus map through a random process. We consider a link function $F(\mathbf{p}) \in \mathbb{R}^K$, $\mathbf{p} \in S^{K-1}$ mapping probability S^{K-1} space into the Euclidean space and its inverse $F^{-1}(\mathbf{r})$ such that $F^{-1}(F(\mathbf{p})) = \mathbf{p}$. We write $\tilde{\mathbf{D}}_n^p = F(\mathbf{D}_n^p)$ and $\tilde{\mathbf{T}}_n = F(\mathbf{T}_n)$.

In [15,16], the observed probability maps $\tilde{\mathbf{D}}^p$ were supposed to be Gaussian distributed. In order to get a robust estimate of the consensus, i.e. to be able to discard locally the influence of outliers, we replace the Gaussian assumption by a Student's t-distribution written as a Gaussian scale mixture:

$$p(\tilde{\mathbf{D}}_n^p|\tilde{\mathbf{T}}_n) = \int_0^\infty \mathcal{N}(\tilde{\mathbf{D}}_n^p; \tilde{\mathbf{T}}_n + \mathbf{b}_p, \frac{\mathbf{\Sigma}_p}{\tau_n^p})\ \mathrm{Ga}(\tau_n^p; \frac{\nu_p}{2}, \frac{\nu_p}{2})\ d\tau\,, \tag{1}$$

where the bias $\mathbf{b}_p$ and covariance $\mathbf{\Sigma}_p$ characterize the performance of the rater p, and where $\mathrm{Ga}\left(\tau; \frac{\nu_p}{2}, \frac{\nu_p}{2}\right)$ is the Gamma distribution. The scale factors $\mathcal{T}^p = \{\tau_n^p\} \in \mathbb{R}^{+N}$ are additional latent variables that weight separately each data point $\tilde{\mathbf{D}}_n^p$ allowing to take into account local variations in the performances of rater p. ν_p^{-1} characterizes the amount of data outliers that it is necessary to discard in the estimation of the consensus. Finally, instead of the logit function as in [13], we propose to use the square root function $F_{\mathrm{sqrt}}((\mathbf{p}_1, \mathbf{p}_2)^T) = (\sqrt{\mathbf{p}_1}, \sqrt{\mathbf{p}_2})^T$, and its inverse $F_{\mathrm{sqrt}}^{-1}(\mathbf{r}) = \left(\frac{\mathbf{r}_1^2}{\mathbf{r}_1^2+\mathbf{r}_2^2}, \frac{\mathbf{r}_2^2}{\mathbf{r}_1^2+\mathbf{r}_2^2}\right)^T$ as a link function. By doing so, the probability $p(\mathbf{D}_n^p|\mathbf{T}_n) \propto \exp\left(-\frac{H^2(\mathbf{D}_n^p, \mathbf{T}_n)}{\sigma_p}\right)$ is related to the Hellinger distance $H^2(\mathbf{D}_n^p, \mathbf{T}_n)$ on the space of probability distributions. Maximizing the likelihood reverts to minimizing distances between probability distributions.

Bias prior. In [16] it was showed that if no prior is provided on the bias, its estimation is undetermined. Therefore we define a zero mean Gaussian prior on the bias with precision β, i.e. $p(\mathbf{b}_p|\beta) = \mathcal{N}(\mathbf{b}; 0, \beta^{-1}\mathbf{I}_K)$.

Consensus smoothness prior. A reasonable assumption is that segmentation probability maps are smooth. In [14], for categorical labels, a Markov random field (MRF) was introduced to enforce the connexity of discrete label map. Yet, the MRF hyperparameter β has to be set manually because its inference cannot be done in closed form. For continuous labels, prior work [15] did not include any smoothness prior. We introduce a smoothness prior defined as a generalized linear model of a set of L spatially smooth functions $\{\Phi_l(\mathbf{x})\}$, whose hyperparameters can be estimated. If $\mathbf{x}_n \in \mathbb{R}^d$ is the position of voxel n, then the prior

on the variables $\tilde{\mathbf{T}}_n$ is defined as $p(\tilde{\mathbf{T}}_n|\mathbf{W}_l) = \mathcal{N}(\tilde{\mathbf{T}}_n; \sum_{l=1}^{L} \Phi_l(\mathbf{x}_n)\mathbf{W}_l; \mathbf{\Sigma}_T\mathbf{I}_K)$ where $\mathbf{W}_l$ are vectors of size K and where $\mathbf{\Sigma}_T \in \mathbb{R}^+$ is the prior variance. For computation convenience, we write the prior using $\mathbf{W}_k \in \mathbb{R}^L$, such that $p(\tilde{\mathbf{T}}_{nk}|\mathbf{W}_k) = \mathcal{N}(\tilde{\mathbf{T}}_{nk}; \mathbf{W}_k^T\mathbf{\Phi}_n, \mathbf{\Sigma}_T)$ where $\mathbf{\Phi}_n^T = (\Phi_1(\mathbf{x}_n), \cdots, \Phi_L(\mathbf{x}_n))$. The weights $\mathbf{W}_k$ are gathered in a weight matrix $\mathbf{W} \in \mathbb{R}^{K\times L}$ such that we can write $p(\tilde{\mathbf{T}}_n|\mathbf{W}) = \mathcal{N}(\tilde{\mathbf{T}}_n; \mathbf{W}\mathbf{\Phi}_n; \mathbf{\Sigma}_T\mathbf{I}_K)$. The weights $\mathbf{W}_k$ are equipped with a zero mean Gaussian prior and precision α: $p(\mathbf{W}_k|\alpha) = \mathcal{N}(\mathbf{W}_k; 0, \alpha^{-1}\mathbf{I}_L)$. The graphical model of the framework is shown in Fig. 1a.

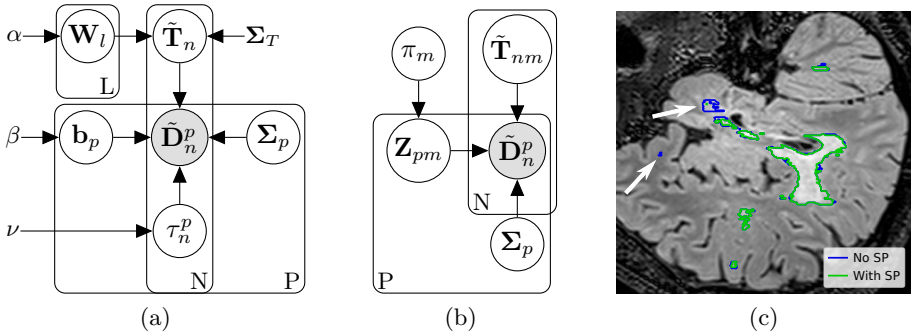


Fig. 1. Graphical model of the robust fusion framework (1a) and of the mixture of consensuses (1b). Effect of the spatial prior on the consensus smoothness (1c).

Generation of probabilistic maps. Probabilistic maps are typically outputted by segmentation algorithms, such as neural networks. They may be generated from binary masks using log-odds maps [13] computed as the sigmoid of signed distance maps from each binary structure. Yet, this approach ignores the underlying intensity image. We propose to compute a signed geodesic distance instead of Euclidean distance in order to take image intensity information into account, hence addressing a known shortcoming of STAPLE [4,10]. It is defined as a combination of the Euclidean distance and intensity gradient information [8].

2.2 Bayesian inference

To estimate the consensus and learn the parameters governing raters' performances, we want to maximize the marginal log likelihood:

$$\log p(\tilde{\mathbf{D}}|\beta, \nu_p, \mathbf{\Sigma}_p) = \sum_{n=1}^{N} \log \left(\int_{\mathbb{R}^K} \prod_{p=1}^{P} \left[\int_{\mathbb{R}^K} p(\tilde{\mathbf{D}}_n^p, \tilde{\mathbf{T}}_n, \mathbf{b}_p|\beta, \nu_p, \mathbf{\Sigma}_p)\ d\mathbf{b}_p \right]\ d\tilde{\mathbf{T}}_n \right). \tag{2}$$

Previous approaches maximized this quantity using an EM algorithm requiring to compute the posterior probability $p(\tilde{\mathbf{T}}_n|\tilde{\mathbf{D}}_n^p, \mathbf{b}_p, \mathbf{\Sigma}_p)$. It cannot be computed

in closed form when replacing Gaussians with Student's t-distributions. Instead, we use a VB approach where a factorized posterior over all latent variables is assumed: $p(\tilde{\mathbf{T}}, \mathbf{b}, \tau|\tilde{\mathbf{D}}) \approx q_{\tilde{\mathbf{T}}}(\tilde{\mathbf{T}})q_{\mathbf{b}}(\mathbf{b})q_{\tau}(\tau)$. Those approximation functions are estimated through a mean field approach which leads to closed form expressions.

The posterior approximation for the consensus map $\tilde{\mathbf{T}}_n$ can be written as a Gaussian distribution $q_{\tilde{\mathbf{T}}_n}(\tilde{\mathbf{T}}_n) = \mathcal{N}(\tilde{\mathbf{T}}_n; \mu_{\tilde{\mathbf{T}}_n}, \mathbf{\Sigma}_{\tilde{\mathbf{T}}_n})$ with $\mathbf{\Sigma}_{\tilde{\mathbf{T}}_n} = \big[\sum_{p=1}^{P} \hat{\tau}_n^p (\mathbf{\Sigma}_p)^{-1} + \mathbf{\Sigma}_T^{-1}\mathbf{I}_K\big]^{-1}$ and $\mu_{\tilde{\mathbf{T}}_n} = \mathbf{\Sigma}_{\tilde{\mathbf{T}}_n}\big[\sum_{p=1}^{P} \hat{\tau}_n^p \mathbf{\Sigma}_p^{-1}\big(\tilde{\mathbf{D}}_n^p - \mu_{\mathbf{b}_p}\big) + \mathbf{\Sigma}_T^{-1}\mu_{\mathbf{W}}\mathbf{\Phi}_n\big]$, where $\hat{\tau}_n^p = \mathbb{E}[\tau_n^p]$, $\mu_{\mathbf{b}_p} = \mathbb{E}[\mathbf{b}_p]$ and $\mu_{\mathbf{W}} = \mathbb{E}[\mathbf{W}]$. The consensus is now computed as a weighted mean of raters' values, where the weights vary spatially through $\hat{\tau}_n^p$ according to the rater's local performances. Likewise, $q_{\mathbf{b}_p}$ is found to be Gaussian distributed with covariance $\mathbf{\Sigma}_{\mathbf{b}_p} = \big[\beta\mathbf{I}_K + \sum_{n=1}^{N} \hat{\tau}_n^p \mathbf{\Sigma}_p^{-1}\big]^{-1}$ and mean $\mu_{\mathbf{b}_p} = \mathbf{\Sigma}_{\mathbf{b}_p}\sum_{n=1}^{N} \hat{\tau}_n^p \mathbf{\Sigma}_p^{-1}\big(\tilde{\mathbf{D}}_{np} - \mathbb{E}[\tilde{\mathbf{T}}_n]\big)$. Update formula for the other variables are reported in the supplementary material.

3 Mixture of consensuses

We now assume that the raters' maps are derived from not a single but M consensus maps. We introduce for each rater a new binary latent variable $\mathbf{Z}_{pm} \in \{0, 1\}$, $\sum_m \mathbf{Z}_{pm} = 1$, specifying from which consensus a rater map is generated. The associated component prior is given by the mixing coefficients π_m such that $p(\mathbf{Z}_{pm} = 1) = \pi_m$. We simplify the model by replacing the Student's t by Gaussian distributions and removing the bias, i.e. $p(\tilde{\mathbf{D}}_p|\tilde{\mathbf{T}}) = \prod_m \mathcal{N}(\tilde{\mathbf{T}}_m, \mathbf{\Sigma}_p)^{\mathbf{Z}_{pm}}$. The graphical model is presented in Fig. 1b.

Like in previous section, we use a VB to infer the consensus and model parameters. A naive solution would compute the posterior component probabilities r_{pm} (responsibilities) as a classical Gaussian mixture clustering problem with multivariate Gaussians of dimension N thus leading to dubious results due to the curse of dimensionality (high dimension, few samples). Instead, we propose first to reduce the dimension of each map by applying a principal component analysis (PCA) and then to cluster the maps in this low dimensional space. The resulting consensus maps are obtained by applying the inverse mapping from the components weights to the original space.

Variational calculus leads again to a Gaussian distribution for $q_{\tilde{\mathbf{T}}_{nm}}(\tilde{\mathbf{T}}_{nm})$, with covariance $\mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nm}} = \big[\sum_{p=1}^{P} r_{pm}(\mathbf{\Sigma}_p)^{-1}\big]^{-1}$ and mean $\mu_{\tilde{\mathbf{T}}_{nm}} = \mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nm}}\sum_{p=1}^{P} r_{pm}\mathbf{\Sigma}_p^{-1}\tilde{\mathbf{D}}_n^p$. The raters' contributions to each consensus are now weighted by the responsibilities r_{pm}. Other update formulas are reported in the supplementary material.

This approach has been found experimentally to be very sensitive to the initial values. To increase its stability we follow an incremental scheme inspired by variational boosting [12]. We introduce one consensus map at a time and the distribution parameters of components included in the previous iterations are not updated. Initialization is performed at each iteration by summing the absolute value of the residuals $\text{res}_p = \sum_{n,m} |\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm}|$ and setting the responsibility for the new component to $\frac{\text{res}_p}{\sum_p \text{res}_\text{p}}$ for rater p. Other responsibilities are uniformly

initialized such that $\sum_m r_{pm} = 1$. In practice, the algorithm is stopped when no rater is added to the newly introduced component after convergence.

4 Results

4.1 Datasets

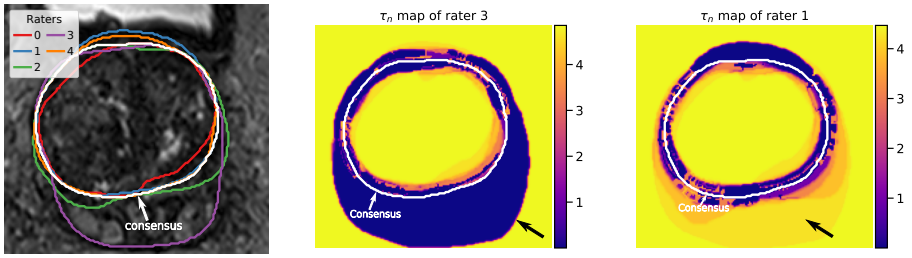
The proposed method was tested on 3 datasets: the MICCAI 2016 dataset of MS lesions segmentations [7], prostate segmentations from a private database and lung nodule segmentations from the LIDC dataset [2]. The first two datasets include 7 (resp. 5) raters' binary delineations for 15 (resp. 18) subjects. The LIDC dataset comprises nodules delineations drawn by 4 radiologists on 888 CT images. Only nodules annotated by at least 3 radiologists were considered. The image set was split into 10 folds, one being kept separated for testing while the rest was used to train 9 different segmentation networks by 9-fold cross validation (CV). On the test set, only nodules of size above 10 mm were kept corresponding to 34 nodules. Ground truth segmentations were defined as a majority voting among raters.

4.2 Robust consensus estimate

We first demonstrate the estimation of a single consensus from 5 prostate delineations produced by human experts. The input binary masks are converted to probabilities using the geodesic distance transform and the sigmoid function. Fig. 2a shows the 5 raters' segmentations and the associated consensus as estimated by our approach. It can be seen that rater 3 seems to be an outlier with respect to the other raters at the bottom of the image, although they agree elsewhere. This local variation of the rater's performance is captured by the scale factor τ_n^p that modulates spatially the contribution of each rater to the consensus. In areas of poor rater's performance, τ exhibits lower values which correspond to larger rater's variance. Locally, raters with weak confidence will not contribute as much as others to the consensus. This is shown in Fig. 2b and 2c, where rater 3 has smaller τ_n values than rater 1 at the bottom of the image (black arrows).

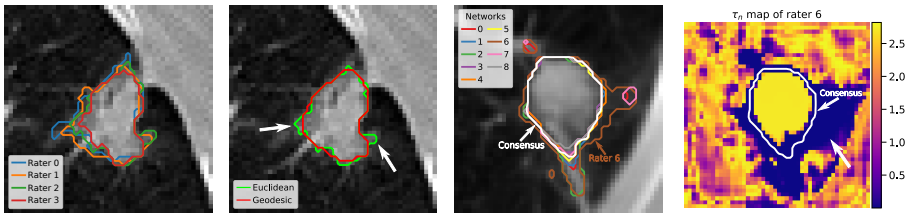
Converting binary masks to the continuous domain using a geodesic distance allows to take image intensity information into account and leads to consensus estimates more consistent with intensity boundaries (Fig. 3b). Moreover, the introduction of a spatial prior over the consensus allows to control the smoothness of the output (Fig. 1c). In practice, we use a dictionary of Gaussian bases centered on a regular staggered grid. Key parameters are the spacing between the bases centers, the standard deviations and the position of the origin basis. Larger spacing and scales induce smoother contours in the final map.

Our algorithm is of specific interest when fusing probability maps outputted by segmentation algorithms. For instance, we consider lung nodules probabilistic segmentations given by 9 neural networks with a same U-Net architecture and



(a) (b) (c)

Fig. 2. Fusion of prostate segmentation binary masks (2a). The outlier rater 3 exhibits locally a higher variability linked to lower values of τ_n pointed by the black arrow (2b), whereas rater 1 (2c) shows higher τ_n values in the same area.



(a) (b) (c) (d)

Fig. 3. Geodesic vs. Euclidean distance on radiologists' delineations from LIDC (3a and 3b). Networks probability maps fusion (3c). τ_n map for the network 6 with local variability highlighted by the arrow (3d).

trained with 9-fold CV. Fig. 3c shows the 9 segmentations for a case with the estimated consensus segmentation. Large discrepancies can be observed locally between network 6 and the others, which is also captured by the scale factor (Fig. 3d). To assess the performance of our method, we performed a comparison study with prior works. Dice scores and Hausdorff distances were computed between the estimated consensus and the ground truth defined as a majority vote between the 4 radiologists (Fig. 4). Our proposed approaches are highlighted in bold. Out of the 9 tested methods, STAPLE and majority vote use binary masks and do not exploit the image content. Both are giving poorer results than continuous methods. PSTAPLE (resp. continuous STAPLE) correspond to the approach proposed in [1] (resp. [15]). Gaussian models correspond to the same framework as ours, with Gaussian distributions replacing the Student's t. Models with Student's t or Gaussian are all implemented with a spatial prior unless stated otherwise. Consensuses produced with spatial regularization lead to clearly better results. In PSTAPLE [1] the regularization is done by a MRF for which the results were found to be sensitive to its parameters. Instead, our approach allows to estimate automatically the spatial prior hyperparameter.

As shown in Fig. 4, our proposed robust approach with Student's t-distribution leads to competitive results, with higher Dice scores, and lower Hausdorff distances, illustrating the relevance of our method.

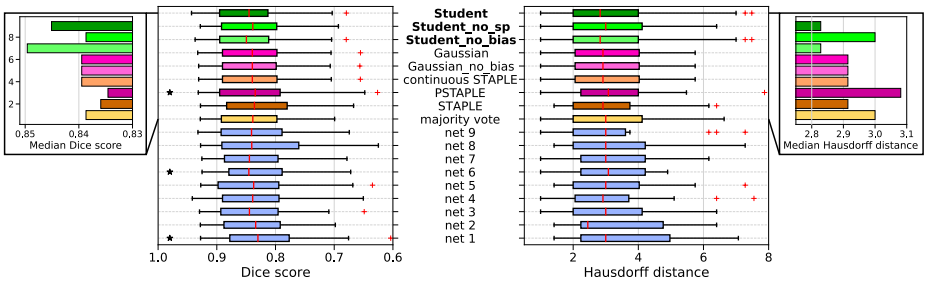


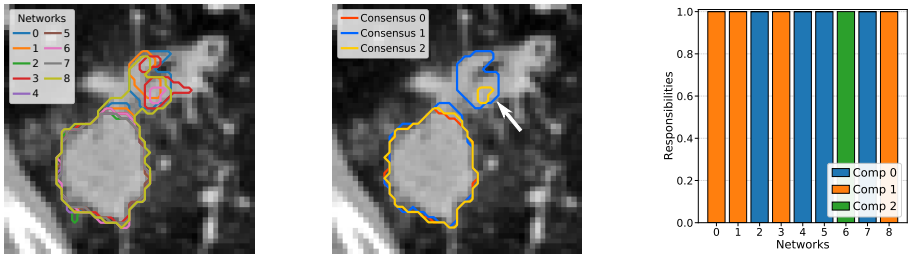
Fig. 4. Dice score and Hausdorff distance distributions over the nodule test set. Leftmost values are the best results. Distributions marked with a $\star$ are found significantly different from the one given by our approach ("Student") with the Wilcoxon signed-rank test and p-value 0.05.

4.3 Mixture of consensuses

We assume here that raters' masks can be derived from possibly several underlying ground truths rather than one. An example of mixture estimated from networks probability maps is given in Fig. 5. Three relevant components are selected which differ in the region highlighted by the white arrow in Fig. 5b. Without the mixture approach, only one consensus corresponding to the first component would have been obtained and the region pointed by the arrow would have been ignored. Thus, the mixture allows to enrich the representation and propose several possible patterns by taking into account the residuals. A case where only one component is retained in the model is shown in the supplementary material.

5 Conclusion

We presented a novel framework for the robust fusion of probabilistic segmentation masks. Our method relies on Student's t-distributions which allow to take rater's spatial uncertainty into account. All parameters of the model are estimated automatically using Bayesian inference. Furthermore, the concept of mixture of consensuses was explored, which allows to consider several patterns among raters. The approach was tested on several datasets and produced competitive results in comparison with other methods. We believe our method can be a useful tool to combine probabilistic masks generated by different segmentation algorithms.



(a) (b) (c)

Fig. 5. Mixture of consensuses for a lung nodule. Input probabilistic masks (5a). Estimated consensuses (5b). Responsibilities with 3 relevant components (5c).

Acknowledgments

This work was partially funded by the French government, through the UCA^{JEDI} "Investments in the Future" project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01. It was partially supported by the Clinical Data Warehouse of Greater Paris University Hospitals and by the Inria Sophia Antipolis - Méditerranée,"NEF" computation cluster.

References

1. Akhondi-Asl, A., Warfield, S.K.: Simultaneous Truth and Performance Level Estimation Through Fusion of Probabilistic Segmentations. IEEE Transactions on Medical Imaging **32**(10), 1840–1852 (Oct 2013)
2. Armato III, S.G., McLennan, G., Bidaut, L., et al.: The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. Medical Physics **38**(2), 915–931 (2011)
3. Asman, A.J., Landman, B.A.: Formulating Spatially Varying Performance in the Statistical Fusion Framework. IEEE Transactions on Medical Imaging **31**(6), 1326–1336 (June 2012)
4. Asman, A.J., Landman, B.A.: Non-local statistical label fusion for multi-atlas segmentation. Medical Image Analysis **17**(2), 194 – 208 (2013)
5. 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)
6. Commowick, O., Warfield, S.K.: A Continuous STAPLE for Scalar, Vector, and Tensor Images: An Application to DTI Analysis. IEEE Transactions on Medical Imaging **28**(6), 838–846 (June 2009)
7. Commowick, O., Istace, A., Kain, M., et al.: Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. Scientific Reports **8**(1), 13650 (2018)
8. Criminisi, Antonio and Sharp, Toby and Blake, Andrew: GeoS: Geodesic Image Segmentation. In: Proc. European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 5302, pp. 99–112. Springer (January 2008)

9. Langerak, T.R., van der Heide, U.A., Kotte, A.N.T.J., et al.: Label Fusion in Atlas-Based Segmentation Using a Selective and Iterative Method for Performance Level Estimation (SIMPLE). IEEE Transactions on Medical Imaging **29**(12), 2000–2008 (Dec 2010)
10. Liu, X., Montillo, A., Tan, E.T., Schenck, J.F.: iSTAPLE: improved label fusion for segmentation by combining STAPLE with image intensity. In: Ourselin, S., Haynor, D.R. (eds.) Medical Imaging 2013: Image Processing. vol. 8669, pp. 727 – 732. International Society for Optics and Photonics, SPIE (2013)
11. Menze, B.H., Jakab, A., Bauer, S., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (Oct 2015)
12. Miller, A.C., Foti, N.J., Adams, R.P.: Variational Boosting: Iteratively Refining Posterior Approximations. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 2420–2429. PMLR, International Convention Centre, Sydney, Australia (06–11 Aug 2017)
13. Pohl, K.M., Fisher, J., Bouix, S., et al.: Using the logarithm of odds to define a vector space on probabilistic atlases. Medical Image Analysis **11**(5), 465 – 477 (2007), special Issue on the Ninth International Conference on Medical Image Computing and Computer-Assisted Interventions - MICCAI 2006
14. 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 on Medical Imaging **23**(7), 903–921 (July 2004)
15. Warfield, S.K., Zou, K.H., Wells, W.M.: Validation of image segmentation by estimating rater bias and variance. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **366**(1874), 2361–2375 (2008)
16. Xing, F., Prince, J.L., Landman, B.A.: Investigation of Bias in Continuous Medical Image Label Fusion. PLOS ONE **11**(6), 1–15 (June 2016)

Robust Fusion of Probability Maps
— Supplementary material —

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

[1] Université Côte d'Azur, Inria, Epione project-team, Sophia Antipolis, France
[2] Département de radiologie, CHU La Pitié Salpétrière/Tenon, Sorbonne Université, Paris, France

1 Variational updates

1.1 Robust estimate of consensus

The posterior approximation of τ_n^p is a Gamma distribution $q_{\tau_n^p}(\tau_n^p) = \mathrm{Ga}(\tau_n^p; a_{np}, b_{np})$ with $a_{np} = \frac{\nu_p + K}{2}$ and $b_{np} = \frac{\nu_p}{2} + \frac{1}{2}(\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p})^T \mathbf{\Sigma}_p^{-1}(\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p}) + \frac{\mathrm{Tr}(\mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_{\mathbf{b}_p}) + \mathrm{Tr}(\mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_{\tilde{\mathbf{T}}_n})}{2}$.

If the spatial prior is included in the model, the posterior approximation of $\mathbf{W}_k$ needs to be computed and is given by $q_{\mathbf{W}_k}(\mathbf{W}_k) = \mathcal{N}(\mathbf{W}_k; \mu_{\mathbf{W}_k}, \mathbf{\Sigma}_{\mathbf{W}_k})$, where $\mathbf{\Sigma}_{\mathbf{W}_k} = \left[\mathbf{\Sigma}_T^{-1}\left(\sum_{n=1}^N \mathbf{\Phi}_n \mathbf{\Phi}_n^T\right) + \alpha \mathbf{I}_L\right]^{-1}$ and $\mu_{\mathbf{W}_k} = \mathbf{\Sigma}_{\mathbf{W}_k}\left[\sum_{n=1}^N \mathbf{\Phi}_n \mathbf{\Sigma}_T^{-1} \mu_{\tilde{\mathbf{T}}_{nk}}\right]$.

The posterior approximations q_β and $q_{\mathbf{\Sigma}_p}$ are assumed to be Dirac distributions. The update formula for these parameters are found by deriving the lower bound and equalling to zero, which gives $\beta = PK\left(\sum_{p=1}^P \mathrm{Tr}(\mathbf{\Sigma}_{\mathbf{b}_p}) + \mu_{\mathbf{b}_p}^T \mu_{\mathbf{b}_p}\right)^{-1}$ and $\mathbf{\Sigma}_p = N^{-1}\left(\sum_{n=1}^N (\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p})\hat{\tau}_n^p (\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p})^T + \hat{\tau}_n^p(\mathbf{\Sigma}_{\tilde{\mathbf{T}}_n} + \mathbf{\Sigma}_{\mathbf{b}_p})\right)$.

The posterior approximations q_α and $q_{\mathbf{\Sigma}_T}$ are also assumed to be Dirac distributions, leading to the update formula $\alpha = LK\left(\sum_{k=1}^L \mathrm{Tr}(\mathbf{\Sigma}_{\mathbf{W}_k}) + \mu_{\mathbf{W}_k}^T \mu_{\mathbf{W}_k}\right)^{-1}$ and $\mathbf{\Sigma}_T = (NK)^{-1}\sum_{n=1}^N \sum_{k=1}^K \left((\mu_{\tilde{\mathbf{T}}_{nk}} - \mathbf{\Phi}_n^T \mu_{\mathbf{W}_k})^2 + \mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nk}} + \mathrm{Tr}(\mathbf{\Phi}_n \mathbf{\Phi}_n^T \mathbf{\Sigma}_{\mathbf{W}_k})\right)$.

Deriving likewise the lower bound with respect to the degrees of freedom ν_p, they are found to be the solution of the following equation $\sum_{n=1}^N -\psi(\frac{\nu_p}{2}) + \log\frac{\nu_p}{2} + 1 + \mathbb{E}[\log\tau_n^p] - \mathbb{E}[\tau_n^p] = 0$, with ψ being the digamma function. In practice, the ν_p are updated by solving the equation numerically.

1.2 Mixture of consensuses

The label posterior is updated with $q_{\mathbf{Z}_{pm}}(\mathbf{Z}_{pm}) = r_{pm} = \frac{\rho_{pm}}{\sum_{m=1}^M \rho_{pm}}$, where $\log\rho_{pm} = \log\pi_m + \sum_{n=1}^N \left(-\frac{K}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}_p| - \frac{1}{2}D_{\mathbf{\Sigma}_p}(\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_{mn}}) - \frac{1}{2}\mathrm{Tr}(\mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_{\tilde{\mathbf{T}}_{mn}})\right)$. $D_{\mathbf{\Sigma}}(x)$ is the squared Mahalanobis distance, i.e. $D_{\mathbf{\Sigma}}(x) = (x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)$.

The mixing coefficients π_m are updated with $\pi_m = \frac{1}{P}\sum_{n=1}^P r_{pm}$. Finally, we update the covariance $\mathbf{\Sigma}_p$ of each rater with the following formula $\mathbf{\Sigma}_p = \frac{1}{N}\left(\sum_{n=1}^N \sum_{m=1}^M r_{pm}\left((\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_{nm}})(\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_{nm}})^T + \mathbf{\Sigma}_{T_{nm}}\right)\right)$.

2 Supplementary figures

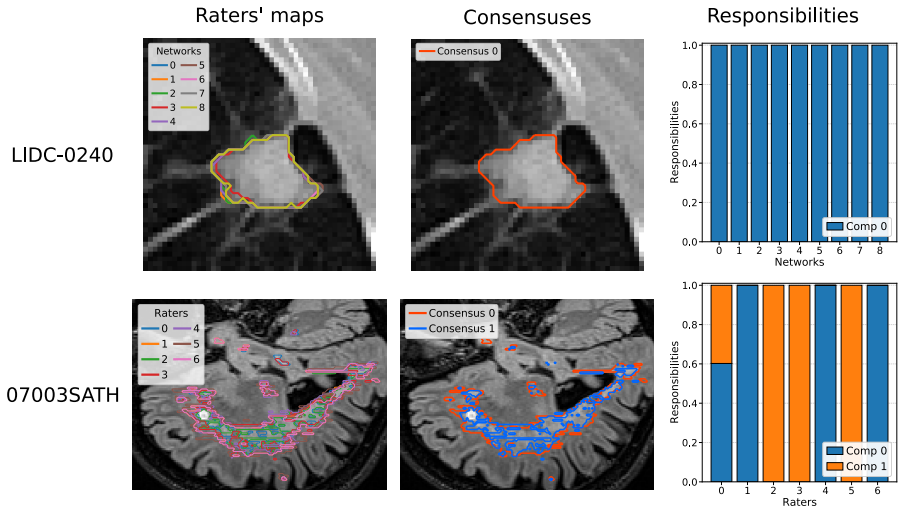


Fig. 1. Mixture of consensuses applied on a lung nodule (top row) and on MS lesions (bottom row). 1 (resp. 2) components are found relevant on the top (resp. bottom) row.

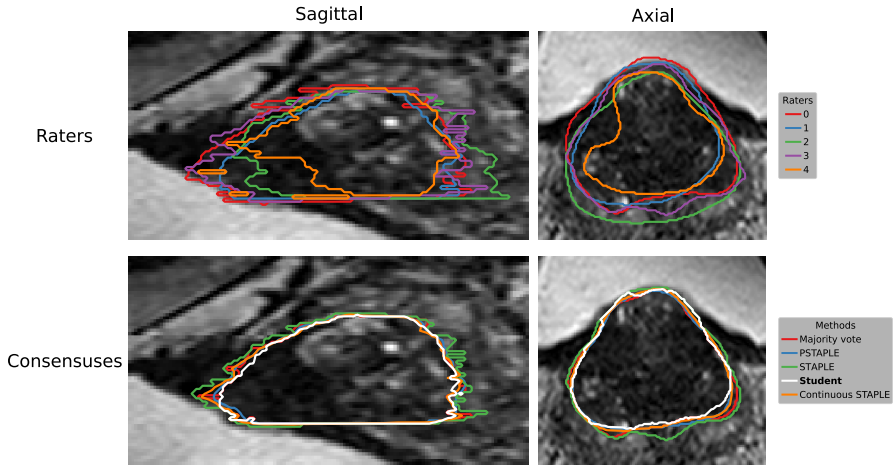


Fig. 2. Consensuses given by different methods on human raters' prostate delineations. Methods with spatial regularization (PSTAPLE and ours) produce smoother contours.