Robust supervised segmentation of neuropathology whole-slide microscopy images

Michel Vandenberghe, Yael Balbastre, Nicolas Souedet, Anne-Sophie Herard, Marc Dhenain, Frédérique Frouin, Thierry Delzescaux

Robust Supervised Segmentation of Neuropathology Whole-Slide Microscopy Images

Michel E. Vandenberghe, Yaël Balbastre, Nicolas Souedet, Anne-Sophie Hérard, Marc Dhenain, Frédérique Frouin, Thierry Delzescaux

Abstract— **Alzheimer's disease is characterized by brain pathological aggregates such as Aβ plaques and neurofibrillary tangles which trigger neuroinflammation and participate to neuronal loss. Quantification of these pathological markers on histological sections is widely performed to study the disease and to evaluate new therapies. However, segmentation of neuropathology images presents difficulties inherent to histology (presence of debris, tissue folding, non-specific staining) as well as specific challenges (sparse staining, irregular shape of the lesions). Here, we present a supervised classification approach for the robust pixel-level classification of large neuropathology whole slide images. We propose a weighted form of Random Forest in order to fit nonlinear decision boundaries that take into account class imbalance. Both color and texture descriptors were used as predictors and model selection was performed via a leave-one-image-out cross-validation scheme. Our method showed superior results compared to the current state of the art method when applied to the segmentation of Aβ plaques and neurofibrillary tangles in a human brain sample. Furthermore, using parallel computing, our approach easily scales-up to large gigabyte-sized images. To show this, we segmented a whole brain histology dataset of a mouse model of Alzheimer's disease. This demonstrates our method relevance as a routine tool for whole slide microscopy images analysis in clinical and preclinical research settings.**

I. Introduction

Aβ plaques (AP) and neurofibrillary tangles (NFT), two forms of misfolded protein aggregates, are the histopathological signatures of Alzheimer's disease [1]. These neuropathological markers are extensively studied in human brain samples and animal models of Alzheimer's disease and they represent important therapeutic targets. Thus, their precise quantification is a critical issue for both physiopathological research and drug development. Quantification is commonly performed by segmenting histology whole slide microscopy images and computing the proportion of positively stained pixels relative to the remaining brain tissue.

In this context, global and adaptive thresholding are popular segmentation approaches for they are simple and fully automated [2, 3]. However these methods are prone to errors. Indeed, histological procedures often lead to the presence of artifacts, such as debris and tissue folding, which have similar color properties to the marker of interest. In addition, histological staining can lead to important background non-specific staining. This hinders methods based solely on color to provide optimal segmentation results and, as neuropathological markers represent a tiny portion of the brain tissue, minor segmentation errors can strongly impact the overall quantification.

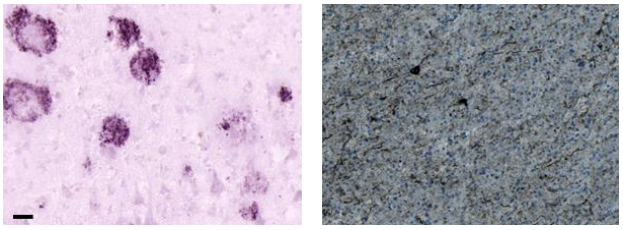

Figure 1. Immunohistochemistry of a brain sample from a patient with Alzheimer's disease. Left: Aβ plaques appear as dark purple clusters. Right: neurofibrillary tangles appear in black over a blue Nissl counterstaining. Scale bar: 20 µm.

To overcome these limitations, Chubb *et al*. [4] developed a supervised classification method called BioVision which uses color and local intensity information (mean of the *R*, *G* and *B* channels in a 16-pixels diagonal neighborhood). This latter feature is particularly helpful to account for noise in neuropathology images. For each class, BioVision estimates the joint distribution of the predictors using Gaussian Mixture Models. In the segmentation step, each pixel is classified with a Bayesian classifier. Class imbalance between neuropathological markers and the rest of the tissue can be accounted for by injecting prior probabilities of each class in the decision. Although numerous contextual features have been shown to be efficient for image segmentation, they are typically high-dimensional and because of the curse of dimensionality, density estimation for these features with Gaussian Mixture Models is hazardous. Thus, one limitation of BioVision is that it cannot adequately incorporate complex contextual features, such as texture descriptors, which could be beneficial to the classification task. The authors propose to use a post-processing step to remove incorrectly classified pixels based on shape and size features. However, pathological aggregates are not well-defined objects. As shown in Figure 1, they present irregular shapes and disparate sizes. Hence, misclassified pixels are hard to detect with morphological features and a robust one-step segmentation approach would be preferable.

Here, we propose a supervised classification approach that incorporates color and texture features in order to better discriminate markers of interest from noise. We propose to use Weighted Random Forest (WRF) for robust classification and parallel computing to handle large whole

M.E.V., Y.B., N.S., A.S.H., M.D. and T.D. are with the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) – Molecular Imaging Research Center (MIRCen), Fontenay-Aux-Roses, France (corresponding author: +33146548236; thierry.delzescaux@cea.fr).

F.F. is with Sorbonne Universités UPMC Paris 06, Inserm, CNRS, LIB, Paris France

978-1-4244-9270-1/15/$31.00 ©2015 EU

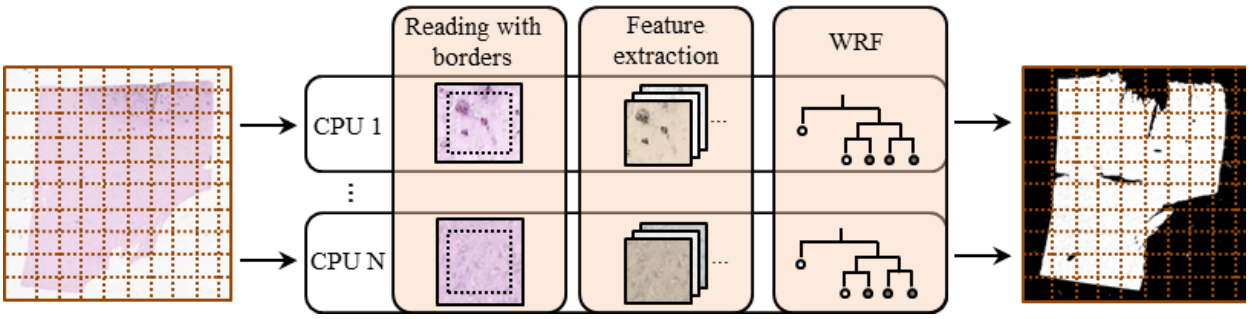

Figure 2. Parallel processing of a whole slide microscopy image.

slide images. We show that this approach is more robust to artifacts and background signal than BioVision. Furthermore, we demonstrate our approach usability on large datasets by segmenting a mouse high-resolution 3D whole-brain histology dataset.

II. Materials and Methods

A. Datasets

The first two datasets were used as benchmarks for AP and NFT detection. A cortical brain tissue sample (GIE NeuroCEB brain bank) from a patient with confirmed diagnosis of Alzheimer's disease was sectioned. Ten tissue sections were stained for AP detection (4G8 immunohistochemistry). Two sections were stained for NFT detection (AT8 immunohistochemistry with a Nissl counterstaining). All sections were digitized using a Zeiss Axio ScanZ.1 at a resolution of 0.44µm. At this resolution, each image has a size of approximately 15000×15000 pixels. Fifty representative image patches (200×200 pixels) for each staining were then extracted from whole-slide images and each patch was manually segmented into 3 classes: marker of interest, tissue and glass slide background. For each staining, patches were split so that half of them were used for learning and model selection and the other half was kept for final validation and comparison between algorithms.

The third dataset is a 3D reconstruction of the whole brain of an APP/PS1 mouse model of Alzheimer's disease. To generate 3D whole brain AP histology, a set of 78 sections were stained for AP detection (6E10 immunohistochemistry). Sections were digitized at a resolution of 0.44µm and the 2D images were reconstructed in 3D with an inter-section distance of 125 µm using a protocol described by Vandenberghe *et al.* [5]. The histological volume has a size of 24000×16000×78 voxels.

B. Feature extraction

Our approach includes color, local intensity and texture features. For color features, HSV was chosen over RGB as it is generally assumed that HSV color space is closer to human perception. Local intensity was computed for each pixel as the mean of *R*, *G* and *B* values in a disk-shaped neighborhood of a given radius. Images were convolved with a family of Gabor filters to extract texture information. Gabor filtering is among the most popular approach for texture classification and it has been shown to model the function of simple cells in the mammalian visual system [6]. The Gabor filter kernel consists of a sinusoidal wave multiplied by a Gaussian function. The Gabor filter response has a real and an imaginary component. Here, we used a family of filters with 4 orientations and 4 frequencies [7]. This led to a 36-dimensional feature vector for each pixel. As feature extraction can be particularly time consuming, large images were divided in small chunks that were processed in parallel (Fig. 2) [8]. In order to take into account chunk borders properly, each chunk was processed with additional width equal to the size of the convolution kernel.

B. Weighted Random Forest

Random Forest (RF) is an ensemble of fully grown decision trees [9]. Each classification tree is built using a bootstrap sample of the original learning set and only a subset of randomly selected features at each split of the tree. This allows RF to have a low variance compared to a single decision tree and eventually, a better performance. RF can fit highly nonlinear decision boundaries which makes it particularly useful to discriminate the markers of interest from artifacts. During tree growing, the feature space is partitioned at each node to minimize the cross-entropy, defined by:

$$H = -\sum_{j=1}^{m} p_j \log(p_j), \qquad (1)$$

where p_j is the proportion of pixels of class j at a given node.

One limitation of decision tree learning algorithms is that when the classification problem is highly imbalanced, they tend to be biased toward the majority class. To overcome this problem, we propose to use a weighted form of RF modified from Chen *et al.* [10]. Each class is assigned a weight ω_j. At a given node, we have a population of pixels x_i with $i = 1 \dots n$, $Y \in \mathbb{R}^n$ their corresponding class labels, a vector of weights $W \in \mathbb{R}^n$ such that $W_i = \omega_{Y_i}$ and a matrix of indicator variables $M \in \mathbb{R}^{m \times n}$ such that $M_{j,i} = 1$ if x_i is of class j and 0 otherwise. In (1), p_j is calculated as a weighted proportion:

$$p_j = \frac{M_j W}{\sum_{i=1}^{n} W_i} \quad (2)$$

If class weights are equal, this is equivalent to the classical RF. If the minority class is given a higher weight than other classes, this results in an increase of its influence in the tree building process. After growing an ensemble of B trees $\{T_b\}_1^B$, class prediction for any new pixel x is done by weighted majority voting:

$$\hat{Y}(x) = \arg\max_{j\in[1,m]} \omega_j \sum_{b=1}^{B} I(T_b(x) = j), \quad (3)$$

where $I(.)$ is the indicator function. For a good tradeoff between performance and computational burden, we chose $B = 100$ trees to build the ensemble models.

C. *Model selection via leave-one-image-out cross-validation*

We investigated the effect of feature extraction and learning parameters on classification performance with a full factorial design. We hypothesized that the local intensity neighborhood radius could have an impact on classification. Furthermore, according to Bianconi *et al*. [11], who evaluated the effect of the Gabor filter parameters on texture classification, we considered the effects of the standard deviation of the Gaussian envelope and its spatial aspect ratio. Finally, we considered the effect of increasing the weight of the minority class in the WRF (while the weights of tissue and background classes were kept equal to each other). A factorial design with three levels per parameter was constructed, leading to a total of 81 combinations (Table 1).

TABLE I. LEVELS OF THE FACTORIAL DESIGN

Parameter	Values
Local intensity radius	4, 8, 16 (pixels)
Gaussian envelope standard deviation	1, 2, 3 (pixels)
Gaussian envelope aspect ratio	0.5, 1, 1.5
Minority class weight	1.0, 1.25, 1.5

As RF relies on bootstrapping the original data, it is tempting to evaluate model performance by computing an out-of-the bag estimate for each model. However, alike cross-validation, out-of-the bag estimation is valid if the observations used for estimating classification performance are independent of those used for constructing the model. In our case, neighboring pixels are correlated, hampering pixel-level resampling to provide independent samples. In contrast, the 25 patches in the training set are nearly independent since they are not neighboring and were sampled from various tissue sections. We used these patches as blocks for cross-validation as follows. Let $Z = (X, Y)$ be the learning set with X the feature vectors for all pixels and Y their corresponding class labels. For each image $i = 1, \ldots, k$ included in the learning set, let $Z_i \subset Z$ be its subset in the learning set and let $Z'_i = Z \setminus Z_i$ its complement. A model $Q_i(Z'_i)$ is constructed for each i and used to make predictions $\hat{Y}_i$ over X_i. This leads to a leave-one-image-out cross-validation scheme. Precision and recall were calculated for each class j as, respectively, $P(Y = j \mid \hat{Y} = j)$ and $P(\hat{Y} = j \mid Y = j)$. A mean cross-validation f1 score was calculated for each parameters combination t as:

$$\bar{f1}(t) = \frac{1}{n}\sum_{j=1}^{n} 2\,\frac{Precision_{t,j} \times Recall_{t,j}}{Precision_{t,j} + Recall_{t,j}} \quad (4)$$

Using $\bar{f1}$ ensures that all classes are equally important. Finally, the optimal combination corresponding to $\arg\max \overline{f1}(t)$ was chosen to construct the final model using the whole learning set. The final model was compared to our implementation of the BioVision algorithm using test set images.

III. RESULTS

Leave-one-image-out cross-validation scores for the AP and NFT datasets remained relatively stable ($\overline{f1}$ between 0.87 and 0.90 for the AP dataset and between 0.89 and 0.90 for the NFT dataset) using the previously defined parameter levels. Similarly to Bianconi *et al*. [11], we found that the smoothing parameter of the Gabor filter has a significant influence on classification. As shown in Figure 3a, there is a systematic bias between precision and recall when class weights are equal (i.e. for RF). Figure 3b, shows how this shift can be compensated by adjusting the minority class weight. Recall increases with weight which leads to an inevitable decrease of precision. Overweighting the minority class provided little improvement of the $\overline{f1}$ score but the equilibrium between precision and recall is valuable to ensure that the overall quantification is unbiased.

Table 2 shows the comparison of our approach with BioVision in term of classification performance on the 25 test images for the AP and the NFT datasets. Both methods appropriately classified tissue and glass slide background. Our approach showed the best score for every class in both datasets. While both methods achieved good AP (Fig. 4a) and NFT segmentations (Fig. 4b), our approach is clearly

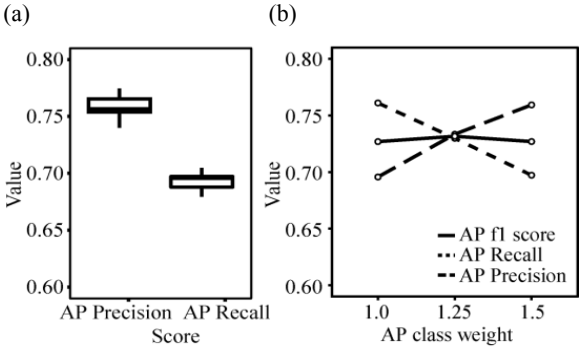

Figure 3. (a) Systematic imbalance between precision and recall for AP classification with Random Forest. (b) Effect of adjusting AP class weight.

TABLE II. F1 SCORES FOR Aβ PLAQUES (AP), NEUROFIBRILLARY TANGLES (NFT), TISSUE (TS), BACKGROUND (BK) AND MEAN F1 SCORE (MN) FOR THE TWO BENCHMARK DATASETS.

Methods	AP dataset f1 scores				NFT dataset f1 scores			
	AP	Ts	Bk	Mn	NFT	Ts	Bk	Mn
BioVision	0.68	0.96	0.95	0.87	0.49	0.95	0.97	0.80
Proposed approach	**0.76**	**0.98**	**0.96**	**0.90**	**0.56**	**0.96**	**0.98**	**0.83**

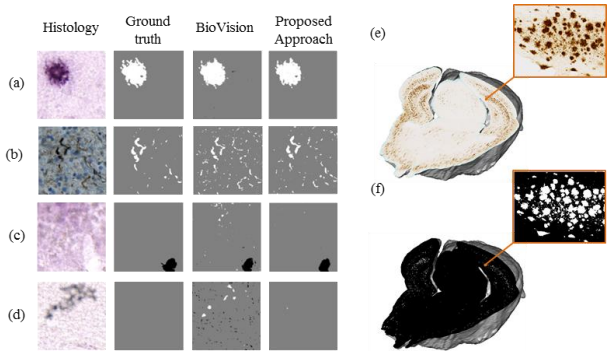

Figure 4. Segmentation results. (a-d) Comparisons of BioVision and our approach on test set images for AP and NFT detection: AP and NFT classes appear in white, tissue class in gray and background class in black. (e) 3D reconstruction of a mouse brain with AP appearing in brown and (f) the corresponding segmentation.

more robust than BioVision in the presence of high background noise (Fig. 4c) and artifacts (Fig. 4d).

The 3D whole brain dataset and its associated segmentation are shown in Figure 4e,f. Using parallel computing on a 16-core workstation, we were able to extract features and segment the whole slide image (24000×16000 pixels) of a mouse brain section in 25 minutes. Computing time speed-up was nearly linear with the number of CPUs which indicates that segmentation time for large images can be significantly decreased using computer clusters.

IV. Discussion and Conclusion

Our contribution to neuropathology image analysis is to provide an accurate and highly scalable segmentation method which is directly applicable for clinical and preclinical research. This one-step classification approach prevents the need for automatic or heavy manual post processing to remove incorrectly classified pixels. However, it should be noted that, in order to get a robust model, the learning set has to be as much representative as possible of the different structures in the neuropathology images including noise and artifacts. We believe that our approach could be applied more generally for microscopy image segmentation when objects of interest are sparse and images are noisy. Future work could be undertaken to compare various contextual descriptors, such as Gabor features variants [12] and co-occurrence matrices [13] in term of classification performance and computing time.

Acknowledgment

We would like to thank Dr Charles Duyckaerts from Pitié-Salpêtrière Hospital and the GIE NeuroCEB brain bank for providing the brain tissue sample as well as Kelly Herbert and Fanny Petit for their contribution to histology experiments.

References

[1] C. Duyckaerts, B. Delatour, and M.-C. Potier, "Classification and basic pathology of Alzheimer disease," *Acta Neuropathologica,* vol. 118, Jul 2009.

[2] H. D. Samaroo, A. C. Opsahl, J. Schreiber, *et al.,* "High throughput object-based image analysis of beta-amyloid plaques in human and transgenic mouse brain," *Journal of Neuroscience Methods,* vol. 204, pp. 179-188, 2012.

[3] A. Feki, O. Teboul, A. Dubois, *et al.,* "Fully automated and adaptive detection of amyloid plaques in stained brain sections of Alzheimer transgenic mice," *MICCAI*, 2007, vol. 10, pp. 960-8.

[4] C. Chubb, Y. Inagaki, P. Sheu, *et al.,* "BioVision: An application for the automated image analysis of histological sections," *Neurobiology of Aging,* vol. 27, pp. 1462-1476, 2006.

[5] M.E. Vandenberghe, A.S. Hérard, N. Souedet, *et al.*, "High-throughput 3D whole-brain quantitative histopathology in rodents," submitted for publication.

[6] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters" *Journal of the Optical Society of America A*, vol. 2, pp. 1160-1169, 1985.

[7] M. R. Turner, "Texture Discrimination by Gabor Functions," *Biological Cybernetics,* vol. 55, pp. 71-82, 1986.

[8] Y. Balbastre, N. Souedet, D. Rivière *et al.*, "Parallel computing in image analysis using BrainVISA software: application to histopathological staining segmentation in whole slide images," *12th European Congress on Digital Pathology*, 2014.

[9] L. Breiman, "Random forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[10] C. Chen, A. Liaw and L. Breiman, "Using Random Forest to Learn Imbalanced Data," UC Berkely, Tech. Report, 2004.

[11] F. Bianconi and A. Fernandez, "Evaluation of the effects of Gabor filter parameters on texture classification," *Pattern Recognition,* vol. 40, pp. 3325-3335, 2007.

[12] S. E. Grigorescu, N. Petkov, and P. Kruizinga, "Comparison of texture features based on Gabor filters," *IEEE Transactions on Image Processing,* vol. 11, pp. 1160-1167, 2002.

[13] C. Palm, "Color texture classification by integrative Co-occurrence matrices," Pattern Recognition, vol. 37, pp. 965-976, 2004.