Medical Image Analysis 44 (2018) 143–155

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Learning non-linear patch embeddings with neural networks for label fusion

Gerard Sanroma[a,*], Oualid M. Benkarim[a], Gemma Piella[a], Oscar Camara[a], Guorong Wu[b], Dinggang Shen[b,e,*], Juan D. Gispert[c], José Luis Molinuevo[c], Miguel A. González Ballester[a,d], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] Department of Information and Communication Technologies, Universitat Pompeu Fabra, Tànger 122-140, Barcelona 08018, Spain
[b] Department of Radiology and BRIC, University of North Carolina at Chapel Hill, 102 Mason Farm Rd., NC 27599, USA
[c] Barcelonaβeta Brain Research Center, Pasqual Maragall Foundation, Wellington 30, Barcelona 08005 Spain
[d] ICREA, Pg. Lluis Companys 23, Barcelona 08010 Spain
[e] Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea

A R T I C L E I N F O

Article history:
Received 21 March 2017
Revised 5 October 2017
Accepted 27 November 2017
Available online 2 December 2017

Keywords:
Patch-based label fusion
Multi-atlas segmentation
Neural networks
Embedding
Brain MRI
Hippocampus

A B S T R A C T

In brain structural segmentation, multi-atlas strategies are increasingly being used over single-atlas strategies because of their ability to fit a wider anatomical variability. Patch-based label fusion (PBLF) is a type of such multi-atlas approaches that labels each target point as a weighted combination of neighboring atlas labels, where atlas points with higher local similarity to the target contribute more strongly to label fusion. PBLF can be potentially improved by increasing the discriminative capabilities of the local image similarity measurements. We propose a framework to compute patch embeddings using neural networks so as to increase discriminative abilities of similarity-based weighted voting in PBLF. As particular cases, our framework includes embeddings with different complexities, namely, a simple scaling, an affine transformation, and non-linear transformations. We compare our method with state-of-the-art alternatives in whole hippocampus and hippocampal subfields segmentation experiments using publicly available datasets. Results show that even the simplest versions of our method outperform standard PBLF, thus evidencing the benefits of discriminative learning. More complex transformation models tended to achieve better results than simpler ones, obtaining a considerable increase in average Dice score compared to standard PBLF.

1. Introduction

Segmentation of brain structures from magnetic resonance images (MRI) is an important step in many neuroscience applications, including discovery of morphological biomarkers, monitoring disease progression or diagnosis. For example, segmentation is widely used as basic image quantification step in studies of early brain development (Benkarim et al., 2017) and dementia (Chupin et al., 2009; Li et al., 2007).

Multi-atlas segmentation (MAS) is being increasingly used for segmenting brain MRI (Sanroma et al., 2016). In MAS, a set of atlas images are first registered to the image to be segmented (i.e., target) along with their anatomical labelmaps containing the spatial overlay of the anatomical structures. Then, the so-called *label fusion* process, labels each target point using the support of the corresponding atlas labels. Compared to using a single atlas, MAS can potentially fit a wider anatomical variability and has higher robustness to registration errors. Image intensities are often not sufficient for globally discriminating the different structures and therefore, spatial constraints are essential (Colliot et al., 2006). Such spatial constraints are usually implemented by restricting the set of feasible labels for each target point to the set of labels in neighboring atlas points.

Patch-based label fusion (PBLF) is a popular approach that computes each target label as a weighted combination of neighboring atlas labels, where atlas locations with higher local image similarity to the to-be-segmented target point have higher weight in the

* Corresponding authors.
E-mail address: gerard.sanroma@upf.edu (G. Sanroma).

[1] Part of the data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) . As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.media.2017.11.013
1361-8415/

combination (Artaechevarria et al., 2009; Coupé et al., 2011; Wang et al., 2013). Here, the similarity between local image patches around each target and atlas point is taken as a proxy for the local registration accuracy and hence, for anatomical correspondence.

PBLF can potentially be improved by increasing the discriminative capabilities of patch similarity measurements. For example, we proposed to learn discriminative patch embeddings reflecting the latent anatomical similarity between patches (Sanroma et al., 2015a). A similar approach was recently proposed using convolutional neural networks (CNNs) (Yang et al., 2016). Such learned embeddings are then used in standard PBLF. Other supervised approaches for learning optimal fusion rules have been presented. For example, in Sanroma et al. (2015b) we proposed a transductive learning approach, and in Benkarim et al. (2016) we proposed to integrate discriminative learning into probabilistic label fusion. Semi-supervised learning approaches have also been proposed for propagating the anatomical labels from atlases to targets (Guo and Zhang, 2012; Koch et al., 2014). Machine learning techniques such as support vector machines (SVM) (Cortes and Vapnik, 1995) have also been used (Bai et al., 2015; Hao et al., 2013; Sdika, 2015).

In practice, most of these methods learn a different model (i.e., classifier) at each location (Bai et al., 2015; Benkarim et al., 2016; Guo and Zhang, 2012; Hao et al., 2013; Koch et al., 2014; Sanroma et al., 2015a; 2015b; Sdika, 2015). This serves two purposes: (1) it implicitly imposes spatial constraints by restricting the training samples on each model to only neighboring atlas locations; and (2) it divides the difficult problem of finding a single global model into the problem of finding multiple simpler local models. However, this increases the complexity of storage and use of the method due to the high number of local models generated, which can easily reach several hundred thousands, even after restricting the modeling to only the most challenging regions. Another difficulty when using local models is that training images must be in spatial correspondence in order to retrieve the training data for each local model. As a result, some methods opt for training the models in a common template space (Sanroma et al., 2015a). This implies that the target image must be segmented in the template space, incurring in interpolation errors when re-sampling the resulting segmentation to the original target space. Moreover, methods that consider pairwise relationships (Benkarim et al., 2016; Sanroma et al., 2015a; Yang et al., 2016) need pairwise registrations among the training images to evaluate the similarity between the embedded patches. This has memory complexity $\mathcal{O}(N^2)$ during training, with N being the number of atlases, thus limiting the amount of atlases that can effectively be used for training. A related approach uses convolutional neural networks (CNN) for segmenting cardiac images (Yang et al., 2016). For an input image, they obtain a stack of output images by applying the learned convolutional filters. The number of images in the stack is related to the dimensionality of the output embeddings. Thus, memory requirements for label fusion are $\mathcal{O}(N \times d)$, where N is the number of atlases and d the dimensionality of the output embedding. This poses serious limitations on the number of atlases at test time (in fact they only use 5 atlases for each target image). In brain MRI segmentation, usually more than 10 atlases are used (Aljabar et al., 2009; Lotjonen et al., 2010; Sanroma et al., 2014).

To overcome these issues, we propose a method to learn discriminative patch embeddings using neural networks,[2] with the following contributions:

- By incorporating our method into the regular label fusion process, we focus on the problem of learning the model, thus leveraging the capability of the label fusion process of restricting the set of possible labels at each point.
- The previous contribution facilitates that we compute a single model per bilateral structure (i.e., one model for both left and right parts of each structure). We take advantage of stochastic gradient descent (SGD) in order to process the vast amounts of data in small mini-batches. Therefore, our method allows for a practical storage and use.
- We learn the model in the native space of each training atlas instead of using a template. Therefore, models are learned in the same space as they were annotated, thus avoiding interpolation artifacts during training. Another advantage is that models are orientation-invariant and hence target images can directly be segmented in their native space. As consequence of this, the target anatomy can directly be quantified from the resulting segmentation, without need to correct for geometric distortions caused by the transformation to the template space.
- We learn the embeddings using patch relationships *within the same image*, leading to an attractive $\mathcal{O}(N)$ storage complexity at training (with N the number of atlases), compared to more costly approaches (Benkarim et al., 2016; Sanroma et al., 2015a; Yang et al., 2016) that require pairwise atlas registrations in this phase.
- Our method embeds the image patches independently rather than the whole images. Therefore, we can generate output embeddings of arbitrary dimensionality without compromising the number of atlases that can reasonably be handled (memory requirement at segmentation time is $\mathcal{O}(N)$).

We apply our method to segment the whole hippocampus and the hippocampal subfields (see Section 4), a structure targeted by many studies on psychiatric and neurological disorders (Chupin et al., 2009; Li et al., 2007). Accurate segmentation methods are required in order to quantify the subtle morphological changes undergone by these structures, especially in the early stages of the disease (Frisoni et al., 2010; West et al., 2004).

In the next section, we introduce multi-atlas segmentation and how it can be improved by using embedding techniques, before describing our method in Section 3.

2. Multi-atlas segmentation

Let us denote $\hat{X}$ the target image to be segmented and $X_i, i = 1, \ldots, N$ a set of atlas images along with their corresponding labelmaps Y_i containing the anatomical information. Multi-atlas segmentation (MAS) aims at estimating the segmentation on the target image using the atlas images and their labelmaps.

This is implemented by (1) registering the atlas images to the target and (2) computing each target label as a combination of locally corresponding atlas labels, the so-called *label fusion*.

Weighted voting is a popular label fusion approach that computes the target label as a weighted combination of atlas labels (Artaechevarria et al., 2009; Coupé et al., 2011; Wang et al., 2013). More formally, the label $\hat{y}_p$ for a given target point $p \in \Omega$ in the image domain Ω, is computed as:

$$\hat{y}_p = \arg\max_l \sum_{iq} \omega_{iq} \delta\left[y_{iq} = l\right] \tag{1}$$

where y_{iq} is the label in ith atlas at point $q \in \mathcal{N}_p$ in the *spatial* neighborhood of $p \in \Omega$, ω_{iq} is the weight denoting the importance of y_{iq} in determining the target label, and δ is Kronecker's delta (i.e., $\delta[a = b]$ is 1 if $a = b$, 0 otherwise).

One of the earliest label fusion approaches, known as majority voting (Heckemann et al., 2006; Rohlfing et al., 2004) assigns each target label the atlas label occurring most frequently, which is equivalent to using a constant weight, i.e., $\omega_{iq} = c,\ \forall i, q$. This simple strategy already achieves substantial improvement

[2] The code of the method is available at https://github.com/gsanroma/deeplf.