Multi-scale Structured CNN with Label Consistency for Brain MR Image Segmentation

Siqi Bao and Albert C. S. Chung
Lo Kwee-Seong Medical Image Analysis Laboratory,
Department of Computer Science and Engineering,
The Hong Kong University of Science and Technology, Hong Kong.

Abstract. In this paper, a novel method for brain MR image segmentation has been proposed, with deep learning techniques to obtain preliminary labeling and graphical models to produce the final result. A specific architecture, namely multi-scale structured convolutional neural networks (MS-CNN), is designed to capture discriminative features for each sub-cortical structure and to generate a label probability map for the target image. Due to complex background in brain images and the lack of spatial constraints among testing samples, the initial result obtained with MS-CNN is not smooth. To deal with this problem, dynamic random walker with decayed region of interest is then proposed to enforce label consistency. Comprehensive evaluations have been carried out on two publicly available datasets and experimental results indicate that the proposed method can obtain better segmentation quality efficiently.

1 Introduction

Image segmentation is a critical step in quantitative brain image analysis, clinical diagnosis and treatment plan. However, segmenting these sub-cortical structures is difficult because they are small and often exhibit large shape variations. Although manual annotation is a standard procedure for obtaining quality segmentation, it is time-consuming and can suffer from inter- and intra-observer inconsistencies. In recent years, researchers have been focusing on developing advanced non-rigid registration [6] and label fusion methods [12], to segment the target image through fusing the warped label maps from atlases. The segmentation quality with non-rigid registration can outperform that with affine transformation, but at the cost of expensive computation. As the time consumed to label one target image increases along with the number of atlases, the heavy computation burden hinders its application to a large atlas database.

Machine learning is another way to collect atlas priors from training data and can guarantee labeling efficiency during testing. Much attention has been drawn to the deep learning techniques, especially deep convolutional neural networks (CNN) in the last few years. CNN was inspired by visual mechanism in biology and was first introduced for document recognition [8], later widely applied to image classification, scene parsing, acoustic analysis and so on. To improve the performance of CNN, two intuitive ways are generally utilized, i.e., increasing
the depth or the width of layers. Apart from the straightforward enlargement of network architecture, some elegant micro-structures have also been designed to enhance the capability recently. Network in Network (NIN) structure [9] replaces the conventional linear transform in convolution layer with multilayer perceptron, seeking to obtain a better abstraction on each local receptive field. The winner of ImageNet Challenge 2014 introduced the module Inception [14] which settles several parallel convolution layers with different kernel sizes to extract features, rather than placing serial convolution layers along the feature transfer path in the deep network.

Despite the rapid development of CNN in computer vision, it is still infeasible to directly apply the latest methods to brain magnetic resonance (MR) image labeling, mainly due to the complex background and poor contrast in medical images. As such, in this paper we propose a novel network called multi-scale structured CNN (MS-CNN), on top of which label consistency is introduced to refine the preliminary results obtained with deep learning.

2 Methodology

2.1 Background

The basic component in CNN is the convolution and polling layer, which is illustrated in Fig. 1. Three layers are displayed in this figure, including input, convolution and polling, and each layer can be comprised of several maps (channels). The convolution step carries out two successive operations: linear transform and non-linear activation, generating a set of feature maps. For a node in the k-th feature map, with a small sub-region (receptive field, shown as Red Patches) from last layer as input, its convolution response can be estimated as $a^k = f(W^k x + b^k)$, where $x$ is the flattened vector from the receptive field, $W^k$ is the weight vector associated with the k-th feature map and $b^k$ is the corresponding bias. $f(\cdot)$ refers to the non-linear activation, which can be formulated with conventional saturating function (such as sigmoid or tanh) or recently proposed non-saturating unit, like rectified linear unit (ReLU) [10]. Given the significant acceleration on training speed brought by non-saturating unit [7], in this paper, ReLU is chosen as the activation function and the above convolution response can be then rewritten as $a^k = \max(W^k x + b^k, 0)$.

![Fig. 1. Illustration of Convolution and Polling Layer (color image).](image-url)
The polling procedure usually follows the convolution layer, which focuses on a local patch of one feature map each time and slides through the whole map with a certain stride. The strategy used here can be maximum polling, which selects the highest value within the small patch, or average polling, which estimates its mean value. This operation will only make the size of features maps shrunk, while the number of channels remains the same. The output of the polling layer can be used as input for the next convolution layer to construct deep CNN.

### 2.2 Multi-scale Structured CNN

Because of the eminent performance of CNN in computer vision, we apply it to brain MR image segmentation. Although these two applications share some similarity, there still exist several significant differences between them. First, for general image classification, its task is to make an inference about the image category based on achieved abstraction. While for brain image segmentation, it needs to assign an accurate label to each pixel with learned features. Furthermore, as compared with the relatively less noisy background in ordinary images, the brain anatomical structures are adjacent with each other and they may share similar histogram profiles in MR images, which makes the labeling even more challenging. To deal with the above difficulties, a specific architecture called multi-scale structured CNN (MS-CNN) is designed in this paper (Fig. 2).

![Fig. 2. Overview of the proposed multi-scale structured CNN. Red Square: initial input; Purple and Green Squares: generated multi-scale input; Blue Cube: convolution layer; Dashed Line: hidden average polling layer; Origin Dashed Rectangle: feature extraction module; Strip: full connected layer; Blue Node: output neuron (color image).](image)

During training, for each structure taken into consideration, the rest of the brain image can be regarded as background and most parts are meaningless. As such, instead of using the whole image as input, the region of interest (ROI) is determined first and then its inside pixels represented with surrounding 2D patches are extracted as training data, together with their corresponding labels.
As shown in Fig. 2, with a large patch (Red Square, size $20 \times 20 \times 1$) as initial input, multi-scale input is generated by selecting patches around the center pixel (Purple Node) with a set of radii. For each scale, the input is fed into its corresponding module, with several alternate convolution and polling layers. During this stage, there is no interaction among the three modules and the features are extracted independently for each scale. Then these features are concatenated together, followed with several full connected layers to estimate the final output.

The strategy of CNN accompanied with multi-scale input has been utilized by some recent works in scene labeling. In [3], the original image is represented with Laplacian pyramid as input and the classical CNN [8] is utilized to generate features for each pyramid level. In [11], instead of using several convolution layers with different parameter settings, one fixed convolution filter is applied recurrently to last layer output and pyramid level at the homologous scale.

Unlike the above methods, which apply a single (fixed) network to different levels (scale) of image Laplacian pyramid, MS-CNN extracts a set of patches with different radii (scale) and each scale has its own specific network. The benefits are twofold. With a large patch instead of a whole image as input, most of the meaningless background can be removed, which greatly reduces the computation burden. Thanks to the design of independent network for each scale, the abstract properties, like structure shape information can be better captured with the large input patches and the detailed features, like obscure boundaries are more likely to be detected with the small patches at a lower scale.

2.3 Label Consistency

Given the crowding of anatomical structures in brain image, the ROI for a particular structure can not avoid including pixels from adjacent structures. As various structures can still have some similar segments, these pixels may give a high response to the learned features and can be mislabeled as foreground. Moreover, although neighboring pixels are supposed to have consistent labels, due to the lack of spatial constraints on testing samples, some outliers can be generated if only relying on the learning-based method. In Fig. 3, the left sub-figure is manually labeled ground truth for reference and the middle one is the segmentation obtained with MS-CNN. It can be noticed that the result is not smooth and some holes or protuberances have been marked with Red Circles. Therefore, there is a need to enforce label consistency and refine the segmentation result.

![Ground Truth](image1)  ![MS-CNN](image2)  ![MS-CNN with Label Consistency](image3)

**Fig. 3.** Visual results of Hippocampus with Dice values (color image).
For learning-based labeling methods, some graphical models are usually employed in post-processing procedure, such as Markov random fields or conditional random fields. In this paper, a novel label consistency method is proposed under the framework of random walker (RW), which is also formulated on an undirected graph $G = (V, E)$. The node set $V$ is comprised of three subsets: foreground seeds $V_F$, background seeds $V_B$ and candidate nodes $V_c$ whose labels need to be refined. The edge $e_{ij} \in E$ stands for the connection between nodes $v_i$ and $v_j$, with $w_{ij}$ as edge weight. The objective function of RW is given as [4]:

$$
\min \sum_{v_i} [w_{iF}^2(z_i - 1)^2 + w_{iB}^2z_i^2] + \sum_{e_{ij}} w_{ij}^2(z_i - z_j)^2, \quad \text{s.t. } z_F = 1, \ z_B = 0. \quad (1)
$$

$z_i$ represents the probability that one node belongs to the foreground ($0 \leq z_i \leq 1$), with the constraint that $z_F = 1$ and $z_B = 0$. For each node $v_i$, its connections can be divided into two categories: the direct linkage with seeds ($w_{iF}, w_{iB}$) and association with neighboring pixels ($w_{ij}$). The probabilities estimated with MS-CNN that $v_i$ belongs to the foreground and background are encoded as prior $w_{iF}$ and $w_{iB}$ respectively. As for the lattice connection, the edge weight is defined with the Gaussian function:

$$
w_{ij} = \exp(-\beta(I(v_i) - I(v_j))^2), \quad (2)
$$

where $\mathcal{N}(\cdot)$ refers to the 6-nearest neighbors in 3D medical images, $I(\cdot)$ is the intensity value and $\beta$ is a tuning parameter.

As the performance of RW is sensitive to seed positions [13], the foreground and background seeds need to be placed carefully. In this paper, a dynamic RW with decayed region of interest (ROI) is proposed to select nodes and update the labeling result iteratively, which is presented in Fig. 4. The cross section of 3D image is used for illustration. The Blue Curve at Iteration 0 is the surface of initial segmentation result for the target image, which is generated with affine transformation and majority voting. The pixels on the inner (Red) and outer (Black) iso-surfaces with a certain distance $d^0$ to the structural surface are chosen as foreground and background seeds respectively. As for the pixels located between these two iso-surfaces, they are collected as candidate nodes. After the selection of node set $V^0$ for Iteration 0, the connecting edges can be added, with weights estimated using MS-CNN label priors and Equation (2).

![Fig. 4. Dynamic RW with Decayed ROI (color image).](image-url)
Based on the constructed graph $G^0$, by minimizing the energy function in Equation (1), the label for each candidate node can be updated correspondingly, $L(v_i) = 1$ if $z_i \geq 0.5$ and $L(v_i) = 0$ otherwise. With the refined structural surface (Blue Curve at Iteration 1), a new set of nodes can be selected by setting $d^i = \alpha \cdot d^0$, where $\alpha$ is the decayed coefficient. The reasons to employ a shrunk ROI are twofold. On one hand, given a more reliable segmentation result, pixels with a shorter distance can be regarded as seeds and then fewer candidate nodes need to be refined. In this way, computational cost at each iteration can be reduced gradually. On the other hand, the shrunk ROI can strengthen the influence of seeds on candidate nodes, as the path length through image lattice is shortened.

The graph construction and RW optimization process will be carried out alternately until $d^N < 1$. In Fig. 3, the right sub-figure displays the results after enforcing label consistency. Although it still has some defects, as compared with the result of MS-CNN, the boundary is more smooth and the labeling accuracy measured with Dice Coefficients has been improved by 2.6%. More comprehensive evaluations will be presented in the experiment section.

3 Experiments

Experiments have been carried out on two publicly available datasets: IBSR\(^1\) and LPBA40\(^2\). IBSR includes 18 T1-weighted MR brain volumes, with 84 structures manually labeled and LPBA40 has 40 brain MR images with 56 structures delineated. In the experiments, each dataset was randomly divided into two equal parts for training and testing respectively. For each sub-cortical structure, the shortest distance between one pixel and the structural surface was estimated first and then ROI was set to include pixels with a distance smaller than $d^0$. Patches surrounding these pixels and their labels are extracted and fed into MS-CNN for training. With learned discriminative features, the label probability for each pixel in the target image can be estimated efficiently by performing once forward pass on the testing sample. Based on priors estimated with MS-CNN, label consistency was then carried out to further improve the segmentation quality.

To evaluate the performance of the proposed method, the comparison with other methods has been carried out. As several different voxel resolutions exist in each dataset, affine transformations between atlases and the target image were first conducted as pre-processing. The initial structural surface used in label consistency (Blue Curve at Iteration 0) was generated by fusing the warped label maps with majority voting (MV). And the result of MV was also employed as base line for comparison. Although general label fusion methods work well for fusing results with non-rigid registration, in the light of efficiency, most of them can not adapt well to the inferior results with affine transformation. With a large search window, patch-based label fusion (PBL) [2] can still be applied to the latter case with an acceptable time consumption.

---

\(^1\) http://www.nitrc.org/projects/ibsr
\(^2\) http://www.loni.usc.edu/atlases/Atlas_Detail.php?atlas_id=12
Table 1. Results on IBSR and LPBA40 datasets with available sub-cortical structures.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Amygdala</th>
<th>Caudate</th>
<th>Hippocampus</th>
<th>Pallidum</th>
<th>Putamen</th>
<th>Thalamus</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBSR</td>
<td>0.516</td>
<td>0.616</td>
<td>0.550</td>
<td>0.690</td>
<td>0.767</td>
<td>0.811</td>
<td>0.658</td>
</tr>
<tr>
<td>LPBA40</td>
<td>0.614</td>
<td>0.809</td>
<td>0.715</td>
<td>0.730</td>
<td>0.824</td>
<td>0.871</td>
<td>0.760</td>
</tr>
<tr>
<td>MV</td>
<td>0.654</td>
<td>0.849</td>
<td>0.788</td>
<td>0.787</td>
<td>0.875</td>
<td>0.889</td>
<td>0.807</td>
</tr>
<tr>
<td>PBL</td>
<td>-</td>
<td>-</td>
<td>0.817</td>
<td>-</td>
<td>0.882</td>
<td>-</td>
<td>0.822</td>
</tr>
<tr>
<td>MS-CNN</td>
<td>0.672</td>
<td>0.870</td>
<td>0.817</td>
<td>0.795</td>
<td>0.792</td>
<td>0.898</td>
<td>0.843</td>
</tr>
<tr>
<td>LC</td>
<td>-</td>
<td>-</td>
<td>0.837</td>
<td>-</td>
<td>0.843</td>
<td>-</td>
<td>0.843</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.827</td>
<td></td>
<td>0.850</td>
<td></td>
<td>0.850</td>
</tr>
</tbody>
</table>

In the experiments, affine transformation was estimated with the FLIRT toolbox [5] and PBL was performed with source code provided by [1]. The parameter settings are given as follows: $d^0 = 3$, $\alpha = 0.7$ and $\beta = 5$. For each category, the size of training samples is around 0.4 million. In PBL, following the suggestions given in [2], for each pixel, the size of its surrounding patch was set to $7 \times 7 \times 7$ and the search volume was $9 \times 9 \times 9$. The segmentation results measured with Dice Coefficients are listed in Table 1, with the highest value shown in Red. Here LC refers to the complete version of the proposed method, with MS-CNN followed by label consistency. These results indicate that the proposed method can obtain the best segmentation quality as compared with other standard methods.

The detailed analysis of the proposed method was also carried out. To test the effects of multi-scale strategy used in MS-CNN, only the feature extraction module of the largest scale was kept to train a model for Hippocampi on IBSR dataset. The results obtained with multi- and single-scale approaches on left and right Hippocampi are shown in Fig. 5(a). The left two sets of bars are learning-based results and the right two sets of bars are those refined with label consistency. Although the structure of Hippocampus is complicated, the results indicate that with multi-scale strategy, more discriminative features can be captured and the labeling result can be improved.

![Fig. 5. Detailed analysis of components in the proposed method.](image-url)

To check the effects of dynamic RW with shrunk ROI, the anterior result estimated with MS-CNN and that generated by label consistency at each iteration are displayed in Fig. 5(b). It shows that the accuracy can be improved con-
siently with the iteration process. Moreover, as MS-CNN is a learning-based method, which takes little time to estimate label probability for testing samples, and the shrunk ROI strategy employed in label consistency greatly reduces computation burden, the proposed method can segment each structure efficiently.

4 Conclusion

In this paper, a novel method using deep learning techniques has been proposed for brain magnetic resonance image segmentation. A multi-scale structured CNN approach is introduced to capture discriminative features from a large input patch. Due to the lack of constraints among testing patches, embracing learning method alone often leads to a rough boundary and desultory segmentation result. As such, a dynamic random walker method with decayed region of interest is introduced as post-processing to enforce label consistency. Experimental results demonstrate that the proposed method can obtain better performance as compared with other state-of-the-art methods.

References