3D Deep Affine-Invariant Shape Learning for Brain MR Image Segmentation

Zhou He, Siqi Bao, and Albert Chung(✉)

Department of Computer Science and Engineering,
The Hong Kong University of Science and Technology, Kowloon, Hong Kong
achung@cse.ust.hk

Abstract. Recent advancements in medical image segmentation techniques have achieved compelling results. However, most of the widely used approaches do not take into account any prior knowledge about the shape of the biomedical structures being segmented. More recently, some works have presented approaches to incorporate shape information. However, many of them are indeed introducing more parameters to the segmentation network to learn the general features, which any segmentation network is able learn, instead of specifically *shape* features. In this paper, we present a novel approach that seamlessly integrates the shape information into the segmentation network. Experiments on human brain MRI segmentation demonstrate that our approach can achieve a lower Hausdorff distance and higher Dice coefficient than the state-of-the-art approaches.

1 Introduction

A variety of approaches have been adopted to address the challenging problem of 3D medical image segmentation, such as 3D U-Net [1] and V-Net [4], which have been proven to be highly effective. These approaches, however, simply transplant the 2D image semantic segmentation algorithms to a 3D medical image analysis context. They have little awareness to the fact that 3D medical structures of the same class, unlike objects in 2D natural images, in general have similar shapes. For example, for a 2D natural image segmentation task on the class of 'person', different persons could be very different in shape since a person may have different poses when being photographed, e.g., arms opened/closed, sitting/standing, etc. For the segmentation on biomedical structures such as human caudate nucleus, all caudate nuclei have very similar shape with little structural variation. However, this information is rarely used in deep learning-based 3D medical image segmentation. While some recent literature has introduced some approaches to leverage shape information, many of them are merely introducing more hyperparameters to the network to increase its capacity, while not actually using exactly the *shape* information.

In this paper, we present a novel approach which incorporates the information about the shape of the segmentation target into the loss function of a general

D. Stoyanov et al. (Eds.): DLMIA 2018/ML-CDS 2018, LNCS 11045, pp. 56–64, 2018.
https://doi.org/10.1007/978-3-030-00889-5_7

3D segmentation network. This shape information is deep-learned from a fully convolutional network, whose feature map of the final layer (defined as the **shape signature**) captures the important global shape information. We first pre-train this shape-learning network by ground truth label maps that have undergone different affine transformations, and then have the weights of this network fixed. This shape-learning network will then be able to capture the essential shape information that is invariant to affine transformation. Afterwards, when training the segmentation network, the prediction label map and ground truth label map will both be fed into the pre-trained shape-learning network, and the Euclidean distance between their shape signatures will quantify the dissimilarity in shape between the segmentation prediction and ground truth. This shape loss is then added to the loss function of the segmentation network to facilitate the training.

Our main contributions are summarized as follows:

1. Designed a novel shape-learning network that is able to capture the affine-invariant global shape information in the final feature map;
2. Incorporated the shape dissimilarity information to the segmentation network, making it shape-aware;

2 Related Work

We start by reviewing related prior works on general medical image segmentation, and the utilization of shape information.

2.1 Medical Image Segmentation

Deep learning-based image semantic segmentation became highly successful since the emergence of Fully-convolutional Network (FCN) [3]. This approach has later been adapted to a biomedical image segmentation setting with the novel design of U-Net [7], which contains skip connections between the contracting path and expanding path so that the intricate details in biomedical images can be kept. Recently, U-Net has been modified to accommodate 3D volumes by replacing all the 2D convolutions and convolution transposes by their 3D counterparts, as described in 3D U-Net [1]. Apart from the change in network architecture, some other adaptations have been made to make CNNs more compatible with medical image segmentation. For instance, in V-Net [4], the loss function is derived from Dice coefficient which is a common metric in medical image segmentation.

2.2 The Utilization of Shape Information in Segmentation

Some prior works claimed to have leveraged the shape information of biomedical structures for segmentation purpose. [6] introduced an autoencoder known as Shape Regularization Network (SRN) that regularizes the segmentation result to make it conform to the shape it should have. Its functions include eliminating any noisy part from the general shape, or filling up any holes in the preliminary segmentation result. A more recent work Anatomically Constrained Neural

Networks (ACNN) [5] used an autoencoder to learn the shape by training that autoencoder to reconstruct a label map itself, and used the Euclidean distance between the bottleneck layers of the autoencoder to quantify the dissimilarity in shape.

A commonality among these prior works is that they introduce another network which is trained to capture the shape information, and this network is then used to guide the segmentation network. However, the shape learner in SRN and ACNN are both learning the general features of a 3D structure, including position, volume, shape, etc., instead of specifically learning the *shape*. Inspired by these issues, we propose an approach to learn the essential shape information that is invariant to affine transformations.

3 Methodology

3.1 Overview

While the term *shape* may have many definitions in different settings, in the medical image segmentation setting here, we define it to be the intrinsic properties of the 3D biomedical structures that are invariant to spatial affine transformations, including rotation, translation and scaling, etc. The network architecture used in our approach is composed of two parts, where the first part is to capture the shape information, and the second part to use it.

Concretely, the first part, defined as **shape-learning network**, is a 3D fully convolutional neural network (ConvNet), with the input being the raw binary 3D label map of the biomedical structure being segmented, and the output being a one-channel low-resolution feature map (hereafter referred to as **shape signature**). The second part, defined as **shape-guided segmentation network**, is a segmentation network with the architecture modeled after the 3D U-Net [1] and loss function being the sum of Dice loss and shape loss. The role of the shape-learning network is to learn the **shape signature** of a 3D binary label map, and this information would later become a part of the loss function of the segmentation network. The architecture of the shape-learning network is shown in Fig. 1, while the complete illustration showing the full network architecture is attached at Fig. 2.

3.2 Shape-Learning Network

The first step is to train the shape-learning network. In every iteration of training, we feed the shape-learning network with two binary label maps which are the **same structure** from the **same subject** that have gone through **different affine transformations**. Since these two label maps come from the same subject's same structure under different affine transformations, we call them an *affine pair*, and argue that they contain exactly the **same shape information**. If the network was able to capture shape information well, the difference between the shape signatures of these two label maps from the same *affine pair* should

be small. We therefore compute the Euclidean distance between the shape signatures of these two label maps, and use this difference in Euclidean distance as loss and propagate the loss through the entire network and to update the network weights.

Let the ground truth label map be $M \in \mathbb{R}^{w \times h \times d}$ and the shape signature be $\hat{M} \in \mathbb{R}^{w' \times h' \times d'}$ where w', h', d' are much smaller than w, h, d respectively, the shape-learning network is essentially a non-linear mapping from M to $\hat{M}$, namely $\hat{M} = g_\theta(M)$, where θ is the weights in the convolutional layers of this network. Given this shape-learning network g_θ, the shape loss between two binary label maps $M_1, M_2 \in \mathbb{R}^{w \times h \times d}$ is therefore

$$\mathcal{L}_{shape}(M_1, M_2) = \|g_\theta(M_1) - g_\theta(M_2)\|_2$$

Training this shape-learning network therefore essentially means finding the θ that satisfies

$$\theta = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}_{shape}(M_1, M_2)$$

where M_1 and M_2 are two instances of the same structure in the same subject, that have gone through different random affine transformation. After the training is finished, given a label map M, $g_\theta(M)$ gives the shape signature of this label map.

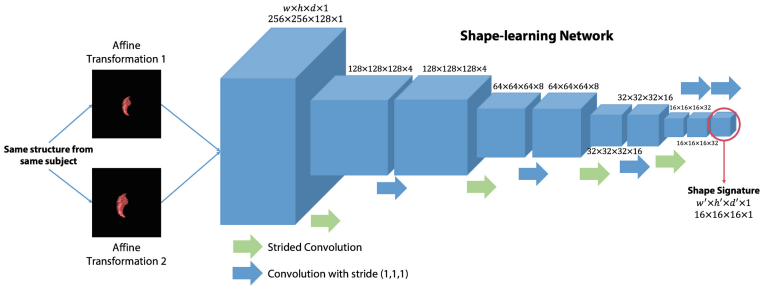

Fig. 1. Architecture of the shape-learning network.

3.3 Shape-Guided Segmentation Network

After the training of the shape-learning network is finished, we then train the segmentation network which is responsible for generating the segmentation label map. The segmentation network is a mapping f from the input (the raw voxels of brain MR image) I to the segmentation result $\tilde{M}$ defined as $\tilde{M} = f_W(I)$ where W is the weights of the segmentation network. The difference between the segmentation result $\tilde{M}$ and ground truth label map M is first measured by the Dice loss defined as

$$\mathcal{L}_{dice}(M, \tilde{M}) = 1 - \frac{2\sum_i M_i \tilde{M}_i}{\sum_i M_i + \sum_i \tilde{M}_i}$$

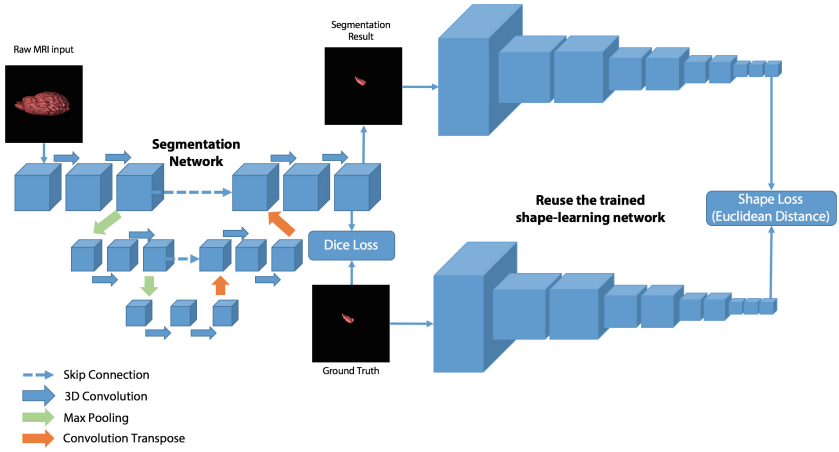

Fig. 2. Diagram of the full network architecture. The segmentation network, following a 3D U-Net architecture, is shown on the left, and the pre-trained shape-learning network that extracts the shape signature is shown on the right. Number of channels is not reflected on this diagram for brevity.

And by the definition given in the previous section, the shape loss between $\tilde{M}$ and the ground truth M is defined as

$$\mathcal{L}_{shape}(M, \tilde{M}) = \|g_\theta(M) - g_\theta(\tilde{M})\|_2$$

After adding the shape loss term to the Dice loss, we obtain the total loss which is

$$\mathcal{L}_{total}(M, \tilde{M}) = \mathcal{L}_{dice}(M, \tilde{M}) + \alpha \mathcal{L}_{shape}(M, \tilde{M})$$

where α is a hyperparameter that balances the weights of Dice loss and shape loss. The weights W of the segmentation network is

$$W = \underset{W}{\operatorname{argmin}}\, \mathcal{L}_{total}(M, \tilde{M})$$

and the segmentation network can be trained by the Stochastic Gradient Descent (SGD) with backpropagation since the entire pipeline is differentiable end-to-end.

4 Experiments

4.1 Experimental Setup

Experiments have been implemented on the human left and right caudate nucleus, as well as left and right hippocampus in the LONI Probabilistic Brain Atlas (LPBA40) dataset [8], which is a publicly available series of maps of human brain anatomic regions. The Magnetic Resonance (MR) images in the native

space are used as raw input, while the label maps in the delineation space are the ground truth labels.

All MRI inputs and their corresponding labels are preprocessed and cropped to a region of $256 \times 256 \times 128$ in size, which is identical for every subject. Raw MRI inputs are preprocessed so that the original 12-bit image representation is normalized to a mean of 0.0 and standard deviation of 1.0. The label maps are further preprocessed for left and right caudate nuclei respectively, so that the label maps of both structures are binary three-dimensional arrays. Data augmentation operations on the training data include randomly rotating the object in 3D space up to 8°, randomly scaling the object from 0.85 times to 1.15 times, as well as randomly translating the object. Left and right caudate nucleus and left and right hippocampus are all processed separately and are ran in separate experiments. Note that these transformations are also used in preparing an affine pair when training the shape-learning network, which requires the same structure to go through two random affine transformations.

4.2 Training the Shape-Learning Network

We first train the shape-learning network, and demonstrate why it is able to capture the essential shape information in the shape signature layer. The shape-learning network was trained with *affine pairs*, where the binary label maps have both gone through a random affine transformation that was employed in the data augmentation step. On each structure, we train for 200 iterations with batch size 1 and learning rate 1×10^{-4} on Adam Optimizer [2]. Experimental results illustrated in Table 1 demonstrate that the average difference in shape signature between affine pairs of the same subject's same structure is much lower than the average shape difference between pairs from different subjects. Therefore, a well-trained shape-learning network is able to capture a structure's essential shape information that is invariant to affine transformations.

Table 1. Average shape loss of 50 random *affine pairs* of the four biomedical structures tested, when the pairs are drawn from the same subject or a different subject.

	Same subject	Different subjects
Left Caudate	**0.120**	0.316
Right Caudate	**0.083**	0.210
Left Hippocampus	**0.251**	0.787
Right Hippocampus	**0.088**	0.277

4.3 Training the Segmentation Network

After finish training the shape-learning network, we freeze its weights and train the segmentation network. The experiments were also run with a batch size

of 1, with the optimizer being Adam Optimizer [2] and the learning rate being 1×10^{-4}. The weight of shape loss α was chosen experimentally to be 0.1, and the models of left and right caudate nucleus and left and right hippocampus are first trained without shape loss for 800 iterations, and then trained with shape loss for another 400 iterations. To prevent the shape loss term from being extremely large, we experimentally set it to be capped at 1.0. As ablation experiments, we also run experiments with the same set of hyperparameters and the same dataset with a 3D U-Net model as a comparison. Note that the 3D U-Net here refers to the U-shape network in [1] trained with only Dice loss. Experimental results of Dice coefficient and Hausdorff distance on left and right caudate nucleus of 3D U-Net and our method are listed in Table 2, while the visual results are shown in Fig. 3.

Table 2. Performance of segmentation, evaluated on both Dice coefficient (Dice) and Hausdorff distance (HD).

Structure	Metric	3D U-Net	Our method
Left Caudate	Dice	0.831	**0.835**
	HD	5.472	**5.299**
Right Caudate	Dice	0.782	**0.820**
	HD	6.369	**5.004**
Left Hippocampus	Dice	0.771	**0.793**
	HD	20.170	**5.843**
Right Hippocampus	Dice	0.732	**0.759**
	HD	54.553	**29.878**

The Dice coefficient and Hausdorff distance in the tables are both metrics to evaluate the similarity between a segmentation result and its ground truth label map. A higher Dice coefficient and a lower Hausdorff distance both means greater similarity. It's shown that our approach achieves better results than 3D U-Net in terms of both Dice coefficient and Hausdorff distance. In the visual results, it is shown that our approach, compared with 3D U-Net, captures the intricate shape details better. In both examples in Fig. 3, 3D U-Net cannot segment the sharp part in the lower part of a caudate nucleus while our method is able to.

Since all experiment settings except loss function are the same for 3D U-Net and our method, the better performance of our method is due to the incorporation of shape information. Concretely, the shape loss measures the difference in shape signature, while shape signature extracted by a network trained to minimize the difference in shape signature between two *affine pairs* of the same subject. Therefore, when the difference in shape signature is used as a part of segmentation network's loss function, it naturally guides the segmentation network to produce segmentation results that comply with the shapes they should have, thus having better results both quantitatively and visually.

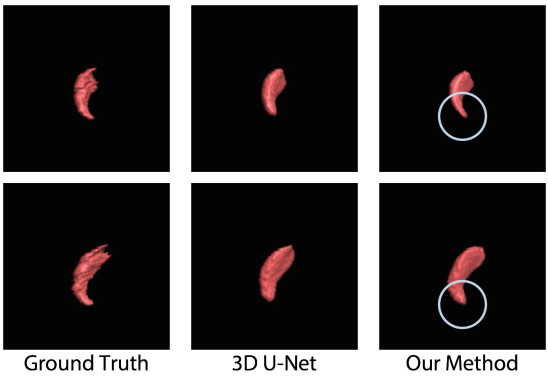

Fig. 3. Visual results of our approach compared with 3D U-Net.

5 Conclusion

We present a novel approach that incorporates shape information into the task of 3D medical image segmentation, by training an shape-learning network that learns the shape signature of the target to be segmented. We run experiments on the public LPBA40 dataset on the brain structure of caudate nucleus and hippocampus. Experimental results show that our approach leads to better results than 3D U-Net in terms of both Dice coefficient and Hausdorff distance.

References

1. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
2. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
3. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE CVPR, pp. 3431–3440 (2015)
4. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 3DV, pp. 565–571. IEEE (2016)
5. Oktay, O., et al.: Anatomically constrained neural networks (ACNN): application to cardiac image enhancement and segmentation. In: IEEE TMI (2017)
6. Ravishankar, H., Venkataramani, R., Thiruvenkadam, S., Sudhakar, P., Vaidya, V.: Learning and incorporating shape models for semantic segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10433, pp. 203–211. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_24

7. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
8. Shattuck, D.W., et al.: Construction of a 3D probabilistic atlas of human cortical structures. NeuroImage **39**(3), 1064–1080 (2008)