2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)
April 4-7, 2018, Washington, D.C., USA

# A DISCRIMINATIVE LEARNING BASED APPROACH FOR AUTOMATED NASOPHARYNGEAL CARCINOMA SEGMENTATION LEVERAGING MULTI-MODALITY SIMILARITY METRIC LEARNING

*Zongqing Ma*[1], *Xi Wu*[2], *Shanhui Sun*[3], *Chaoyang Xia*[2], *Zhipeng Yang*[2], *Shuo Li*[4], *Jiliu Zhou*[1,2*]

[1]College of Computer Science, Sichuan University, China
[2]School of Computer Science, Chengdu University of Information Technology, China
[3]CuraCloud Corporation, USA
[4]Department of Medical Biophysics, University of Western Ontario, Canada

## ABSTRACT

The combination of imaging information from multi-modality images may be highly beneficial for radiotherapy treatment planning in terms of tumor delineation. This paper proposes a discriminative learning based approach for automated nasopharyngeal carcinoma (NPC) segmentation using multi-modality images. Specially, an image-patch-based multi-modality convolutional neural network (CNN) is designed to jointly learn a multi-modality similarity metric and classification of paired image patch of different modalities. The CNN integrates two normal classification sub-networks into a Siamese-like sub-network. With the help of the multi-modality similarity metric learning provided by the Siamese-like sub-network, the classification sub-networks are able to take advantage of each other's multimodal information. Validation of our method was performed on 50 CT-MR subjects. Experimental results demonstrate our method achieves improved segmentation performance compared to its counterpart without multi-modality similarity metric learning and the segmentation method of solely using CT, with a Dice Similarity Coefficient metric of 0.712 compared to 0.659 and 0.636.

***Index Terms***— Convolutional neural networks, multi-modality segmentation, multimodal image similarity, nasopharyngeal carcinoma

## 1. INTRODUCTION

Accurate tumor delineation in medical images is an important but challenging task in image-guided radiotherapy, particularly in head and neck (HN) cancer [1]. Nasopharyngeal carcinoma (NPC) is a type of HN cancer that originates in the nasopharynx. Computed tomography (CT) is usually considered as the basic modality for NPC radiotherapy treatment planning, and magnetic resonance (MR) imaging is another commonly employed modality in NPC diagnosis and treatment. These two modalities combine complementary information. CT is preferred for visualizing bone cortex invasion, while MR is superior to CT in detecting tumor extensions into soft tissue and separating tumor from mucus [2]. Integrating information from different imaging modalities may be beneficial for accurate tumor delineation in radiotherapy treatment planning. Meanwhile, radiologists often need to draw tumor margins manually, which is time-consuming and error-prone. Therefore, it is necessary to develop automated multimodality segmentation methods to accelerate and facilitate clinical applications.

In previous works, research focusing on multimodality segmentation for HN cancer has been limited, and most of them have been developed with the aim of co-segmenting PET-CT or PET-MR data. Song et al. [3] proposed a graph-based segmentation approach on PET-CT images, and the algorithm need to be initiated by the user. Mancas et al. [4] proposed an automated approach using iterative watersheds on CT and PET registered images. Leibfarth et al. [5] developed an automated algorithm for the co-segmentation of HN tumors using both PET and MR data. The research focused on concurrent using CT and MR is scarce. Yang et al. [1] integrated PET, CT and MR information for automated HN cancer segmentation, but it did not focus on NPC. The algorithm in [6] used concurrent CT and MR images for NPC delineation, but it was a semi-automatic method and did not improve accuracy.

Inspired by the recent success of deep convolutional neural network (CNN) [7] and the characteristic of the Siamese network [8], we propose an image-patch-based multi-modality CNN to automatically segment NPC with CT-MR data. Specially, the multi-modality CNN consists of one Siamese-like sub-network and two traditional classification sub-networks, and the three sub-networks share weights of convolutional layers. The multi-modality CNN is trained with two learning objectives: learn a multi-modality similarity metric and classification of paired image patch of different modalities simultaneously. By leveraging multi-modality similarity metric learning provided by the Siamese-like sub-network, the classification sub-networks are able to take advantage of each other's multimodal information.

This paper is organized as follows. The proposed segmentation method is described in Section 2. The experimental setup and obtained results are presented in Section 3. Finally, the paper is concluded in Section 4.

## 2. MATERIALS AND METHODS

### 2.1. Materials and data acquisition

CT and MR images of 50 patients with NPC from the same hospital are used in the experiments. CT images are acquired with a Siemens SOMATOM Definition AS+ system, and have a voxel size of 0.97 × 0.97 × 3 $mm^3$ with a dimension ranging from 512 × 512 × 122 to 512 × 512 × 149. MR images are acquired with a Philips Achieva 3T scanner system. We use T1-weighted (T1W) images with contrast for this study because this protocol provides better tumor visibility than other MRI protocols. The T1W images have a voxel size of 0.61 × 0.61 × 0.8 $mm^3$ with a dimension of 528 × 528 × 290. Manual ground-truth segmentation of the nasopharyngeal tumor is provided by an experienced radiation oncologist and performed slice by slice.

### 2.2. Pre-processing

Considering the acquired CT and MR images include a large scan volume from head to neck and the NPC occupies a small region of the images, to reduce computational complexity, we first select the volume of interest that contain the nasopharyngeal tumor from each image. Then, the CT and corresponding T1W images are resampled to a resolution of 1.0 × 1.0 × 1.0 $mm^3$. As a multi-modality segmentation approach, co-registration between different modal images is a prerequisite step. The T1W image is subsequently registered with the corresponding CT image using rigid and deformable transformation based on the Elastix toolbox [9], and the intensity range of the registered T1W images is normalized to [0-255]. Finally, the CT images are normalized to the same intensity range by adjusting the window width to 350 Hounsfield units (HU) and the window level to 40 HU.

### 2.3. Multi-modality convolutional neural network

In this work, we focus on utilizing the CNN model to automatically segment nasopharyngeal tumor using multi-modality images, and consider the segmentation as a binary classification problem. Although CNNs have been widely used for similar tasks in the literature, it is still challenging to fuse multiple medical image modalities. Inspired by the Siamese network, we present an image-patch-based multi-modality CNN that provides a straightforward solution for combining multi-modality image data by leveraging multi-modality similarity metric learning.

In our multi-modality CNN, we integrate two normal classification sub-networks into a Siamese-like sub-network as shown in Fig. 1. The multi-modality CNN takes paired 2D image patches from different modalities as input and consists of one multi-modality similarity metric learning sub-network $N_M$ and two classification sub-networks $N_{C\text{-}CT}$ and $N_{C\text{-}MR}$. The basic structure of each sub-network is composed of five convolutional layers intertwined two max-pooling layers, finalized with two fully connected (FC) layers. We adopt Batch Normalization [10] to all convolutional layers, and use rectifier linear units (ReLU) as the activation function for every convolutional layer and FC layer. To avoid overfitting, dropout is used in the first FC layer. Note that the two branches of $N_M$ share convolutional layer and FC layer weights, and the three sub-networks share convolutional layer weights, i.e., the parameters of the multi-modality CNN can be divided into three parts according to the architecture: $W_M = \{W_{Con}, W_{FC}\}$, $W_{C\text{-}CT} = \{W_{Con}, W_{FC\text{-}CT}\}$ and $W_{C\text{-}MR} = \{W_{Con}, W_{FC\text{-}MR}\}$.

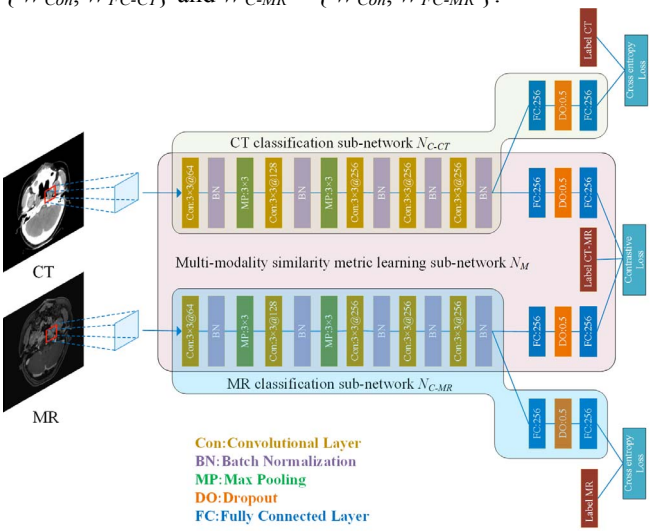



**Fig. 1.** Detailed architecture of the multi-modality CNN

The multi-modality CNN is designed to jointly learn a multi-modality similarity metric and classification for each modality. Given input paired image patch $x^i_{CT}$ and $x^i_{MR}$, and corresponding tumor label $y^i_{CT}$, $y^i_{MR}$ and similarity label $y^i_{CT\text{-}MR}$ that indicates the two image patches are spatial close to each other or not, $N_{C\text{-}CT}$ and $N_{C\text{-}MR}$ are trained to minimize the classification error as:

$$J_{C-CT}(\{W_{Con}, W_{FC-CT}\}) = \sum_i^N l(N_{C-CT}(x^i_{CT}, \{W_{Con}, W_{FC-CT}\}), y^i_{CT}) \quad (1)$$

$$J_{C-MR}(\{W_{Con}, W_{FC-MR}\}) = \sum_i^N l(N_{C-MR}(x^i_{MR}, \{W_{Con}, W_{FC-MR}\}), y^i_{MR}) \quad (2)$$

where $l$ is the cross-entropy loss. While the loss function of $N_M$ is

$$J_M(\{W_{Con}, W_{FC}\}) = \sum_i^N l_c(N_M(x^i_{CT}, x^i_{MR}, \{W_{Con}, W_{FC}\}), y^i_{CT-MR}) \quad (3)$$

where $l_c$ is the contrastive loss. Consequentially, the total loss of the proposed multi-modality CNN is

$$J_{total} = \lambda_1 J_M(\{W_{Con}, W_{FC}\}) + \lambda_2 J_{C-CT}(\{W_{Con}, W_{FC-CT}\}) + \lambda_3 J_{C-MR}(\{W_{Con}, W_{FC-MR}\}) \quad (4)$$

where $\lambda_1$, $\lambda_2$ and $\lambda_3$ are weighting factors, and $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

During the predicting stage, the sub-network $N_{C\text{-}CT}$ and $N_{C\text{-}MR}$ are applied to perform classification for the testing CT and MR image, respectively, and corresponding probability map $P_{MR}$ and $P_{CT}$, which show the likelihood of each voxel assigned to be tumor, are generated. Since the two classification sub-networks have jointly optimized with the multi-modality similarity metric learning sub-network in the training stage, they are able to take advantage of each other's multimodal information in the predicting stage.

In the post-processing stage, in order to remove the misclassified small regions and smooth the boundary, we employ the basic graph-based co-segmentation framework previously proposed in [3], and additionally incorporate probability map $P_{MR}$ and $P_{CT}$ into the region term and context term of the graph cost function. Since the graph-based approach is initialized by the multi-modality CNN, we achieve a fully-automated approach.

## 3. EXPERIMENTS AND RESULTS

### 3.1. Implementation details and parameter settings

In our experiments, we used 40 subjects to train the multi-modality CNN and the other 10 subjects to test the segmentation performance. For each training CT and registered MR images, 10,000 paired image patches were randomly extracted. The tumor label was determined by the central voxel, i.e. patches centered at a voxel inside the tumor were utilized as positive samples, otherwise, as negative ones. The similarity label was set to 1 or 0 according to whether the physical distance of the paired image patches was within 10 or not. The patch size was experimentally set to 31 × 31. The multi-modality CNN was implemented based on the Caffe framework [11] and trained using the SGD method with an initial learning rate of 0.001, momentum of 0.9, and weight decay of 0.0005. We trained for 20 epochs with a batch size of 100 in this study.

### 3.2. Evaluation metrics

Dice Similarity Coefficient (DSC), Average Symmetric Surface Distance (ASSD) were adopted to quantitatively evaluate the segmentation performance in our experiments. The DSC is defined as:

$$DSC = \frac{2 \times size(M \cap A)}{size(M + A)} \quad (5)$$

where $M$ and $A$ indicate the manual ground-truth and the automated segmentation, respectively. ASSD is defined as:

$$ASSD = \frac{1}{2}\left(\frac{\sum_{m \in M_s} \min_{a \in A_s} d(m,a)}{|M_s|} + \frac{\sum_{a \in A_s} \min_{m \in M_s} d(a,m)}{|A_s|}\right) \quad (6)$$

where $M_S$ and $A_S$ denote the surface voxels of the manual ground-truth and the automated segmentation, respectively. $d(a, m)$ represents the Euclidean distance between $a$ and $m$.

### 3.3. Results

*3.3.1. Validation of the proposed method*

To validate the performance of the proposed multi-modality segmentation method, we compared it with the segmentation method of solely using CT, i.e., training and predicting used the classification sub-networks $N_{C\text{-}CT}$. These two methods were applied to all test subjects. Two examples of the qualitative results are shown in Fig. 2. It can be observed that the segmentation results obtained by the proposed method are close to the manual ground-truth, and outperform the results obtained by solely using CT. Quantitative evaluation results averaged over all test subjects are given in Table 1. As can be seen in Table 1, the proposed method makes use of the advantages of CT and MR images and achieves a mean DSC value of 0.712 and a mean ASSD value of 2.129. The CT-only method gives a mean DSC value of 0.636 and a mean ASSD value of 3.529. Meanwhile, the proposed method obtains the minimum standard deviation. These results show that by using complementary information from different modalities, the proposed method exhibits expected improvement in comparison with the CT-only method.

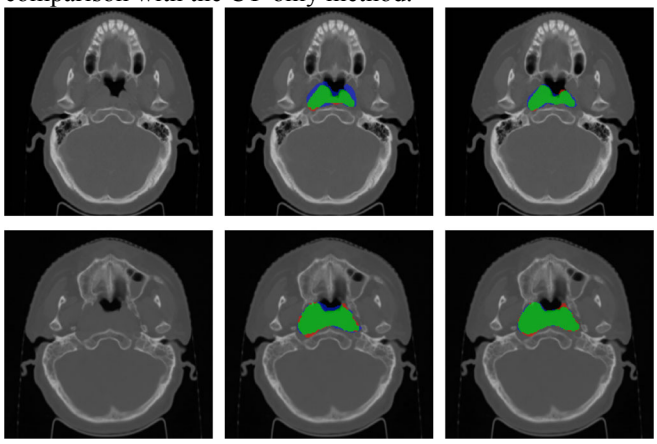


**Fig. 2.** Examples of qualitative segmentation comparison. First to third columns: original CT images, segmentation results by the CT-only method and segmentation results by the proposed method (green: correct voxels, red: unidentified voxels, blue: misidentified voxels).

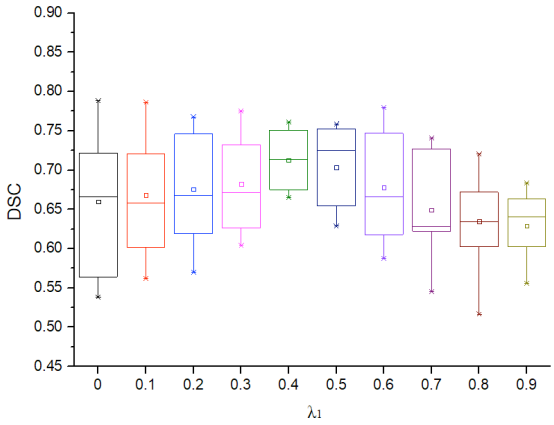
**Table 1.** Quantitative evaluation results achieved by the CT-only method and the proposed method

| | DSC | ASSD |
|---|---|---|
| CT-only | 0.636±0.085 | 3.529±1.584 |
| Proposed | 0.712±0.038 | 2.129±0.600 |

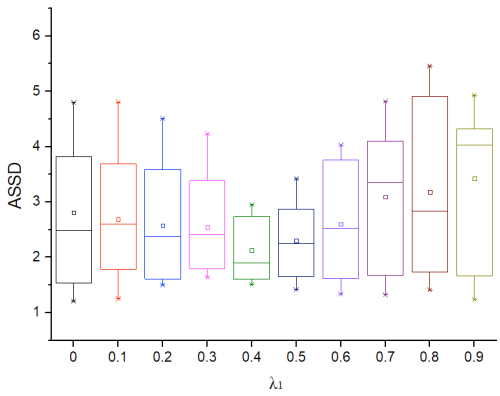
*3.3.2. Study of multi-modality similarity metric learning*

We conducted the experiments to study the effectiveness of the multi-modality similarity metric learning. We compared the performance achieved by different $\lambda_1$ while keeping $\lambda_2$ equal to $\lambda_3$. The influence of the multi-modality similarity metric learning on segmentation performance are presented in Fig. 3. As can be observed from Fig. 3, the method

without multi-modality similarity metric learning, i.e., $\lambda_1 = 0$, achieves a mean DSC value of 0.659 and a mean ASSD value of 2.811. When $\lambda_1$ is less than 0.4, the mean DSC value increases and the mean ASSD value decreases as λ1 increasing, and the dispersion degree of DSC and ASSD value show a downward trend. When $\lambda_1$ is greater than 0.4, the mean DSC value decreases with increased $\lambda_1$, while the mean ASSD value increases. In other words, as $\lambda_1$ increasing, the segmentation performance is gradually improved at first but declined afterward, and the best segmentation performance of our experiments is obtained by $\lambda_1 = 0.4$. The results show that the multi-modality similarity metric learning is effective in improving the segmentation performance of NPC, and we need to seek a balance between the multi-modality similarity metric learning sub-network and the classification sub-networks to make good use of multimodal information.




(a)




(b)

**Fig. 3.** The influence of the multi-modality similarity metric learning on segmentation performance.

## 4. CONCLUSION

In this paper, we have presented a novel deep learning model for automated NPC segmentation using multi-modality images. The model is trained to jointly learn a multi-modality similarity metric and classification for different modalities. By leveraging multi-modality similarity metric learning, the proposed method provides a straightforward way for combining multi-modality image data, and results in noticeable performance improvement. The experimental results show the feasibility and effectiveness of the proposed method.

## 5. REFERENCES


[1] J. Yang, B.M. Beadle, A.S. Garden, D.L. Schwartz, and M. Aristophanous, "A multimodality segmentation framework for automatic target delineation in head and neck radiotherapy," *Medical Physics*, vol. 42, no. 9, pp. 5310-5320, 2015.

[2] C. Rasch, R. Keus, F.A. Pameijer, et al., "The potential impact of CT-MRI matching on tumor volume delineation in advanced head and neck cancer," *International Journal of Radiation Oncology Biology Physics*, vol.39, no. 4, pp. 841-848, 1997.

[3] Q. Song, J. Bai, D. Han, et al. "Optimal co-segmentation of tumor in PET-CT images with context information," *IEEE Transaction on Medical Imaging*, vol. 32, no. 9 pp. 1685-1697, 2013.

[4] M. Mancas, and G. Bernard, "Towards an automatic tumor segmentation using iterative watersheds," in *Proceedings of SPIE*, pp.1598-1608, 2004.

[5] S. Leibfarth, F. Eckert, S. Welz, et al. "Automatic delineation of tumor volumes by co-segmentation of combined PET/MR data," *Physics in Medicine Biology*, vol. 60, no. 14, pp. 5399-5412, 2015.

[6] I. Fitton, S.A. Cornelissen, J.C. Duppen, et al. "Semi-automatic delineation using weighted CT-MRI registered images for radiotherapy of nasopharyngeal cancer," *Medical Physics*, vol. 38, no. 8, pp. 4662-4666, 2011.

[7] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, pp. 1097–1105, 2012.

[8] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 539-546, 2005.

[9] S. Klein, M. Staring, and K. Murphy, M.A. Viergever, J.P. Pluim, "Elastix: a toolbox for intensity-based medical image registration," *IEEE Transaction on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

[10] S. Ioffe, and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:* 1502.03167, 2015**.**

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 675–678, 2014.