Evaluation and Comparison of Automatic Intervertebral Disc Localization and Segmentation methods with 3D Multi-modality MR Images: A Grand Challenge

Guodong Zeng[1(✉)], Daniel Belavy[2], Shuo Li[3], and Guoyan Zheng[1]

[1] Institute for Surgical Technology and Biomechanics, University of Bern, Bern, Switzerland
{guodong.zeng,guoyan.zheng}@istb.unibe.ch
[2] Deakin University, Geelong, Australia
[3] University of Western Ontario, London, Canada

Abstract. The localization and segmentation of Intervertebral Discs (IVDs) with 3D Multi-modality MR Images are critically important for spine disease diagnosis and measurements. Manual annotation is a tedious and laborious procedure. There exist automatic IVD localization and segmentation methods on multi-modality IVD MR images, but an objective comparison of such methods is lacking. Thus we organized the following challenge: Automatic Intervertebral Disc Localization and Segmentation from 3D Multi-modality MR Images, held at the 2018 International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2018). Our challenge ensures an objective comparison by running 8 submitted methods with docker container. Experimental results show that overall the best localization method achieves a mean localization distance of 0.77 mm and the best segmentation method achieves a mean Dice of 90.64% and a mean average absolute distance of 0.60 mm, respectively. This challenge still keeps open for future submission and provides an online platform for methods comparison.

Keywords: Intervertebral disc · MRI · Localization · Segmentation · Multi-modality · Challenge

1 Introduction

Degeneration of intervertebral discs (IVDs) has a strong association with low back pain (LBP) which is one of most prevalent health problems amongst population and a leading cause of disability [1]. Magnetic Resonance (MR) Imaging (MRI) is widely recognized as the imaging technique of choice for the assessment of lumbar IVD abnormalities due to its excellent soft tissue contrast and no ionizing radiation [2]. Thus, automated image analysis and quantification for

G. Zheng et al. (Eds.): CSI 2018, LNCS 11397, pp. 163–171, 2019.
https://doi.org/10.1007/978-3-030-13736-6_14

spinal diseases using MR images have drawn a lot of attention. Localization and segmentation are important steps before analysis and quantification. Previous works on disc degeneration were mainly done by manual segmentation, which is a time-consuming and tedious procedure. Automatic localization and segmentation of IVDs are highly preferred in clinical practice.

However, it is very difficult to directly compare different methods because they are usually evaluated on different datasets. Thus, objective evaluation and comparison are highly desired. For example, Zheng et al. [3] held a challenge on 3D IVD localization and segmentation in MICCAI 2015. But this challenge only investigated on single modality MR images, i.e., T2 MR data. Multi-modality MR images provide complementary information which can help improve recognition accuracy, and therefore have been utilized in many medical image analysis tasks. In this challenge, we investigate different methods working on four-modality IVD MR images acquired with Dixon protocol: fat, in-phase, opposed-phase and water modality MR image. The four multi-modality MR images of the same subject were acquired in the same space and thus are aligned with each other.

How to ensure objective and fair comparison is a big concern in organizing a challenge. In this challenge, all participants are required to submit a docker container of their method. A docker container includes codes and all dependencies so that others can re-run the method quickly and reliably on another computer. By doing this, all results of each participant were generated by running submitted containers on the challenge organizer's machine.

The paper is arranged as follows. We first present the challenge organization, rules for evaluation, image dataset and the established validation framework in Sect. 2. A summary of each submitted method will be described in Sect. 3. The validation results of each participant will be presented in Sect. 4, followed by conclusion in Sect. 5.

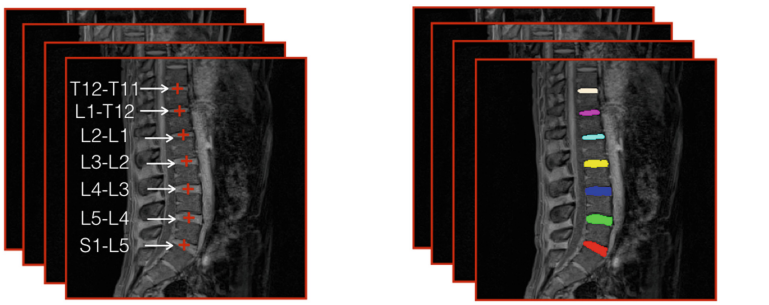


Fig. 1. The 7 defined IVDs to be localized and segmented from each subject.

2 Challenge Setup

The aim of this challenge is to investigate fully automatic IVD localization and segmentation algorithms on a set of 3D Multi-modality MR images and to provide a standardized validation framework. The task for this challenge includes two parts: localization part and segmentation part. The task of localization part is to fully automatic localize the centers of 7 IVDs (T11-S1) for each test subject while the task of segmentation part is to fully automatic segment 7 disc regions T11-S1, which is illustrated in Fig. 1. For localization part, instead of detecting IVDs explicitly as a separate task, the centroids of each IVD generated from segmentation mask are recognized as the localization results.

2.1 Organization

Each participant could download the training data for method development after submitting a scanned copy of the signed registration form. For test data, both 3D MR images and corresponding ground truths will be only known to challenge organizers.

Participants should containerize their methods with Docker[1] and submit them to challenge organizers for evaluation. Containerized methods consist of codes and all dependencies so that challenge organizers can run all participants' methods quickly and reliably without complex development environment setup. By doing this, all prediction results were generated by running methods on challenge organiser's machine so that a fair comparison could be realized. More details about how to do the method containerization and to run the containers could be found at our challenge website[2], where an example in Python script was shown.

In the phase of testing, for each containerized method, it was run on each test subject one by one to get the segmentation result. To guarantee the running of containerized method is correct, challenge organizers sent the segmentation result of the first training subject back to the participants for verification. A desktop with a 3.6 GHz Intel(R) i7 CPU and a GTX 1080 Ti graphics card with 11 GB GPU memory was used to evaluate all submitted methods.

2.2 Description of Image Dataset

There are in total 24 sets of 3D multi-modality MRI data which contains at least 7 IVDs of the lower spine, collected from 12 subjects in two different stages in a study investigating the effect of prolonged bed rest (spaceflight simulation) on the lumbar intervertebral discs [4]. Each set of 3D multi-modality MRI data consists of four modality aligned high-resolution 3D MR images: in-phase, opposed-phase, fat and water images. Thus, in total we have 12 subjects $\times$ 2 stages $\times$ 4modalities $=$ 96 volume data.

[1] https://www.docker.com.
[2] https://ivdm3seg.weebly.com/methods.html.

All MR images were scanned with a 1.5-Tesla MRI scanner of Siemens (Siemens Health-care, Erlangen, Germany) using Dixon protocol [5]: slice thickness = 2.0 mm, pixel Spacing = 1.25 mm, repetition Time (TR) = 10.6 ms, echo time (TE) = 4.76 ms. The ground truth segmentation for each set of data were then manually annotated and were provided in the form of binary mask. All images (four volumes per patient) and binary masks (one binary volume per patient) are stored in the Neuroimaging Informatics Technology Initiative (NIFTI) file format.

During the challenge period, the organizer released training set of IVD challenge (8 subjects $\times$ 2 stages $\times$ 4modalities = 64 volume data). For test data, both MR images and ground truth segmentation will be only known to challenge organizer for independent evaluation and fair comparison (4 subjects $\times$ 2 stages $\times$ 4modalities = 64 volume data).

2.3 Rules for Evaluation

Submitted methods can generate multi-label segmentation or binary-label segmentation. We provide following rules for evaluation:

Multi-Label Prediction. If the prediction segmentation is not binary but with multiple labels, we will directly do the evaluation separately for seven IVDs in one test subject.

Binary-Label Prediction. If the prediction segmentation is binary, we first assign labels to each intervertebral disc based on ground truth segmentation and then do the evaluation. Specifically, the complete image space is spitted into 7 sections, corresponding to 7 intervertebral discs in the ground truth segmentation. Then we can do the evaluation similar as evaluation for multi-label prediction.

2.4 Evaluation Metrics

Three metrics were used to evaluate different methods: Mean Localization Distance (MLD) is used for localization task while Mean Dice Similarity Coefficients (MDSC) and Mean Average Surface Distance (MASD) are used for segmentation task. The details about how these three metrics are computed can be found as follows:

1. Mean Localization Distance (MLD)

For each IVD, we first calculate the localization distance (R) between the centroids of prediction and ground truth.

$$R = \sqrt{(\triangle x)^2 + (\triangle y)^2 + (\triangle z)^2} \tag{1}$$

where $\triangle x$, $\triangle y$ and $\triangle z$ are the distances between the identified IVD centroids calculated from prediction and ground truth in x, y, z axis respectively.

After localization distance (R) was calculated, the MLD can be computed as follows:

$$MLD = \frac{\sum_{i=1}^{N_{images}} \sum_{j=1}^{N_{IVDs}} R_{ij}}{N_{images} N_{IVDs}} \tag{2}$$

where N_{images} is the number of test subjects, and N_{IVDs} is the number of IVDs in each test subject, i.e. 7 in our experiment. MLD indicates the measurement of average localization error for IVDs and lower value means better localization performance.

2. Mean Dice Similarity Coefficients (MDSC)

For each IVD, we first calculate Dice Similarity Coefficients (DSC) between prediction segmentation and ground truth segmentation, which is computed as follows:

$$DSC = \frac{2|A \cap B|}{|A| + |B|} \times 100\% \tag{3}$$

where A and B are foregrounds of prediction and ground truth segmentation respectively. And Mean Dice Similarity Coefficients (MDSC) is computed as below:

$$MDSC = \frac{\sum_{i=1}^{N_{images}} \sum_{j=1}^{N_{IVDs}} DSC_{ij}}{N_{images} N_{IVDs}} \tag{4}$$

where N_{images} is the number of test subjects, and N_{IVDs} is the number of IVDs in each test subject, i.e. 7 in our experiment. MDSC indicates the measurement of average overlap between the prediction and ground truth for IVDs and higher value means better segmentation performance.

3. Mean Average Surface Distance (MASD)

For each IVD, we first calculate Average Surface Distance (ASD) between prediction segmentation and ground truth segmentation. ASD calculation is implemented by medpy toolbox[3]. And Mean Average Surface Distance (MASD) is computed as below:

$$MASD = \frac{\sum_{i=1}^{N_{images}} \sum_{j=1}^{N_{IVDs}} ASD_{ij}}{N_{images} N_{IVDs}} \tag{5}$$

where N_{images} is the number of test subjects, and N_{IVDs} is the number of IVDs in each test subject, i.e. 7 in our experiment. MASD measures average surface distance between the prediction and ground truth for IVDs and lower value means better performance.

For each intervertebral disc, both the localization distance and ASD will be set as maximum value (458.24 mm) if the Dice value is less than 0.1% and additionally the number of segmented voxels assigned to this disc is smaller than 5% of the total voxels of the ground truth segmentation. In such a case, a method is regarded missing the segmentation of the disc completely.

[3] http://loli.github.io/medpy/.

2.5 Evaluation Ranking

The final ranking of all methods is based on three metrics described in Sect. 2.4. For each metric, we sort all methods (in total n methods) from best to worst. The best method will get a ranking score of 1, while the worst method get a ranking score of n. For each method, it will get an overall ranking score, which is the sum of its own ranking scores at each metrics. Lower overall ranking score indicates better performance. Finally, the final ranking of all methods will be in descending order by their overall ranking scores.

3 Methods

In total 8 teams submitted their methods and participated this challenge, but we only received 7 methods description. A brief summary of 7 methods is given below, in alphabetical order. Detailed method description and results of each team can be found at our challenge website[4].

1. changliu: a 2.5D U-Net-like [6] network which utilizes SEBottleneck [7] to achieve channel-wise attention and predicts segmentation mask of one slice from multiple-slice input (11 slices) [8].

2. gaoyunhecuhk: a 2D fully convolutional neural network which uses DenseNet [9] as the backbone network. Their network only down-samples for 2 times and uses Atrous Spatial Pyramid Pooling(ASPP) [10] to ensure a large receptive field [11].

3. livia: a UNet-like architecture which follows the multi-modality fusion strategy presented in [12], and all convolutional blocks are replaced by an Inception-like module and all convolutions are replaced by asymmetric convolutions [13].

4. lrde: the only method not using deep learning, but based on mathematical morphology operators which was driven by shape prior knowledge and their contrast in the different modalities [14].

5. mader: they first applied random forests with conditional random field (CRF) to detect 7 landmarks, i.e. the centroids of 7 IVDs. Then small fixed-size sections around each landmark were cropped and reoriented. At last, a V-Net [15] was trained to perform segmentation of IVDs [16].

6. smartsoft: Three 2D Unet-like neural networks were separately trained on 2d slice images in axial, sagittal, coronal axis respectively. The final segmentation result will be achieved by ensemble from three models [17].

7. ucsf_Claudia: V-Net was trained on full volumes to leverage the spatial context of the whole image [15]. The combination of weighted cross entropy (wce) loss and soft Dice loss was used. A 3D connected component analysis was employed to eliminate predicted volumes of less than 1200 voxels [18].

[4] https://ivdm3seg.weebly.com/miccai2018.html.

4 Experimental Results

The mean performance of each team is shown in Table 1. For each metric, there is an individual ranking and the final ranking is based on the sum of all metrics' ranking. The team changliu achieved best performance on all metrics, with a mean Dice Similarity Coefficients of 90.64%, a mean Average Surface Distance of 0.60 mm and a mean Localization Distance of 0.77 mm.

Table 1. Mean performance and ranking of each team on each metric. Metrics include MDSC, MASD and MLD. The final ranking is based on the sum of ranking on all metrics, in which lower value means better performance. Bold indicates the method performs best on that metric.

Final ranking (#)	TEAM	MDSC(%)	MASD (mm)	MLD (mm)	MDSC ranking value	MASD ranking value	MLD ranking value	Sum of ranking value
1	changliu	**90.64**	**0.60**	**0.77**	**1**	**1**	**1**	**3**
2	gaoyunhe_cuhk	90.58	0.61	0.78	2	2	2	6
3	ucsf_Claudia	89.71	0.74	0.86	3	4	3	10
4	livia	89.67	0.65	0.96	4	3	5	12
5	wanghuan	88.77	0.82	0.92	5	5	4	14
6	smartsoft	81.93	34.03	34.27	6	6	6	18
7	mader	66.42	108.19	108.41	7	7	7	21
8	lrde_01	24.35	319.53	319.81	8	8	8	24

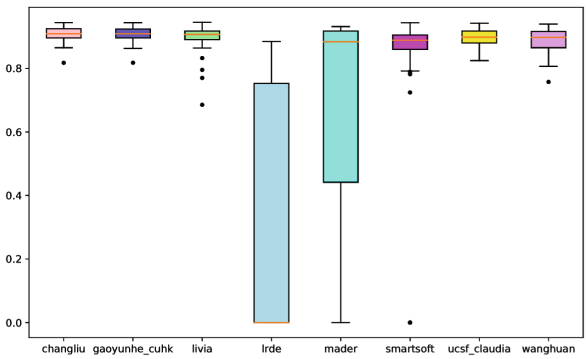
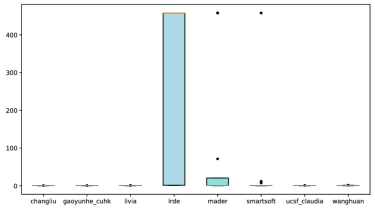


Fig. 2. Dice Similarity Coefficients (DSC) boxplot of 56 IVDs (8 test subjects × 7). The box shows the interquartile range (IQR) and extends from first quartile (Q1) to third quartile (Q3) values of the data, with a line at the median data. The whiskers extend up to 1.5 times of the IQR and those flier points beyond the whiskers are outliers.

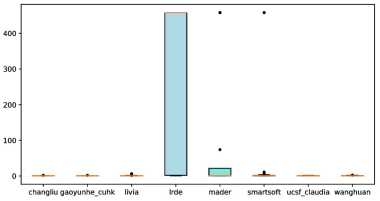
And the following four teams, i.e. gaoyunhe_cuhk, ucsf_Claudia, livia and wanghuan also achieved good performance on both segmentation and localization tasks. Especially for the team gaoyunhe_cuhk, whose results show a very minor difference with the winner team changliu. Specifically, team gaoyunhe_cuhk

reported a mean Dice Similarity Coefficients of 90.58%, a mean Average Surface Distance of 0.61 mm and a mean Localization Distance of 0.78 mm. But for the other three teams of smartsoft, mader, and lrde, they reported poor results on MDSC, MASD and MLD in this challenge.

Figure 2 shows boxplots of in total 56 IVDs (8 test subjects $\times$ 7) of each method on Dice Similarity Coefficients (DSC). As seen in Fig. 2, in terms of segmentation from team of mader and lrde, there are lots of completely failed cases whose DSC value are almost zero. Also, there are several such failed cases in the team of smartsoft. Figure 3 show boxplots of each method on Average Surface Distance (ASD) and Localization Distance (LD). Note that for each IVD, both the ASD and LD will be set as maximum value (458.24 mm) if a method is regarded missing the segmentation completely as mentioned in Sect. 2.4. As observed in Fig. 3, for teams of smartsoft, mader, and lrde, all of them reported some completely failed cases whose ASD and LD values are 458.24 mm.



(a) Average Surface Distance (ASD) boxplot of IVDs



(b) Localization Distance (LD) boxplot of IVDs

Fig. 3. The boxes show the interquartile range (IQR) and extends from first quartile (Q1) to third quartile (Q3) values of the data, with a line at the median data. The whiskers extend up to 1.5 times of the IQR and those flier points beyond the whiskers are outliers.

5 Conclusion

This paper presents an objective comparison of state-of-the-art methods, which were submitted to the MICCAI 2018 challenge on Automatic Intervertebral Disc Localization and Segmentation from 3D Multi-modality MR Images. In total 8 teams submitted their results by docker container. The challenge organisers run their submitted methods on a local machine and then do the evaluation to ensure a fair comparison. The test data and ground truth are only known to the challenge organizers. The top-two ranking methods achieve similar results and the following three methods produce quite good results on both segmentation and localization tasks. The other 3 teams report poor results because their methods completely miss some IVDs. The organizers choose not to disclose the test data and corresponding ground truth, and the Challenge remains open for new submission in the future.

References

1. An, H.S., et al.: Introduction: disc degeneration: summary. Spine **29**(23), 2677–2678 (2004)
2. Emch, T.M., Modic, M.T.: Imaging of lumbar degenerative disk disease: history and current state. Skelet. Radiol. **40**(9), 1175 (2011)
3. Zheng, G., et al.: Evaluation and comparison of 3D intervertebral disc localization and segmentation methods for 3D T2 MR data: a grand challenge. Med. Image Anal. **35**, 327–344 (2017)
4. Belavỳ, D.L., Armbrecht, G., Felsenberg, D.: Incomplete recovery of lumbar intervertebral discs 2 years after 60-day bed rest. Spine **37**(14), 1245–1251 (2012)
5. Li, X., Dou, Q., Chen, H., Fu, C.-W., Heng, P.-A.: Multi-scale and modality dropout learning for intervertebral disc localization and segmentation. In: Yao, J., Vrtovec, T., Zheng, G., Frangi, A., Glocker, B., Li, S. (eds.) CSI 2016. LNCS, vol. 10182, pp. 85–91. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-55050-3_8
6. 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
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507 (2017)
8. Liu, C.: IVDM3Seg Challenge MICCAI 2018: Method Description of Team Changliu (2018). https://ivdm3seg.weebly.com/changliu.html
9. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, vol. 1, p. 3 (2017)
10. Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K.: DenseASPP for semantic segmentation in street scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3684–3692 (2018)
11. Gao, Y.: IVDM3Seg Challenge MICCAI 2018: Method Description of Team gaoyunhe_cuhk (2018). https://ivdm3seg.weebly.com/gaoyunhe_cuhk.html
12. Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ayed, I.B.: HyperDense-net: a hyper-densely connected CNN for multi-modal image segmentation. arXiv preprint arXiv:1804.02967 (2018)
13. Dolz, J., Desrosiers, C., Ayed, I.B.: HD-UNet: hyper-dense UNet with asymmetric convolutions for multi-modal intervertebral disc segmentation (2018). https://ivdm3seg.weebly.com/livia.html
14. Carlinet, E., Géraud, T.: Intervertebral Disc Segmentation Using Mathematical Morphology (2018). https://ivdm3seg.weebly.com/lrde.html
15. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
16. Mader, A.O., Lorenz, C., Meyer, C.: Segmenting Labeled Intervertebral Discs in Multi Modality MR Images (2018). https://ivdm3seg.weebly.com/mader.html
17. Georgiev, N., Asenov, A.: Automatic Segmentation of Lumbar Spine 3D MRI Using Ensemble of 2D Algorithms (2018). https://ivdm3seg.weebly.com/smartsoft.html
18. Iriondo, C., Girard, M.: Vesalius: VNet-based fully automatic segmentation of intervertebral discs in multimodality MR images (2018). https://ivdm3seg.weebly.com/ucsf_claudia.html