Pelvis Segmentation Using Multi-pass U-Net and Iterative Shape Estimation

Chunliang Wang[1(✉)], Bryan Connolly[2], Pedro Filipe de Oliveira Lopes[3], Alejandro F. Frangi[3], and Örjan Smedby[1]

[1] Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Stockholm, Sweden
chunwan@kth.se
[2] Radiology Department, Karolinska Institute, Solna, Sweden
[3] Center for Computational Imaging and Simulation Technologies in Biomedicine, The University of Sheffield, Sheffield, UK

Abstract. In this report, an automatic method for segmentation of the pelvis in three-dimensional (3D) computed tomography (CT) images is proposed. The method is based on a 3D U-net which has as input the 3D CT image and estimated volumetric shape models of the targeted structures and which returns the probability maps of each structure. During training, the 3D U-net is initially trained using blank shape context inputs to generate the segmentation masks, i.e. relying only on the image channel of the input. The preliminary segmentation results are used to estimate a new shape model, which is then fed to the same network again, with the input images. With the additional shape context information, the U-net is trained again to generate better segmentation results. During the testing phase, the input image is fed through the same 3D U-net multiple times, first with blank shape context channels and then with iteratively re-estimated shape models. Preliminary results show that the proposed multi-pass U-net with iterative shape estimation outperforms both 2D and 3D conventional U-nets without the shape model.

Keywords: Deep learning · Multi-pass U-net · Pelvis segmentation
Shape context · Statistic shape model

1 Introduction

Improving resolution of computed tomography (CT) scanners and recent advances in three-dimensional (3D) printing technology make image-guided custom treatment planning and custom implant design more feasible than ever before. However, the timely creation of accurate models of a patient's anatomical structures from 3D high-resolution medical images remains a non-negligible challenge. This challenge limits the clinical applicability of new personalised image-based approaches. As a crucial step in surgery planning, radiotherapy

T. Vrtovec et al. (Eds.): MSKI 2018, LNCS 11404, pp. 49–57, 2019.
https://doi.org/10.1007/978-3-030-11166-3_5

planning and quantitative disease evaluation, pelvis segmentation is often done manually in clinical practice. It can take half an hour to several hours for a trained radiologist to segment the complete pelvis, which is not acceptable in most public hospitals [1].

Several automated and semi-automated segmentation methods have been developed to address this problem. The methods range from intensity thresholding-based approaches, to clustering-based approaches and deformable-model-based approaches [1–4]. However, most existing methods have achieved moderate success only on a few cases. In this study, we tried to evaluate a new deep-learning-based pelvis segmentation method on a relatively large dataset of 90 patients. In recent years, deep-learning-based methods have gained more and more attention due to their superior performance in several image segmentation challenges [5,6]. The deep neural networks are often trained on a large number of training samples in a more or less black-box manner. The neural network learns task-specific image features and rules by minimizing the objective function often correlated with the segmentation error. However, these learned features and rules are difficult to interpret. One can only imagine that the classification decision is based on local appearance and global shapes represented by a stack of convolution kernels and mixed through weighted non-linear rules. The consensus from existing literature is that shape prior knowledge is crucial for accurate and robust pelvis segmentation [1]. The most promising conventional pelvis segmentation methods are based on either statistical shape models or multi-atlas registration [1,3,7]. How to enforce shape prior in a deep learning framework and whether that helps to improve segmentation performance remains an open question. In a previous study [8], Wang et al. proposed a method where statistical shape models were combined with 2.5D U-net to segment multiple structures of the heart in 3D MRI and CT images. In their setup, three two-dimensional (2D) U-nets were first trained to segment the multiple heart structures in three orthogonal views, and then the probability maps were combined and used to estimate 3D shape models of the heart and ventricles. The estimated shapes combined with input images are fed to a second set of 2D U-nets to deliver a refined segmentation result. In this study, a similar strategy was adopted and 3D U-net was trained to take both the 3D CT image and estimated volumetric shape models of the targeted structures as input to generate the segmentation results. However, instead of training two U-nets, the same 3D U-net was retrained multiple times, first with blank images as shape context channels, and then with volumetric shape representation estimated on the preliminary segmentation results. During the testing phase, the input image was passed through the trained 3D U-net several times, with iteratively re-estimated shape context information. In the preliminary experiments performed, the proposed method delivered better segmentation results than the conventional methods using 2D or 3D U-net or statistical shape model methods.

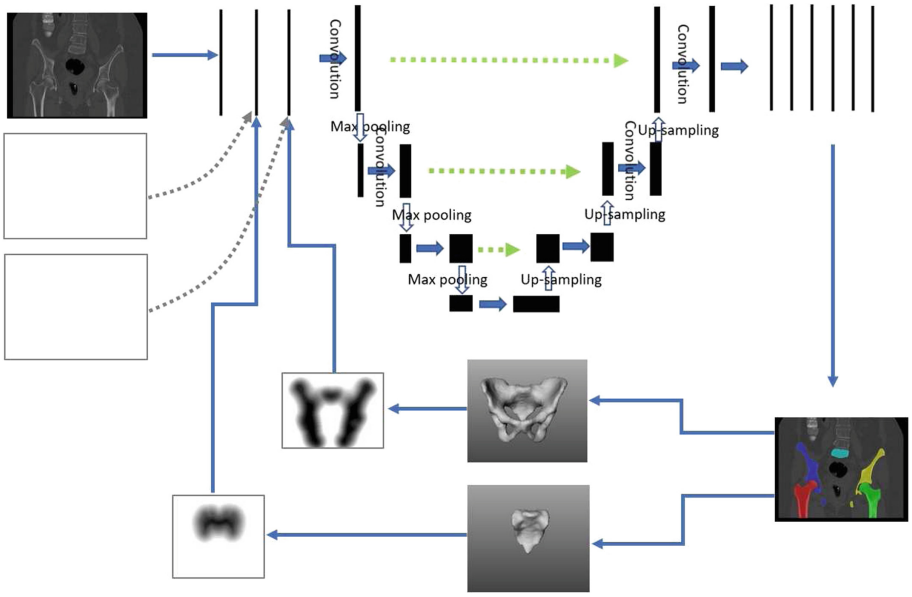


Fig. 1. Overview of the proposed multi-pass three-dimensional U-net with iterative shape model estimation.

2 Methods

The proposed framework consists of a multi-pass 3D U-net with iterative shape model estimation as summarized in Fig. 1. The core module of this framework is a 3D U-net that takes 3 channels as input: one containing the 3D CT image, two other with the shape context images. The network outputs 6 probability maps for the background, left femur, right femur, left hip bone, right hip bone and sacrum, respectively. The U-net architecture was initially proposed by Ronneberger et al. [6]. In this study we simply extend it to process 3D images instead of 2D images. The 3D U-net consists of 3 max pooling layers and 3 up-sampling layers, and two convolutional layers with kernel size of $3 \times 3 \times 3$ located between max pooling or/and up-sampling layers. Due to limited GPU memory, the input volume size to the U-net is set to $128 \times 128 \times 72$. Original CT data are down-sampled to 3 mm isotropic resolution before being split into overlapping 3D patches of required size. The outputs of the U-net are passed on to a level-set-based volumetric statistical shape estimation module where the shapes of the whole pelvis (including hip bones and sacrum as a whole) and the sacrum bone are estimated. The volumetric shape images are then added to the corresponding input channels and trigger the 3D U-net to re-run. The segmentation and shape estimation loop can be repeated several times until no significant changes occur.

2.1 Data Set and Ground Truth Generation

For this work 90 abdominal and pelvis CT scans were selected from public databases (50 from the CT Colonography study [9], 40 from the Lymph Nodes study [10]). Imaging protocols for these studies are available at [9] and [10] respectively. The scans selected were checked to guarantee full coverage of the pelvis, i.e. all pelvic bones are contained in the scan without being partially cut off. Ground-truth segmentation masks were created by an experienced radiologist, using an interactive segmentation tool based on fuzzy connectedness and the level set method. Besides the interactive segmentation, all segmentation masks were carefully inspected and manually edited by the radiologist using the manual segmentation tool in ITKSnap 3.6.0 [11]. On average, each case took between 30 to 50 m to complete the interactive segmentation and manual mask curation procedure.

2.2 Statistical Shape Model Creation

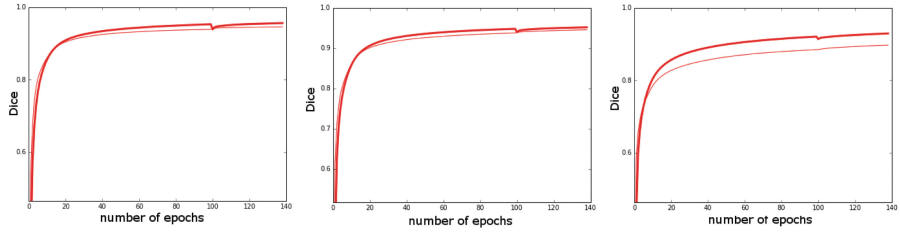
The so-called shape context, or shape image, is a volumetric representation of the subject's shape, which is a signed distance map from the surface of the segmented object. However, instead of performing distance transform on the segmentation result directly, a statistical shape model is fit to it to eliminate irregularity that may present in the segmentation results. A statistical model is created by averaging of the signed distance maps of several segmented subjects (training subjects) and computing the main variations using principal component analysis (PCA) [12]. In this study, the shape model is created using 20 randomly selected training subjects, in which the top 10 principal components are computed via PCA. As suggested by Leventon et al., the shape model M that matches the current segmentation is estimated by solving a level set function

$$\frac{\partial \phi}{\partial t} = \alpha F(x) + \beta M(T(x)) + \gamma \kappa(x) \left| \nabla \phi \right|, \tag{1}$$

where F is the image force related to the probability output by the U-net, M is the statistical model as a weighted sum of the mean shape and modes of variation, T is the global transformation and κ is the mean curvature. The transformation T and the weighting factors of modes of variation are updated iteratively by minimizing the squared distance between the model and the level set function, which is also a signed distance map. α, β and γ are weighting factors that can be determined empirically.

2.3 Training Phase

In the proposed framework, the statistical shape model training is performed only once, but the multi-pass 3D U-net must be trained in two steps. The first step is to train the net by feeding the 3D CT volumes and blank shape context volumes (all voxels are set to zero) to the 3D U-net, which will force the network



(a) (b) (c)

Fig. 2. Dice coefficient of (a) the right femur, (b) right hip bone and (c) sacrum bone during the two-step training (first step: 0 – 100 epochs, second step: 100 – 140 epochs). Thick line: training set. Thin line: testing set.

to learn the segmentation using only CT images. In the second step, the pre-trained U-net is retrained with the 3D CT volumes and the real shape context volumes generated from fitting the statistical shape models to the output of the pre-trained 3D U-net. As the weights of the network were already initialized to recognize anatomical structures from the CT channel, the U-net is expected to learn gradually to use the context information where it helps to improve the segmentation results, instead of heavily relying on the context layer. For the U-net training, categorical cross entropy was used as the loss function, stochastic gradient descent as optimizer and the learning rate was set to 0.01. The network was first trained for 100 epochs in the first-step training and 40 epochs in the second-step training (Fig. 2). Sample augmentation was used, where random translating, rotating and scaling are added when creating the training images.

2.4 Testing Phase

During testing, the input images were first sent to the trained 3D U-net with blank shape context layers, and then patient-specific shape models were created by fitting the statistical shape model to the preliminary segmentation results. The shape models were added as context layers to the 3D U-net to re-generate the segmentations. The process can be repeated several times. Besides image resampling and intensity normalization, no pre-processing steps are required for the proposed multi-pass U-net segmentation.

3 Results

To test the performance of the proposed method, a 5-fold cross validation was performed using the 90 cases. In each fold, 72 subjects were used for training and 18 subjects were used for testing. Both the 3D U-net and the statistical shape models are retrained in every fold. For comparison, the hierarchical statistical shape model based segmentation method reported in [13] was implemented. Plain 2D and 3D U-nets were also trained on the same dataset and compared with

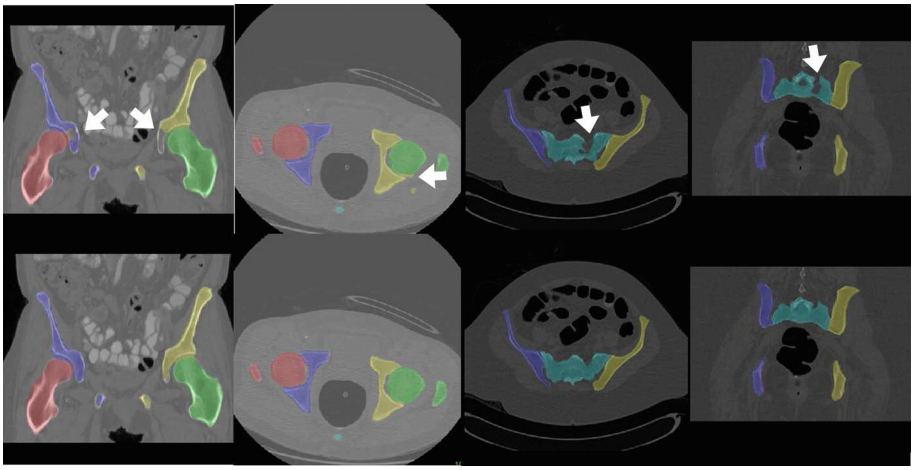


Fig. 3. Comparison of results from the first-pass segmentation (top row) and the second-pass segmentation (bottom row). Arrows indicate examples of segmentation errors.

the proposed method. The average segmentation accuracy of 5 bone structures of the pelvis from the 90 subjects is summarized in Table 1. On average, running a 2-pass 3D U-net delivers better results than a 3D U-net. The performance gain is more visible on the sacrum than on other bone structures. Figure 3 shows several examples where adding the shape context component helped to improve the segmentation results. Running the multi-pass U-net for the third time will slightly improve the segmentation accuracy, but no significant improvement is observed when run over 3 iterations. For each fold, the training took 60 h on a NVIDIA GTX 1080ti GPU. For testing, the U-net prediction takes 20–30 s to process a 3D volume, and the shape model estimation takes 2–3 m for each pass.

Table 1. Segmentation accuracy measured as dice coefficient of the proposed method and alternative methods.

Methods	Statistical shape model	2D U-net axial view	3D U-net	1st-pass 3D U-net with blank context	2nd-pass 3D U-net with shape context	3rd-pass 3D U-net with shape context
Left femur	-	0.925 ± 0.178	0.949 ± 0.039	0.937 ± 0.047	0.958 ± 0.032	0.958 ± 0.031
Right femur	-	0.939 ± 0.122	0.953 ± 0.019	0.942 ± 0.026	0.961 ± 0.018	0.962 ± 0.018
Left hip	0.915 ± 0.065	0.940 ± 0.079	0.947 ± 0.016	0.947 ± 0.020	0.957 ± 0.012	0.958 ± 0.013
Right hip	0.908 ± 0.059	0.943 ± 0.054	0.947 ± 0.016	0.944 ± 0.021	0.957 ± 0.011	0.957 ± 0.011
Sacrum	0.850 ± 0.082	0.894 ± 0.056	0.905 ± 0.032	0.909 ± 0.029	0.921 ± 0.028	0.924 ± 0.027

4 Discussion and Conclusion

The main contribution of the proposed method is to explicitly integrate shape information into a 3D deep neural network, while previously shape context has only been tested in 2D neural networks. The results suggest that adding shape context information to a deep neural network seems to improve the segmentation accuracy, especially for relatively challenging anatomical structures like the sacrum. Our previous paper [8] reported similar findings when adding shape context to 2D U-net. Moving to 3D U-net is important, since even though the previous study showed that shape context helped to improve the segmentation accuracy, it is unclear whether that is due to the 3D shape model providing 3D information that is not accessible to 2D U-nets, or to it providing useful context information that will help segmentation.

In comparison with other context approaches such as auto-context, the shape context can eliminate irregularities in the preliminary segmentation results, such as isolated regions outside the targeted bone or holes inside the bone, while the conventional auto-context that will be based only on the output of the first U-net. In another study, we compared several types of context information for deep neural networks and found shape context produced the most accurate results. These findings will be reported in a future publication.

Another contribution of the proposed method is to replace the two sets of U-nets with a single U-net that can take blank shape context channels and introduce an iterative training and testing scheme, which was not reported in the previous work. The iterative scheme not only reduces the file size for deploying the trained net, saving the time of loading and switching network into computer memory, but also makes it possible to run the shape model estimation and shape-context-based segmentation multiple times in an iterative manner. This will hopefully further improve the segmentation accuracy. (However, the benefit of running over 2 passes was not very evident in our experiments.) It is worth noticing the two networks setup proposed in [8], can also be run in an iterative mode by repeating the testing process using the second set of 2D U-nets. Similar performance gain is expected in the 2D cases.

Wang et al. reported overfitting on certain structures when applying their 2.5 U-net with shape context to the heart segmentation, i.e. the segmentation accuracy drops when the shape context is added [8]. The design of trying to use a single 3D U-net to handle cases with and without shape context will force the network to not rely on the shape context too much, avoiding overfitting.

Several strategies were tested to make sure that, after the second-round training, the single 3D U-net can perform well on samples with only the input CT image (with blank shape context channels) and samples with both CT image and shape context channels. These strategies included mixing the training samples with and without shape context, alternating training between two types of training samples, and complete retraining while gradually adding one group into another. However, it was found that simply continuing to train the U-net with samples with shape context works best. As shown in Table 1, the segmentation

accuracy of the 1st pass on the retrained 3D U-net is only slightly inferior to the results of the 3D U-net trained on samples without shape context.

One limitation of this study was the lack of diseased cases in the image sample used. The performance must be evaluated where fracture or other abnormality exists. Another limitation was that the down-sampling of the input images due to hardware limitations which introduced additional errors. Finally, the ground-truth segmentation was generated by a single doctor, no inter-observer variation information is available. Future research activities have been planned to address these limitations.

In conclusion, a multi-pass 3D U-net framework with iteratively estimated shape models as context information was proposed. Preliminary results show that the proposed method outperforms both 2D and 3D conventional U-nets in 3D pelvis segmentation.

Acknowledgements. This study was supported by the Swedish Heart-lung foundation (grant no. 20160609), Swedish Medtech4Health AIDA research grant, and the Swedish Childhood Cancer Foundation (grant no. MT2016-00166).

References

1. Seim, H., Kainmüller, D., Heller, M., Lamecker, H., Zachow, S., Hege, H.C.: Automatic segmentation of the pelvic bones from CT data based on a statistical shape model. In: Proceedings 1st Eurographics Conference on Visual Computing for Biomedicine - EG VCBM 2008, pp. 93–100 (2008). https://doi.org/10.2312/VCBM/VCBM08/093-100
2. Kang, Y., Engelke, K., Kalender, W.: A new accurate and precise 3-D segmentation method for skeletal structures in volumetric CT data. IEEE Trans. Med. Imaging **22**(5), 586–598 (2003). https://doi.org/10.1109/TMI.2003.812265
3. Chu, C., Chen, C., Liu, L., Zheng, G.: FACTS: fully automatic CT segmentation of a hip joint. Ann. Biomed. Eng. **43**(5), 1247–1259 (2015). https://doi.org/10.1007/s10439-014-1176-4
4. Chu, C., Bai, J., Wu, X., Zheng, G.: MASCG: multi-atlas segmentation constrained graph method for accurate segmentation of hip CT images. Med. Image Anal. **26**(1), 173–184 (2015). https://doi.org/10.1016/j.media.2015.08.011
5. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition - CVPR 2015, pp. 3431–3440. IEEE (2015). https://doi.org/10.1109/CVPR.2015.7298965
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. Yokota, F., Okada, T., Takao, M., Sugano, N., Tada, Y., Sato, Y.: Automated segmentation of the femur and pelvis from 3D CT data of diseased hip using hierarchical statistical shape model of joint structure. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5762, pp. 811–818. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04271-3_98

8. Wang, C., Smedby, Ö.: Automatic whole heart segmentation using deep learning and shape context. In: Pop, M., et al. (eds.) STACOM 2017. LNCS, vol. 10663, pp. 242–249. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75541-0_26
9. Johnson, C., et al.: Accuracy of CT colonography for detection of large adenomas and cancers. N. Engl. J. Med. **359**(12), 1207–1217 (2008). https://doi.org/10.1056/NEJMoa0800996
10. Roth, H.R., et al.: A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8673, pp. 520–527. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10404-1_65
11. Yushkevich, P., et al.: User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage **31**(3), 1116–1128 (2006). https://doi.org/10.1016/j.neuroimage.2006.01.015
12. Leventon, M., Grimson, W., Faugeras, O.: Statistical shape influence in geodesic active contours. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition - CVPR 2000, pp. 316–323. IEEE (2000). https://doi.org/10.1109/CVPR.2000.855835
13. Wang, C., Smedby, Ö.: Automatic multi-organ segmentation in non-enhanced CT datasets using hierarchical shape priors. In: Proceedings of 22nd International Conference on Pattern Recognition - ICPR 2014, pp. 3327–3332. IEEE (2014). https://doi.org/10.1109/ICPR.2014.574