Short Papers

Vesalius: VNet-Based Fully Automatic Segmentation of Intervertebral Discs in Multimodality MR Images

Claudia Iriondo[1,2](✉) and Michael Girard[2,3]

[1] Department of Bioengineering, University of California, Berkeley,
Berkeley, CA, USA
iriondo@berkeley.edu
[2] Department of Radiology and Biomedical Imaging, University of California,
San Francisco, San Francisco, CA, USA
michael.girard@ucsf.edu
[3] Center for Digital Health Innovation, University of California, San Francisco,
San Francisco, CA, USA

Abstract. Named after Andreas Vesalius (1914–1964) for his landmark description of the intervertebral discs, Vesalius is a VNet-based method for fully automatic segmentation of intervertebral discs T11/T12 to L5/S1 in sagittal Dixon MR sequences. Our method uses aggressive data augmentation, transfer learning from a T2 weighted MR dataset, a fully convolutional VNet architecture trained on full resolution image volumes, and model ensembling to smoothly segment intervertebral discs with up to 0.9285 Dice in our preliminary tests.

Keywords: Spine · Intervertebral disc · Segmentation · VNet

1 Data Splitting

As mentioned in the IVDM3seg challenge description, data released for method development consisted of 8 patients scanned at two timepoints. Leakage between training and testing data should be avoided; the network should be discouraged from "memorizing" specific patients' intervertebral discs during training. Volumes were paired based on structural similarity and restricted to the same training/testing group. Augmentations of these volumes were also contained to the same group. Train test split was 14/2 and leave-one-out-cross validation was performed by shuffling training and testing groups while respecting patient divisions, creating a total of 8 unique data folds.

2 Data Augmentation and Preprocessing

The full dataset was augmented 38X using a combination of 3D rotation, 3D affine transforms, and 3D elastic deformations and stored offline. These deformations were designed to mimic variable patient positioning, spinal curvature, disc size, and disc shape.

G. Zheng et al. (Eds.): CSI 2018, LNCS 11397, pp. 175–177, 2019.
https://doi.org/10.1007/978-3-030-13736-6

Volumes and segmentation masks were resampled to isotropic dimensions, augmented, then resampled back to their original dimensions. Volumes and masks were interpolated using cubic interpolation and masks defined by a 0.5 threshold. Intensities were normalized to zero mean, unit variance on a per volume, per channel basis.

3 Network Structure and Training Details

A 3D VNet architecture was implemented in Tensorflow using Python (algorithm [1], graph [2]) and trained end-to-end. VNet is a fully convolutional neural network architecture consisting of sequential 3D strided convolutional downsampling units and a transpose convolution upsampling path with skip connections concatenated at each resolution. By training the network with our full volumes, instead of patches, we were able to leverage the spatial context of the whole image to predict our binary segmentation mask. Our final network took a 4 channel input, one channel per modality, expanded to 16 channels at the first layer, and doubled in channels every subsequent level for a total of 256 channels at the bottom of the network. The networks 4 levels had 1,2,3,3 convolutions respectively and 3 convolutions at the bottom level with ReLu activations.

Our loss metric combined weighted cross entropy (wce) loss and soft Dice loss. Although the contribution of weighted cross entropy in the combined loss function was relatively small (wce scaled by 0.017 and added to soft Dice), our combined loss metric was successful in addressing the imbalance of foreground to background voxels.

Finally, each network was trained for 25 epochs (approximately 8 h, although convergence was seen within 30 min) on a single Nvidia Titan X GPU using gradient descent optimizer with exponential learning rate decay.

4 Transfer Learning

Networks underwent supervised pre-training for 25 epochs on a single-channel T2 weighted dataset from a previous MICCAI competition [3]. The T2w dataset was cropped to match the field of view and resolution of the Dixon sequence, augmented using the techniques described above, and broadcast to four input channels to match the dimensions. Learned weights were used for weight initialization of our VNet.

5 Hyperparameter Tuning and Ensembling

A random search of 60 unique combinations of hyperparameters was performed. Due to computational restrictions, the search was only performed on 1 of the 8 data folds. The hyperparameter set with the highest and most stable test Dice accuracy and visually smoothest segmentation was selected. Finally, 8 models were trained, each on a unique data fold, using this hyperparameter set (initial learning rate = 0.029, decay steps/decay rate = 3500/0.0700, background voxel weight wce = 0.014, foreground voxel weight wce = 1.0, wce contribution to loss = 0.017, batch size = 1, dropout = 0.80).

The input image is loaded, each channel normalized to zero mean unit variance, and it is run through inference of models 1 through 8. The logits of all models are averaged and used for prediction. In the case of a "missing" prediction for T11/T12 disc, the input is flipped across axis = 0 and run through the inference again, and auxiliary predictions are used for segmentation.

6 Post-processing

A 3D connected component analysis was used to eliminate predicted volumes of less than 1200 voxels. Based on our observation of the manual segmentation ground truth, segmentations appeared to be processed slice-wise in the sagittal plane. Partial volume effect is a known problem in determining boundaries for segmentation of "bookend" slices. To address this issue, a 2D connected component analysis was performed on the bookend sagittal slices to remove any segmentations smaller than 25 pixels (size of smallest manually drawn ROI). Finally, 3D connected components were labeled bottom up with background assigned a value of 0, disc L5/S1 assigned a value of 1, L4/L5 2 and so on. The center of the disk is defined as the centroid of each 3D connected component.

References

1. Milletari, F., Navab, N., Ahmadi, S.: V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation (2016). https://arxiv.org/abs/1606.04797
2. Monteiro, M.: https://github.com/MiguelMonteiro/VNet-Tensorflow. Accessed May 2018
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. (2017)

Segmenting Labeled Intervertebral Discs in Multi Modality MR Images

Alexander Oliver Mader[1,2,3(✉)], Cristian Lorenz[3], and Carsten Meyer[1,2,3]

[1] Institute of Computer Science,
Kiel University of Applied Sciences, Kiel, Germany
[2] Department of Computer Science, Faculty of Engineering,
Kiel University, Kiel, Germany
[3] Department of Digital Imaging,
Philips Research Hamburg, Hamburg, Germany
alexander.o.mader@fh-kiel.de

1 Method

The task it so segment seven well-defined intervertebral discs (IVDs) in multi modality MR images. For this we propose a method specifically designed to be trained on a very small training set. The key idea is to reorient sections around the individual IVDs to a standard orientation in order to be efficiently segmented by an IVD-agnostic V-Net [2]. This leads to the following four step approach, as illustrated in Fig. 1.

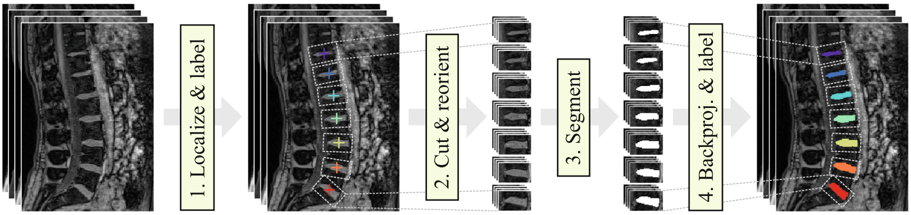

Fig. 1. Illustration of our four step approach to predict labeled IVD segmentations.

1.1 Localizing and Labeling IVDs

First, we use our approach proposed in [1] to localize and label the IVD's center of mass positions. It is a general method to localize and label arbitrary key points by applying landmark-specific localizers (e.g., random forests or FCNs, here random forests) followed by a conditional random field (CRF) to model the global shape. It has been applied to different dimensionalities (2D, 3D) and modalities (X-ray, CT) already, but has neither been applied so far to MR images nor to a multi modality setup. Here, we extend our method to work with multi modality (i.e., multi-channel) MR images by correspondingly increasing the depth of

G. Zheng et al. (Eds.): CSI 2018, LNCS 11397, pp. 178–180, 2019.
https://doi.org/10.1007/978-3-030-13736-6

the image volume (4 in this case instead of a single channel volume). An additional modification compared to our previous work applies to the CRF: Instead of using the proposed binary potentials, we use ternary potentials to increase the rotation and scaling invariance in combination with unary potentials related to the localizers. I.e., we use a Gaussian distribution to model the ratio between two distances and a von Mises distribution to model the relative angle between two vectors projected to one plane. Applying this method we obtain a labeled localization hypothesis for each IVD.

1.2 Sampling Reoriented IVD Sections

Given the IVD locations predicted by the previous step, we sample small reoriented fixed-size sections around each prediction. The size ($6 \times 5 \times 3\,\mathrm{cm}$) is chosen such that the classes are balanced and the sections are reoriented such that the IVDs are level inside the sections (see second step in Fig. 1). PCA was applied to the training segmentations to find the standard orientation of each IVD.

1.3 Segmenting IVDs

As third step, we perform the actual segmentation of the disc tissue using the fully convolutional network V-Net [2]. We use the standard architecture and train it using the setup proposed by the authors. A mini-batch size of 7 is used and the optimization is carried out for 5000 epochs. To tackle the problem of few training cases, we train one label agnostic model to segment all seven IVD sections, effectively using the network to discriminate disc tissue from non-disc tissue (2-class problem instead of 8-class problem). To further accelerate the performance, we increase the training set size even more by a factor of 10 using data augmentation in the form of translation and rotation. Note that histogram matching is performed prior segmentation as data normalization.

Finally, the resulting segmentations are back-projected into the original label space and relabeled according to the label predicted by the CRF in first step.

1.4 Evaluation

We used an 8-fold cross validation setup (14 training images, 2 test images) to estimate essential parameters and to evaluate the training performance. On average, our method achieves a Dice coefficient of 0.894 and a mean surface distance of 0.45 mm, while processing one image in (on average) less than 10 s. For testing, we use an ensemble of our 8 models to improve robustness.

Acknowledgements. This work has been financially supported by the Federal Ministry of Education and Research under the grant 03FH013IX5. The liability for the content of this work lies with the authors.

References

1. Mader, A.O., et al.: Detection and localization of landmarks in the lower extremities using an automatically learned conditional random field. In: Cardoso, M.J., et al. (eds.) GRAIL/MFCA/MICGen -2017. LNCS, vol. 10551, pp. 64–75. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67675-3_7
2. 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)

Author Index