Multiclass CBCT Image Segmentation for Orthodontics with Deep Learning

H. Wang, J. Minnema, K.J. Batenburg, T. Forouzanfar, F.J. Hu, and G. Wu

Appendix

CBCT scan information

All CBCT scans were acquired using a NewTom VGi scanner, with a tube voltage of 75 kV - 110 kV and tube current of 1 mA - 32 mA. All CBCT scans contained axial slices of 512 x 512 voxels with a size of 0.3 mm.

CNN architecture

In this study we employed a MS-D network that was originally developed by Pelt and Sethian (Pelt and Sethian 2018). This MS-D network uses dilated convolutional filters to capture relevant patterns at different image scales. In addition, all layers of the MS-D network are densely connected, which means that relevant patterns can be directly passed to deeper layers in the network. As a result, the MS-D network consists of far fewer trainable parameters than alternative CNN architectures such as U-Net or ResNet. This reduces the risk of overfitting on the training data (Pelt and Sethian 2018), without suffering from lower segmentation performances (Minnema et al. 2019). Moreover, the MS-D network has demonstrated strong performance in improving the quality of tomographic data (Pelt et al. 2018). A schematic overview of an MS-D network with a depth of 3 and a width of 1 is presented in Figure 1A. A detailed description of the MS-D network can be found in (Pelt and Sethian 2018).

Implementation and training details

The depth and the dilation factors of the network were adopted from the study by Minnema et al. (Minnema et al. 2019). Specifically, the depth was 100, and the dilation factor was 1 for the first convolutional layer and increased by 1 for each subsequent layer. After 10 layers, the dilation factor was reset to 1, and the same scheme was applied. The network width was chosen as 1.

Three different experiments were designed to evaluate the MS-D network's segmentation performance. The first experiment was multi-class segmentation, in which the MS-D network was trained to simultaneously segment 3 labels: (1) jaw, (2)

teeth, and (3) background. The second and third experiments were binary segmentation, where the MS-D network segmented either jaw, or teeth, respectively.

In this study, training the MS-D network was performed following a modified version of the k-fold cross-validation scheme. The k-fold cross-validation is typically used to tune the hyper-parameters of the network (Anguita et al. 2012). In the standard procedure, the data sets are split into k folds. The data in k-1 folds are used for training and 1 remaining fold is used for validation. This process is repeated until all folds are used exactly once as validation set. The hyper-parameters of the network can then be chosen based on the highest possible performance on the validation set. The model is subsequently tested on an independent hold-out test set.

However, this typical k-fold cross-validation scheme can lead to unreliable results when the hold-out test set consists of few CBCT scans, as it heavily depends on the properties of the randomly chosen CBCT scans in the hold-out test set. In order to overcome the limitation of available data for testing, we applied a 4-fold cross-validation scheme on the test set (Fig. 1B), while using a hold-out validation set. More specifically, 28 CBCT scans were divided into 4 subsets (S1, S2, S3, and S4), each containing 7 scans. The number of slices was 2226, 2214, 2216, and 2187 in S1, S2, S3, and S4 respectively. Each experiment followed a 4-fold cross-validation scheme, which means that 3 subsets were used for training and 1 subset was used for testing. This process was repeated 4 times such that each CBCT scan was used for testing exactly once. Performing a 4-fold cross-validation scheme on the test set allowed us to evaluate the segmentation performance of the MS-D network on all CBCT scans (28 in total), thus making the evaluation robust to differences between the CBCT scans and insensitive to the random choice of test set.

An independent hold-out validation set was used to determine the optimal number of epochs for training. This validation set consisted of 2 CBCT scans which were not included in the 4-fold cross-validation scheme. The number of epochs was chosen as 20 for all training iterations, as the segmentation performance on the validation set did not improve when trained longer. It should be noted that the validation set consisted of relatively few CBCT scans. However, because the MS-D network has a low risk of overfitting the training data (Pelt and Sethian 2018), and only a single hyper-parameter was tuned (i.e., number of epochs), 2 CBCT scans were sufficient to reliably determine

the number of epochs in our study.

The MS-D network was implemented by Hendriksen (Hendriksen 2019) and the python code for training the MS-D network is publicly available at https://github.com/ahendriksen/msd_pytorch. Implementation of the MS-D network was performed using the deep learning platform PyTorch (version 0.3.1) in Python (version 3.6.1). Training and testing were performed on 2D axial CBCT slices using a batch size of 1 and the default Adam optimizer (Kingma and Ba 2014) on a Linux desktop computer (HP Workstation Z840) with 64 GB RAM, a Xeon E5-2687 v4 3.0 GHZ CPU and a GTX 1080 Ti GPU card. Each training epoch took approximately 1 hour.

CNN performance evaluation

The segmentation performance of the MS-D network was evaluated using the Dice similarity coefficient (DSC) which is a well-known metric in the medical image segmentation domain (Zou et al. 2004). DSCs were calculated on the patient level, which means that a single DSC was calculated for each segmented CBCT volume.

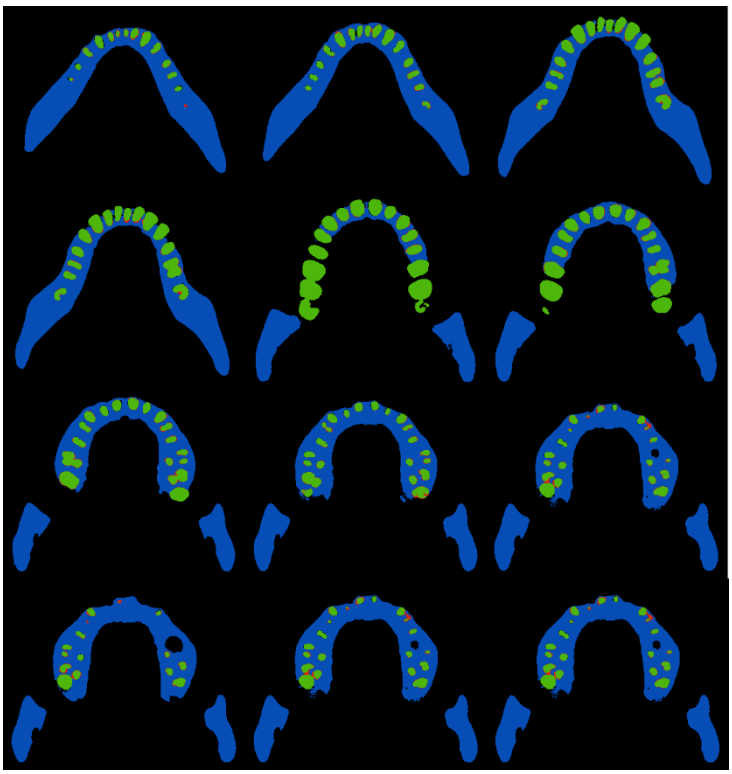
All segmented CBCT scans (i.e., MS-D network segmentations and gold standard segmentations) were also converted into 3D models using 3D Slicer software (3D Slicer). Surface deviations between the MS-D network-based 3D models and the gold standard 3D models were calculated to evaluate the accuracy of the MS-D segmentation around the edges of bony structures. These surface deviations were analyzed within the range of -5.0 mm and +5.0 mm using GOM Inspect software (GOM Inspect 2018, GOM GmbH, Braunschweig, Germany). Additionally, mean absolute deviations (MADs) were calculated between all the MS-D network-based 3D models and the corresponding gold standard 3D models.

After the 4 iterations of the cross-validation scheme, the performance of the MS-D network was averaged over the 28 segmented CBCT scans. All results are presented as means ± standard deviation (SD). The data analysis was performed using GraphPad Prism 8 (GraphPad). Equivalence tests were performed with a threshold difference of 0.005. If the 90% CIs were within (-0.005, 0.005), the two groups were considered to be equivalent with a confidence of 95%.

Appendix Table. Mean absolute surface deviation of jaw and teeth 3D models

Patient ID	Jaw segmentation		Teeth segmentation	
	Multiclass (mm)	Binary (mm)	Multiclass (mm)	Binary (mm)
P1	0.407	0.403	0.184	0.170
P2	0.443	0.450	0.238	0.211
P3	0.484	0.538	0.230	0.203
P4	0.540	0.527	0.227	0.197
P5	0.364	0.408	0.251	0.208
P6	0.371	0.372	0.132	0.126
P7	0.345	0.340	0.133	0.141
P8	0.438	0.434	0.168	0.132
P9	0.297	0.327	0.152	0.124
P10	0.401	0.432	0.189	0.178
P11	0.369	0.324	0.297	0.149
P12	0.615	0.630	0.258	0.267
P13	0.330	0.345	0.244	0.194
P14	0.524	0.516	0.199	0.197
P17	0.375	0.416	0.178	0.136
P18	0.446	0.483	0.159	0.118
P19	0.415	0.454	0.280	0.156
P20	0.308	0.333	0.255	0.179
P21	0.371	0.368	0.110	0.090
P22	0.268	0.265	0.146	0.132
P23	0.469	0.488	0.352	0.321
P24	0.273	0.296	0.168	0.118
P25	0.387	0.434	0.202	0.157
P26	0.370	0.336	0.186	0.160
P27	0.378	0.349	0.291	0.168
P28	0.198	0.244	0.100	0.077
P29	0.490	0.669	0.166	0.134
P30	0.257	0.305	0.206	0.128
min	0.198	0.244	0.100	0.077
max	0.615	0.669	0.352	0.321
Mean ± SD	0.390±0.093	0.410±0.103	0.204±0.061	0.163±0.051

Each fold contained 7 CBCT scans. S1: Patient 1-7; S2: Patient 8-14; S3: Patient 17-23; S4: Patient 24-30.



Appendix Figure. The conflicting labels induced by the binary segmentation (12 CBCT slices from patient 1). The conflicting labels are marked in red. The blue color represents the segmented jaw and the green color represents the segmented teeth.

References

3D Slicer. http://www.slicer.org.

Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S. 2012. The 'K' in K-fold Cross Validation. ESANN 2012 proceedings. 441-446.

Hendriksen AA. 2019. ahendriksenh/msd_pytorch: v0.7.2. Zenodo.

Kingma DP, Ba J. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980.

Minnema J, van Eijnatten M, Hendriksen AA, Liberton N, Pelt DM, Batenburg KJ, Forouzanfar T, Wolff J. 2019. Segmentation of dental cone-beam CT scans affected by metal artifacts using a mixed-scale dense convolutional neural network. Med Phys. 46(11):5027-5035.

Pelt DM, Batenburg KJ, Sethian JA. 2018. Improving tomographic reconstruction from limited data using mixed-scale dense convolutional neural networks. J. Imaging. 4(11):128.

Pelt DM, Sethian JA. 2018. A mixed-scale dense convolutional neural network for image analysis. PNAS. 115(2):254-259.

Zou KH, Warfield SK, Bharatha A, Tempany CM, Kaus MR, Haker SJ, Wells III WM, Jolesz FA, Kikinis R. 2004. Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. Acad Radiol. 11(2):178-189.