Multi-dimensional Gated Recurrent Units for Automated Anatomical Landmark Localization

Simon Andermatt[*,1], Simon Pezold[*,1], Michael Amann[2,3,4], and
Philippe C. Cattin[1]

[1]Department of Biomedical Engineering, University of Basel, Allschwil, Switzerland
[2]Department of Neurology, University Hospital Basel, Basel, Switzerland
[3]Department of Radiology, University Hospital Basel, Basel, Switzerland
[4]MIAC AG, Basel, Switzerland
[*]S. Andermatt and S. Pezold contributed equally.

Abstract. We present an automated method for localizing an anatomical landmark in three-dimensional medical images. The method combines two recurrent neural networks in a coarse-to-fine approach: The first network determines a candidate neighborhood by analyzing the complete given image volume. The second network localizes the actual landmark precisely and accurately in the candidate neighborhood. Both networks take advantage of multi-dimensional gated recurrent units in their main layers, which allow for high model complexity with a comparatively small set of parameters. We localize the medullopontine sulcus in 3D magnetic resonance images of the head and neck. We show that the proposed approach outperforms similar localization techniques both in terms of mean distance in millimeters and voxels w.r.t. manual labelings of the data. With a mean localization error of 1.7 mm, the proposed approach performs on par with neurological experts, as we demonstrate in an interrater comparison.

1 Introduction

Localizing anatomical landmarks is a common task in many medical applications. Finding matching anatomical points in images may be necessary for seeding a segmentation algorithm, for registration problems, or for providing points of reference for quantitative measurements. Although finding landmarks in volumetric images is error-prone and time-consuming, the task is often still carried out manually. Using a fully automated approach mitigates the inter and intra-rater variability through an objective and efficient process without manual interference. Therefore, many automated localization methods have been proposed, with varying degrees of robustness, reliability, and generalization potential. Some of the methods, such as Bhanu Prakash et al. [2] or Elattar et al. [3], use very basic image processing techniques, but many others rely on concepts from machine learning: for example, for localizing landmarks in the brain, Guerrero et al. [6] use manifold learning and O'Neil et al. [9] use random forests; for cardiac landmark localization, Karavides et al. [7] use Adaboost and Lu and Jolly [8] use

probabilistic boosting trees; Xue et al. [11] use boosting for localizing landmarks on the knee joint. For a recent overview, also see Zhou et al. [15].

In recent years, ground-breaking advancements using neural networks have been achieved in various domains, allowing for automatic learning of discriminative features for the problem at hand and avoiding the need for manually designed (often called handcrafted) features. Consequently, these techniques have also found their way into landmark localization. Examples are Zheng et al. [14], who use two neural networks successively to localize the carotid bifurcation in 3D CT images, Ghesu et al. [4], who propose a so-called artificial agent for localizing various anatomical landmarks in 2D and 3D images of different modalities, and Yang et al. [12], who apply convolutional neural networks for landmark localization on the femur in MR images.

Existing approaches based on convolutional neural networks (CNNs) are capable of detecting very delicate structure, yet are limited to the local neighborhood of the filters used in each layer of the network. Using a recurrent neural network (RNN) for this task allows for flexible feature relationships of varying length and scale. This is especially useful given a localization task, where the surrounding tissues structure can take a number of different shapes and sizes. Tackling volumetric data with RNNs for *segmentation* has been recently demonstrated by Andermatt et al. [1] with multi-dimensional gated recurrent units (MD-GRUs). To our knowledge, neither multi-dimensional RNN nor MD-GRUs have been applied to the task of *landmark localization* so far.

In this paper, we propose to apply MD-GRUs in a two-stage approach to the task of anatomical landmark localization. In the first stage, the anatomical region of interest is roughly located in the given image volume. We then determine the actual landmark coordinate in a subvolume in the second stage. We apply the proposed method to 3D MR images of the head and neck, in which we locate the medullopontine sulcus, and compare the found coordinates to those of manual labels. Our results from an interrater comparison suggest that the proposed method cannot be distinguished from a clinical expert.

2   Methods

For the accurate localization of landmarks, we propose to use two separate localization networks of similar structure, to both accelerate the process and allow for a decently complex network. Both localization networks work on the same number of voxels – in our case we fixed it to 64^3 voxels – and find the coordinate in said volume which lies closest to the true landmark. The first network is provided data subsampled to such a degree, that the full original volume can be represented inside of it. The network will then approximate a location, which will in turn be used to sample a subvolume at the original resolution from the image data around the found location. In our case, the first network is provided with 4-fold subsampled data and the second processes data at the original resolution, centered at the location which was found by the first network.
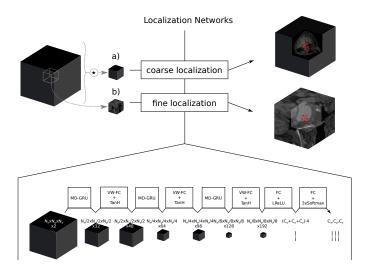
Fig. 1. Localization network. a) Coarse approximation of landmark coordinates in sub-sampled low resolution representation of full data. b) Fine approximation of landmark coordinates in extracted window around detected coarse location in a second localization network. Both networks use the architecture depicted at the bottom.

Subsampling MD-GRU Layer We propose to adapt the MD-GRU layer [1], which was introduced to handle segmentation problems, to the application of landmark localization. In order to do so, we implement the ability to subsample at each MD-GRU layer and hence at each convolutional gated recurrent unit (C-GRU) which it consists of. This effectively reduces the spatial problem size, allowing a multi-resolution processing approach. We adjust the original C-GRU equations as follows:

$$f^j(t, \alpha, \beta) = \sum_i^I x_t^i \star \alpha^{i,j} + \beta^j, \qquad g^j(t, \alpha) = \sum_k^J h_{t-1}^k * \alpha^{k,j}, \tag{1}$$

$$r_t^j = \sigma(f^j(t, w_r, b_r) + g^j(t, u_r)), \qquad z_t^j = \sigma(f^j(t, w_z, b_z) + g^j(t, u_z)), \tag{2}$$

$$\tilde{h}_t^j = \phi(f^j(t, w, b) + r_t^j \odot g^j(t, u)), \qquad h_t^j = z_t^j \odot h_{t-1}^j + (1 - z_t^j) \odot \tilde{h}_t^j, \tag{3}$$

where $x_t^{\cdot}$, $h_t^{\cdot}$ denote the input and state of the C-GRU at time t, and i, j, k denote the respective channels. The operator $\odot$ denotes elementwise multiplication, as in [1]. Variables u, w, and b are trainable weights. We call $\tilde{h}$ in Eq. (3) the proposal and r and z in Eqs. (2) the reset and update gate.

We accomplish subsampling by introducing strided convolutions, which are denoted as $\star$ in Eq. (1). The size of the state as well as of all the gates and the proposal will be reduced by the factor of the chosen stride S per spatial dimension. Each C-GRUs' output is then subjected to one-dimensional average pooling, compressing the time dimension by stride S. The sum of all d compressed

C-GRU results $\hat{h}$ yields the MD-GRU output H:

$$H^j = \sum_d \hat{h}^j, \qquad \hat{h}^j_{t'} = \frac{1}{S} \sum_{s=0}^{S-1} h^j_{St'+s}. \qquad (4)$$

Localization Network At the core, we use the same localization network for all experiments. We use three subsequent compositions of a subsampling MD-GRU layer, a voxelwise fully connected layer, and a tanh activation function. The subsampling MD-GRU layers are provided with 32, 64, and 128 channels, respectively. All of them use strides of 2 along spatial dimensions, the volume is hence subsampled 8-fold at each composition. We use DropConnect [10] with a drop rate of 0.5 on the input convolution filters of both gates r^j, z^j and the proposal $\tilde{h}$. The voxelwise fully connected layers are realized through convolution layers with spatial filters of 1^3, with 48, 96, and 192 channels each.

The resulting subvolume is of size $N_x/8 \times N_y/8 \times N_z/8$, given the input shape was $(N_x \times N_y \times N_z)$. The subvolume is reshaped into a vector, in which we process each coordinate by two fully connected layers of $(C_x + C_y + C_z) \cdot 4$ and $(C_x + C_y + C_z)$ layers, which are connected through a leaky rectifying unit defined as $\mathrm{lrelu}(x) = \max\{0.01\,x,\ x\}$. The resulting vector is split into three separate vectors of sizes C_x, C_y, and C_z, where $C_.$ gives the number of possible coordinate positions along the respective dimension. These are then fed into individual softmax activation functions to estimate the probabilities for each coordinate in each vector. We use the sum of all cross entropy losses as loss function for the entire network. Figure 1 shows an overview of the network architecture.

Subsampling In the first stage, we use a strided convolution on the input to match the localization networks input resolution. We pad the input, such that the shape of the volume is a multiple of the required shape for the localization network. In our case, we padded the data to 256^3 and used strides S of 4 with a filter size of $S \cdot 2 + 1$ and 16 channels for the convolution layer.

Superresolution Our method, as explained so far, is restricted to voxel coordinates, since we estimate with our method discrete instead of continuous coordinates. In the following, we explain two extensions to our idea to yield superresolution results.

The first extension takes advantage of the coordinate resolution-independent formulation in the *Localization Network* paragraph above. Instead of estimating as many classes for each of the three coordinates as there are voxels in the respective dimension in the volume, we estimate n times the amount. This allows us to estimate values which are $1/n$ voxels apart and hence allow for a more fine-grained localization. In our experiments, we use $n = 4$ resulting in 256 classes.

Our second idea exploits neighborhood information in our coordinate probability vectors by fitting a parabola to the largest probability and its two neighbors per coordinate. The maxima of these functions can then be interpreted as our
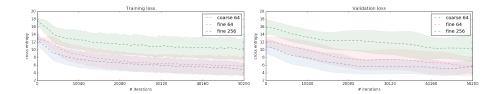
Fig. 2. Cross entropy loss. Mean ± one standard deviation on training and validation set for the 3 trained networks, smoothed using a gaussian for visualization.
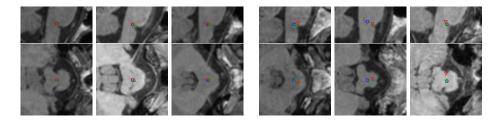

Fig. 3. Localization results for rater 1 *(red ▽)*, rater 2 *(green △)*, and proposed method *(blue ○)*. Shown are the best three *(left)* and worst three *(right)* localizations of the proposed method wrt. rater 1, both in sagittal *(top)* and transverse *(bottom)* view.

coordinate location. This allows for an even finer localization, but is based and hence limited on the chosen number of coordinate probabilities.

Optimization We trained each localization network together with their sub-sampling addition individually. All networks were trained for a total of 50 epochs, where one epoch comprised one random sample from each training subject, which led to a total of 50 200 iterations. We used AdaDelta [13] with a learning rate of 0.001. We initialized all weights of the convolutions with the method of Glorot and Bengio [5], the biases with zero and the fully connected layers at the end of the localization network with random values from $[-\sqrt{3}/N_i, +\sqrt{3}/N_i]$, where N_i is the number of input units. For the first network, we sampled from the center of the padded volume with a random offset in the range of $[-100, 100]$ voxels per coordinate; for the second network, we just required that the training landmark was within the volume. The training loss is visualized in Fig. 2.

For preprocessing, we apply a high-pass filter on the input, the results of which we use together with the original data as input to our networks. Additionally, we normalize to zero mean and a standard deviation of one for each of the input volumes. Apart from this, no preprocessing is required.

3 Results

To evaluate the proposed approach, we located the medullopontine sulcus, a distinct cavity in the brainstem, in MR images of the head and neck (see Fig. 3).

Table 1. Localization accuracy and precision. a) Localization error on the test set when using only the first network (*top row*) and both networks with a varying number of coordinate classes, with or without parabola fitting (*bottom row:* proposed combination); b) localization error on the test set in comparison to two human raters; c) localization errors reported in the literature.

a)	Error [mm]			b)	Error [mm]		
	Median	Mean	Std.		Median	Mean	Std.
Coarse localization	4.83	5.02	2.22	Rater 1 vs. rater 2	**1.39**	**1.59**	0.98
Fine, 64 classes	1.74	1.97	1.02	Proposed vs. rater 1	1.40	1.69	1.02
Fine+parab., 64 cl.	1.77	1.89	**0.98**	Proposed vs. rater 2	1.65	1.73	**0.87**
Fine, 256 classes	1.47	1.72	1.03	Proposed vs. both	1.50	1.71	0.95
Fine+parab., 256 cl.	**1.40**	**1.69**	1.02				

c)	Error [mm]				
Method	Median	Mean	Std.	Voxel size [mm^3]	Target landmark
Proposed	1.50	1.71	0.95	$1.00 \times 1.00 \times 1.00$	medullopontine sulcus
Zheng et al. [14]	1.21	2.64	4.98	$0.46 \times 0.46 \times 0.50$	carotid bifurcation
Ghesu et al. [4]	0.8	1.8	2.9	$1.00 \times 1.00 \times 1.00$	carotid bifurcation
Yang et al. [12]	—	4.13	1.70	$0.37 \times 0.37 \times 0.70$	femoral medial distal point
Xue et al. [11]	—	1.41	0.91	$0.3 \times 0.3 \times [0.6, 3]$	knee joint (23 landmarks)
Guerrero et al. [6]	—	0.45	0.22	—	anterior commissure

Images were acquired with a T1-weighted MPRAGE sequence, having a resolution of 1 mm^3 and a size between $160 \times 240 \times 256$ voxels and $192 \times 256 \times 256$ voxels. Altogether, we had 1218 images of 265 subjects, with a median number of 5 images per subject (minimum: 1, maximum: 8), which we randomly assigned to a training set (1004 images of 213 subjects), a validation set (114 images of 26 subjects), and a test set (100 images of 26 subjects), making sure that all images of each subject were assigned to the same set.

For training and evaluation of the localization, we used manual labels of the landmark. These labels were provided by clinical expert raters who placed them on a graphical user interface enabling them to zoom in and out of the imaged volumes as necessary. To allow for interrater comparisons, we had two raters place the landmark in all images of the test set.

Training 50 epochs for the coarse and fine networks took around 41 and 34 hours, respectively. Testing, on the other hand, requires less than 2 seconds for either network, resulting in a total of around 3–4 seconds for localization. Using our extension of estimating 256 class probabilities instead of 64 per coordinate requires only 2.5 hours more training time and took around 2.5 seconds per volume for testing, which results in around 4 seconds in total for localization.

Figure 3 shows our three best and worst localization results. Note that our largest error (rightmost column in Fig. 3) is actually produced by a mislabeling of a clinical expert, as can be seen by the off-center position of the red marker.

Table 1a shows the localization errors when using only the first network as compared to using both. The second network increases the localization accuracy notably, as does using more coordinate classes and fitting a parabola.

Table 1b shows the results from comparing both human raters with the proposed approach. The listed values indicate that our approach almost reaches human performance: comparing our results to those of a human rater produces approximately the same error as two human raters compared to each other.

Table 1c shows results for landmark localization reported in the literature.

4 Discussion and Conclusion

Our results, as listed in Table 1c, appear competitive: compared to other neural network approaches [4,12,14], mean error and standard deviation are better in terms of millimeters and voxels. When comparing to Xue et al. [11], one has to keep in mind their notably higher in-plane resolution. While Guerrero et al. [6] achieve higher accuracy and precision, a comparison appears difficult: apart from not stating the voxel size, their method requires images with similar field of view, which cannot be guaranteed in our case, as parts of our images are centered on the neck while others are centered on the head. In any case, caution has to be taken when comparing these results: on the one hand, evaluated anatomical landmarks, imaging modalities, and image resolutions differ. On the other hand, our interrater comparison (recall Table 1b) suggests that there is a lower bound for the achievable accuracy, which might be well above a given image resolution and might depend on the particular anatomical landmark. Determining the limit of actually achievable accuracy of our method would require evaluating data with lower interrater variability. The results of Xue et al. [11] allow a similar conclusion, in that their method's error is similar to the error from their interrater comparison, as well. Unfortunately, the other authors do not provide interrater comparisons.

We have shown two ideas that improved our localization results. The combination of both even surpassed the accuracy of each of them applied separately. Considering interrater variability, we are still slightly less accurate than a human rater. We think that this is partly based on the discrete probability distribution and our sampling technique when training the algorithm. We randomly sampled subvolumes using integer coordinates during training since this process does not require interpolation. But this also means that each training sample could only get mapped on a subset of all possible coordinate classes.

Conclusion We have shown that the localization of the medullopontine sulcus is successfully possible using our proposed automated technique, which adapts MD-GRUs to the task of landmark localization. We introduced a number of improvements, which all led to even more accurate results without significantly increasing the training time. Future work will focus on evaluating our localization approach on multiple anatomical landmarks in different imaging modalities.

References

1. Andermatt, S., Pezold, S., Cattin, P.: Multi-dimensional Gated Recurrent Units for the Segmentation of Biomedical 3D-Data. In: International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. pp. 142–151. Springer (2016)
2. Bhanu Prakash, K.N., Hu, Q., Aziz, A., Nowinski, W.L.: Rapid and Automatic Localization of the Anterior and Posterior Commissure Point Landmarks in MR Volumetric Neuroimages. Academic Radiology 13(1), 36–54 (Jan 2006)
3. Elattar, M., Wiegerinck, E., van Kesteren, F., Dubois, L., Planken, N., Vanbavel, E., Baan, J., Marquering, H.: Automatic aortic root landmark detection in CTA images for preprocedural planning of transcatheter aortic valve implantation. The International Journal of Cardiovascular Imaging 32(3), 501–511 (Mar 2016)
4. Ghesu, F.C., Georgescu, B., Mansi, T., Neumann, D., Hornegger, J., Comaniciu, D.: An Artificial Agent for Anatomical Landmark Detection in Medical Images. In: MICCAI 2016. pp. 229–237. Springer, Cham (Oct 2016)
5. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Aistats. vol. 9, pp. 249–256 (2010)
6. Guerrero, R., Wolz, R., Rueckert, D.: Laplacian Eigenmaps Manifold Learning for Landmark Localization in Brain MR Images. In: MICCAI 2011. pp. 566–573. Springer, Berlin, Heidelberg (Sep 2011)
7. Karavides, T., Leung, K.Y.E., Paclik, P., Hendriks, E.A., Bosch, J.G.: Database guided detection of anatomical landmark points in 3D images of the heart. In: 2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. pp. 1089–1092 (Apr 2010)
8. Lu, X., Jolly, M.P.: Discriminative Context Modeling Using Auxiliary Markers for LV Landmark Detection from a Single MR Image. In: Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges. pp. 105–114. Springer, Berlin, Heidelberg (Oct 2012)
9. O'Neil, A., Dabbah, M., Poole, I.: Cross-Modality Anatomical Landmark Detection Using Histograms of Unsigned Gradient Orientations and Atlas Location Autocontext. In: Machine Learning in Medical Imaging. pp. 139–146. Springer, Cham (Oct 2016)
10. Wan, L., Zeiler, M., Zhang, S., Cun, Y.L., Fergus, R.: Regularization of neural networks using dropconnect. In: Proceedings of the 30th International Conference on Machine Learning (ICML-13). pp. 1058–1066 (2013)
11. Xue, N., Doellinger, M., Ho, C.P., Surowiec, R.K., Schwarz, R.: Automatic detection of anatomical landmarks on the knee joint using MRI data. Journal of Magnetic Resonance Imaging 41(1), 183–192 (Jan 2015)
12. Yang, D., Zhang, S., Yan, Z., Tan, C., Li, K., Metaxas, D.: Automated anatomical landmark detection ondistal femur surface using convolutional neural network. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI). pp. 17–21 (Apr 2015)
13. Zeiler, M.D.: ADADELTA: An Adaptive Learning Rate Method. arXiv:1212.5701 [cs] (Dec 2012)
14. Zheng, Y., Liu, D., Georgescu, B., Nguyen, H., Comaniciu, D.: 3D Deep Learning for Efficient and Robust Landmark Detection in Volumetric Data. In: MICCAI 2015. pp. 565–572. Springer, Cham (Oct 2015)
15. Zhou, S.K.: Discriminative anatomy detection: Classification vs regression. Pattern Recognition Letters 43, 25–38 (Jul 2014)