Regressing Heatmaps for Multiple Landmark Localization Using CNNs

Christian Payer[1](✉), Darko Štern[2], Horst Bischof[1], and Martin Urschler[2,3]

[1] Institute for Computer Graphics and Vision,
Graz University of Technology, Graz, Austria
christian.payer@icg.tugraz.at
[2] Ludwig Boltzmann Institute for Clinical Forensic Imaging, Graz, Austria
[3] BioTechMed-Graz, Graz, Austria

Abstract. We explore the applicability of deep convolutional neural networks (CNNs) for multiple landmark localization in medical image data. Exploiting the idea of regressing heatmaps for individual landmark locations, we investigate several fully convolutional 2D and 3D CNN architectures by training them in an end-to-end manner. We further propose a novel SpatialConfiguration-Net architecture that effectively combines accurate local appearance responses with spatial landmark configurations that model anatomical variation. Evaluation of our different architectures on 2D and 3D hand image datasets show that heatmap regression based on CNNs achieves state-of-the-art landmark localization performance, with SpatialConfiguration-Net being robust even in case of limited amounts of training data.

1 Introduction

Localization of anatomical landmarks is an important step in many medical image analysis tasks, e.g. for registration or to initialize segmentation algorithms. Since machine learning approaches based on deep convolutional neural networks (CNN) outperformed the state-of-the-art in many computer vision tasks, e.g. ImageNet classification [1], we explore in this work the capability of CNNs to accurately locate anatomical landmarks in 2D and 3D medical data.

Inspired by the human visual system, neural networks serve as superior feature extractors [2] compared to hand-crafted filters as used e.g. in random forests. However, they involve increased model complexity by requiring a large number of weights that need to be optimized, which is only possible when a large amount of data is available to prevent overfitting. Unfortunately, acquiring large amounts of medical data is challenging, thus imposing practical limits on network complexity. Additionally, working with 3D volumetric data further increases the number of required network weights to be learned due to the added dimension of filter kernels. Demanding CNN training for 3D input was previously investigated

C. Payer—This work was supported by the province of Styria (HTI:Tech_for_Med ABT08-22-T-7/2013-13) and the Austrian Science Fund (FWF): P 28078-N33.

S. Ourselin et al. (Eds.): MICCAI 2016, Part II, LNCS 9901, pp. 230–238, 2016.
DOI: 10.1007/978-3-319-46723-8_27

for different applications [3–5]. In [3], knee cartilage segmentation of volumetric data is performed by three 2D CNNs, representing xy, yz, and xz planes, respectively. Despite not using full 3D information, it outperformed other segmentation algorithms. In [4], authors differently decompose 3D into 2D CNNs by randomly sampling n viewing planes of a volume for detection of lymph nodes. Finally, [5] presents deep network based 3D landmark detection, where first a shallow network generates multiple landmark candidates using a sliding window approach. Image patches around landmark candidates are classified with a subsequent deep network to reduce false positives. This strategy has the substantial drawback of not being able to observe global landmark configuration, which is crucial for robustly localizing locally similar structures, e.g. fingertips in Fig. 1. Thus, to get rid of false positives, state-of-the-art localization methods for multiple landmark localization rely on local feature responses combined with high level knowledge about global landmark configuration [6], in the form of graphical [7] or statistical shape models [8]. This widely used explicit incorporation has proven very successful due to strong anatomical constraints present in medical data. When designing CNN architectures these constraints could be used to reduce network complexity and allow training even on limited datasets.

In this work, we investigate the idea of directly estimating multiple landmark locations from 2D or 3D input data using a single CNN, trained in an end-to-end manner. Exploring the idea of Pfister et al. [9] to regress heatmaps for landmarks simultaneously instead of absolute landmark coordinates, we evaluate different fully convolutional [10] deep network architectures that benefit from constrained relationships among landmarks. Additionally, we propose a novel architecture for multiple landmark localization inspired by latest trends in the computer vision community to design CNNs that implicitly encode high level knowledge as a convolution stage [11]. Our novel architecture thus emphasizes the CNN's capability of learning features for both accurately describing local appearance as well as enforcing restrictions in possible spatial landmark configurations. We discuss benefits and drawbacks of our evaluated architectures when applied to two datasets of hand images (2D radiographs, 3D MRI) and show that CNNs achieve state-of-the-art performance compared to other multiple landmark localization approaches, even in the presence of limited training data.

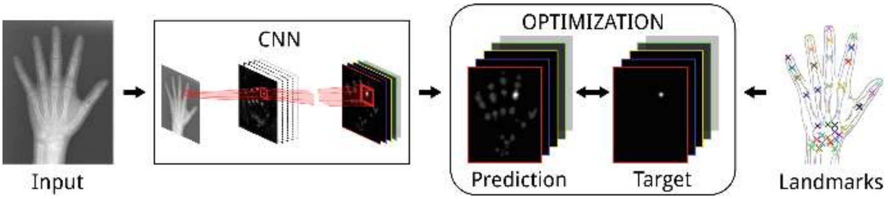
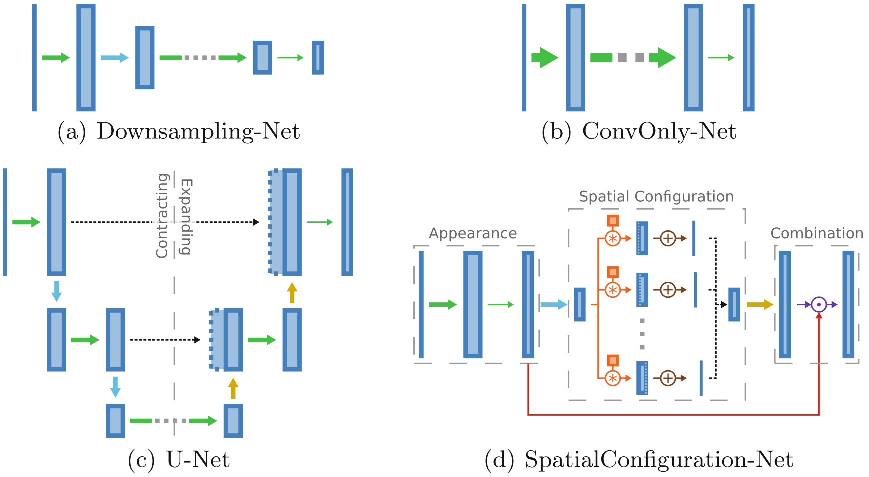


Fig. 1. Multiple landmark localization by regressing heatmaps in a CNN framework.



(a) Downsampling-Net

(b) ConvOnly-Net

(c) U-Net

(d) SpatialConfiguration-Net

Fig. 2. Schematic representation of the CNN architectures. Blue boxes represent images, → convolution, → pooling, ⇢ concatenation, → upsampling, ⊛→ channel-by-channel convolution, ⊕→ channel-by-channel addition, ⊙→ element-wise multiplication. Arrow thickness illustrates kernel sizes.

2 Heatmap Regression Using CNNs

As shown in Fig. 1, our approach for multiple landmark localization uses a CNN framework to regress heatmaps directly from input images. Similarly to [9], we represent heatmaps H_i as images where Gaussians are located at the position of landmarks L_i. Given a set of input images and corresponding target heatmaps, we design different fully convolutional network architectures (see Fig. 2 for schematic representations), all capable of capturing spatial relationships between landmarks by allowing convolutional filter kernels to cover large image areas. After training the CNN architectures, final landmark coordinates are obtained as maximum responses of the predicted heatmaps. We propose three different CNN architectures inspired by the literature, which are explained in more detail in the following, while Sect. 2.1 describes our novel SpatialConfiguration-Net.

Downsampling-Net: This architecture (Fig. 2a) uses alternating convolution and pooling layers. Due to the involved downsampling, it is capable of covering large image areas with small kernel sizes. As a drawback of the low resolution of the target heatmaps, poor accuracy in localization has to be expected.

ConvOnly-Net: To overcome the low target resolution, this architecture (Fig. 2b) does neither use pooling layers, nor strided convolution layers. Thus, much larger kernels are needed for observing the same area as the Downsampling-Net which largely increases the number of network weights to optimize.

U-Net: The architecture (Fig. 2c) is slightly adapted from [12], since we replace maximum with average pooling in the contracting path. Also, instead of learning

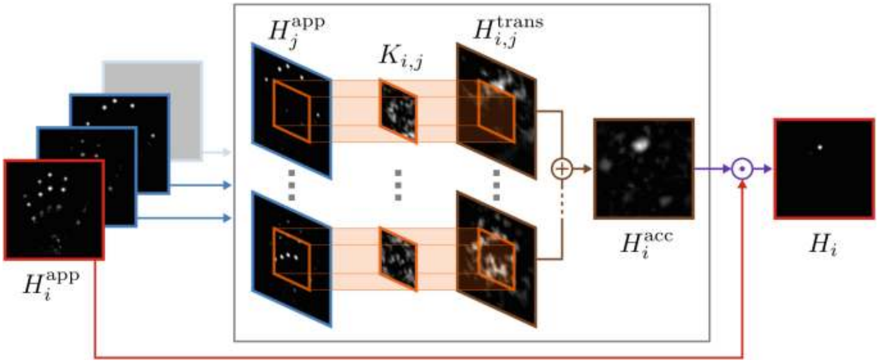


Fig. 3. Illustration showing the combination of local appearance from landmark L_i with transformed heatmaps $H_{i,j}^{\text{trans}}$ from all other landmarks.

the deconvolution kernels, we use fixed linear upsampling kernels in the expanding path, thus obtaining a fully symmetric architecture. Due to the contracting and expanding path, the net is able to grasp a large image area using small kernel sizes while still keeping high accuracy.

2.1 SpatialConfiguration-Net

Finally, we propose a novel, three block architecture (Fig. 2d), that combines local appearance of landmarks with the spatial configuration to all other landmarks. The first block of the network consists of three convolutional layers with small kernel sizes, that result in local appearance heatmaps H_i^{app} for every landmark L_i. Although these intermediate heatmaps are very accurate, they may contain ambiguities due to similarly looking landmarks, e.g. fingertips, as the kernels do not grasp information on the larger surrounding area. This ambiguity is reduced by combining the output of the appearance block with the prediction of our spatial configuration block. A sample heatmap calculation for one landmark is visualized in Fig. 3. The position of each landmark predicted by this block is based on the estimated locations of all remaining landmarks as obtained from the appearance block. Thus, the relative position of landmark L_i according to L_j is learned as a convolution kernel $K_{i,j}$, transforming the response of heatmap H_j^{app} into a heatmap $H_{i,j}^{\text{trans}}$, that predicts the position of L_i. The transformed heatmap is defined as

$$H_{i,j}^{\text{trans}} = H_j^{\text{app}} * K_{i,j}, \tag{1}$$

where $*$ denotes a convolution of H_j^{app} with $K_{i,j}$. Note that by not having any restriction on the kernels $K_{i,j}$, the net is able to learn the spatial configuration between landmarks on its own. For each L_i, the responses of the transformed heatmaps $H_{i,j}^{\text{trans}}$ are accumulated resulting in a location estimation obtained from all other landmarks. This accummulated heatmap is defined as

$$H_i^{\text{acc}} = \sum_{j=1}^{n} H_{i,j}^{\text{trans}}. \tag{2}$$

The final heatmap, which combines local appearance and spatial configuration between all other landmarks, is calculated as

$$H_i = H_i^{\text{app}} \odot H_i^{\text{acc}}, \tag{3}$$

with $\odot$ the element-wise product. This suppresses locations from local appearance predictions that are infeasible due to the spatial configuration of landmarks.

The spatial configuration block is calculated on a lower resolution, as kernels $K_{i,j}$ have to be very large to capture the spatial landmark configuration. However, a high resolution is not necessary for the spatial configuration block, as it is solely used to remove ambiguities. To preserve accuracy of the local appearance block, the outputs of the spatial configuration block are upsampled and the final heatmaps H_i are calculated on the same resolution as the input.

3 Experimental Setup and Results

Materials: We evaluated localization performance of the network architectures on two different datasets. The first one consists of 895 publicly available X-ray images of hands[1] with an average size of 1563×2169, acquired with different X-ray scanners. 37 characteristic landmarks, e.g. finger tips or bone joints, were annotated manually by an expert. As the images do not contain information about physical pixel resolution, we assume a wrist width of 50 mm, defined by two annotated landmarks. The second dataset consists of 60 T1-weighted 3D gradient echo hand MR scans with 28 annotated landmarks. The average volume size was $294 \times 512 \times 72$ with a voxel resolution of $0.45 \times 0.45 \times 0.9\,\text{mm}^3$.

Experimental Setup: The 2D hand radiographs were acquired with various different X-ray scanners, resulting in large intensity variations. Histogram equalization was performed to adjust intensity values. Additionally, we preprocessed pixels by subtracting mean intensities and setting standard deviation equal to 1. For the 3D data, we only subtracted the mean since intensity variations were negligible. To augment the datasets, nine additional synthetic images were created for each image by applying rotation (up to 30°), translation (up to 1 cm), and intensity scaling/shifting (up to 10 % difference, only used in 2D). Heatmaps were created by centering Gaussians at landmark positions, normalized to maximum value of 1 and with σ ranging from 1.5 to 3 depending on heatmap size.

To achieve best possible performance, we tuned each network architecture regarding kernel and layer size as well as number of outputs. All networks consist of standard layers (Fig. 2), i.e., convolution, pooling (average), concatenation, addition, multiplication, and fixed deconvolution (linear upsampling). In each network, the final convolution layer has the same number of outputs as landmarks

[1] Digital Hand Atlas Database System, http://www.ipilab.org/BAAweb/.

and a kernel size of 1 without an activation function. All other convolution layers have a ReLU activation function and produce 128 intermediate outputs in 2D and 64 in 3D (except 3D ConvOnly-Net with 32 outputs, due to memory limitations). Additionally, all pooling layers use averaging and halve the output size in every dimension, while all linear upsampling layers double output size. The networks are structured as follows: The 2D ConvOnly-Net consists of 6 convolution layers with 11 × 11 kernels (3D: 6, 5 × 5 × 5). The Downsampling-Net is composed of multiple blocks containing two convolution layers followed by pooling. After the last block, two additional convolution layers are appended. In 2D we use 5 × 5 kernels and 2 downsampling blocks (3D: 3 × 3 × 3, 1 block). The U-Net consists of a contracting path, being equivalent to Downsampling-Net, and an expanding path, consisting of blocks of upsampling, concatenation with the same level output from the contracting path, and finally two convolution layers. In 2D we use 3 × 3 kernels with 4 down- and upsampling blocks (3D: 3 × 3 × 3, 3 blocks). The 2D SpatialConfiguration-Net consists of 3 convolution layers with 5 × 5 kernels, followed by the spatial configuration block, using 15 × 15 kernels with a downsampling factor of $\frac{1}{8}$ (3D: 3 × 3 × 3, 3, and 9 × 9 × 5, factor $\frac{1}{4}$).

We evaluated the 2D dataset with three-fold cross-validation and additionally compared to results obtained with two other state-of-the-art localization methods of Lindner et al. [8], who applied their code to our dataset, and of Ebner et al. [13]. The 3D dataset evaluation used five cross-validation rounds splitting the dataset randomly into 43 training and 17 testing images, respectively, and we also compared our results to Ebner et al. [13]. We additionally evaluated the performance of U-Net and SpatialConfiguration-Net on a dataset with largely reduced number of images, to show the limits of these architectures in terms of required training data. Here, for the same three cross-validation setups as in the original 2D experiment, we used only 10 of the 597 annotated images and tested on the remaining 298. By excessive data augmentation on these 10 images we get to the same number of training images as used in the original experiment.

Results: All networks were trained from scratch using the Caffe framework [14]. We did not fine-tune networks pre-trained on large image databases, as no such

Table 1. Localization results on 2D dataset containing 895 images with 37 landmarks, grouped as full and reduced dataset. #*w* shows the relative number of network weights.

Set	Method	Image height		Localization error (in mm)		#Outliers > 10 mm	#*w*
		Input	Target	Median	Mean ± SD		
Full	Downsampling-Net	256	64	1.85	1.96 ± 1.14	12 (0.036 %)	1.8
	ConvOnly-Net	128	128	1.13	1.29 ± 1.13	**9** (0.027 %)	8.7
	U-Net	256	256	0.68	0.87 ± 1.05	15 (0.045 %)	2.0
	SpatialConf-Net	256	256	0.91	1.13 ± 0.98	12 (0.036 %)	**1.0**
	Lindner et al. [8]	1250	1250	0.64	**0.85 ± 1.01**	20 (0.060 %)	-
	Ebner et al. [13]	1250	1250	**0.51**	0.97 ± 2.45	228 (0.689 %)	-
Red	U-Net	256	256	1.24	3.29 ± 11.78	1175 (3.548 %)	2.0
	SpatialConf-Net	256	256	1.14	1.61 ± 3.43	120 (0.362 %)	**1.0**

Table 2. 3D localization results on 85 images with 28 landmarks per image.

Method	Image height		Localization error (in mm)		#Outliers > 10 mm
	Input	Target	Median	Mean ± SD	
Downsampling-Net	128	64	1.91	2.21 ± 2.82	16 (0.672 %)
ConvOnly-Net	128	128	1.10	8.17 ± 23.62	360 (15.126 %)
U-Net	128	128	**1.01**	**1.18** ± **1.31**	**3** (0.126 %)
SpatialConf-Net	128	128	**1.01**	1.19 ± 1.48	**3** (0.126 %)
Ebner et al. [13]	512	512	1.27	1.44 ± 1.51	6 (0.252 %)

network exists for 3D and converting 2D kernels to 3D is not straightforward. Our networks were optimized using stochastic gradient descent with L2-loss, momentum of 0.99, and a batch size of 5 for 2D and 2 for 3D inputs, respectively. The learning rate was set to 10^{-5} for the ConvOnly- and Downsampling-Nets, and 10^{-6} for the U- and SpatialConfiguration-Nets, with weight decay of 0.0005. The network biases were initialized with 0, the weight values drawn from a Gaussian distribution with a standard deviation of 0.01. Networks were trained until the testing error reached a plateau, which took between 15000 and 40000 iterations, depending on the architecture. We did not observe overfitting to the datasets as also the test error remained at a plateau. Even after decreasing learning rate, results did not improve any further. Training time was similar for all architectures, between 5 and 10 h per cross-validation round on a 12 GB RAM NVidia Geforce TitanX. Testing per image takes below 10 s, with down- and upsampling of in- and output consuming most of the time. Detailed localization results for 2D and 3D datasets are shown in Tables 1 and 2, also comparing performance of CNN architectures with the state-of-the-art.

4 Discussion and Conclusion

Results of our experiments show, that fully convolutional CNN architectures trained in an end-to-end manner are able to achieve state-of-the-art localization performance by regressing heatmaps. Despite using a much lower input and target heatmap resolution, still the best-performing U-Net architecture achieves the same accuracy as the method of Lindner et al. [8] on the 2D dataset, while all architectures have less outliers (see Table 1). On the 3D dataset (see Table 2) with the U-Net and SpatialConfiguration-Net architectures we achieve even better results than the method of Ebner et al. [13]. With medium number of network weights, Downsampling-Net is capable to capture spatial configuration of the landmarks, however, since it involves downsampling, accuracy is worst among the compared architectures. ConvOnly-Net improves the accuracy, nevertheless it requires a high number of network weights, leading to worst performance in terms of outliers when used for 3D data due to memory restrictions preventing large enough kernel sizes. We found that localization performance corresponds with target heatmap size, as emphasized by U-Net and SpatialConfiguration-Net showing the best results among compared architectures. In future work, we plan to also evaluate datasets with sparser landmarks and more spatial variation.

By both accurately describing local appearance and enforcing restrictions in possible spatial landmark configurations, our novel SpatialConfiguration-Net architecture is able to achieve accurate localization performance with a low amount of outliers, despite requiring the lowest number of network weights. While achieving the same result for the 3D dataset, in the 2D experiment, we found that U-net performance is slightly better than SpatialConfiguration-Net, however, U-net requires more network weights. When evaluating SpatialConfiguration-Net on the augmented training dataset where anatomical variation is defined from only 10 input images, it reveals its capability to model spatial landmark configuration, outperforming U-Net significantly in number of outliers. Thus, explicitly encoding spatial landmark configuration as in our novel SpatialConfiguration-Net proves to be a promising strategy for landmark localization in the presence of limited training data, as is usually the case when working with medical images.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
2. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2323 (1998)
3. Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., Nielsen, M.: Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8150, pp. 246–253. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40763-5_31
4. Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M.: 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, Heidelberg (2014). doi:10.1007/978-3-319-10404-1_65
5. Zheng, Y., Liu, D., Georgescu, B., Nguyen, H., Comaniciu, D.: 3D deep learning for efficient and robust landmark detection in volumetric data. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 565–572. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24553-9_69
6. Liu, D., Zhou, K.S., Bernhardt, D., Comaniciu, D.: Search strategies for multiple landmark detection by submodular maximization. In: CVPR, pp. 2831–2838 (2010)
7. Donner, R., Menze, B.H., Bischof, H., Langs, G.: Global localization of 3D anatomical structures by pre-filtered hough forests and discrete optimization. Med. Image Anal. **17**(8), 1304–1314 (2013)
8. Lindner, C., Bromiley, P.A., Ionita, M.C., Cootes, T.: Robust and accurate shape model matching using random forest regression-voting. IEEE Trans. Pattern Anal. Mach. Intell. **37**, 1862–1874 (2015)
9. Pfister, T., Charles, J., Zisserman, A.: Flowing convnets for human pose estimation in videos. In: ICCV, pp. 1913–1921 (2015)

10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)
11. Liu, Z., Li, X., Luo, P., Change, C., Tang, L.X.: Semantic image segmentation via deep parsing network. In: ICCV, pp. 1377–1385 (2015)
12. 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, Heidelberg (2015). doi:10.1007/978-3-319-24574-4_28
13. Ebner, T., Stern, D., Donner, R., Bischof, H., Urschler, M.: Towards automatic bone age estimation from MRI: localization of 3D anatomical landmarks. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8674, pp. 421–428. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10470-6_53
14. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of ACM International Conference on Multimedia (MM 2014), pp. 675–678 (2014)