Anatomical landmark detection in medical applications driven by synthetic data

Gernot Riegler[1] Martin Urschler[2] Matthias Rüther[1] Horst Bischof[1] Darko Stern[1]
[1]Graz University of Technology [2]Ludwig Boltzmann Institute for Clinical Forensic Imaging
{riegler, ruether, bischof, stern}@icg.tugraz.at martin.urschler@cfi.lbg.ac.at

Abstract

An important initial step in many medical image analysis applications is the accurate detection of anatomical landmarks. Most successful methods for this task rely on data-driven machine learning algorithms. However, modern machine learning techniques, e.g. convolutional neural networks, need a large corpus of training data, which is often an unrealistic setting for medical datasets. In this work, we investigate how to adapt synthetic image datasets from other computer vision tasks to overcome the under-representation of the anatomical pose and shape variations in medical image datasets. We transform both data domains to a common one in such a way that a convolutional neural network can be trained on the larger synthetic image dataset and fine-tuned on the smaller medical image dataset. Our evaluations on data of MR hand and whole body CT images demonstrate that this approach improves the detection results compared to training a convolutional neural network only on the medical data. The proposed approach may also be usable in other medical applications, where training data is scarce.

1. Introduction

Reliable anatomical landmark detection is an important first step for many medical image algorithms. Since manual labeling of anatomical locations is often tedious and time consuming to be performed by medical experts, especially in three-dimensional (3D) computer tomography (CT) and magnetic resonance (MR) images, different automatic methods were developed for localization of anatomical structures in medical data [3]. Most reliable landmark detection algorithms are based on machine learning approaches that are highly dependable on the quality of the training data [4, 5]. Unlike for other computer vision applications, where images are "cheap" to obtain with e.g. a personal photo camera, or available on the Internet for free, building a representative medical image database is expensive and a time consuming process. Thus, acquisition of a single 3D MR or CT image may last more than half an hour on a highly expensive scanner and is only available to specially qualified personnel in the facilities designated to provide such a service. Under-representation in the medical image datasets is also caused by large variation of the human anatomy. Increasing the number of seen anatomical structures could be made by scanning healthy volunteers, but in case of e.g. radiographic or CT scans, it is forbidden in most countries due to the harmful ionizing radiation. The patient privacy rights are another important issue that have to be taken into account when creating a publicly available dataset. To make datasets available to institutions without clinical imaging systems, the patient has to give its written consent and his personal information has to be hidden before making the image publicly available.

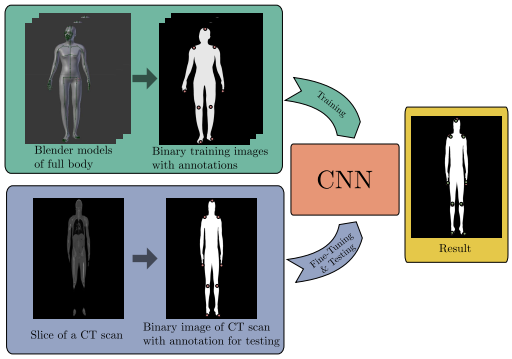

Figure 1. Adaptation of body pose estimation data for anatomical landmark detection in medical data. We transform both image domains to a common domain, *i.e.* a binarization of the data. Then we can use the large training datasets to optimize the parameters of a CNN. Finally, we fine-tune and evaluate the method on binarized MR or CT data.

2. Anatomical Landmark Detection

To significantly increase the size of training datasets without the additional costs of collecting medical images, synthetic images have been introduced into learning paradigms [6]. Used to train machine learning algorithms, the large quantity of synthetic images can be quickly generated and made freely available. An additional advantage of synthetic images that make them very attractive for the medical community is having an exact knowledge of the ground-

Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6	Layer 7	Layer 8
conv5-32	maxpool2	conv5-32	maxpool2	conv3-32	fc2048	fc2048	fc$k$

Table 1. Network architecture for anatomical landmark detection for hand and whole body images. conv$w$-c stands for a convolutional layer with filter size $w \times w$ and c output channels, maxpool$p$ for max-pooling with pooling-width p, and fc$o$ for a fully-connected layer with o outputs. The number of outputs in the last layer depends on the number of anatomical landmarks.

truth positions of the represented structure. Nevertheless, capturing the position and variation of under-investigated anatomies, as well as its image intensity variation in different image modalities still remains the main challenge in building realistic synthetic images.

When talking about detection of human or hand pose, synthetic images have been successfully applied in computer vision applications. In [18] and [19] the authors used video-game images to train a holistic and a deformable part-based model, respectively. Most famous is the work by Shotton *et al*. [12], where the authors train a Random Forest with a huge amount of synthetic and real training data for the purpose of human pose estimation. Recently, such data-driven approaches got also popular for articulated hand pose estimation. Tang *et al*. [14] and Tompson *et al*. [16] utilized a semi-automatic method to label a corpus of real depth-sensor data to train their models. In contrast, Riegler *et al*. [11] showed that convolution neural networks (CNN) [9] can also successfully be trained on synthetic depth images of the hand for articulated hand pose estimation.

The latter work was our inspiration to use synthetic images to overcome the limitation in the number of images, especially labeled ones that are used in training an algorithm for detection of body and hand pose in medical images. Similarly as in [11], we used a MakeHuman [1] model and Blender [2] to generate a depth image training data set, while casting the 3D medical image, such as MR or CT data, to a depth image is straight forward. However, we can further simplify the problem of mapping the different data domains in training and testing by thresholding the image data for both domains to obtain binary image data. With this simple processing step, we can learn anatomical landmark estimators for medical images from synthetic data that belong to a different training domain (see Fig. 1). Following the idea in [20] we further adapt the trained CNN to the medical image data set by performing a fine-tuning of the network on a subset of the medical data.

2.1. Data Generation

The success of state-of-the-art CNN based approaches for computer vision tasks depends on the depth of the CNN (*i.e.* the number of layers) [13] and the amount of training data to avoid over-fitting. In the original work of [11] a Blender model of a hand is articulated to sample from a large space of different poses. For each pose a depth image is rendered where the depth is measured relative to a predefined camera and also the perfect ground-truth annotation can be obtained in 3D. This avoids the cumbersome and error-prone manual annotation process.

We modified the publicly available model in two ways to fit our needs: First, we created new pose and shape spaces for the anatomical landmark detection in hand and whole body images. The key difference is that we need less variation in the articulation, but much more shape variation, especially for the whole body images. This is due to the fact that the pose of the hand and body is relatively fixed during a medical scan, but the shape of different people can vary significantly, especially for the whole body scans (due to differences in *e.g.* gender, weight and size). The second modification to the model is the rendering itself, as we render the binary image of the object along with the depth image. We expect that the silhouette contains already sufficient shape information for the detection of the anatomical landmarks of interest in 2D. For the depth coordinate we can train the regression model on a common depth domain, as we will show in the experiments. Alternatively, we could determine the depth information later with *i.e.* the mean position over the data set, because the depth coordinate varies the least in 3D medical images In Fig. 1 we visualize one example of our binary renderings for the whole body.

By marking the position in 2D binary image where image intensity value is above the air threshold along the depth direction of the 3D medical image, the 3D medical images were mapped to the same domain as the 2D synthetic images in the training data set. Represented in 3D CT image as two thin plates normal to the depth direction, CT scanner tables were eliminated while generating the 2D binary images by accumulating the number of high image intensity values along the depth direction. Namely, the body in a 3D CT image has more constant high intensity values than the scanner table in the depth direction.

2.2. Inference Method

Provided with enough data, we can train a CNN for anatomical landmark detection. We follow recent work in body pose estimation [15, 17] and hand pose estimation [11, 16] and formulate the task as a regression problem. For each training image s_n, we have an associated annotation vector $\mathbf{y}_n = (x_1, y_1, \ldots, x_k, y_k)^T$ that contains the k ground-truth landmark locations. To aid generalization, we scale the input image of the hand or whole body to a unit size of $p_h \times p_w$. Additionally, we also transform the annotation vector similar to [17], such that the coordinates are in the interval of $[-1, 1]$:

$$\mathbf{y}^p = \left(\frac{x_1 - \frac{p_w}{2}}{p_w}, \frac{y_1 - \frac{p_h}{2}}{p_h}, \ldots \right)^T \quad . \tag{1}$$

We train a network ϕ with parameters $\mathbf{w}$ using stochastic gradient descent with a mini-batch size of 100 that min-

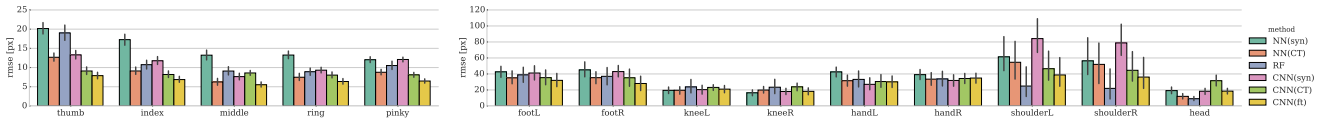

Figure 2. Anatomical landmark detection results on MR hand data (left) and CT scans of the whole body (right). We report the mean and standard deviation of the root mean squared error (RMSE) between estimated and ground-truth anatomical landmark location in pixel (*px*). The different colors depict the various methods.

imizes the L_2 loss over the N training samples:

$$\mathbf{w} = \arg\min_{\mathbf{w}} \frac{1}{2} \sum_{n=1}^{N} ||\phi(s_n; \mathbf{w}) - \mathbf{y}_n^P||^2 \quad . \tag{2}$$

The architecture of the network ϕ is listed in Table 1. We use a *ReLU* [10] as activation after each convolutional and fully-connected layer. Further, in the fully-connected layers we use Dropout [8] to regularize the network.

Although we created the synthetic dataset with attention to the pose variations in the MR data, we observed that fine-tuning the model [7, 20] gives a significant performance improvement. Therefore, we use the small subset of binarized MR/CT data to adapt the network weights to the original domain, *i.e.* we use a new set of training data and re-train the network on it. This follows the intuition that the synthetic data is used to learn a rich feature representation and the real data is used to re-parameterize the output space.

3. Experimental Set-Up

Datasets We evaluated our approach on two different medical datasets, one of left hands from MR scans and one of whole body CT scans. The left hand T1-weighted 3D gradient echo MR images consists of scans from $N = 132$ healthy male volunteers with an age between 13 and 24 years. The volunteers were asked to put their hand into the MR scanner with slightly spread fingers and a half-kilo bag of sand was placed on the top of their hand. In most of the images, contraction of the fingers could be ignored, nevertheless the distances between the fingers, especially to the thumb, varies among the volunteers. In the hand dataset, five landmarks are located at the top the fingers.

The whole body dataset consists of $N = 20$ CT images of cancer patients, of which 14 are female and 6 are male. All patients are in a lying position with hands close to the body and have slightly spread legs. Legs have the highest variation in articulation and also the feet could be in an arbitrary position. The nine landmarks estimated in this dataset are located at the top of the head, and at the left and right shoulder, middle finger, knee and big toe, respectively.

Anatomical Landmark Detection of the Hand For the landmark detection in MR images of the left hand we generated around 2.2M synthetic images. We scale the images to a patch size of $p_w = p_h = 96$ and transform the annotations accordingly such that they are in the range of $[-1, 1]$.

We train the network for 30 epochs with a learning rate of 0.01 and a momentum term of 0.9. The learning rate is decreased after each epoch by a factor of 0.7943. After the training on the synthetic data, we fine-tune the network on the binary MR images by performing a three-fold cross-validation. Therefore, we train the network for 700 epochs on this data (note that the number of training instances is much smaller for fine-tuning). The learning rate and momentum term stay the same, but instead of a continuous decrease of the learning rate, we decrease the learning rate once after 500 iterations by a factor of 0.1.

Anatomical Landmark Detection of the Whole Body For the landmark detection in CT images of the whole body we generated $600,000$ synthetic images, where the human model varies in age, gender, size, body weight and pose. We scale the images to a patch size with $p_w = 76, p_h = 160$ and transform the annotations accordingly such that they are in the range of $[-1, 1]$.

The network is also trained for 30 epochs with a learning rate of 0.01 and a momentum term of 0.9. After each epoch we decrease the learning rate by a factor of 0.7943. After the training on the synthetic data, we again fine-tune the network on the binarized CT data by performing a leave-one-out cross-validation. Therefore, we train the network for 700 epochs on 19 out of the 20 images and keep the remaining one for testing. The learning rate and momentum term stay the same, but instead a continuous decrease of the learning rate, we decrease the learning rate once after 500 iterations by a factor of 0.1.

Infer Depth of Anatomical Landmark Detections On binarized input data one can only expect to determine the 2D location of anatomical landmarks. However, the variation in the depth coordinate can be expected to be rather small, because the hand for example is always in the same resting position when scanned in the MR. In fact, if we analyze the depth of the individual landmarks for the hand dataset, we observe a standard deviation from the mean depth in the range from 3.137 voxels to 3.183 voxels for the pinky and the thumb, respectively.

To refine these results, we can train a CNN on normalized depth images, instead of binarized images. We normalize the depth values of pixels that belong to the hand to a zero mean and a unit standard deviation and set the back-

ground pixels to 3. In the same way we normalize the depth annotation for the training samples. We do this for the synthetic data to pre-train the CNN, as well as for the MR data to fine-tune the network. The rest of the training follows the same steps as described in the previous experiments.

4. Results and Discussion

The results of the MR hand landmark detection are depicted in Fig. 2. We compute for each landmark of the 263 images the root mean squared error (*rmse*). The plot shows the mean and standard deviation of each landmark for all methods. We compare the fine-tuned network (*CNN(ft)*) to the network trained solely on the synthetic data (*CNN(syn)*) and the binarized MR data (*CNN(MR)*). Further, we use a nearest neighbor approach as baseline, where we compute the location estimates by taking the mean annotation of the three nearest patches. We denote the methods as *NN(syn)* and *NN(MR)* if the nearest neighbour search is performed on the synthetic data or the MR data, respectively. Finally, we compare our method to the random regression forest *RF* based state-of-the-art approach of [3].

First, we can observe that the *NN(syn)* approach performs much worse than *NN(MR)*. We assume that this is due to the modeling of the synthetic data. Although we took care to replicate the poses and shapes of the MR data, they are still far off. The same explanation might hold for the jump from *CNN(syn)* to *CNN(ft)*. However, the fine-tuning is still better than training the network on binary MR images only. The reason for the improvement is that the CNN needs a lot of training data to tune all parameters, which is performed with the synthetic data. The output parameters are then adjusted with the fine-tuning of the network. Finally, the method based on the random regression forest [3] is inferior to the *NN(MR)*, *CNN(MR)* and *CNN(ft)* methods. This is especially emphasized on the landmark with the most variation, the thumb. Our explanation is that the *RF* has the tendency to regress towards the mean anatomical landmark locations.

We visualize the results of the CT whole body cross-validation in Fig. 2. The results are similar, but not as pronounced as in the first evaluation. The *NN(syn)* is not as good as the *NN(CT)*. However, the *NN(CT)* fails for the shoulder and head landmarks. This may be due to the small dataset. What remains as an open question is why the *RF* outperforms other approaches for the same three landmarks.

Qualitative results of our approach are depicted in Fig. 3. One can observe for the whole body estimates that for the landmarks left and right shoulder and hand the CNN learns a certain offset from the ground-truth location that causes the most increase in the error. Additionally, palms and fingers are often not clearly visible in the images due to movement artifacts. For the other landmarks, the detections are centered around the ground-truth location.

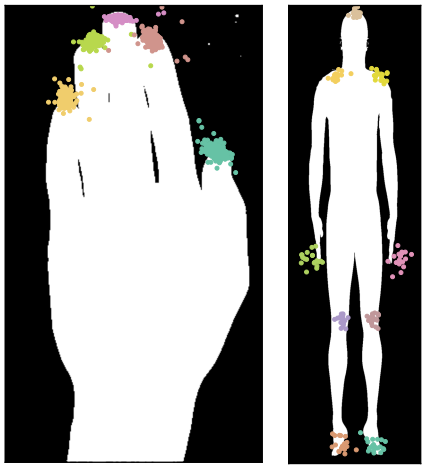

Figure 3. Qualitative results for binarized MR hand (left) and CT whole body (right) images. The estimated anatomical landmark localizations are projected into one common image.

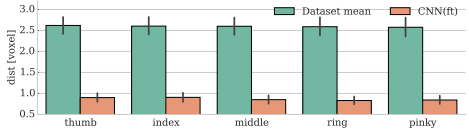

Figure 4. Mean error of the estimated depth coordinate if we just use the mean depth of the training samples in green, or if we train a CNN on the depth representation of the data.

The results for the depth estimations for the hand dataset are depicted in Fig. 4. We compare the mean absolute error of the estimated depth to the ground-truth for all five fingertips. In the first method we simple take the mean depth of all training MR annotations. As expected we get a mean error of ~ 3 voxels, which correspondence to the ~ 3 voxels standard deviation present in the dataset, as we stated in the previous section. However, the CNN trained on the depth representation of the data can reduce the mean error to 1 voxel in the mean. We speculate that this can mainly attributed to the consideration of the mean depth of the hand.

5. Conclusion

We demonstrated in this work that we can facilitate the training of state-of-the-art machine learning methods for medical landmark localisation by exploiting larger datasets from a different domain. This is achieved by transforming the training data from a different training domain and the one from medical MR and CT images to one common domain, *e.g.* thresholding the depth values and MR/CT volume data to create binary images. Such a simple transformation is already sufficient to train a feature representation of a CNN. In future work we want to explore more complex common domains, so we can train even richer feature representations (*e.g.* deeper CNNs).

Acknowledgment: This work was supported by the Austrian Science Fund (FWF): P 28078-N33 (FAME)

References

[1] M. Bastioni, S. Re, and S. Misra. Ideas and Methods for Modeling 3D Human Figures: The Principal Algorithms Used by MakeHuman and Their Implementation in a New Approach to Parametric Modeling. In *COMPUTE*, 2008.

[2] Blender Online Community. *Blender - A 3D Modelling and Rendering Package*. Blender Foundation, 2015.

[3] A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu. Regression Forests for Efficient Anatomy Detection and Localization in CT Studies. In *MICCAIW*, 2011.

[4] R. Donner, B. H. Menze, H. Bischof, and G. Langs. Global localization of 3D anatomical structures by pre-filtered hough forests and discrete optimization. *Med Image Anal*, 17(8):1304–1314, 2013.

[5] T. Ebner, D. Stern, H. Bischof, and M. Urschler. Towards automatic bone age estimation from MRI: Localization of 3d anatomical landmarks. In *MICCAI*. 2014.

[6] A. C. Evans, A. L. Janke, L. D. Collins, and S. Baillet. Brain templates and atlases. *Neuroimage*, 62(2):911–922, Aug 2012.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[8] G. E. Hinton, S. Nitish, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[10] V. Nair and G. E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *ICML*, 2010.

[11] G. Riegler, D. Ferstl, M. Ruether, and H. Bischof. A Framework for Articulated Hand Pose Estimation and Evaluation. In *SCIA*, 2015.

[12] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time Human Pose Recognition in Parts from Single Depth Images. In *CVPR*, 2011.

[13] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556, 2014.

[14] D. Tang, H. J. Chang, A. Tejani, and T.-K. Kim. Latent Regression Forest: Structured Estimation of 3D Articulated Hand Posture. In *CVPR*, 2014.

[15] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. 2014.

[16] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks. *TOG*, 33(5):169, 2014.

[17] A. Toshev and C. Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks. In *CVPR*, 2014.

[18] D. Vazquez, A. M. Lopez, J. Marin, D. Ponsa, and D. Geronimo. Virtual and Real World Adaptation for Pedestrian Detection. 36(4):797–809, 2014.

[19] J. Xu, S. Ramos, D. Vazquez, and A. M. Lopez. Domain Adaptation of Deformable Part-Based Models. 36(12):2367–2380, 2014.

[20] J. Xu, D. Vazquez, A. M. Lopez, J. Marin, and D. Ponsa. Learning a Part-based Pedestrian Detector in a Virtual World. *Intelligent Transportation Systems, IEEE Transactions on*, 15(5):2121–2131, 2014.