A Cascade Regression Model for Anatomical Landmark Detection

Zimeng Tan[1,2,3], Yongjie Duan[1,2,3], Ziyi Wu[1,2,3], Jianjiang Feng[1,2,3](✉), and Jie Zhou[1,2,3]

[1] Department of Automation, Tsinghua University, Beijing, China
jfeng@tsinghua.edu.cn
[2] State Key Lab of Intelligent Technologies and Systems, Tsinghua University, Beijing, China
[3] Beijing National Research Center for Information Science and Technology, Beijing, China

Abstract. Automatic anatomical landmark detection is beneficial to many other medical image analysis tasks. In this paper, we propose a two-stage cascade regression model to make coarse-to-fine landmark detection. Specifically, in the first stage, a Gaussian heatmap regression model customized from U-Net is exploited to make primary prediction, which takes the downsampled entire image as input. In the second stage, we develop a CNN to regress displacements from the primary prediction to the landmarks, using patches in original resolution centered at the previous localization as input. Owing to the different sizes and resolutions of inputs in two stages, the global context information and local appearance can be integrated by our algorithm. The spacial relationships among landmarks can also be exploited by predicting all the landmarks simultaneously. In evaluation on the coronary and aorta CTA images, we show that our proposed method is widely applicable and delivers state-of-the-art performance even with limited training data.

Keywords: Anatomical landmark detection · Heatmap regression · Cascade model

1 Introduction

Anatomical landmark detection plays an important assisted role in many medical image analysis tasks, such as organ segmentation, registration and vessel extraction [1]. However, for accurate landmark detection, there still remain many challenges: (a) anatomical differences between patients are widespread, (b) while detecting multiple landmarks simultaneously, spatial constrains among landmarks should be taken into account, (c) detection of 3D anatomical landmarks aggravates the computational cost intensively, making real-time application challenging, (d) limited annotated training data available restricts algorithmic design typically. Although many methods have been proposed [2–5], there is still room for improvement. Among these methods, our method is more related to [3,4].

M. Pop et al. (Eds.): STACOM 2019, LNCS 12009, pp. 43–51, 2020.
https://doi.org/10.1007/978-3-030-39074-7_5

For landmark detection, an intuitive patch-based approach is to regress displacements from patches center to the target landmark [3]. Then the landmark position is calculated by these displacements following a majority/average voting strategy. Trained by numerous patches, it is possible to design deep networks which can capture discriminative information and perform better than the shallow ones. Nonetheless, these methods always focus on local appearance merely and global information is not well utilized. The large number of patches also leads to a heavy computational burden. For improvement, Noothout et al. [6] proposed a model performing classification and regression jointly, in which only displacements of patches classified as containing landmarks contributed to the final result.

Another interesting method is based on regressing heatmaps [4]. With entire image as input, these models are supposed to output synthetic heatmap, denoting the probability of each voxel belonging to the target landmark. The prediction position is simply chosen to be the output voxel with the maximum temperature. Apparently, they can utilize global context information and have good spatial generalization. However, the input volume shrinks in methods using FCN [7], which causes theoretical lower bound of prediction error. For instance, output heatmap of size 128 with input of size 512 leads to 3 voxels error at most. Furthermore, the total number of network weights for 3D medical images increases intensively, making the training difficult with limited training data at hand.

Combining the advantages of the two methods above, we propose a cascade regression model combining heatmap regression and displacement regression. The proposed method makes coarse-to-fine prediction, taking entire image in lower resolution and patches in higher resolution as input respectively, which combines global information and local appearance. The spatial relationships among landmarks are also taken into account by learning long-range context, which improves overall performance. The cascade structure is similar to the method of He et al. [9], in which the facial landmark localizations were refined via finer and finer modeling. In contrast, instead of the deep CNN, a carefully designed heatmap regression model is exploited to make initial prediction in our method. Besides, the local patches are extracted as input in the subsequent stage [10], rather than entire image in [9].

We evaluated our method on the coronary and aorta CTA images by detecting 5 and 9 anatomical landmarks respectively. These landmarks are of great clinical significance: cardiac landmarks contribute to diagnosis, prognosis, and therapy of cardiovascular diseases [1]; detection of aortic landmarks is an effective assistant tool in aortic vascular modeling [6]. The results demonstrate our method is competent for the cardiac and aortic landmark detection task and achieves performance comparable to the state-of-the-art approach [6].

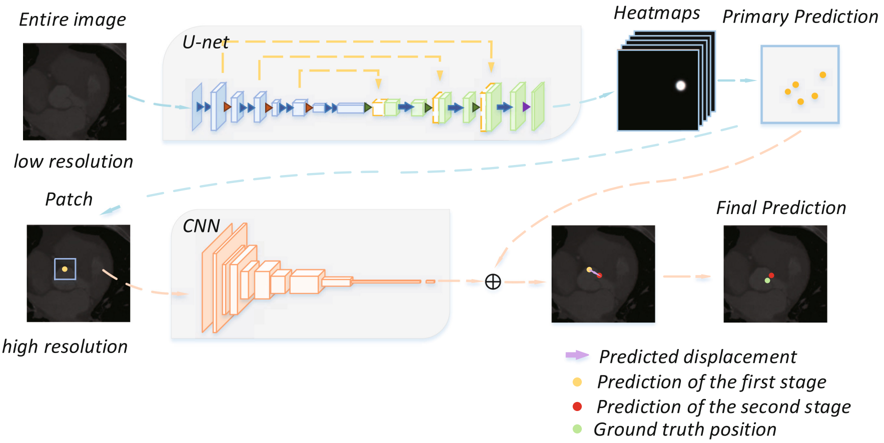


Fig. 1. The overview of our cascade regression model.

2 Proposed Method

Figure 1 illustrates the overall cascade regression model framework for single landmark detection. We show the 2D case for clarity but the model works similarly in 3D. In the first stage, a modified U-Net is employed to get a relatively accurate initial localization, taking the entire image in lower resolution as input and heatmaps as output. Owing to the skip architecture, this module can capture multi-scale knowledge. Aiming to learn more precise context information, in the second stage, the patch centered at initial localization in higher resolution is extracted and fed to the displacement regression model. The CNN adjusts the initial localization by moving it toward ground truth position. The different sizes and resolutions of two stages emphasize that they focus on long-range context and local appearance, respectively.

2.1 Primary Prediction

We exploit heatmap regression to make the first stage prediction. In this scheme, each landmark has a separate output channel where a Gaussian heat spot is centered at its location. During inference, the predicted position is simply determined by the maximum response. Following the principle of classification, for N_l landmarks, the model is trained for $N_l + 1$ channels, where the first N_l channels describe the probability belonging to the corresponding landmark and the last channel belonging to background. Particularly, considering that softmax operation may influence the status of landmark positions in heatmap ground truth (e.g. for 5 landmarks, the values of 1th landmark in 6 channels are changed from $(1, 0, ..., 0)$ to $(0.35, 0.13, ..., 0.13)$ after softmax, which can be smaller than its neighbors), we adjust the sum of all channels to 1 by fixing the background channel and scaling the others.

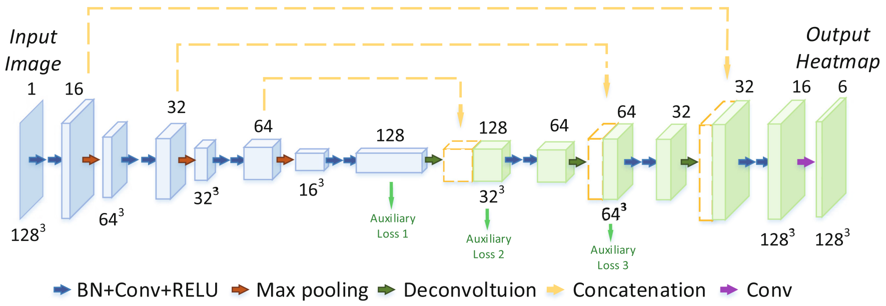


Fig. 2. The architecture of the proposed model in the first stage.

The temperature t_i for ith landmark (i.e. ith channel) can be defined as:

$$f(x) = \begin{cases} k\exp(\frac{-(v-p_i)^2}{2\sigma^2}), & i = 1, 2, 3, ..., N_l, \\ 1 - k\exp(\frac{-(v-p_{closest})^2}{2\sigma^2}), & i = N_l + 1. \end{cases} \tag{1}$$

The heatmaps of first N_l channels are determined by the distance from the voxel v to the landmark position p_i, while the heatmap of background channel is according to the closest landmark position $p_{closest}$. σ is standard deviation and k is Gaussian height.

As shown in Fig. 2, our model realizes this scheme by customizing the original 3D U-Net [8]. Similar to its standard version, the network is comprised of 3D convolution, max-pooling, deconvolution (up sampling) and short-cut connections from layers in contracting path to the ones in expansive path with equal resolution. Each convolution layer follows 'same mode' (i.e. ouput has the same size as input) and uses RELU activation function. The model takes entire downsampled image as input and outputs heatmap volumes. Benefiting from the natural superiority of U-Net, the model can capture long-range context information, where the spatial relationships among landmarks can also be taken into account, increasing overall accuracy.

Aiming to tackle the problem of class imbalance, namely heat spot only occupies a small proportion of volume, we employ a weighted mean squared error (MSE) loss function between the predicted and ground truth heatmaps. The weights are chosen to be the exponential powers of the predicted values in the output. On the other hand, to deal with gradient vanishing problem, we shorten the backpropagation path of gradient flow signals by incorporating three side-paths auxiliary loss. The final formulation of loss function is expressed as:

$$\mathcal{L}(P; H^{GT}) = \mathcal{L}_{mse}(P; H^{GT}) + \sum_{s=1,2,3} \beta_s \mathcal{L}^s_{mse}(p^s; H^{GT}) \tag{2}$$

where H^{GT} is the ground truth heatmap, P is the final output, β_s is the weight of different side-path p^s and set as 0.3, 0.6, 0.9 corresponding s as 1, 2, 3.

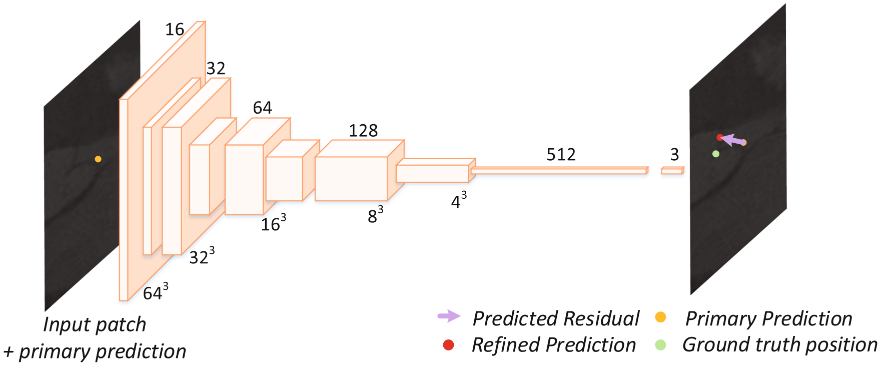


Fig. 3. The architecture of the displacement regression module in the second stage.

2.2 Refinement Strategy

In the second stage, we propose a CNN model to refine the primary prediction. Given the first stage model taking the entire image as input, we assume that landmarks should be distributed around the initial prediction. The CNN takes patches in original resolution centered at the inital prediction to capture more precise local information. Considering that local appearance of certain landmarks may be ambiguous (e.g. locally similar vascular structures), we restrict this stage model to change the initial prediction in a small range.

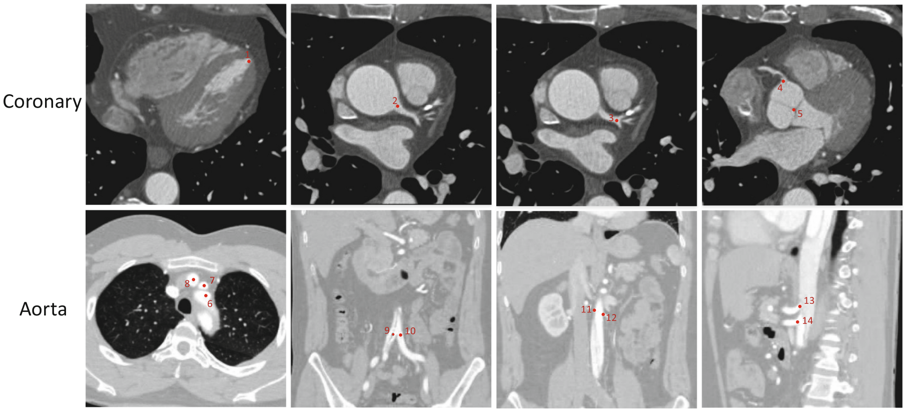
The CNN is trained to predict the displacement vector $\triangle S$ from the primary prediction S_0 to the true landmark position S^{GT}. Given a volume V, a training sample is represented by $(\varGamma(V,q), \triangle S^{GT})$ where q is a point randomly sampled around S_0 in a small range from V and $\varGamma(V,q)$ is its associated patch. The ground truth displace vector $\triangle S^{GT}$ is given by $\triangle S^{GT} = S^{GT} - S_0$. During inference, patch $\varGamma(V,S_0)$ is fed to the model and the final prediction is obtained by $S = S_0 + \triangle S$. The CNN is trained by minimising Euclidean loss between the predicted and the true displacement vector.

As shown in Fig. 3, the CNN model contains 4 convolutional layers followed by max-pooling layers, and 2 fully-connected layers. Each layer except the last one employs RELU activation function. Considering that certain landmarks may have distinct appearance than the others (e.g. the apex cordis), we refine them separately. That is, we train a refinement network per landmark. Since the CNN is trained by patches, a small number of training data is sufficient in this stage.

3 Experiments and Results

3.1 Data and Experiment Settings

We evaluated the proposed method on the two datasets of coronary and aorta CTA images. As shown in Figs. 4 and 5 cardiac landmarks and 9 aortic landmarks



Cardiac: 1. apex cordis 2. left coronary ostium 3. LM Bifurcation
4. right coronary ostium 5. origin of the non-coronary aortic valve commissure
Aorta: 6. Left subclavian artery 7. Left common carotid artery 8. brachiocephalic trunk
9. left Iliac artery 10. right Iliac artery 11/12. left/right renal artery
13/14. superior/inferior mesenteric artery

Fig. 4. Landmarks defined on the coronary and aorta CTA images.

are annotated manually by a expert. For both datasets, we do not apply data augmentation such as scaling and rotation, which may increase the complexity of landmark distribution.

Coronary dataset is randomly divided into training data with 75 scans and test data with 40 scans. All volumes were zero-padded to $512 \times 512 \times 512$ voxels with isotropic voxel size 0.4 mm. Then they were downsampled 4 times and fed into the model in the first stage. In the second stage, patches size 64 in the original resolution were extracted and the batch size was set to 4. The model was trained using Adam with a learning rate of 0.001 for 11,250 and 45,000 iterations in the two stages, respectively.

Aorta dataset consists of training data with 25 scans and test data with 23 scans. which has an average size of $512 \times 512 \times 777$ voxels, with a voxel size of $0.71 \times 0.71 \times 0.81\,\text{mm}^3$. The annotated landmarks are located at the bifurcation of the aorta and its main branches. Considering that aortic landmark detection is more challenging due to its low resolution and complex organ distribution, the volumes were manually cropped first and downsampled 2 times to fed into the first stage model. The rest of the training process is similar.

3.2 Results

Summary metrics obtained by different networks on the coronary dataset are listed in Table 1. We use average Euclidean distance between ground truth and estimated landmark positions as evaluation measure. We first compared two-stage cascade model and only the first stage model. After refinement, the detection accuracy improves significantly, demonstrating the benefit of our cascade architecture.

Table 1. Average Euclidean distance errors expressed in mm, for the detection of 5 cardiac landmarks on the coronary dataset. The results are obtained by the two-stage model and the first stage model only, which takes either patches or entire image as input, comparing with the algorithms of Noothout et al. [6].

	1	2	3	4	5	Mean
First Stage						
Patch-based	3.14	1.45	8.43	17.12	1.12	6.25
Entire image	3.18	4.79	3.23	6.17	2.60	3.99
Two-stage model						
Our proposed	**2.58**	**1.48**	**1.37**	**3.51**	**1.46**	**2.08**
Noothout et al. [6]	-	2.88	2.19	3.78	2.10	-

Table 2. Average with standard deviation Euclidean distance errors in mm for the detection of 14 landmarks on the two datasets by the proposed algorithm.

Landmarks	1	2	3	4	5					Overall
Cardiac	2.58±1.33	1.48±1.87	1.37±0.84	3.51 ±2.1	1.46±0.86					2.08±1.71
Landmarks	6	7	8	9	10	11	12	13	14	Overall
Aortic	5.62±3.53	7.43±4.72	6.38±4.34	5.60±4.71	7.38±4.17	6.22±3.68	7.82±5.99	4.40±3.35	4.39±1.57	6.14±4.33

To demonstrate that integrating spatial relationships among landmarks can improve overall performance, we adjusted the model in the first stage to take patches size of 48 as input instead of entire image. In this way, the network can only utilize the context information around one landmark at a time. It was trained to predict heatmap patch according to the input. The predicted position was determined by the maximum response in the volume composed of predicted patches. The experiment results show our method in the first stage performs better overall. Specifically, the patch-based network is superior in detecting the left coronary ostium and the origin of the non-coronary aortic valve commissure, which may be more dependent on precise context information. On the other hand, our proposed model performs much better in detecting the right coronary ostium and the bifurcation of the LM, where the relationships among landmarks are probably necessary for accurate detection (e.g. the position of the left coronary ostium is important for localizing the bifurcation of the LM).

Furthermore, we compared our model with the method of Noothout et al. [6], which detected 6 anatomical landmarks in cardiac CT scans (4 of them are the same as us). The metrics are quoted directly from [6] since that dataset is not publicly available. Although our dataset is different from that in [6], we can conclude that the performance of the proposed algorithm is at least comparable to [6].

Table 2 lists more detailed metrics of detection for each landmark on the two datasets using our algorithm. The high detection error of aortic landmarks is due to the low resolution of aortic images. In the model design, we do not utilize unique atlas information related to coronary or aorta, which guarantees the method capable for the anatomical landmark detection tasks in different regions of the human body. Some visual results are shown in Fig. 5.

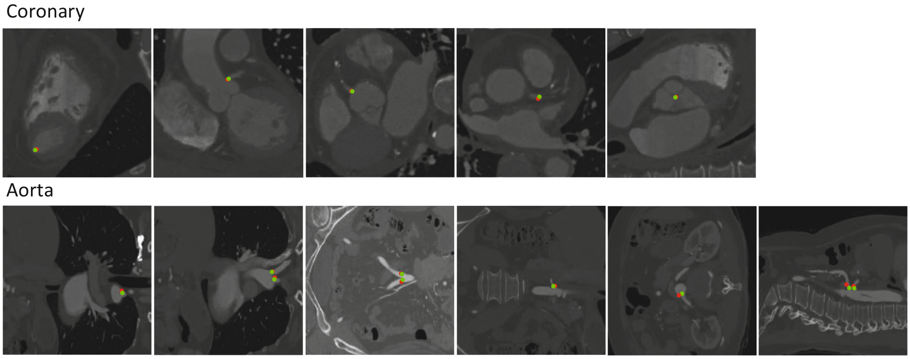


Fig. 5. Visualisation of landmark detection in coronary and aorta images by the cascade regression model. The ground truth and predictions are indicated by green and red dots, respectively. (Color figure online)

4 Conclusion

We have proposed a two-stage cascade regression model for detecting anatomical landmarks in coronary and aorta CTA images. Owing to different sizes and resolutions of input in two stages, the model combines the global information and local appearance. By learning long-range context, the spatial relationships among landmarks are also taken into account, increasing overall performance. The experiment results demonstrate that our method achieved performance comparable to the state-of-the-art algorithm [6]. Limited by memory and computation time, we used downsampled image in the first stage. It is foreseeable that the model would gain better performance with images of higher resolution as input. Another limitation is we only have one annotator, which makes it impossible to assess inter-observer error for landmarks. It is also worthwhile to apply multi-stage refinement to capture more precise information. The experiment results have demonstrated that our method is generic for anatomical landmarks detection and the next step is to extend it to other medical images.

References

1. Zhou, S.K.: Discriminative anatomy detection: classification vs regression. Pattern Recogn. Lett. **43**, 25–38 (2014)
2. Yang, D., et al.: Automated anatomical landmark detection ondistal femur surface using convolutional neural network. In: IEEE ISBI (2015)
3. Gao, Y., Shen, D.: Context-aware anatomical landmark detection: application to deformable model initialization in prostate CT images. In: Wu, G., Zhang, D., Zhou, L. (eds.) MLMI 2014. LNCS, vol. 8679, pp. 165–173. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10581-9_21

4. Payer, C., Štern, D., Bischof, H., Urschler, M.: Regressing heatmaps for multiple landmark localization using CNNs. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 230–238. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_27
5. Ghesu, F.C., Georgescu, B., Mansi, T., Neumann, D., Hornegger, J., Comaniciu, D.: An artificial agent for anatomical landmark detection in medical images. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9902, pp. 229–237. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46726-9_27
6. Noothout, J.M.H., de Vos, B.D., Wolterink, J.M., Leiner, T., Isgum, I.: CNN-based landmark detection in cardiac CTA scans. In: MIDL (2018)
7. O'Neil, A.Q., et al.: Attaining human-level performance with atlas location autocontext for anatomical landmark detection in 3D CT data. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11131, pp. 470–484. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11015-4_34
8. 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, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
9. He, Z., Kan, M., Zhang, J., Chen, X., Shan, S.: A fully end-to-end cascaded CNN for facial landmark detection. In: IEEE FG (2017)
10. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: CVPR (2013)