2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Anatomical Landmark Detection using Deep Appearance-Context Network

Pavan Kumar Reddy, Aparna Kanakatte, Jayavardhana Gubbi, Murali Poduval,
Avik Ghose, Balamuralidhar Purushothaman
Embedded Devices and Intelligent Systems, TCS Research, Bangalore

Abstract— **Accurate identification of anatomical landmarks is a crucial step in medical image analysis. While deep neural networks have shown impressive performance on computer vision tasks, they rely on a large amount of data, which is often not available. In this work, we propose an attention-driven end-to-end deep learning architecture, which learns the local appearance and global context separately that helps in stable training under limited data. The experiments conducted demonstrate the effectiveness of the proposed approach with impressive results in localizing landmarks when evaluated on cephalometric and spine X-ray image data. The predicted landmarks are further utilized in biomedical applications to demonstrate the impact.**

I. Introduction

Anatomical landmarks are distinct point on anatomical structures. Accurate and reliable anatomical landmark identification is an important first step for many medical imaging and surgical planning applications. These landmarks play a major role in registering multi-modality images [1]. The interpretation of a skeletal image by a radiologist begins with identifying relevant landmarks. This is followed by recognizing the identified anatomical structures and the respective relations and measurements [2]. Traditionally, anatomical landmarks are identified manually by an expert during treatment planning, which is exhaustive, time consuming and subjective.

Automatic landmark detection in X-ray images is a much tougher problem as the images are formed based on the bone and tissue absorption. X-ray images are not very well detailed and there may not be enough local information in the images to detect landmarks. Another important challenge is the patient specific variations and ambiguous anatomical structures or deformations affecting correct prediction and localization of landmarks. Further, obtaining a large amount of data is difficult in medical imaging hindering the use of deep neural networks. Coming up with a reliable and robust algorithm to detect landmarks and structures across different modalities even in the presence of artifacts helps in better diagnosis, accurate pre-operative and therapeutic planning. In this paper, we propose a single unified deep architecture that can be used to automatically detect landmarks for different organs even with deformations.

II. Literature Review

Landmark detection in general-purpose image processing has been in use for over three decades. Popular algorithms such as Scale-Invariant Feature Transform (SIFT) [3] built on corner detection algorithms have been used for multiple applications covering navigation to object detection. However, their use is limited to the detection of corners and edges irrespective of the structures. Further, they do not depend on global features and are built on RGB information [4] unlike biomedical imaging.

Cephalometric radiography is a standard tool in orthodontic treatment planning and corrective surgery planning. Landmarks from the cephalometric radiographs are used for several orthodontic analysis where several linear and angular measurements are calculated from their positions. In Lindner *et al.* [5], a combination of random forest regression-voting and statistical shape analysis was shown to provide good accuracies on cephalometric radiographs. In Payer *et al.* [6], one dedicated network learns the locally accurate but ambiguous predictions, whereas a second network that operates at much lower resolution learns the global context. It has been shown that methods using global information perform better compared to the ones which use only local information [7]. We propose a framework where separate sub-networks learn the local appearance and global context, and interact together to detect landmarks similar to Payer *et al.* [6]. Instead of regressing the landmark locations directly, the network is trained to predict the probable landmark location as a heatmap.

The spine is one of the most important parts of the human body which carries the weight of the body. Accurate detection and labeling of vertebral levels is the first step in diagnosis on spine-related ailments. Spine images are found to be more challenging comparatively due to the presence of multiple similar landmarks and also due to the reduced image quality in the lower spine region. The presence of deformations due to conditions such as scoliosis also make it difficult to localize. Wu *et al.* [8] detect the four landmarks corresponding to four corners of the vertebra in anterior-posterior X-ray images using a combination of CNNs and statistical outlier detection methods. In this work, we apply the new network end-to-end and show better outcomes without the aid of statistical methods.

We propose an attention mechanism while learning local appearance, and show that this helps in better localization of the landmarks. While learning global context, we include features extracted from input layer, thus allowing the network to learn global context from intensity variations in input

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

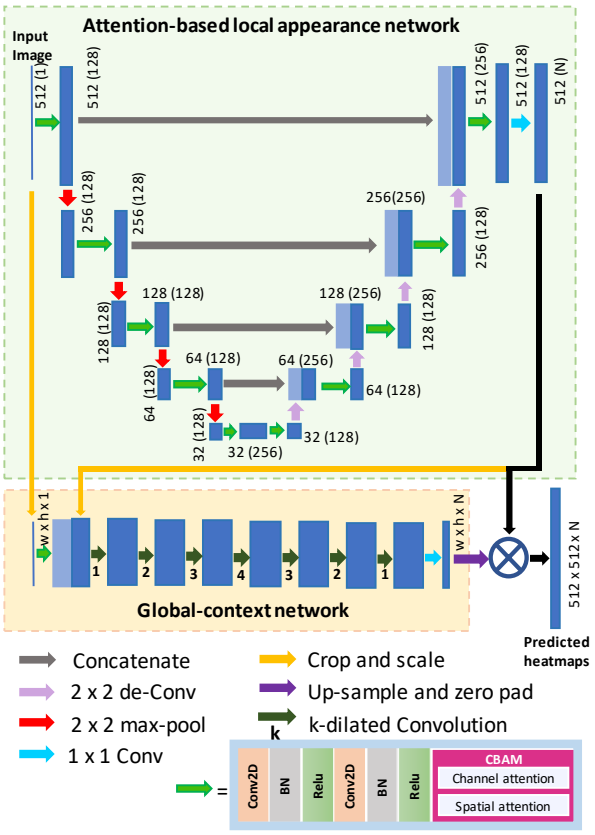

Fig. 1. Proposed architecture. The local-appearance and global-context networks interact with each other as shown. All convolution operations are followed by batch-normalization and *Relu* activation. For cephalometric data, we use $w = h = 32$, and for spine X-ray images, $w = 32$ and $h = 64$. The global-context network has 128 filters at each layer. N indicates the number of landmarks.

image. We have tested the performance of the proposed algorithm on two different anatomical regions - skull and spine using limited dataset. Further, biomedical markers such as linear and angular measurements are identified to prove the efficacy of the proposed method.

III. Proposed Architecture

Consider a d-dimensional input image $I \in \mathbb{R}^d$. Let L_i be the i^{th} with $i \in \{1, 2, ...N\}$ where N being the number of landmarks to be detected. For the landmark L_i, the heatmap $h_i \in \mathbb{R}^d$ can be interpreted as a probability distribution of a given landmark. Rather than allowing the network to generate the probability heatmap using a sliding window approach [9], we create the heatmap apriori and train a regression network to predict the heatmap as an image. Assuming the distribution of the landmark to follow a Gaussian distribution, we create a heatmap such that the value at any point is given by the Gaussian kernel centered at that location of the landmark. For a given 2D image $I \in \mathbb{R}^2$ of size $W \times H$ a heatmap for the landmark L_i at location $\hat{\mathbf{x}}_{L_i}$ is given by:

$$h_i(\mathbf{x}) = \frac{\rho}{2\pi\sigma} \exp\left(-\frac{\|\mathbf{x} - \hat{\mathbf{x}}_{L_i}\|^2}{2\sigma^2}\right), \qquad (1)$$

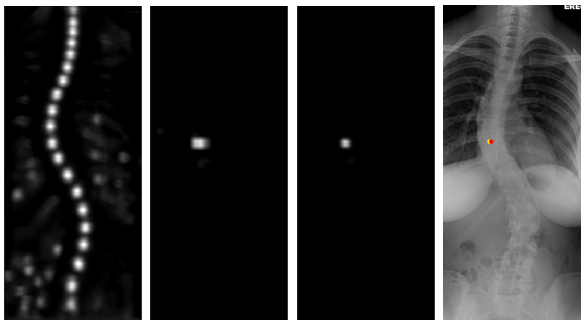

Fig. 2. Output of different stages of the network for one of the target landmarks: (from left to right) local-appearance output which shows multiple peaks, global-context output with coarse location prediction, the final output obtained by point-wise multiplying the two responses, and predicted landmark overlaid on the input image.

where σ is the standard deviation. The parameter ρ is used to scale the heatmap between 0 and 1. Each heatmap is of the same size as the input image. For N number of landmarks to be predicted, N such heatmaps are produced. The target for the network is of size $W \times H \times N$.

The architecture of the proposed network is shown in Figure 1. The network consists of two main parts: (a) a local-appearance network; and (b) a global-context network. These two sub-networks interact with each other through a point-wise multiplication ($\otimes$) as shown in the architecture. If $h_i^{LA}(\mathbf{x})$ is the output of local-appearance network and $h_i^{GC}(\mathbf{x})$ is the output of global-context network, then the final response of the network is given by:

$$h_i(\mathbf{x}) = h_i^{LA}(\mathbf{x}) \otimes h_i^{GC}(\mathbf{x}). \qquad (2)$$

Because of tasks getting split, the optimization process forces local-appearance network to use all its capacity to learn the location of the landmark more precisely, while leaving the task of figuring out the global information to the global-context network. Due to the point-wise interaction between the two networks, the ambiguous candidate landmarks are eliminated as shown in Figure 2.

For the local appearance learning, we use a modified version of U-Net architecture [10]. At each level in both encoder and decoder side, we add convolutional block attention modules (CBAM) [11] which helps to adaptively refine features along channels as well as spatial domain. To realize the global-context network, we use a fully convolutional network consisting of seven convolutional layers with different dilation factors. We use a deeper network with smaller dilated filters of size 3×3 (for cephalometric) or 5×3 (for spine images). Due to the use of smaller filters with dilation, we achieve the desired receptive field with just seven layers, and the number of parameters to train was reduced by four times.

Another important aspect of the architecture is the addition of a skip connection from the input layer to the global-context network *i.e.,* the input layer is down-scaled and concatenated to the down-scaled output of the local-appearance network. This helps the network to better understand the intensity variation in the input image. We use mean squared error as the loss function and minimize the loss function

given by:

$$\mathcal{L}(\mathbf{x},\theta) = \sum_{i=1}^{N} \sum_{\mathbf{x}} \parallel \hat{h}_i(\mathbf{x},\theta) - h_i(\mathbf{x}) \parallel^2, \qquad (3)$$

where θ is the network parameter, $\hat{h}_i$ is the predicted heatmap and h_i is the target heatmap.

IV. Results and Discussion

For evaluation, we compare with the detection accuracy with state-of-the-art methods for cephalometric and spine datasets. Further, we chose derived biomedical applications for each and assess the performance of pathology detection. The performance is evaluated based on successful detection rates (SDR) defined as the percentage of landmarks detected within a specified distance from the ground truth, and mean radial error (MRE) defined as $MRE = 1/n \sum_{i=1}^{n} R_i$, where R_i is the Euclidean distance between the actual and predicted location, and n is the number of test images.

A. Cephalometric

Cephalometric X-Ray landmark dataset is an open-source dataset that has 400 cephalometric X-ray images out of which 150 are for training and remaining for testing [12]. The ground truth of 19 landmarks is manually marked and reviewed by two experienced medical doctors. The mean position of these two annotations are considered as ground truth. Each pixel has a resolution of 0.1mm both along x and y directions and image size is 2400 $\times$ 1935 pixels.

An arbitrary image is selected from the training dataset and cropped such that all the desired landmarks are covered within this region. This region is used as template. The input images are registered to this template. The registration is performed using mutual information as metric, under the affine warping model. The detected landmarks are transformed back to get the actual landmarks. While training, we use the ADAM optimizer with learning rate of 10^{-4}. The input images were augmented with random rotations of $\pm 2^{\circ}$, translation of ± 10 pixels along both x and y directions, and scaling between 0.9 to 1.1 times the original size. We set the standard deviation for Gaussian kernel as $\sigma = 5$.

TABLE I

Comparison of results for cephalometric data. Best performance is highlighted in **bold** and second best in blue.

Method	MRE (mm)	SDR(%) 2mm	2.5mm	3mm	4mm
Test Data - 1 & 2					
Ibragimov *et al.* [13]	-	68.13	74.63	79.77	86.87
Lindner *et al.* [5]	-	74.95	80.28	84.56	89.68
Urschler *et al.* [14]	-	70.21	76.95	82.08	89.01
Payer *et al.* [6]	-	73.33	78.76	83.24	89.75
Chen *et al.* [15]	1.29	**82.03**	**88.74**	**92.74**	**97.14**
Proposed	**1.26**	81.85	87.73	92.06	96.51
Test Data - 1					
Chen *et al.* [15]	1.17	**86.67**	**92.67**	**95.54**	**98.53**
Proposed	**1.14**	86.28	91.12	94.81	97.58
Test Data - 2					
Chen *et al.* [15]	1.48	75.05	**82.84**	**88.53**	**95.05**
Proposed	**1.44**	**75.21**	82.65	87.95	94.89

TABLE II

Comparison of anatomical types in term of SCR on *Test 1 Data*. The methods 1 to 4 in table are Lindner *et al.* [5], Ibragimov *et al.* [13], Wang *et al.* [16] and Chen *et al.* [15]

Method	ANB	SNB	SNA	ODI	APDI	FHI	FMA	MW	Avg.
1	64.99	84.52	68.45	84.64	82.14	67.92	75.54	82.19	76.30
2	59.42	71.09	59.00	78.04	80.16	58.97	77.03	83.94	70.96
3	58.61	78.85	59.86	76.59	83.49	82.44	77.18	83.20	75.03
4	-	-	-	-	-	-	-	-	79.05
Prop.	**76.67**	**90.67**	**68.67**	**88.00**	**84.67**	**84.67**	**82.67**	**92.67**	**83.59**

Network performance analysis: The cephalometric dataset consists of two test sets with *Test Data - 1* containing 150 images and *Test Data - 2* with 100 images. Table I compares the performance of the proposed with other methods reported on MRE and SDR evaluated at 2mm, 2.5mm, 3mm and 4mm. It can be seen from that the proposed provided good accuracies with close to 82% of the landmarks within 2mm when both the test sets were combined. Although Chen *et al.* [15] shows slightly better performance in terms of SDR, our method has lesser MRE, indicating higher accuracy in prediction. When tested separately, proposed method outperforms on both test sets in terms of MRE, and also shows better performance in SDR at 2mm for *Test Data - 2*, which is a tougher dataset.

Detection of anatomical types using predicted landmark: For the evaluation of the performance of landmark detection, eight standard cephalometric measurements are used to classify anatomical types that are derived from the predicted landmarks [12]. The performance of our method on *Test Data - 1* is compared with the state-of-the-art as shown in Table II. It can be seen that the proposed outperforms other methods for all anatomical types as well as the average. In spite of low results in SDR values, the proposed method outperforms all other methods in biomedical task at hand.

B. Spine

The dataset [8] consists of spinal anterior-posterior X-ray images with various stages of scoliosis. 17 vertebrae are selected mostly from thoracic and lumbar spine to characterize the shape of the spine. For each vertebrae, four landmarks are annotated at four corners. It was observed that the annotations are not consistent and hence 281 training and 57 test images where thoracic and lumbar vertebrae are correctly annotated were selected. The images are scaled such that the height is equal to 512 while maintaining the original aspect ratio and then zero-padded so that the image size is 512 $\times$ 512.

Given the large aspect ratio of input images and because of padding, there is a large portion of the image with no information. It was observed that the global-context network with a size 32 $\times$ 32 was not sufficient. This is because the landmarks are in very close proximity and the local appearance of corresponding landmarks for each vertebra is quite similar. Hence, to minimize the number of parameters, the padded region is first removed by cropping 106 pixels on either side making the remaining image of size 512 $\times$ 300. This is then resized to 64 $\times$ 32. We use a filter size of

TABLE III

PERFORMANCE COMPARISON ON SPINE DATASET.

Method	SDR(%)			
	2mm	3mm	4mm	6mm
Payer et al. [6]	35.91	54.28	66.98	80.26
Proposed	39.2	48.35	73.9	81.23

TABLE IV

SMAPE SCORE COMPARISON FOR SPINE DATASET.

Method	SMAPE%
Bidur et al. [17]	25.69
Kang et al. [18]	7.84
Wu et al. [8]	23.44
Chen et al. [19]	23.59
Proposed	15.85

5×3 making the overall receptive field of 65×33. All other experimental setup is similar to the cephalometric data.

Network performance analysis: Since the physical distance of the pixel is not available, we use the relative scaling factor for each image by assuming that the physical distance between the L4 and L5 vertebra to be 35mm. The results in Table III indicates that our method performs better when compared with the results of Payer *et al.* [6].

Measurement of lateral curvature of spine in Scoliosis: The detected spine landmarks are critical in evaluating the Cobb angle, which is the measure of lateral curvature of the spine. It is defined by the angle between two lines parallel to the upper plate of the superior end vertebra and the lower plate of the inferior end vertebra. Three Cobb angles, namely, the proximal-thoracic, main thoracic, and thoracic-lumbar angles, are needed for scoliosis assessment [8]. Table IV compares the Symmetric Mean Absolute Percentage Error (SMAPE) score of the proposed with other methods in the literature. It can be seen that the performance of our method is better than all the methods that employ four corner prediction. Kang *et al.* [18] use vertebral tilt field and obtain better results on the dataset. However, this cannot be generalised for unknown conditions and hence, recent works have focused on four corner prediction. Further, vertebrae detection makes the system more reliable and explainable.

V. CONCLUSION

An end-to-end landmark detection system which uses local and global information that can be used for multiple body parts for 2D X-ray images is presented. The performance of the proposed method has been tested on two anatomically different datasets. We achieve a SDR of close to 82% for cephalometric data (at 2mm) and excellent results in terms of MRE. On the spine dataset, our method outperformed other methods. We have also shown the usefulness of the detected landmarks by deriving clinical relevance and also outperforming the state-of-the-art with an average score of 83.59% in classifying anatomical types from cephalometric images and a healthy SMAPE score of 15.85% using the Cobb angles derived from spine landmarks. Our future work includes extension of landmark detection to 3D volume data.

REFERENCES

[1] T. Lange, N. Papenberg, S. Heldmann, J. Modersitzki, B. Fischer, H. Lamecker, and P. M. Schlag, "3D ultrasound-CT registration of the liver using combined landmark-intensity information," *Int. J. of comput. assist. radiol. and surgery*, vol. 4, no. 1, pp. 79–88, 2009.

[2] R. Leonardi, A. Annunziata, and M. Caltabiano, "Landmark identification error in posteroanterior cephalometric radiography: A systematic review," *The Angle Orthodontist*, vol. 78, no. 4, pp. 761–765, 2008.

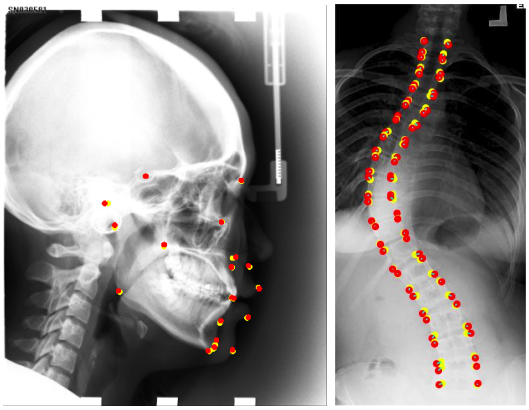

Fig. 3. Sample images from two datasets overlaid with landmarks. Yellow points indicate the annotations, and red indicates the predicted landmarks.

[3] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] Alberto Albiol, David Monzo, Antoine Martin, Jorge Sastre, and Antonio Albiol, "Face recognition using HOG–EBGM," *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1537–1543, 2008.

[5] Claudia Lindner and Tim F Cootes, "Fully automatic cephalometric evaluation using random forest regression-voting," in *IEEE Int. Symposium on Biomed. Imag.* Citeseer, 2015.

[6] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Integrating spatial configuration into heatmap regression based CNNs for landmark localization," *Med. Imag. Anal.*, vol. 54, pp. 207–219, 2019.

[7] N. Lay, N. Birkbeck, J. Zhang, and S. K. Zhou, "Rapid multi-organ segmentation using context integration and discriminative models," in *Int. Conf. Inf. Process. Med. Imag.* Springer, 2013, pp. 450–462.

[8] Hongbo Wu, Chris Bailey, Parham Rasoulinejad, and Shuo Li, "Automatic landmark estimation for adolescent idiopathic scoliosis assessment using boostnet," in *MICCAI*. Springer, 2017, pp. 127–135.

[9] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Advances in NIPS*, 2014, pp. 1799–1807.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[11] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. on Comput. Vis.*, 2018, pp. 3–19.

[12] Ching-Wei Wang, Cheng-Ta Huang, Jia-Hong Lee, Chung-Hsing Li, Sheng-Wei Chang, Ming-Jhih Siao, Tat-Ming Lai, Bulat Ibragimov, Tomaž Vrtovec, Olaf Ronneberger, et al., "A benchmark for comparison of dental radiography analysis algorithms," *Med. Imag. Anal.*, vol. 31, pp. 63–76, 2016.

[13] Bulat Ibragimov, Boštjan Likar, F Pernus, and Tomaž Vrtovec, "Computerized cephalometry by game theory with shape-and appearance-based landmark refinement," in *Proc. ISBI*, 2015.

[14] Martin Urschler, Thomas Ebner, and Darko Štern, "Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization," *Med. Imag. Anal*, vol. 43, pp. 23–36, 2018.

[15] Runnan Chen, Yuexin Ma, Nenglun Chen, Daniel Lee, and Wenping Wang, "Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting," in *MICCAI*, 2019, pp. 873–881.

[16] Shumeng Wang, Huiqi Li, Jiazhi Li, Yanjun Zhang, and Bingshuang Zou, "Automatic analysis of lateral cephalograms based on multiresolution decision tree regression voting," *J. of Healthcare Eng.*, 2018.

[17] Bidur Khanal, Lavsen Dahal, Prashant Adhikari, and Bishesh Khanal, "Automatic Cobb angle detection using vertebra detector and vertebra corners regression," in *Int. Workshop and Challenge on Comput. Methods and Clinical Appl. for Spine Imag.* Springer, 2019, pp. 81–87.

[18] K. C. Kim, H. S. Yun, S. Kim, and J. K. Seo, "Automation of spine curve assessment in frontal radiographs using deep learning of vertebral-tilt vector," *IEEE Access*, vol. 8, pp. 84618–84630, 2020.

[19] Bo Chen, Qiuhao Xu, Liansheng Wang, Stephanie Leung, Jonathan Chung, and Shuo Li, "An automated and accurate spine curve analysis system," *IEEE Access*, vol. 7, pp. 124596–124605, 2019.