Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Facial Landmark Detection with Learnable Connectivity Graph Convolutional Network

LE QUAN NGUYEN[1], VAN DUNG PHAM, YANFEN LI[1], HANXIANG WANG [1], LIEN MINH DANG, HYOUNG-KYU SONG[2], AND HYEONJOON MOON[1]

[1]Department of Computer Science and Engineering, Sejong University, Seoul 05006, Korea; quannl71290@sju.ac.kr,
[2]Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, Korea (e-mail: songhk@sejong.ac.kr)

Corresponding author: Hyeonjoon Moon (e-mail: hmoon@sejong.ac.kr).

ABSTRACT The conventional heatmap regression with deep networks has become one of the mainstream approaches for landmark detection. Despite their success, these methods do not exploit the overall landmarks structure. We present a new landmark detection which is capable to capture the overall structure of landmarks by modeling these landmarks as a graph structure. Our method combines a deep heatmap regression network with Graph Convolutional Network (GCN) into an end-to-end differentiable model. The proposed method can utilize both visual information and overall landmarks structure to localize landmarks from an image. The ad hoc spatial relationships between landmarks are learned naturally with GCN network. Experiments on multiple datasets show the robustness of the proposed method.

INDEX TERMS Face alignment, Graph Convolutional Network, High Resolution Net, Heatmap

I. INTRODUCTION

Recently, facial landmark detection that is conducted to detect multiple predefined points of human facial components and contour has become increasingly important in various facial analysis tasks like pose estimation [1], face recognition [2], [3], and face alignment [4], [5]. Nevertheless, it is still considered a challenging task in the real world principally because different poses and facial expressions can easily influence the accuracy and reliability of landmark detection. As a result, there is a pressing need to develop a framework that can precisely and robustly detect facial landmarks.

In order to address this problem, the existing approaches are mainly separated into three different categories, which include coordinate regression methods [6]–[8], heatmap regression methods [9]–[11], and graph learning methods [10], [12], [13]. The difference among them is the ways to make use of the information of face appearance. The coordinate regression methods directly learn the mapping relationship between discriminative features and coordinates vectors of landmarks, which have drawn lots of attention. Many previous methods [6], [14] reached satisfactory performances, while the results of coordinate regression methods are sensitive to face occlusion. Besides, the heatmap regression approach creates a probability heatmap for all target landmarks, which achieved state-of-the-art performances in the studies of landmark detection for multiple views [5]. In addition, the landmark detection methods with graphs also have the potential to represent the predefined landmarks as a graph. The landmark detection with graphs makes the landmarks learnable, and it is robust against appearance variations [12], [13].

In this paper, a landmark detection framework is proposed to locate landmarks on facial images efficiently and accurately by leveraging the overall facial landmark structure with graph convolutional network. The proposed method is the combination of both heatmap regression method and graph neural network so is capable of utilizing the advantages of both approaches. The main contributions of this study can be listed as follows.

- We proposed a novel learnable algorithm based on per image graph connectivity, which accounts for both landmarks' class and prediction likelihood. Besides, it allows the weights of messages from neighbour nodes to change in order to adapt to different scenarios.
- Our method allows the reuse of pre-trained heatmap models to obtain powerful landmark preliminaries and visual features, which are then used to construct the

node features of our graph model.

- Our method achieves satisfactory and robust performance on three main metrics NME, FR0.1, and AUC0.1, on a highly challenging WFLW dataset. Experiments demonstrated that the proposed method is balance in utilizing local visual information and the global structure of landmarks.

II. RELATED WORK

There are many algorithms that have been reported in the field of facial landmark detection over the years, including coordinate-based methods [6]–[8], [14], [15], heatmap-based methods [9], [11], [16]–[18], and graph-methods [10], [12], [13], [19].

Coordinate regression methods. This deep learning-based approach directly maps the input images to the landmark coordinates, which are applied to lots of landmark detection works. For example, the Mnemonic Descent Method (MDM) that adopts the combination of CNN and RNN to detect landmark locations was firstly proposed in [6]. Zhang et al. [15] applied multi-task learning methods to obtain more auxiliary information (like gender and expression) to improve the accuracy of face alignment. Experiments showed that the proposed method performed better than other face alignment methods, especially in handling the scenarios of pose changes and severe occlusion [15]. Zhu et al. presented a coarse-to-fine shape searching method to improve the robustness of the convolutional neural network (CNN) [7]. Besides, the authors in [14] proposed an end-to-end model based on deep learning and a new loss function (LUVLi) to focus on the locations and effectiveness of landmarks. In [8], Cascaded Regression and De-occlusion (CRD) algorithm was proposed to remove the occluded part of the face to obtain more accurate locations of landmarks. Even the above coordinate regression methods obtained state-of-the-art performances, they lack the ability of spatial generalization, and it is easy to lose the spatial information on feature maps.

Heatmap regression methods. Another category of methods predicts likelihood heatmaps of landmarks and performs well on facial landmark detection. Chandran et al. presented the first fully convolutional regional network for landmark prediction on high-resolution images [9], which performed well on the images with different resolutions. In another work, a framework combining unsupervised learning and fully supervised learning was designed to generate the heatmaps with landmarks, which can reduce the overfitting problem in the training process [16]. The global heatmap correction unit (GHCU) was designed to correct the detected anomalous points to improve the accuracy of landmark detection in low-quality or partially occluded images. Experiments demonstrated that the method achieved encouraging results on different databases [17]. Wu et al. introduced a novel algorithm to estimate the heatmap of the facial boundary and then locate the key points of the face using the boundary information [18]. According to the style and shape transformation of different regions in the facial image, an image enhancement method is proposed to improve the robustness of the face landmark detection algorithm [11]. Unfortunately, the heatmap regression method is not an end-to-end differential model. By using the soft argmax algorithm to convert the heatmap information into coordinate values, the values obtained are integers, which results in the loss of part of the accuracy and the offset of the coordinate position predicted by the model in the case of low resolution. Besides, the heatmap regression method has a slow training speed and large memory consumption because it requires a large output feature map.

Graph learning methods. Graph learning methods construct graphs by learning global and local features, which were applied in the landmark detection and expression recognition tasks. Li et al. designed a novel deep graph neural network to learn the relationship between human facial landmarks so as to detect landmarks accurately [12]. Similarly, a three-dimensional network was built to generate proposals for the facial area, and then the graphic prior knowledge is used to improve the performance of facial landmark detection [13]. In [10], a graph-based CNN was introduced to extract and fuse different features in order to improve expression recognition performance. Ngoc et al [19]. applied a graph neural network to obtain facial expression information by fusing image features and landmark images. Experimental results verified that the presented network obtained promising performance on various datasets.

Unlike the previous studies that focus on graph classification from the predefined graph, our approach learns the relationship between nodes from data, and it performed well without explicitly labelling occlusion landmarks. In addition, the proposed method combining heatmap information and graph neural networks to detect facial landmarks with a top-performing result. Furthermore, the presented method is robust for many challenging scenarios like noise, occlusion, poor illumination, etc.

III. METHODOLOGY

In this work, we develop an end-to-end landmark detection model which combine heatmap keypoint detection and GCN landmark regression. Given an input image, our model first predict a coarse landmark location and likelihood with heatmap model. The landmark prediction then is refined with GCN landmark regression model.

A. PRELIMINARY

Let $I \in R^{H\times W\times 3}$ be an input image of size (W, H). Our model first take in image I as input and produce a heatmap $\hat{Y} \in [0,1]^{\frac{H}{R}\times\frac{W}{R}\times C}$, where R is a downsampling factor and C is the number of landmark types. We employ the HRNet18 [20] as a CNN backbone to generate heatmap. We also use features extracted from backbone to construct node features for landmark regression model.

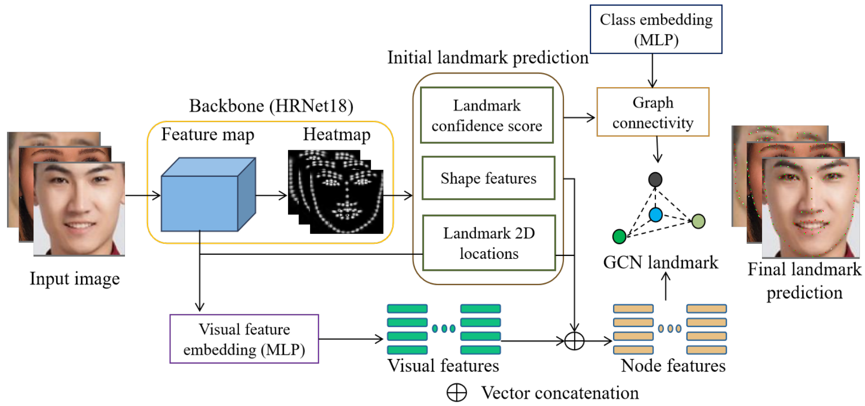

FIGURE 1. Overview of the proposed method. A heatmap is generated from input image by a CNN backbone. Initial landmark predictions and feature map is then used for constructing a graph representation of landmarks structure. The landmark graph representation is fed to the GCN landmark model to produce the final landmark prediction.

B. GRAPH CONVOLUTION

Let $G = (V, E)$ be a graph, where the vertex of graph $V = \{v_i\}$ denotes landmarks and edges $E = \{e_{ij}\}$ represents the learned connectivity between landmarks. Similar to [12], we also use graph convolutional networks (GCN) [21] for handling the information exchanging between nodes.

Let h_i^l be the hidden of vertex v_i at iteration l and e_{ij} be the learned connectivity of between node, Information is propagated through the graph G as follow:

$$h_i^{l+1} = W_1^l h_i^l + \sum_j e_{ij} W_2^l h_j^l \tag{1}$$

C. NODE FEATURES

Follow the work of li et al. [12], we also enrich Node features with visual features and shape features. We think that visual information can provide some useful information such as boundary constraint while shape feature provide the information on overall landmark structure explicitly. These information is very helpful to GCN landmark regression model for refining initial landmarks prediction.

Visual features is taken feature map from the final layers of HRNet18 [20] just before the heatmap layer. The size of visual feature vector taken from feature map is 270 (from HRNet18 [20]), which is much larger than the 2D location vector $[x_i, y_i]$. Therefore, we use a small multi-layer perceptron network (MLP) as an embedding layer to reduce the size of visual feature vector. The node feature of a node v_i is constructed by concatenating the embedded visual feature vector f_i with landmark 2D location

$$h_i^0 = [x_i, y_i] \oplus f_i \tag{2}$$

Shape features: similar to [12], We also use displacement between two nodes as shape features $q_{ij} = [x_i - x_j, y_i - y_j]$. The shape features q_{ij} are concatenated to node features of neighbor nodes v_j before aggregate information to target node v_i. The shape features are added to hidden of neighbor nodes for every iteration to ensure overall shape information persist as the graph model progressively update.

$$h_j^l \leftarrow h_j^l \oplus q_{ij} \tag{3}$$

For simplicity, we can combine equation 1 and 3 as follow:

$$h_i^{l+1} = W_1^l h_i^l + \sum_j e_{ij} W_2^l (h_j^l \oplus q_{ij}) \tag{4}$$

D. LEARNABLE GRAPH CONNECTIVITY

The graph connectivity illustrate the relationship between a pair of landmarks and determine the impact of an incoming signal from a neighbor node to a target node in GCN. As analyzed in [12], using hand-crafted graph connectivity may introduce some biases, which could lead to sub-optimal performance. In their work, li et al. [12] treat graph connectivity $E = \{e_{ij}\}$ as a learnable adjacency matrix and is trained in end-to-end manners. Therefore, the graph connectivity is remain same for a given task and independent to input images.

We argue that may not be the optimal way to handle some challenging situations like occlusion or blurry, where the prediction of some landmarks may be highly uncertain. If the initial prediction of 2 nodes is not reliable, even if their location is highly correlated, it would be better if we use other nodes in which has a more reliable prediction to estimate the landmarks which are visually challenging for prediction. Therefore, we think that the graph connectivity should depend on both the landmark types and confidence scores of the initial prediction from heatmap

For a pair of landmark (v_i, v_j) with corresponding landmark types (l_i, l_j) and confidence scores (c_i, c_j), we first compute a class embedding:

$$l_{ij} = MLP(g(l_i) \oplus g(l_j)) \tag{5}$$

where g is one-hot encoding operation. Then the graph connectivity is computed from class embedding and nodes confidence score:

$$e_{ij} = MLP([c_i, c_j] \oplus l_{ij}) \tag{6}$$

A softmax function is apply to the graph connectivity to normalize the signal from neighbors nodes.

E. TRAINING

GCN landmark: we use L1 loss on all predicted landmark coordinates to learn precise localization:

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N}|\hat{v}_i - v_i| \tag{7}$$

where $v_i = (x_i, y_i)$ and $\hat{v}_i = (\hat{x}_i, \hat{y}_i)$ are predicted and ground truth landmark coordinates respectively, and N is the number of landmarks in an image.

Heatmap model: There are two potential problems may arise when training heatmap model. Firstly, there's an extreme imbalance between the foreground landmarks and background in the heatmap. The second problem is that the heatmap influence the constructions of node features and edge features for GCN landmark model. So a dramatically change in heatmap may cause a large variation in the output of the GCN landmark model. These two problems will lead to an unstable optimization behaviour. As suggested in [22], [23], we employ the Focal loss [24] to stabilize the training process:

$$\mathcal{L}_2 = \frac{-1}{N}\sum\begin{cases}\left(1-\hat{Y}\right)^{\alpha}\log\left(\hat{Y}\right) & \text{if } Y=1\\ (1-Y)^{\beta}\,\hat{Y}^{\alpha}\log\left(1-\hat{Y}\right) & \text{otherwise}\end{cases} \tag{8}$$

where α and β are hyper-parameters of the focal loss, and N is the number of landmarks in an image. We pick $\alpha = 2$ and $\beta = 4$ in all experiments as following [22], [23]. The overall loss function to train the model in end-to-end manners is the combination of the two above loss functions:

$$\mathcal{L} = \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 \tag{9}$$

where λ_1 and λ_2 are the weights for each loss.

IV. EXPERIMENTS

A. DATASET

We evaluate our proposed method on two public datasets:

WFLW [18] dataset consist of 7500 facial images for training and 2500 facial images for testing. All these images are manually annotated with 98 landmarks and 6 attributes: pose, expression, illumination, make-up, occlusion and blur. These attributes depict different difficult scenarios to test the robustness of landmark detection method.

300W [25] dataset includes 5 face datasets: LFPW, AFW, HELEN, XM2VTS and IBUG. All images are annotated with 68 landmarks. Following the common setting in [7], [26], [27], the training set size is 3148 images which are taken from training set of LFPW, HELEN and the full set of AFW. 554 images from LFPW and HELEN testing form common set and 135 images from IBUG are regarded as the challenging subset. The full set is the combination of common and challenging subsets. The official test set has 600 face images which split into 300 indoor images and 300 outdoor images.

B. IMPLEMENTATION DETAILS

Following the preprocessing step in [28], all the face images are cropped according to the center location and resized to 256×256. The HRNet18 is selected as backbone because its network design allows us to extract deep semantic features from high resolution feature maps. The GCN landmark model consist of 3 GCN blocks with hidden sizes of 64, 16 and 2 respectively. The final GCN block predicts the final 2D landmark coordinates. We choose $\lambda_1 = \lambda_2 = 1$ for different part of the overall loss function. The learning rate is set to 10^{-4} and 10^{-3} for CNN backbone and GCN landmark model accordingly. For data augmentation, we used: rotate image with a random angle [-30, 30], scale image with a random scale factor in [0.8, 1.2], random translation in range [0.9, 1.1], random horizontal flip and color jitter.

C. EXPERIMENTAL RESULTS

WFLW is a challenging dataset with multiple difficult detection scenarios. Testing result is reported in Table 1. Following previous research, we evaluate our method with 3 metrics: normalized mean error (inter-occular), AUC@0.1 and FR@0.1. Our method is among the top performers, achieves 4.24% NME (second best), 2.68% FR0.1 (best), and 0.5892 AUC0.1.

300W: We also compare our approach with several top performing methods on 300W dataset. Results on common, challenge and full sets are evaluated using NME(%). We use AUC@0.1 and FR@0.1 for testing set. Our method achieves competitive result compare. As shown in table 2, our method achieves competitive results to previous methods.

D. LEARNED CONNECTIVITY VISUALIZATION

To study the graph structure, we draw the landmark connection based on the learned edge weight. For ease of comparison, each column is in figure 3 shows the connection of a landmark to its neighbor. As can be seen from figure 3, the graph structure varies from image to image. This behavior is intended and we believe the flexibility of the graph structure gives a boost to GCN landmark model performance.

E. ABLATION STUDY

In this section, we exam the performance of our proposed method for learning the graph connectivity by comparing it to the learnable task-specific graph connectivity which is proposed by Li et al. [12]. We experiment with both WFLW and 300W datasets. For fair comparison, we use the same HRNet18 backbone which is pretrained WFLW and 300W datasets. The performance of the backbone on WFLW and 300W dataset is 4.72 and 3.92 respectively, and the backbone weights are freezed in this experiment. The other part of GCN model is kept the same as described in section III. The GCN landmark model is trained with learning rate of 10^{-3} for 2

FIGURE 2. Visualization of Landmark detection result. Image pairs are displayed side by side for comparison. **Left images:** result from heatmap model (HRNet18). **Right images** result from GCN landmark model. **Green dot:** predicted landmark location. **Red dot:** groundtruth landmark location.

TABLE 1. Evaluation on the WFLW dataset (98 Landmarks). Top-2 results are highlighted with colors (1st, 2rd)

	NME(%)							
Method	Year	Test	Pose	Expr.	Illum.	Make-up	Occlu.	Blur
LAB	2018	5.27	10.24	5.51	5.23	5.15	6.79	6.32
SAN	2018	5.22	10.39	5.71	5.19	5.49	6.83	5.80
WING	2018	5.11	8.75	5.36	4.93	5.41	6.37	5.81
HRNet18	2020	4.60	7.94	4.85	4.55	4.29	5.44	5.42
STYLE	2019	4.39	8.42	4.68	4.24	4.37	5.60	4.86
AWING	2019	4.36	7.38	4.58	4.32	4.27	5.19	4.96
li et al.	2020	4.21	7.36	4.49	4.12	4.05	4.98	4.82
AnchorFace	2020	4.32	7.51	4.69	4.20	4.11	4.98	4.82
DSCN	2021	5.66	10.43	6.06	5.48	7.97	14.44	9.96
Our	2022	4.24	7.57	4.47	4.20	4.01	5.03	4.83
	FR@0.1							
LAB	2018	7.56	28.83	6.37	6.73	7.77	13.72	10.74
SAN	2018	6.32	27.91	7.01	4.87	6.31	11.28	6.60
WING	2018	6.00	22.70	4.78	4.30	7.77	12.50	7.76
HRNet18	2020	4.64	23.01	3.50	4.72	2.43	8.29	6.34
STYLE	2019	4.08	18.10	4.46	2.72	4.37	7.74	4.40
AWING	2019	2.84	13.50	2.23	2.58	2.91	5.98	3.75
li et al.	2020	3.04	15.95	2.86	2.72	1.45	5.29	4.01
AnchorFace	2020	2.96	16.56	2.55	2.15	2.43	5.30	3.23
DSCN	2021	8.36	34.36	7.96	5.87	7.97	14.44	9.96
Our	2022	2.68	15.03	2.23	2.44	0.97	5.03	3.36
	AUC@0.1							
LAB	2018	0.5323	0.2345	0.4951	0.5433	0.5394	0.4490	0.4630
SAN	2018	0.5355	0.2355	0.4620	0.5552	0.5222	0.4560	0.4932
WING	2018	0.5504	0.3100	0.4959	0.5408	0.5582	0.4885	0.4932
HRNet18	2020	0.5237	0.2506	0.5102	0.5326	0.5445	0.4585	0.4515
STYLE	2019	0.5913	0.3109	0.5490	0.6089	0.5812	0.5164	0.5513
AWING	2019	0.5719	0.3120	0.5149	0.5777	0.5715	0.5022	0.5120
li et al.	2020	0.5893	0.3150	0.5663	0.5953	0.6038	0.5235	0.5329
AnchorFace	2020	0.5769	0.2923	0.5440	0.5865	0.5914	0.5193	0.5286
DSCN	2021	0.4784	0.1827	0.4354	0.4653	0.4980	0.3965	0.4220
Our	2022	0.5892	0.3226	0.5615	0.5951	0.6083	0.5258	0.5390

epochs then the learning rate is reduce to 10^{-4} for 30 epochs. As the result is shown on table 3, our proposed method make a significant improvement on WFLW dataset while the performance on 300W dataset is near identical to task-specific learnable graph connectivity method. As the WFLW dataset is considered more challenging than the 300W dataset, we conclude that our method improve the final prediction results significantly when the encounter with challenging scenarios which are quite common in WFLW dataset.

V. DISCUSSION

Exploiting landmark structure to improve prediction is a well-studied approach for facial landmark detection. Wu et al. [18] propose to use heatmap boundary to explicitly group highly correlated landmarks together. In AnchorFace [30], the authors proposed to use a set of anchors as a template to model landmark positions. While the GNN is widely used in some other computer vision tasks such as pose estimation, research on application of GNN for facial landmark detection is still quite lacking despite its potential.

To the best of our knowledge, besides the work of Li et al. [12], our paper is the only work that applies GNN for learning the facial landmark structure. The main difference between our work and [12] is that in how the initial landmark position is obtained. In [12], the initial landmark prediction is from the mean average of 2D locations of landmarks while

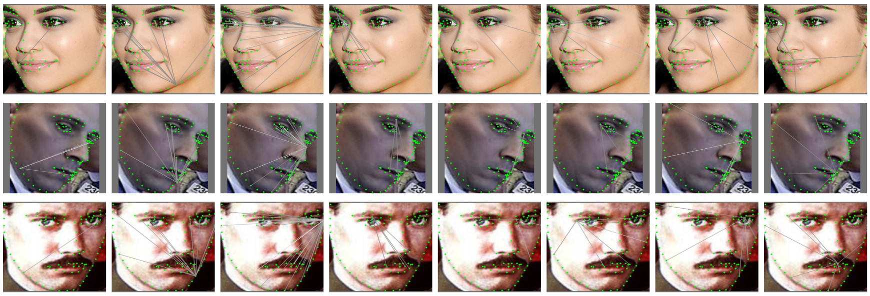

FIGURE 3. Visualization of node connectivity. Each column shows the connection of a landmark to its neighbor. Only edges with value larger than a certain threshold are shown

TABLE 2. Evaluation on the 300W dataset (68 Landmarks)

Method	Year	NME(%)			AUC@0.1	FR@0.1
		Common	Challenge	Full		
LAB [18]	2018	2.98	5.19	3.49	0.5885	0.83
STYLE [11]	2019	3.21	6.49	3.86	-	-
AWING [29]	2019	2.72	**4.52**	3.07	**0.6440**	**0.33**
li et al. [12]	2020	**2.62**	4.77	**3.04**	0.6361	**0.33**
AnchorFace [30]	2020	3.12	6.19	3.72	-	-
HORNet [31]	2020	3.38	6.36	3.96	-	-
DSCN [32]	2021	3.58	5.36	3.85	-	-
Our	2022	2.95	5.15	3.38	0.6024	0.50

TABLE 3. Ablation study on graph connectivity

Method	NME (%)	
	300W full	WFLW
backbone	3.92	4.72
li et al. [12]	**3.23**	4.40
our	**3.23**	**4.23**

in our work the initial landmark prediction results from a heatmap model. By utilizing heatmap model, we can access the confidence score for each landmark for constructing the graph connectivity. While in [12], the graph connectivity is modeled as a learnable adjacency matrix. The comparison of these two approaches is analyzed in ablation study. Another advantage of using heatmap for initial landmark prediction is that we only need a single stage GCN for landmark regression, while [12] method requires a 2-stages cascaded GCN regression model for coarse-to-fine prediction because the mean average 2d location is not good enough for coarse prediction. Our method can reuse pre-trained landmark detection directly which eases the training process. We can just freeze the heatmap model during training and can still reach a reasonable result and simplify the training process.

Other methods such as WING [33] and AWING [29] is about loss function so orthogonal to our approach and can be used in conjunction with our work to further improve landmark detection.

As our method is built on top of a heatmap model, its performance is align with the quality of the used heatmap model. Even though we only test our method with HRNet18, our method can be plugged to any kind of heatmap model and enjoy the boost in accuracy as analyzed in ablation study section.

VI. CONCLUSION

We propose a novel landmark detection model based on graph convolutional network. Our method utilizes the overall landmark structure by modeling them as a graph. The graph structure varies depending on the input images for adapting to different situations. The experimental results show that our approach result is competitive to some the recent state-of-the-art methods. The proposed method can be applied to any heatmap model to get a boost in landmark prediction accuracy

ACKNOWLEDGMENT

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540) and National Research Foundation of Korea (NRF) grant funded by the Korea government, Ministry of Science and ICT (MSIT) (2021R1F1A1046339) and by a grant(20212020900150) from "Development and Demonstration of Technology for Customers Bigdata-based Energy Management in the Field of Heat Supply Chain" funded by Ministry of Trade, Industry and Energy of Korean government.

REFERENCES

[1] A. Bulat and G. Tzimiropoulos, “Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3706–3714.

[2] K.-K. Huang, D.-Q. Dai, C.-X. Ren, Y.-F. Yu, and Z.-R. Lai, “Fusing landmark-based features at kernel level for face recognition,” *Pattern Recognition*, vol. 63, pp. 406–415, 2017.

[3] M. Asad, A. Hussain, and U. Mir, "Low complexity hybrid holistic–landmark based approach for face recognition," *Multimedia Tools and Applications*, pp. 1–14, 2020.

[4] A. Kumar and R. Chellappa, "Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 430–439.

[5] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou, "Joint multi-view face alignment in the wild," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3636–3648, 2019.

[6] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4177–4187.

[7] S. Zhu, C. Li, C. Change Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4998–5006.

[8] J. Wan, J. Li, Z. Lai, B. Du, and L. Zhang, "Robust face alignment by cascaded regression and de-occlusion," *Neural Networks*, vol. 123, pp. 261–272, 2020.

[9] P. Chandran, D. Bradley, M. Gross, and T. Beeler, "Attention-driven cropping for very high resolution facial landmark detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5861–5870.

[10] Y. Liu, X. Zhang, Y. Lin, and H. Wang, "Facial expression recognition via deep action units graph network based on psychological mechanism," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 2, pp. 311–322, 2019.

[11] S. Qian, K. Sun, W. Wu, C. Qian, and J. Jia, "Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 153–10 163.

[12] W. Li, Y. Lu, K. Zheng, H. Liao, C. Lin, J. Luo, C.-T. Cheng, J. Xiao, L. Lu, C.-F. Kuo *et al.*, "Structured landmark detection via topology-adapting deep graph learning," *arXiv preprint arXiv:2004.08190*, 2020.

[13] C. Chen, X. Yang, R. Huang, W. Shi, S. Liu, M. Lin, Y. Huang, Y. Yang, Y. Zhang, H. Luo *et al.*, "Region proposal network with graph prior and iou-balance loss for landmark detection in 3d ultrasound," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 1–5.

[14] A. Kumar, T. K. Marks, W. Mou, Y. Wang, M. Jones, A. Cherian, T. Koike-Akino, X. Liu, and C. Feng, "Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8236–8246.

[15] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 918–930, 2015.

[16] B. Browatzki and C. Wallraven, "3fabrec: Fast few-shot face alignment by reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6110–6120.

[17] Z. Liu, X. Zhu, G. Hu, H. Guo, M. Tang, Z. Lei, N. M. Robertson, and J. Wang, "Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3467–3476.

[18] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2129–2138.

[19] Q. T. Ngoc, S. Lee, and B. C. Song, "Facial landmark-based emotion recognition via directed graph neural network," *Electronics*, vol. 9, no. 5, p. 764, 2020.

[20] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019.

[21] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[22] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 734–750.

[23] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[25] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403.

[26] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 379–388.

[27] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou, "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3317–3326.

[28] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[29] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6971–6981.

[30] Z. Xu, B. Li, M. Geng, and Y. Yuan, "Anchorface: An anchor-based facial landmark detector across large poses," *AAAI*, 2021.

[31] X. Zhen, M. Yu, Z. Xiao, L. Zhang, and L. Shao, "Heterogenous output regression network for direct face alignment," *Pattern Recognition*, vol. 105, p. 107311, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320320301151

[32] J. Ma, J. Li, B. Du, J. Wu, J. Wan, and Y. Xiao, "Robust face alignment by dual-attentional spatial-aware capsule networks," *Pattern Recognition*, vol. 122, p. 108297, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320321004775

[33] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2235–2245.

NGUYEN LE QUAN received the B.S. degree in Food Engineering in 2015 from the Hanoi University of Science and Technology. He is currently pursuing PhD degree in Computer Science from Sejong University, Seoul, South Korea. He joined Computer Vision Pattern Recognition Laboratory (CVPR Lab) at the beginning of 2020. His current research interests include computer vision, graph neural network and video understanding.

YANFEN LI received the B.S. degree in Software engineering in 2018 from the LinYi University. She is currently pursuing PhD degree in Computer Science from Sejong University, Seoul, South Korea. She joined Computer Vision Pattern Recognition Laboratory (CVPR Lab) at the beginning of 2018. Her current research interests include computer vision, deep learning, image processing and video coding.

HANXIANG WANG received the B.S. degree in Software engineering in 2018 from the LinYi University. He is currently pursuing PhD degree in Computer Science from Sejong University, Seoul, South Korea. He joined Computer Vision Pattern Recognition Laboratory (CVPR Lab) at the beginning of 2018. His current research interests include computer vision, video coding and artificial intelligence.

DANG LIEN MINH received the B.S. degree in information systems from the University of Information Technology, VNU HCMC, Vietnam, in 2016. He is currently pursuing the Ph.D. degree in computer science with Sejong University, Seoul, South Korea. In 2017, he joined the Computer Vision Pattern Recognition Laboratory. His current research interests include computer vision, natural language processing, and artificial intelligence

HYOUNG-KYU SONG received the B.S., M.S., and Ph.D. degrees in electronic engineering from Yonsei University, Seoul, South Korea, in 1990, 1992, and 1996, respectively. From 1996 to 2000, he had been a Managerial Engineer with the Korea Electronics Technology Institute (KETI), South Korea. Since 2000, he has been a Professor with the Department of Information and Communication Engineering, and Convergence Engineering for Intelligent Drone, Sejong University, Seoul. His research interests include digital and data communications, information theory, and their applications with an emphasis on mobile communications

HYEONJOON MOON received the B.S. degree in Electronics and Computer Engineering from Korea University in 1990. He received the M.S. and the Ph.D. degrees from Electrical and Computer Engineering at State University of New York at Buffalo in 1992 and 1999, respectively. From January 1996 to October 1999, he was senior research in Electro-Optics/Infrared Image Processing Branch at U.S. Army Research Laboratory (ARL) in Adelphi, MD. He developed a face recognition system evaluation methodology based on the Face Recognition Technology (FERET) program. From November 1999 to February 2003, he was a principal research scientist at Viisage Technology in Littleton, MA. His main interest is on research and development is on real-time facial recognition system for access control, surveillance, and big database applications. He has extensive background on still image and real-time video-based computer vision and pattern recognition. Since March 2004, he has joined the Department of Computer Science and Engineering at Sejong University, where he is currently a professor and chairman. His current research interests include image processing, biometrics, artificial intelligence and machine learning.

• • •