LARGE-POSE FACE ALIGNMENT VIA SHAPE-AWARE HEATMAP

Jiaxin Si, Fei Jiang, and Ruimin Shen

Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
sijiaxin@sjtu.edu.cn, jiangf@sjtu.edu.cn, rmshen@sjtu.edu.cn

ABSTRACT

In this paper, we focus on dealing with problems of large-pose face alignment. Recently proposed heatmap-based algorithms have made promising performance on this problem. However, the traditional heatmap is constructed based on Gaussian model with fixed variance, which is inconsistent with the local shape of faces. In this paper, we propose a shape-aware heatmap to efficiently solve the problems of large-pose face alignment. Specifically, we design a novel heatmap based on Gaussian mixture model, where positions of several adjacent landmarks are utilized to construct different components. Thus the probability distribution is modified to fit the shape of the local region. The experimental results on Menpo-3D and AFLW2000-3D databases show that the proposed method outperforms the state-of-the-art algorithms.

Index Terms— Face alignment, Landmark, Shape-aware Heatmap

1. INTRODUCTION

Face alignment is the process of detecting facial landmarks, which is widely used in other facial analysis tasks, such as face recognition [1, 2], facial expression recognition [3, 4] and head pose estimation [5]. However, it is still a challenging task due to various head poses. As shown in Fig.1, some landmarks may be invisible and the appearances are significant different with the change of the head pose. Therefore it is hard to train a robust model to localize facial landmarks.

Recently, heatmap regression is widely used in both human pose estimation algorithm [6, 7] and face alignment [8]. Belagiannis et al. [6] propose a ConvNet model to regress a heatmap for each key point, and the ground-truth label is a heatmap synthesised by placing a Gaussian with fixed variance at the ground-truth position of the key point. Similar to [6], Bulat et al. [8] stacked four HG nets with the hierarchical, parallel and multi-scale blocks and also regress a set of heatmaps to predict the positions of facial landmarks. The proposed HG net has made a great progress in the field of large-pose face alignment. Note that both [6] and [8] utilize a Gaussian with fixed variance to construct a heatmap for each key points. However, since each value of the heatmap represents the probability that the corresponding pixel is the target key point, the probability distribution of the heatmap should base on the shape of the local region around the key point. In fact, Belagiannis et al. [6] proposed body part heatmaps to assist in the key points detection. The variance of the heatmap is based on the Euclidean distance between two key points. Such heatmap contains the local shape information according to the distribution of the Gaussion model. But for facial landmark detection, positions of adjacent points are relatively complex, so the problems are that how to define a effectively shape-aware heatmap.

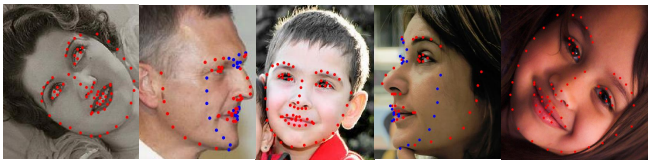

Fig. 1. Faces with various head poses. The red dots show the visible facial landmarks, and the blue dots represent the invisible facial landmarks.

To overcome the above-mentioned problem, we propose a novel shape-aware heatmap, which modifies the probability distribution of the heatmap according to the outline of a facial component in the local region. The maximum value of the shape-aware heatmap is still at the ground truth position, while values of surrounding points should be determined by the local shape. Taking the heatmap of landmark A in Fig.2 for example, the distance between point B and A is same with that point C and A. But the value of B in this heatmap is larger than that of C, since B is along the eyebrow while C is not. To realize such idea, we design a Gaussian mixture model, where Gaussians with different scales are placed according to the positions of target landmark (e.g., A) and its adjacent points (e.g., B) to preserve the local shape information. At last, we demonstrate the efficiency of the shape-aware heatmap on two experiments. The first one is training on Menpo-3D [8] and testing on AFLW2000-3D [9], and the second one is training on 300W-LP [9], and testing on both

The work was supported by NSFC (No. 61671290), the Key Program for International S&T Cooperation Project of China (No. 2016YFE0129500), Shanghai Committee of Science and Technology (No. 17511101903), and China Postdoctoral Science Foundation (No. 2018M642019).

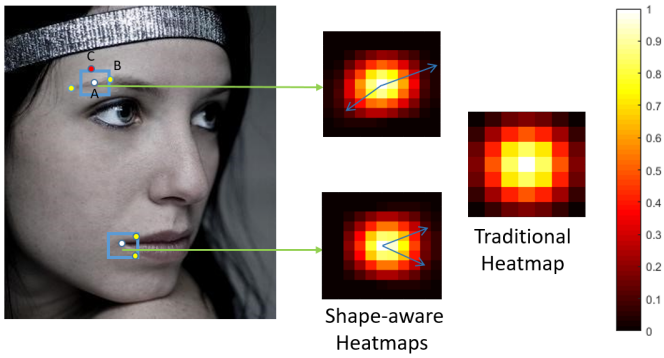

Fig. 2. Comparison between the shape-aware heatmap and the traditional heatmap. We only show the local region part of the heatmap for a better observation. For each heatmap, the brighter the pixel on the heatmap, the higher the probability that the corresponding position is the landmark. And we use the blue arrow to show the direction in which we expect the probability to slowly decline.

Menpo-3D and AFLW2000-3D. These two experiments show that our algorithm based on the shape-aware heatmap is effective, and the results on the large-pose faces are significantly improved.

To sum up, our contributions are as follows:

1. We are the first to propose the shape-aware heatmap, which preserves both the position and local shape information. While the original heatmap only utilizes the position of the landmark. And an effective method is proposed, which constructs the shape-aware heatmap by a careful designed Gaussian mixture model.

2. Experimental results on public databases have demonstrated the effectiveness of the proposed shape-aware heatmap, especially on the large-pose faces.

2. OUR APPROACH

In this section, we focus on the improvement of the heatmap, which has been demonstrated to be effective on facial alignment. We first introduce the proposed shape-aware heatmap, followed by the adjacent points chosen. Finally, we present how to embed the proposed heatmap to the existing network of landmark localization.

2.1. Shape-aware Heatmap

Heatmaps are frequently applied to key point localisation, they are a set of two-dimensional matrixes, which represent the per-pixel energy for the presence of the corresponding point at that pixel location. Previous methods utilized the Gaussian model to build the traditional heatmap, where the maximum point corresponds to the position of the key point, as shown in Fig. 2. However, besides the position information, heatmaps can also preserve the local region information. Such shape information can be gathered by the relative positions of landmarks since they are labeled on the facial components.

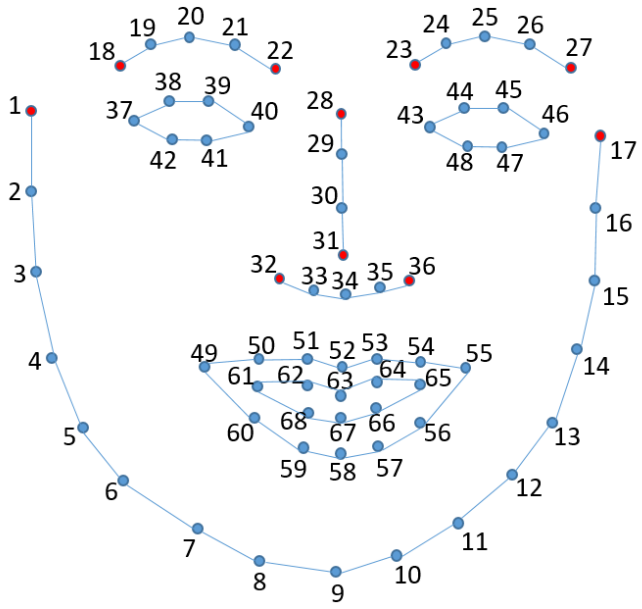

Fig. 3. Adjacent points of each landmark are the connected points. Note that the red points only have one adjacent point.

To preserve the shape information, we adapt a Gaussian mixture model to change the probability distribution of each heatmap and fit the local facial outline. Three constraints are proposed to ensure the effectiveness of the shape-aware heatmap. First, the probability values of the points along the local outline should be enhanced. For example, for the point on the middle of the eyebrow, its probability should decrease slowly along the eyebrows, but fast in other directions, the blue arrows in Fig. 2 show the expected enhanced directions. In order to achieve the first principle, the positions of the adjacent landmarks can be utilized to build our Gaussian mixture model. Thus when we add another component at the connection between target landmark and its adjacent landmarks, the probability of the pixels on the expected directions will increase. The selection of adjacent points will be shown in next Subsection. Second, it is obvious that we can not change the position of the maximum probability, which should still on the ground truth position. Such deviation may occur when we construct the corner landmarks, i.e. mouth corner and eye corner. Third, the probability value from the target landmark to the adjacent points should be monotone decreasing, which means the farther to the target landmark, the smaller probability value.

Following the three constraints, we next show how we construct the shape-aware heatmap. For the $i-th$ landmark, l_i at the position (x_i, y_i), L_i denotes indexes of its adjacent

landmarks. The shape-aware heatmap can be defined by Equation (1).

$$H_i = \mathcal{N}(x_i, y_i) + \sum_{j \in L_i} scale * \mathcal{N}(x + \frac{(x_j - x_i) * step}{len}, y_i + \frac{(y_j - y_i) * step}{len}) \quad (1)$$

where $\mathcal{N}(x, y)$ is the standard Gaussian distribution method, the center is placed at the position (x, y). len is Euclidean distance between l_j and l_i, we drop the adjacent points whose positions are same with (x_i, y_i) to ensure the len is not equal to 0. $step$ is the expected distance between different component of the Gaussian mixture Model. And the fixed value $scale$ is the mixture coefficient of our model. We will explain the effect of the parameters following the order of the three principles. First, the addition of the $\mathcal{N}(x_i, y_i)$ and other components changes the probability distribution according the local shape, which slow down the probability decline along the local outline. Second, $scale$ is less than one to ensure the joint components will not change the position of the maximum point. At last, we make a transform on (x_j, y_j) to prevent the valley on the new probability distribution. It is achieved by limiting the distance between component centers and ground truth position (x_i, y_i). Some examples of shape-aware heatmaps are shown in Fig. 2.

2.2. Adjacent Points

Since the shape information is obtained from the locations of the adjacent landmarks, the choice of the adjacent point set L_i for the shape-aware heatmap H_i is important. In order to express local shapes relatively accurately, the adjacent landmark (x_j, y_j) should be close to landmark (x_t, y_t) and the connection of (x_t, y_t) and (x_j, y_j) is expected to be placed on the outline of facial component. In general, L_i is the set $\{i - 1, i + 1\}$, the landmark and its adjacent points are almost in a straight line. But there are two special cases. The boundary points only have one adjacent points, and the adjacent points of corner landmarks are special, whose angles are acute. Fig.3 use the 68 landmarks as an example to show the adjacent points of different landmarks. The adjacent points of each landmark are the connected ones.

2.3. Overall Architecture

In this paper, we focus on the improvement of the heatmap, we choose 3D-FAN [8] as the common architecture to make a comparison with the traditional heatmap and shape-aware heatmap. 3D-FAN stacks four hour-of-glass(HG) networks. Each HG network ends with a set of heatmaps to predict the location of the landmarks, and all of four sets of heatmaps are involved in the calculation of loss while training, but only the last set of heatmaps is used when testing. And we replace traditional heatmaps by shape-aware heatmaps as the ground-truth label.

Fig. 4. Results visualization of AFLW2000-3D [9] (row 1-2, training on Menpo-3D [8]) and Menpo-3D (row 3-4, training on 300W-LP [9]) database. For each pair of images, the bottom image is predicted by the proposed algorithm, which is based on the shape-aware heatmap, while the top image is predicted by the baseline. For a better comparison, we use the blue rectangles to mark the areas with large differences.

3. EXPERIMENTAL RESULTS

In this section, we firstly introduce the experimental setup in subsection 3.1, including introductions to three databases, metric methods and the parameter setting. Second, the comparision between the shape-aware heatmap, the baseline and other state-of-art algorithms will be shown in subsection 3.2.

3.1. Experimental Setup

To validate the performance of the proposed method on large-pose faces, we choose three public face databases with wide range of head poses and 68 labeled landmarks.

Menpo-3D is re-annotated by Bulat et al. [8]. The number of the landmarks is unified to 68. Menpo-3D have 8955 face images (2300 images are profile faces). **AFLW2000-3D** [9] is built by the first 2000 images from AFLW [14]. Both Menpo-3D and AFLW2000-3D contain a wide range of poses (yaw from $-90°$ to $-90°$). To further validate the shape-aware heatmap on large-pose faces, we manually select 365 images with profile faces from AFLW2000-3D and build a challenging test set. **300W-LP** [9] is a large scale face database, which standardized several databases with 68 landmarks, including AFW [15], LFPW [16], HELEN [17], IBUG [18] and XM2VTS[19]. 300W-LP contains 61,225 images and uses the profiling method to get the large-pose faces, ranging from $-90°$ to $-90°$.

Table 1. **Comparison of NME (%) on the AFLW2000-3D [9] dataset. The first value of baseline and shape-aware heatmap is calculated by the whole dataset same with other algorithms, while the second value is calculated by the profile faces.**

Method	SDM [10]	3DDFA [9]	3DSTN [11]	DHM+RHG [12]	JVCR [13]	Baseline [8]	Shape-aware Heatmap
NME	6.12	5.42	4.49	3.85	3.64	2.13 / 3.53	**2.11/3.35**

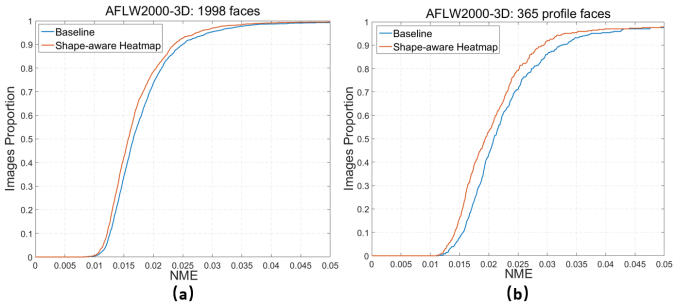

Fig. 5. CED diagrams of different methods. (a) Results on whole AFLW2000-3D [9] database. (b) Results on 365 profile faces in AFLW2000-3D. Both subfigures show the proposed method based on shape-aware heatmap outperforms baseline [8] on AFLW2000-3D.

We choose 3D-FAN [8] as the baseline to make a comparison with the proposed method based on the shape-aware heatmap. And the Normalized Mean Error (NME) is used as the evaluation metric to measure the quality of predicted results. Following [8], the NME is normalized by square root of the face size. As for the parameters, the $step$ is equal to 1 to prevent the valley in the probability distribution. After the rounding of the coordinates, eight positions may be utilized to place components of Gaussian mixture model. And the $scale$ is equal to 0.22 to ensure the position of the maximum probability point is unchanged when we add different components. For a fair comparison, we use the same batch size (15) and learning rates when we training the baseline and the proposed method.

3.2. Comparison

We conduct two experiments to verify the effectiveness of the shape-aware heatmap.

First, we train both the proposed method and baseline [8] on Menpo-3D [8] and test on AFLW2000-3D [9]. The comparison of Cumulative Errors Distribution (CED) diagrams is displayed in the left of Fig.5. Our method outperforms the strong baseline. To validate the performance of the proposed method on profile faces, we select 365 profile faces from AFLW2000-3D and their experimental results are shown in the right subfigure of Fig.5. Our method performs better on the large-pose faces. The NMEs are reduced from 1.91 to 1.78 on the whole database and reduced from 2.76 to 2.42 on profile faces. some examples are displayed in Fig. 4.

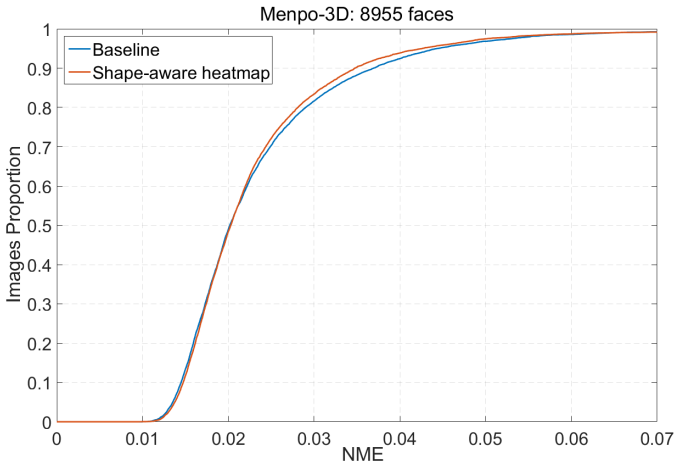

Fig. 6. CED diagrams of different methods on Menpo-3D [8], which is trained on 300W-LP [9]. The proposed method outperforms baseline [8].

To further validate the effectiveness of the shape-aware heatmap on large scale databases, we train the proposed method and baseline on a large scale database 300W-LP, and test on both Menpo-3D and AFLW2000-3D. Fig.6 shows that our method still outperforms the strong baseline on Menpo-3D, some detection results can be found in Fig. 4. And the comparison between the proposed method and the state-of-art methods can be found in Tabel. 3, the NMEs are taken from Tabel 1 of [12] and Tabel 4 of [13], NME results on above 10000 images show that the shape-ware heatmap is effective especially on large-pose faces.

4. CONCLUSION

In this paper, we propose a shape-aware heatmap, which is constructed by a carefully designed Gaussian mixture model according to the shape of local region. Adjacent points are defined and utilized to construct the Gaussian mixture model and make the probability distribution fit the local shape. We are the first to propose a method to build a shape-aware heatmap and the experimental results show that the proposed method based on shape-aware heatmap outperforms the strong baseline and the state-of-art algorithms, especially on faces with large head poses. The shape-aware heatmap can be further applied to other heatmap-based algorithms.

5. REFERENCES

[1] Xiangyu Zhu, Zhen Lei, Junjie Yan, Yi Dong, and Stan Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 787–796.

[2] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys (CSUR)*, pp. 399–458, 2003.

[3] Kaili Zhao, Wen Sheng Chu, and Honggang Zhang, "Deep region and multi-label learning for facial action unit detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3391–3399.

[4] Vinay Bettadapura, "Face expression recognition and analysis: The state of the art," *Computer Science*, 2012.

[5] Yue Wu, Chao Gou, and Qiang Ji, "Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5719–5728.

[6] Vasileios Belagiannis and Andrew Zisserman, "Recurrent human pose estimation," in *IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 2017, pp. 468–475.

[7] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*, 2016, pp. 483–499.

[8] Adrian Bulat and Georgios Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017, pp. 1021–1030.

[9] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 146–155.

[10] Xuehan Xiong and Fernando De La Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 532–539.

[11] Chandrasekhar Bhagavatula, Chenchen Zhu, Khoa Luu, and Marios Savvides, "Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses," in *The IEEE International Conference on Computer Vision*, 2017, p. 7.

[12] Bin Sun, Ming Shao, Siyu Xia, and Yun Fu, "Deep evolutionary 3d diffusion heat maps for large-pose face alignment," in *British Machine Vision Conference*, 2018, p. 256.

[13] Hongwen Zhang, Qi Li, and Zhenan Sun, "Joint voxel and coordinate regression for accurate 3d facial landmark localization," *arXiv preprint arXiv:1801.09242*, 2018.

[14] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *IEEE International Conference on Computer Vision Workshops*, 2011, pp. 2144–2151.

[15] Xiangxin Zhu and Deva Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2879–2886.

[16] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 2930–2940, 2013.

[17] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *IEEE International Conference on Computer Vision Workshops*, 2013, pp. 386–391.

[18] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403.

[19] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre, "Xm2vtsdb: The extended m2vts database," in *Second International Conference on Audio and Video-based Biometric Person Authentication*, 1999, pp. 965–966.