A Robust Facial Landmark Detection in Uncontrolled Natural Condition

Ping Wang, Qianyu Zhou, Yucong Zhao, Shumin Zhao, Jiquan Ma

Computer Science and Technology, Heilongjiang University, Harbin 150080, China
majiquan@hlju.edu.cn

Abstract. Facial landmark detection in uncontrolled natural condition is a challenging task due to illumination, occlusion, expression, posture, blur and other factors. In order to make a stable and accurate prediction, a novel data augmentation strategy based on random mask is used to obtain an efficient training data set. Our work is motivated by the ability of artist’s spatio imagination from part to whole. The artist can find the corresponding key points according to the incomplete image, so as to realize the whole from the part. In the experiment, the corresponding incomplete image is first generated based on the random mask and the complete image, and then the incomplete image is fed into the neural network to fit the mapping between the incomplete image and the facial landmarks. In order to evaluate the performance, a simple regression convolutional neural network based on the improved VGG16 [1] was trained, and the latest technical results were obtained. Furthermore, our method is tested over both complete and incomplete face images. Extensive experiments reveal that our model can work well on different human face data set.

1. Introduction

Facial landmark (FL) detection aims to identify some key points as a sparse representation of a face. FL provides critical clues for 3D face pose estimation, facial expression analysis or transfer. For instance, head pose estimation[2], face tracking[3], face recognition[4-5]. Although existing methods have achieved exciting performance, FL detection is still a challenging task in uncontrolled condition, where occlusion, posture, lighting and other external factors lead to incomplete facial regions.

Recent years, some works have been focusing on this problem and got some improvements. For instance, a unified robust cascading regression framework was proposed by Wu et al[6]. In[6], visibility of landmarks was gradually updated by considering appearance, current shape information and occlusion consistency. Xing et al. proposed a regression model to learn the mapping between face appearance and shape utilizing sparse coding[7]. In Qian’s work, an augmentation based on image style transformation is used to improve the robustness and quality[8]. In[9], the occlusion probability was inferred by the distribution of face appearance and shape. At present, the method with the most extensive application and the highest effect accuracy is based on deep learning. Therefore, this paper mainly studies the application of deep learning in face critical point detection.

We are motivated by the phenomenon that in real life, artists draw a face contour according to their own experience and imagination even by observing partial information. In order to train a FL detection network with imaginative ability, various kinds of incomplete images are generated by binary masks. In this way, our model demonstrated validity and feasibility over CelebA[11] and WFLW[17] dataset.
2. Method
Our framework is composed of two parts: Firstly, we generate incomplete facial images by using random binary masks. Then, those binary masked residual images are fed into a CNN regression network for training.

We formulate the problem of FL detection as an image-based regression. The input is a three-channel image, and the output contains the coordinates of the landmark points. Theoretically, as long as the regression performance is good enough, the model can return satisfactory results. Different from the existing methods, our method just uses the data enhancement to improve the performance by random binary mask, which expand the distribution of face imaging to the uncontrolled scene. Even though, our extended training data is synthetic, it has more complex structure than the natural facial image. Based on our more stronger training data set, a clear improvement for FL detection can be obtained by just using a simple regression network, such as a modified VGG16[1].

2.1. Masked Residual Facial Image
Given the original complete image $\text{img}$ and the random binary mask $M$, the masked residual image $M_{\text{img}}$ can be obtained by $M_{\text{img}} = \text{img} \odot (1-M)$, where the pixel value of binary mask is 0 and 1, the pixel value of image is between 0 and 255, and the pixel value of masked image is also between 0 and 255. And the $\odot$ represents the element-wise multiplication. We use the masks published in[12], where there are six categories according to the ratio of missing areas.

Due to the randomness of the mask, the incomplete area may be concentrated in the middle of the face or the edge of the face. With the increase of the proportion of missing area, the destruction of face structure is more serious. This randomness can also mimic the randomness of face occlusion in the natural scene. Several masks of each category are selected to generate corresponding masked images.

2.2. Regression Network
For the regression model, we use the VGG16[1] as the backbone of landmark detection as shown in the third part of Fig.1. Specifically, we modify the last full convolution as the output layer with size of $2N$, where $N$ is the number of landmarks. The input image size is resized to $256*256*3$. The overall architecture of the network is shown in Fig.1. Top row on the left is data enhancement part. The Landmark Detection is a modified VGG16. During training, the data after data enhancement is used as input, with binary incomplete image.

2.3. Loss Function
In this paper, we use CNN to achieve the regression of landmark. The final output of the network is the predicted coordinate value of landmark. In order to calculate the error between the predicted coordinate value and the real coordinate value. Here we use L2 loss, also known as MSE. For the CelebA dataset, we use 68 annotation points detected by Dlib as the real landmarks of the image. For WFLW dataset, the 98 annotation points are directly used as the real value. The ground-truth landmarks $X := \{x_1, ..., x_N \} \in \mathbb{R}^{2N}$, value of network prediction output $Y := \{y_1, ..., y_N \} \in \mathbb{R}^{2N}$, $N$ is the number of landmarks.

Mathematically, the loss can be written in the following general form:

$$L(x_i, y_i) = \frac{1}{M} \sum_{i=1}^{M} (x_i - y_i)^2$$

From (1), where $N$ is the pre-defined number of landmarks per face to detect, $M = 2N$.

3. Experiments
3.1. Datasets
CelebA [11] dataset is an open dataset provided by the Chinese University of Hong Kong. The images cover large pose changes and background clutter. There are 10177 identities and 202599 facial images.
For each face image, there are 5 landmark positions and 40 binary attribute annotations. We selected 10000 complete images and divided them into training set, verification set and test set with the proportion of 8:1:1. WFLW [17] dataset is a new face key point detection data set with wider facial landmarks in the wild, which contains 10000 face image data with 98 key points and 6 face attribute labels including the transformation of pose, expression, lighting, makeup, occlusion and blur.

Fig. 1. The architecture of our framework using data enhancement and CNN regression network.

In order to get the binary masks, we use the mask dataset provided by NVIDIA [12], the testing mask contains 12000 masks. According to the proportion of missing areas, this masks can be divided into 6 categories of masks with different hole-to-image area ratios:(0.01,0.1], (0.1, 0.2], (0.2, 0.3], (0.3, 0.4], (0.4, 0.5], (0.5, 0.6], Each category contains 1000 masks with and without border constraints. To prevent insufficient training, we randomly selected 20 masks from each category and repeatedly expanded them to 10000 for training. When generating a mask image, randomly select several images from the original face image data set and the mask data set, a masked facial image will be computed by a pixel-wise product between complete facial image and mask. In this way, we have 8000 * 20 = 160000 masked images for training in every epoch.

3.2. The Ground Truth of the Landmarks
Because there are only five manually labeled landmarks in the CelebA dataset, in order to get 68 landmarks for detection, we use the Dlib[13] library to generate 68 landmarks of each image as the ground-truth landmarks of the image. However, the Dlib library can’t mark the landmark well for some face images where the face angle is too large, the color is similar to the background or the face is occluded. In order to prevent the impact of such data on training, we re-cleaned the data set to ensure that each face image can be correctly marked with landmarks by Dlib library. For the set of data which cannot be correctly marked by the Dlib library, we add them to the testing dataset. For WFLW dataset, there are 98 annotation points as face landmark.
3.3. Experimental Setting

Evaluation Metrics: The performance of our model and other models in comparison for facial landmark detection are evaluated by normalization mean error (NME) and failure Rate (%) following protocols used in [17]. For WFLW dataset, we normalized the error using inter-ocular distance in order to compare with the algorithms [14, 8, 15] which did so as the normalizing factor. For other datasets, we applied inter-pupil distance for normalization.

Implementation Details: The original images are resized to 256 * 256 * 3. The batch size is 64, and Adam optimizer is used with a learning rate of 10^-4. and observe the test that we train on the incomplete images generated by six different categories of mask. The model is implemented by Tensorflow framework. The entire network is trained on a NVIDIA GTX 1080Ti GPU. The maximum number of iterations was trained for three days.

3.4. Comparison with State-of-the-arts

Our model is compared with SSST [8] over WFLW. The experimental results are shown in Table 1. The proposed model is also compared with CFSS[18], DVLN[16], LAB[17], SAN[14], WING[15]and Res-18 method on the NME (%) and Failure Rate (%) following protocols used in [8].

Our model achieved 4.9% NME and 5.47% failure rate over the whole testing dataset, while SSST obtained 5.25% and 7.44% in terms of NME and failure rate respectively. When comparing with six different categories, our method has improved in the vast number of categories and more stable for most categories images in terms of NME and failure rate. And our method also achieves 6.13% improvement to SAN model, 7.02% boost to LAB from 5.27% NME to 4.9%, and 4.11% improvement to WING. The experimental results for landmark detection from structural incomplete images are given in Fig.2. Compared with SSST after data enhancement via style transform. As shown, but also closer to the manually labeled landmark in vision.

| Metric      | Method       | Fullset | Pose | Expression | Illumination | Make_up | Occlusion | Blur |
|-------------|--------------|---------|------|------------|--------------|---------|-----------|------|
| Mean Error (%) | CFSS [13]    | 9.07    | 12.36| 10.09      | 8.30         | 8.74    | 11.76     | 9.96 |
|             | DVLN [16]    | 6.08    | 11.54| 6.78       | 5.73         | 5.98    | 7.33      | 6.88 |
|             | LAB [17]     | 5.27    | 10.24| 5.51       | 5.23         | 5.15    | 6.79      | 6.32 |
|             | SAN [14]     | 5.22    | 10.39| 5.71       | 5.19         | 5.49    | 6.83      | 5.80 |
|             | WING [15]    | 5.11    | 8.75 | 5.36       | 4.93         | 5.41    | 6.37      | 5.81 |
|             | Res-18       | 6.09    | 10.76| 6.97       | 5.83         | 6.19    | 7.15      | 6.67 |
|             | SSST w. Res-18 [8] | 5.25 | 9.10 | 5.83 | 4.93 | 5.47 | 6.26 | 5.86 |
|             | Ours w. CNN  | 4.9     | 6.32 | 5.5       | 3.43         | 5.67    | 5.75      | 5.06 |

| Metric      | Method       | Fullset | Pose | Expression | Illumination | Make_up | Occlusion | Blur |
|-------------|--------------|---------|------|------------|--------------|---------|-----------|------|
| Failure Rate (%) | CFSS [13]    | 20.56   | 66.26| 23.25      | 17.34        | 21.84   | 32.88     | 23.67 |
|             | DVLN [16]    | 10.84   | 46.93| 11.15      | 7.31         | 11.65   | 16.30     | 13.71 |
|             | LAB [17]     | 7.56    | 28.83| 6.37       | 6.73         | 7.77    | 13.72     | 10.74 |
|             | SAN [14]     | 6.32    | 27.91| 7.01       | 4.87         | 6.31    | 11.28     | 6.60 |
|             | WING [15]    | 6.00    | 22.70| 4.78       | 4.30         | 7.77    | 12.50     | 7.76 |
|             | Res-18       | 10.92   | 43.87| 13.38      | 7.31         | 11.17   | 16.30     | 11.90 |
|             | SSST w. Res-18 [8] | 7.44 | 32.52| 8.60 | 4.30 | 8.25 | 12.77 | 9.06 |
|             | Ours w. CNN  | 5.47    | 11.28| 6.6        | 4.33         | 11.61   | 7.48      | 4.89 |

The training was carried out on six categories of masks, and the NME (%) and Failure Rate (%) of the complete image with structural incomplete on each model were tested. It is not that the larger the missing area is, the more helpful it is to train the image detection of structural defect. Only by using the appropriate mask to enhance the data can we help the detection of structural incomplete images better. When the missing rate of mask is more than 40%, the whole face part of the generated mask image is likely to be missing. Such image data is difficult to learn features.

We also calculated the experimental results of different size masks on CelebA dataset after data enhancement. Table 2 shows the test results of six kinds of mask data enhanced. CelebA dataset is not too complex, and the detection accuracy is relatively high. When mask = 1,2,3, NME (%) of the test can be reduced to the lowest value of 2.39%, and the failure rate is only 0.45%.
Fig. 2. Visual comparison results on WFLW test set are compared between manually labeled landmark, SSST [8] and CNN enhanced by mask.

Table 2: Normalized mean error (%) and Failure Rate (%) on celebA test set.

| Epoch =400 |
|----------------|
| NME(%) | Failure Rate(%) |
|----------------|
| Mask =1 | 2.39 | 0.45 |
| Mask =2 | 2.68 | 0.45 |
| Mask =3 | 2.76 | 0.45 |
| Mask =4 | 2.77 | 0.48 |
| Mask =5 | 2.89 | 0.52 |
| Mask =6 | 3.01 | 0.54 |

4. Conclusion and Future Work
In the experiment, we noticed that when the missing area of binary mask is too large, the accuracy is general, which may be because in this case, the network cannot extract enough effective information according to the surrounding area of the missing area, resulting in the general accuracy of the results. In addition, for some face images with too large face angle rotation, there are still some difficulties in detection. We plan to solve these problems in the future work, and further improve the accuracy of face marker point detection.

In this paper, we generate binary masked images corresponding to the complete image to “simplify” the complete image with structural mask to further improve the quality of facial landmark point detection, especially for the low-quality face image generated in complex environment. The experimental results verify the effectiveness of learning from residual images to detect facial landmarks from incomplete structured facial images.

References
[1] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
[2] Wu, Y., Gou, C., Ji, Q.: Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 3471–3480
[3] Khan, M.H., McDonagh, J., Tzimiropoulos, G.: Synergy between face alignment and tracking via discriminative global consensus optimization. In: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE (2017) 3811–3819
[4] Zhu, X., Lei, Z., Yan, J., Yi, D., Li, S.Z.: 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) 787–796

[5] Liu, Y., Wei, F., Shao, J., Sheng, L., Yan, J., Wang, X.: Exploring disentangled feature representation beyond face identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2080–2089

[6] Wu, Y., Ji, Q.: Robust facial landmark detection under significant head poses and occlusion. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 3658–3666

[7] Xing, J., Niu, Z., Huang, J., Hu, W., Zhou, X., Yan, S.: Towards robust and accurate multi-view and partially-occluded face alignment. IEEE transactions on pattern analysis and machine intelligence 40 (2017) 987–1001

[8] Qian, S., Sun, K., Wu, W., Qian, C., Jia, J.: Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 10153–10163

[9] Zhu, M., Shi, D., Zheng, M., Sadiq, M.: Robust facial landmark detection via occlusion-adaptive deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 3486–3496

[10] Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. In: Bmvc. Volume 1., Citeseer (2006) 3

[11] Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. (2015) 3730–3738

[12] Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 85–100

[13] Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 1867–1874

[14] Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Style aggregated network for facial landmark detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 379–388

[15] Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2235–2245

[16] Wu, W., Yang, S.: Leveraging intra and inter-dataset variations for robust face alignment. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. (2017) 150–159

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

[18] Zhu, S., Li, C., Loy, C.C., Tang, X.: Unconstrained face alignment via cascaded compositional learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 3409–3417