Multi-task convolution network for face alignment

Yang Sun\textsuperscript{1}, Xuan Zhang\textsuperscript{2} and Chongrong Li\textsuperscript{3}

\textsuperscript{1}Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing, China, Sunyang15@mails.tsinghua.edu.cn
\textsuperscript{2}Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing, China, zhangx@cernet.edu.cn
\textsuperscript{3}Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing, China, licr@cernet.edu.cn

Abstract. In order to solve face alignment more effectively, we proposed a multi-task deep convolution network for face alignment, which achieves good performance even in the case of large pose variations and severe occlusion. Instead of dealing with face alignment as a single task, we jointly trained the auxiliary task of pose estimation together with face alignment to guide the distribution of facial points. By doing so we are able to 1) avoid trapping in the local optimum due to the inaccurate face boxes, 2) improve the robustness in dealing with faces with pose variation and severe occlusion. Compared with the traditional methods, our method also improves the accuracy by providing better initialization instead of mean shape. Extensive experiments show that our method has great performance on various benchmarks.

1. Introduction

Face alignment \cite{2,3,4,5,22} aims to locate facial key points automatically, which plays an important role in many facial tasks, such as face recognition \cite{6}, face attributes inference \cite{7}. Although scholars have put forward many methods in the past decades, face alignment is still a challenging problem especially in the case of heavy occlusion and large pose variations.

However, both template fitting approaches \cite{1} and regression-based methods \cite{2} take face alignment as an independent problem and ignore the inherent relationship between face alignment and other auxiliary tasks. Researches have proved the effectiveness of multi-task learning. It’s meaningful to explore the inherent correlation between face alignment and pose estimation (the angles of roll, yaw and pitch of head). Reference \cite{20} takes the age estimation and gender classification as auxiliary tasks, but we should point out that the most helpful information is encoded in poses which related to the distribution facial points.

In this paper, we proposed a new framework to integrate these two tasks by using deep convolution network. We built a shared feature extraction model with two branches on the top to predict face landmarks and pose angles. Fig 1. Shows the results of our method in the case of large pose and occlusion. Thanks to the multi-task training, the deep multi-task convolution network gets the state-of-the-art performance.
Furthermore, we also evaluated the feasibility of using our network’s output as the initial input in the
traditional cascaded regression methods. As known to all, an obvious shortcoming of traditional cascaded
regression method is the heavily dependence on initialization [8]. Particularly, it easily trapped into local
optimum when the real shape is far from the mean shape. Besides this, the occlusion confuses the
regression to estimate the next shape. To solve this problem, we used our network’s output as the robust
initialization. The landmark is closer to the ground truth shape by several iterates and can avoid being
trapped into local optimum.

The following are our contributions: We proposed a multi-task convolution network for face
alignment task which could works well even in challenging test datasets. Besides this, we took the deep
convolution network’s output as initialization of other methods to improve the regression results.

2. Related Work
Many face alignment methods have been proposed in recent years. The Regression-based methods and
deep-learning-based methods are the most popular methods. The regression-based methods regress the
face landmark explicitly by learning the features in the image, e.g. [18] use random forest [9] to achieve
good performance with local binary features, [2] regresses features with random ferns. Convolution
networks have also been used in face alignment task successfully, [19] detects face landmarks with three
level networks.

Multi-task learning [23] has been used in many computer vision tasks successfully. Reference [20]
proposes the method of TCDCN which first introduced multi-task learning into facial landmark detection.
But TCDCN only works for five points with the auxiliary task of head pose estimation, gender
classification and age estimation. Some tasks are not closely related to face alignment, in other words,
the age doesn’t have much relationship with where the facial points located in. Besides this, the most
related task pose estimation only output right, left, or frontal but not the real angle. Focus on these
problems, we removed the irrelevant tasks, and describe the pose by using three float numbers, which
represent roll, yaw and pitch of a head to build our network.

In addition, many studies are concentrated on the initial problem. Reference [2] proposed to use
different initializations and take the mean results as the final pose. Reference [17] solved the initialization
problem by predicting the rough estimation from global image patch. These methods solve the problem
by some assumptions which alleviate the problem but not solve the problem. We used our method to
improve the robustness of initialization.

3. The Proposed Approach
3.1. Multi-Task CNN
In the field of machine learning, there have been some principles to guide the multi-task formulation.
Now suppose we want to train T tasks together, and there are N samples to be trained, the training set can
be represented as \( \{x_i, y_i\}_{i=1}^{N} \), where \( x_i \) is the i-th image and \( y_i \) is the corresponding labels. The purpose of the multi-task network is to minimize:

\[
\arg\min_w \sum_{i=1}^{T} \sum_{i=1}^{N} l(f(x'_i, w), y'_i) + \Phi(w)
\]

where \( f(.) \) is the function to compute the output using the parameter \( w \). And the \( l(.) \) is the loss function which represents the difference between the output and the real label. In this task we choose the Euclidean Loss for each sample:

\[
l'_i = \| y'_i - f(x'_i, w) \|^2_2
\]

But these two tasks have different levels of importance since our main goal is to optimize the face alignment with the help of auxiliary task. As a result, we use different coefficient to represent the level of importance. In our method, we share the same input for the task of landmark detection and pose estimation. We denoted the inputs as \( \{x_i\}_{i=1}^{N} \) and their responding labels as \( \{y'_i, y''_i\}_{i=1}^{N} \), where \( y'_i \) is the real landmark and \( y''_i \) is the pose angle. \( y'_i \in R^{136} \) represent 68 landmarks and \( y''_i \in R^{3} \) represent the yaw, roll and pitch angle of head. The problem is converted to following formula:

\[
\arg\min_w \sum_{i=1}^{2} \sum_{i=1}^{N} \lambda_i l'_i(f(x'_i, w), y'_i) + \Phi(w)
\]

Our network use five convolution layers and four full-connected layers to predict the landmarks and pose. And the last full-connected layers output the results. Compared with the cascaded networks such as TCDCN, our model is deeper which could achieves robust performance.

3.2. Our Network Structure

In [20], multi-task CNN has been designed for face alignment and other relevant tasks. However, we find that maybe the results are limited by the following aspects: (1) There are so many irrelevant tasks such as gender estimate which may not help the accuracy improving. (2) The network is not deep enough so that couldn’t capture sufficient information about the landmark location. (3) The input is only 40*40 gray-scale face image, and it is too small to get the state-of-the-art performance when dealing with 68 landmarks detection. What’s more, the gray-scale image may loss some information, so we use the color face image as the network input.

Fig 2. Shows the structure of our network. Our network contains two part. The first part is a shared feature extraction model, which contains five convolutional layers, three polling layers, four fully connected layers and one dropout layer to avoid over-fit. The second part outputs the landmarks position and pose angles using the same feature vectors. In back propagation, the pose information is encoded into the network. The network’s input is 224*224 BGR face image detected from the open library dlib. We resize the face image into 224*224 and then minus the mean value. We use the Rectified Linear Units (ReLU) as the activation function. The pooling layer we all use the max-pooling. The landmark location layer and pose direction layer are connected to one shared layer to predict the results, so they can share most of the computes to speed up. Compared with TCDCN, our network has more filters and deeper structures to capture more features to achieve better performance.
3.3. Training Matters

Our model is trained with the LFPW [12], Helen [13], AFW [24] and 300W [11] dataset. These datasets contain train and test images with 68 landmarks. In order to make the net more robust we enlarged the datasets to 22000 with data augmentation. In the training phase, we used the dlib to detect the face boxes, and transform the points to [-1 1]. Then in test phase, we need to translate the [-1 1] to the real point with the information of boxes. The pose angles are labeled with the software of intraface which can use the point’s information to generate 3D head pose. In training phase, base learning rate is set to 0.001 and weight decay is 0.0005, the max iterations is 900000. The tasks of face alignment and pose direction share the same weights during the phase of feature extraction.

3.4. Experiment and Result

In our experiment, we test the performance of our method using the mean error which is measured by the distances between ground truth shape and estimated shape. Then the result was normalized by inter-ocular distance. The equation can be calculated as follows:

$$e_i = \frac{|y - S|}{n \times D_i}$$ (4)

where $D_i$ is the inter-ocular distance determined by the center of the two eyes, $n$ is the number of landmarks, $y$ is the output shape and $S$ is the ground truth shape.

3.4.1. The Effectiveness of Related Task. Many works have proved that auxiliary tasks can help us improve the performance of main task. But there is no such theory to tell us which auxiliary task will help us or not. Because of this, we carefully choice the auxiliary task which is beneficial to landmark detection. Intuition tells us, compared with age or gender, the task of pose estimation is more relevant to face alignment. The reasons are as follows: Firstly, the pose angles can guide the landmarks’ distribution, especially for the large pose variation. Secondly, it provides the constraint to avoid the landmark detection trapped into local optimum caused by the inaccurate of the face detection or large pose variation.

To evaluate the effectiveness of auxiliary task pose estimate, we tested the performance of two different nets (joint pose estimate learning and do not joint). The two networks have the similar constructs except the output layer of pose estimation. For simplicity, we call the two net as S-Net (Single net) and M-Net (Multi-task net) and evaluate them on the 300W dataset. The mean error of S-Net is 8.62 while M-Net is 7.39. This demonstrates that pose estimation task is beneficial to landmark detection.

3.4.2. Compare with CNN-Based Method. We use the CNN-based method TCDCN [20] as the baseline. Moreover, we also compared our method with the Cascade CNN which detect the landmark with three cascade networks. Thanks to the multi-task and deeper network, our method can achieve better performance than these above methods. These two works only detect five point (the center of eyes, nose and the corner of mouse) and the 68 landmarks don't conclude the points of center of eyes. We only compare the performance with the points of nose, left mouth corner and right mouth corner. TABLE I. shows the mean error calculated by (4), we can draw the conclusion that our method has better

Figure 2. Structures of our Multi-Task Convolution Network
performance than the above two CNN-based method. Without much code optimization, our C++ implementation cost about 7ms for one image to pass the network on a NVIDIA GTX970 GPU.

Table 1. Mean error of Different methods

| Method        | Key Points |         |         |
|---------------|------------|---------|---------|
|               | Nose       | Left mouth corner | Right mouth corner |
| Cascaded CNN  | 10.3       | 8.4     | 7.4     |
| TCDCN         | 8.5        | 7.8     | 7.1     |
| OURS          | 7.7        | 7.1     | 6.5     |

3.4.3. Evaluate With State-of-the-Art Methods. Besides the CNN-based methods, we also compared our method with some state-of-the-art methods on LFPW, Helen and 300W datasets. It's worth to point out that we use the dlib to detect the face boxes. The state-of-the-art methods includes DRMF [14], ESR [2], PCPR [15], SDM [16], CFAN [17], LBF [18] and CFSS [10]. TABLE II. Shows that our method has good qualities in different dataset.

Table 2. Evaluate with other Methods

| Method | TEST DATASET |         |         |
|--------|--------------|---------|---------|
|        | LFPW         | Helen   | 300W    |
| DRMF   | 6.57         | 6.70    | 9.22    |
| ESR    | --           | --      | 7.58    |
| PCPR   | 6.56         | 5.93    | 8.35    |
| SDM    | 5.67         | 5.50    | 7.50    |
| CFAN   | 5.44         | 5.53    | --      |
| LBF    | --           | --      | 6.32    |
| CFSS   | 4.87         | 4.63    | 5.76    |
| OURS   | 5.63         | 5.49    | 7.39    |

4. Good Initial for Regression

Besides the multi-task learning methods, regression-based methods are also very popular in the field of face alignment. Regression-based methods learn the regression function to approach the target shape.

4.1. Regression-Based Methods

Regression-based methods predict the landmark shape S from a given face image I, and we denote \([x_1, y_1, ..., x_N, y_N]\) for the N landmarks, in this paper the N is 68. These methods consist of T iterations to update the initial shape to approach the ground truth shape:

\[
S' = S^{t-1} + f(I, S^{t-1})
\]

where \(S'\) is the shape after \(t\) iterations, and \(f(., .)\) is a function to calculate the \(\Delta s\).

Although these methods work fast and accurate on frontal face, it usually fails to solve the faces with large pose well. The most important reason is these methods relied heavily on the initial shape. So faces with large pose are always trapped into local optimum when the initial shape is far away from the real landmarks. Many attempts were done to solve this problem and the most common method is use the mean shape as the initial shape. Burgos [25] proposed the smart initialization while the change between iterate is large then stop it and restart with another initial shape. Xiao [26] present progressive initialization with k-means of obvious points to handle challenging poses robustly. These methods solve the problem based on some assumptions which can mitigate the problem instead of solving it.
4.2. Initialize Using Our Method

To solve this problem, we used the result from our network as the initial shape of regression-based method. We implemented the ESR [2], and then compared the performance of these two initial methods using test images from 300W dataset.

**Table 3. Compare of Two Networks**

| Table Head      | Initialization type | Mean shape | Our result shape |
|-----------------|---------------------|------------|------------------|
| Mean error      |                      | 8.33       | 6.98             |

TABLE III. Demonstrates that the performance can be improved using our robust initialization. We find that this robust initialization has great influence especially on large pose faces and inaccurate detection boxes. For large pose faces, the mean shape usually far away from the ground truth shape, so it is easily to drop into local optimization during the regression. As for the inaccurate detection boxes, the mean shape is inaccurate too because it is calculated from the inaccurate boxes, so the performance is usually poor. In a word, using our method as initialization can give the regression a right start to obtain better result, which can avoid dropping into local optimize caused by the bad initialization. Fig 3 shows that our method is robust to the situations of inaccurate detection box, large pose variations and occlusion.

**Figure 3.** Compares the result with different initialization. The first row use the mean shape as initial shape, and the second row use our method to improve the robustness. And the green box is the detected face box.

5. Conclusion and Future Work

We proposed a multi-task convolution network that can works well for face alignment even in the case of challenging datasets. The experimental results proved the effectiveness of learning with pose estimate. Our method can also give good initialization instead of mean shape to improve the accuracy of traditional methods. In the future, we will focus on multi-task learning in other vision works.

6. References

[1] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. TPAMI, 23(6):681–685, 2001.

[2] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. International Journal of Computer Vision, 107(2):177–190, 2014.

[3] D. Miller, I. Kemelmacher-Shlizerman, and S. M. Seitz. Megaface: A million faces for recognition at scale. arXiv preprint arXiv:1505.02108, 2015.

[4] A. Datta, R. Feris, and D. Vaquero. Hierarchical ranking of facial attributes. In AFGR, pp 36–42. IEEE, 2011.

[5] Cootes, T.F., Ionita, M.C., Lindner, C., Sauer, P.: Robust and accurate shape model fitting using random forest regression voting. In: ECCV, pp. 278–291 (2012)

[6] C. Chen, A. Dantcheva, and A. Ross. Automatic facial makeup detection with application in face recognition. In ICB, pp 1–8, 2013.
[7] A. Datta, R. Feris, and D. Vaquero. Hierarchical ranking of facial attributes. In AFGR, pp 36–42. IEEE, 2011
[8] B. M. Smith, J. Brandt, Z. Lin, and L. Zhang. Nonparametric context modeling of local appearance for pose- and expression-robust facial landmark localization. In CVPR, 2014.
[9] L. Breiman. Random forests. Machine learning, 45(1):5–32, 2001.
[10] S. Zhu, C. Li, C. C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4998–5006, 2015.
[11] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In ICCVW, pp 397–403, 2013.
[12] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In CVPR, pp 545–552, 2011.
[13] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In ECCV, pp 679–692, 2012
[14] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In CVPR, pp 3444–3451, 2013.
[15] X. P. Burgos-Artizzu, P. Perona, and P. Dollar. Robust face landmark estimation under occlusion. In ICCV, pp 1513–1520, 2013.
[16] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In CVPR, pp 532–539, 2013.
[17] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine autoencoder networks (CFAN) for real-time face alignment. In ECCV, pp 1–16, 2014
[18] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In CVPR, 2014
[19] Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: CVPR. pp. 3476–3483 (2013)
[20] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In ECCV, pp 94–108, 2014.
[21] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. arXiv preprint arXiv:1604.02878, 2016.
[22] Kostinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: ICCV Workshops. pp. 2144–2151 (2011)
[23] Caruana, R.: Multitask learning. Machine learning 28(1), 41–75 (1997)
[24] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In CVPR, pages 2879–2886, 2012.
[25] X. P. Burgos-Artizzu, P. Perona, and P. Dollar. Robust face landmark estimation under occlusion. In ICCV, pp 1513–1520, 2013
[26] S. Xiao, S. Yan, and A. Kassim. Facial landmark detection via progressive initialization. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 33–40, 2015