Face Alignment Based on Two-Stage Localization

Xuan Zhang, Yang Sun and Wenduo He

1Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing, China, Email: xuanzhang@tsinghua.edu.cn
2Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing, China, Email: sunyang15@mails.tsinghua.edu.cn
3Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing, China, Email: hewd@cernet.edu.cn

Abstract. The traditional cascade regression model could easily drop into local optimum when the initial localization is far from real shape. This paper proposes a face alignment scheme based on two-stage localization. In the face detection stage, the multi-task CNN is adopted to get the information of landmarks and pose as the initial localization for second stage. In the second stage, the cascade regression based on random forest is implemented for face alignment, and the localization from first stage is used as the initial coarse localization for regression model to solve the problem of initialization sensitivity. The scheme can achieve excellent performance even in the case of large pose variations and occlusion. Compared with the traditional method, experiments show that our algorithm can get state-of-art results in real-time on benchmark of 300W dataset.

1. Introduction

Face alignment [1, 2, 3, 4, 5] is the process of getting the position of face landmarks through image processing and analysis. It is not only the most important part of face recognition [6], but also play an important role in 3D face reconstruction, attribute analysis [7], and the quality of face alignment will greatly affect the follow-up tasks. Although many algorithms were proposed in the past decades, face alignment still has much work to do due to the influence of large pose variations and occlusion.

The traditional scheme based on cascade regression model is easy to fall into local optimum when the initial localization shape is not so well, especially when the face is in large pose variations and occlusion. That means the quality of initial localization will affect the accuracy of face alignment when using regression model. Usually the face alignment is carried out after the face detection, the two tasks are treated separately and the face detection could not provide help for face alignment. Inspired by the multi-task learning in CNN [20, 23], we explore if we can get auxiliary information in face detection as the initial localization which can improve the accuracy of face alignment.

In this paper, we proposed a new two-stage scheme to establish a connection between face detection and face alignment. In face detection stage, we use multi-task convolutional neural network to get the information of face landmarks as the initial localization for second stage. In the second stage the random forest [9] is introduced to implement a cascade regression model for face alignment, and first stage provide a better initial face localization for the regression model.

Fig 1 shows the different results of using the mean face shape and using the information of landmark and pose as the initial localization. The first line uses the average position as the initial localization. It can be seen that when such an initial localization is far from the actual localization, the features near the key points may be deviated which make it difficult to regress to the real face shape. When the coarse information of landmark and pose from first stage is used for second stage, we can...
get better initialization close to the real localization which can avoid being trapped into local optimum and makes it easier to get a better performance.

![Figure 1. The different results of using mean shape and coarse initial position](image)

The contributions are as follows: We proposed a two-stage method to improve the performance of face alignment even in challenging benchmarks. The multi-task CNN is adopted in detection stage to get the initial localization for regression stage to solve the problem of initialization sensitivity [8]. We can get start-of-art performance even in the case of heavy pose variations and occlusion.

2. Related Work
In the past decades, many methods for face alignment have been proposed. We can divide the existing methods for face alignment into three categories based on their principle: traditional methods based on active shape models, methods based on cascade regression models, and methods based on deep learning. The cascade regression model is the most popular method because of its fast and accurate. The cascade regression model learns the relationship between the human face shape and the real shape through image features, and regress to the real location coordinates. Piotr et al. [21] presents the concept of a cascading regression model at the 2010 CVPR. At the same time, many similar methods have also been proposed. For example, RCPR [15] solves the problem of target occlusion on the basis of regression; Ren proposes an efficient LBF [18] feature, which combines local binary features into global features and more efficient than previous methods.

Although the cascade regression model [22, 24] has achieved good results in face alignment, it still has a fundamental flaw: sensitive to the initial shape. When the given initial shape is far from the real shape, it usually fall into local optimum. At present, many scholars pay attention to this problem and propose many solutions. SDM [16] uses the mean shape as the initial shape. Although it performs well under most conditions, it still cannot handle the case of large angles. RCPR proposes to run the first 10% regression function for initialization problems, and judges it according to the regression variance and threshold, or it will be re-initialized randomly. ESR [2] uses multiple initialization shapes to run the algorithm multiple times during the test, and finally uses the average of the output results as the predicted shape. These methods only reduce the initialization-sensitive problem and not fundamentally solved. Therefore, the initialization-sensitive problem still needs people to continue to explore.

In addition, Multi-task learning has been proved that it can improve the task performance effectively. Reference [14] proposes the method of MTCNN which train the face detection and face alignment jointly. Inspired by this work, we change the output of network and get the information of face landmarks and pose together with the task of face detection. So we can use the coarse landmarks and pose information to solve the problem of initialization sensitivity which build a bridge in face detection and face alignment.

3. The Face Alignment Based on Two-Stage Localization
The two-stage framework for face alignment is introduced as Fig 2. In the first stage the multi-task convolutional neural network is proposed to predict the coarse facial landmarks and pose as the
auxiliary task of face detection. In the second stage a cascade regression model based on random forest is implemented for face alignment, and the landmark and pose information from the first stage is provided as the initial coarse localization for the regression model. Our method combines the two tasks to provide better alignment results.

3.1. Multi-Task Learning: Face, Offset, Landmarks and Pose

In the process of face detection, we have four tasks to train: face detection, offset regression, landmarks location and pose angle.

Face detection is the main task of CNN, which can be described as a two-classification problem, and the loss function is described as following:

$$L^{cls} = -\sum_{i=1}^{m} (y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)})(1 - \log h(x^{(i)})))$$  \hspace{1cm} (1)$$

Where $L^{cls}$ is the loss of face detection, $y^{(i)}$ is the label of the training example, $h(x^{(i)})$ is the sigmoid function.

The other three tasks are all auxiliary tasks, which are treated as regression problem; we use the Euclidean Loss to minimize the loss between the label and prediction.

$$L = \frac{1}{2N} \sum_{i=1}^{N} \| y_i - \hat{y}_i \|^2$$  \hspace{1cm} (2)$$

We reorganize the loss function of the four tasks described in the previous and obtain the equation (3). Given training sample $\{x^m_i, x^n_i\}$, which $a \in A$ represents each auxiliary task, including offset regression, landmark and angle prediction, it can be seen that the training samples required for each task may be different, $\{y^m_i, y^n_i\}$ are the actual labels of the main task and each auxiliary task. It should be noted that the learning rates of the tasks are different due to their different complexity. And the main task has the bigger value than auxiliary tasks.

$$\arg \min_{\omega^m, \omega^n} (-\sum_{i=1}^{N} \log(p(y_i^m | x_i^m, W^m)) + \frac{1}{2} \sum_{e \in A} \sum_{i=1}^{N} \gamma^e \| y_i^e - f(x_i^e; W^e) \|^2 + \sum_{j=1}^{p} \| W \|^2)$$  \hspace{1cm} (3)$$

In response to this situation, we propose a three-step training method called “auxiliary tasks early-stop and refine”.

At the beginning, all tasks are trained and the mutual suppression of multi-task will prevent the training process falling into overfitting. As the training goes on, the auxiliary tasks start to converge quickly. After the convergence of auxiliary task training, the continuing training of auxiliary task may cause trouble to main task. So we need to stop the auxiliary task training in the early stage after the auxiliary tasks convergence.

After stopping the auxiliary tasks training, we continue to train the main task. At this stage, the back propagation training only follows the loss function of face classification, so the training of the main task can be converged quickly. When the main task converged, the main task's training process is
completed. However, the auxiliary task did not participate in training process. During this process, the parameters of the shared part of network were changed, so the parameters on the branch part of the auxiliary tasks did not match with shared parameters. To solve this problem, we adopted a "refine" strategy that make the parameters in branch part better match the shared parameters. In this process, we don't want to change the shared parameters because the main task has been trained. Therefore, in this process we fixed the shared parameters and only updated the branches of the following auxiliary tasks. After a small number of iterations, the auxiliary task can achieve better performance.

3.2. Network Structure

We designed two convolutional neural networks to realize the multi-task face detection. Inspired by the idea of coarse-to-fine, the two stage network (C-Net and F-Net) is designed to speed up the face detection and auxiliary tasks. C-Net aims to filter out most of the non-face parts in the image and provide the offset of the box. F-Net aims to judge the candidate windows accurately, and output the information of landmarks and pose angle. To reduce the influence of occlusion, lighting and other effects, we use large scale datasets to train networks. In C-Net, we add offset regression to provide more stable windows, and in F-Net we use multi-task learning to improve the accuracy of detection, and also output the information of landmarks and angle as the initial localization for face alignment.

![Figure 3. the structure of C-Net](image)

As Fig 3 shows, C-Net’s input size is 12*12 and contains three convolutional layers. All the kernel size is 3*3 and followed by the ReLU activation function. There is only one pooling layer in the network, because we use full convolutional network to predict the candidate position. The final output layer is our tasks: face detection and offset of the candidate box. The digital under each feature map is the characteristic dimension of each layer after forward process. When testing, we take an RGB image as input, subtract the mean value and normalize it to the range of [0, 1], then output the probability value and offset of each candidate box.

![Figure 4. the structure of F-Net](image)

F-Net is a deeper convolutional network based on multi-task learning and outputs three tasks simultaneously. F-Net has larger input size which can extracted the information more effectively. And we add dropout layer to avoid overfitting. The tasks of pose and landmark use different branch layer. F-Net is much deeper than C-Net and more accurate to output the tasks, but it will take more time to
compute. So the two networks play their respective strengths and realize face detection in real time and output the information of landmarks and pose. The structure is shown as Fig 4.

3.3. Cascade Regression Model Based on Random Forest

Cascade regression model learns the relationship between image features and regression targets by training large number of weak classifiers, and then combine these weak classifiers in to a powerful regression model. We use many random forests to implement the regression model. Each iterative will predict the offset relative to current shape and we update shape through all forests to return to an optimal position. Formula (4) shows the update process of landmark shape.

\[ S_{t+1} = S_t + R(I, S_t) \]  

(4)

Where \( S_{t+1} \) is the predict shape in \( t+1 \) iterations, and \( R \) is the offset function.

In order to train an efficient and rapid regression model, we chose the pixel index feature which compute the difference of two pixel values. When giving enough examples, this feature is fast and efficient. For example, the pixel between eye and nose can capture the change of light and darkness.

But in test image, face shape will be different with the change of angles and expressions. In order to keep the geometric invariance, we transform the current shape to mean shape through scaling and rotation which is much efficient than transform the full image.

We randomly select 400 point in the image as the pixel pool and chose the best pair among 20 pairs as the branch parameters in tree nodes. In this process, we give greater probability to the pairs with small distance.

In order to train the tree separately, we randomly select the example and give them a random initialization. In one tree node, we expect to find a best parameter to separate the examples:

\[ E(\Gamma, \phi) = \sum_{z \in \{l, r\}} \sum_{t \in \Gamma_{\phi, z}} \| d_t - \mu_{\phi, z} \|^2 \]  

(5)

Where \( \Gamma_{\phi, z} \) is these examples that separated to the left child node using parameter \( \phi \), \( d_t \) is the offset between real shape and predict shape, \( \mu_{\phi, z} \) is the mean value of all example’s offset in the left child node. And we use different examples to train every node to get an efficient regression model.

3.4. Initialize with Our Coarse Location

In order to solve the problem of initialization sensitivity, we adopt the output of multi-task convolutional neural network as the initial shape for face alignment. The output can provide the coarse localization that near to the real face shape, which can avoid falling into local optimum. The information of angle was also be used as judgment of the initialization. If the difference between them is too large, the prediction may be inaccurate, the position obtained by the previous frame may be used as initial position. We can use the result of previous frame to regress more accurately.

4. Experiments and Analysis

4.1. Training Matters

The training examples are come from 300-W [11] dataset which includes HELEN [13], LFPW [12], IBUG and AFW [25] subsets. These datasets provide the image with the label of 68 landmarks and we use 3148 images as training set, and increase dataset use the methods of rotation and panning. Left 689 images divide into test set. Angle example is come from AFLW dataset with three angle values. And the examples for face comes from WIDER FACE which concludes 393,703 face images.

For the landmark task, we transform this position to \([-1, 1]\) and use the inter-ocular distance to normalize the error between estimated shape and real shape.
4.2. The Effectiveness Of Strategies: Early-Stop and Refine
To verify the effectiveness of the early-stop and refine strategy, we traced the loss curves of face tasks and auxiliary tasks in the training process.

![Figure 5](image1.png)

**Figure 5.** loss change curve of face detection in early-stop process

Fig 5 shows the decline curve of the loss function in face detection task. After 10,000 iterations, the falling speed slows down which indicate that the auxiliary task can’t provide help anymore, so we stop the auxiliary task and the network refocuses on the training of the main task, we can find that the loss value continues to decrease and achieve good result.

After the training process of main task is completed, the auxiliary task branch does not match the shared part, so the auxiliary task works not well. To solve this problem, we will fix the shared part and refine the branch parameters. Fig 6 is the loss curve of refine process. Fig 7 shows the performance of multi-task learning, although the landmark is not so accurate, but it gives a coarse initialization.

![Figure 6](image2.png)

**Figure 6.** loss change curve of angle in refine process.

![Figure 7](image3.png)

**Figure 7.** the output information of landmarks and angle
4.3. Face Alignment Analysis

In order to verify the effectiveness of this algorithm, we tested the proposed algorithm on 300W test set. Table 1 is a comparison of the normalized error based on the coarse positioning cascading method and other mainstream methods proposed. The data are from the error published by the paper.

| Method      | Common | Challenging | Full  |
|-------------|--------|-------------|-------|
| SDM         | 5.57   | 15.40       | 7.50  |
| ESR         | 5.28   | 17.0        | 7.58  |
| LBF         | 4.95   | 11.98       | 6.93  |
| DRA[17]     | 4.36   | 7.56        | 4.99  |
| RAR[19]     | 4.12   | 8.35        | 4.94  |
| OUR Method  | 4.83   | 8.68        | 5.40  |

Table 1. Evaluate with other methods.

Our method improves the performance compared with the mainstream cascade regression models especially on challenging datasets. The coarse initial localization gives the model lots of information to regress to the real shape more accurately. In addition, compared with more complex model such as RAR, our scheme can achieve faster processing speed in real time by taking the task of detection and alignment together. Fig 8 represents the regression results in 300W dataset, it shows that our scheme performs excellent in different scenarios.

![Figure 8](image_url)

Figure 8. face alignment results on 300W dataset

5. Conclusion and future Work

We propose a two-stage scheme which integrates face detection and alignment together. In detection stage the multi-task convolutional neural network is adopted to get the landmarks and pose which will be used as the initial localization for the second stage. In the second stage, regression model is implemented for face alignment by using the initial coarse localization from the first stage. Experiment shows our scheme can achieve excellent performance in various complex scenes. In the future, we will focus on 3D models of the face alignment.

6. Reference

[1] T. F. Cootes, G. J. Edwards. 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. Megaf ace: 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 1234567890 ISAI2017 IOP Publishing IOP Conf. Series: Journal of Physics: Conf. Series 887 (2017) 012079 doi :10.1088/1742-6596/887/1/012079

[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 CVPR, 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] Kaipeng Zhang, Zhanpeng Zhang. Joint Face Detection and Alignment using Multi-task Cascade Convolutional Networks. IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, 2016.

[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] Lv J, Shao X, Xing J, et al. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection [C]// CVPR. 2017.

[18] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In CVPR, 2014.

[19] Xiao S, Feng J, Xing J, et al. Robust facial landmark detection via recurrent attentive-refinement networks[19]//European conference on computer vision. Springer, Cham, 2016: 57-72.

[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] Dollár P, Welinder P, Perona P. Cascaded pose regression[21]//Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010: 1078-1085.

[22] Kostinger, M., Wohlhart, P., Roth, P: 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. P. Burgos-Artizzu, P. Perona, and P. Dollar. Robust face landmark estimation under occlusion. In ICCV, pp 1513–1520, 2013.

[25] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In CVPR, pages 2879–2886, 2012.