Fast Multi-view Face Alignment via Multi-task Auto-encoders

Qi Li, Zhenan Sun and Ran He
Center for Research on Intelligent Perception and Computing
National Laboratory of Pattern Recognition, Institute of Automation
CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences
Beijing 100190, China
{qli, znsun, rhe}@nlpr.ia.ac.cn

Abstract

Face alignment is an important problem in computer vision. It is still an open problem due to the variations of facial attributes (e.g.*, head pose, facial expression, illumination variation). Many studies have shown that face alignment and facial attribute analysis are often correlated. This paper develops a two-stage multi-task Auto-encoders framework for fast face alignment by incorporating head pose information to handle large view variations. In the first and second stages, multi-task Auto-encoders are used to roughly locate and further refine facial landmark locations with related pose information, respectively. Besides, the shape constraint is naturally encoded into our two-stage face alignment framework to preserve facial structures. A coarse-to-fine strategy is adopted to refine the facial landmark results with the shape constraint. Furthermore, the computational cost of our method is much lower than its deep learning competitors. Experimental results on various challenging datasets show the effectiveness of the proposed method.*

1. Introduction

Face alignment (or facial landmark detection) is a fundamental problem in computer vision. It refers to locating the semantic structural facial landmarks such as eyes, nose and mouth. It plays an important role in many applications, *e.g.*, face detection and recognition, face clustering, facial expression analysis[6, 12]. Although tremendous efforts have been devoted to the development of accurate and robust face alignment algorithms, it is still one of the most challenging problems due to facial variations of expressions, illumination and head poses. Face alignment has received an increasing attention in recent years. It can be roughly divided into three categories: Parameterized Appearance Models, Constrained Local Models and Regression based Models.

The first type is Parameterized Appearance Models. Active Appearance Model (AAM) [3] is the representative one. It uses Principle Component Analysis (PCA) to generate a statistical model of shape and texture variations, respectively. Specifically, face images are first warped to a common coordinate frame. Then the shape basis and appearance basis can be learned by performing PCA on the warped images. Consequently, the face shape can be estimated by utilizing the shape basis and appearance basis. One of the key problems of AAM based models is how to optimize a non-linear objective function. The first is to learn it via regression [3, 17]. Another way of AAM fitting is through non-linear least squares, which can be solved iteratively using Gaussian-Newton optimization [14]. One limitation of Parameterized Appearance Models is the limited expressive ability to capture complex and subtle face variations. Recently, some works try to provide more efficient solutions so that they are applicable in the wild [7, 20, 21].

Another type of face alignment is Part-based Deformable Models. Usually, a holistic appearance model is not enough for finding exact landmark locations. So Part-based Deformable Models have been proposed. Constrained Local Model (CLM) [4] is one of them for face alignment. It first samples some regions and extracts some local features from the image around the current facial landmark locations. Then some "patch experts" are trained based on the extracted local features. The landmark locations can be obtained by optimizing the local likelihood of each part times a global prior for a testing image. One advantage of Part-based Deformable Models is that it is discriminative, and well generalized to unseen appearance variations, global illumination variations and occlusions. Different "patch experts" and optimization methods have been proposed in recent years [18, 31, 24].

Regression Based Models learn a mapping from image features to landmark locations directly. The representative approaches are Explicit Shape Regression (ESR) [2], Supervised Descend Method (SDM) [23]. Different from

AAM that only learns a linear representation between the increment of model parameters and appearance differences, ESR learns a vertical regression function to infer the whole face shape, and minimizes alignment errors explicitly. The shape constraint is incorporated into the regressor in a cascaded learning framework. The above properties enable ESR to learn a flexible model with strong expressive power from large-scale training data. Compared with ESR, SDM tries to learn gradient descend directions to approximate the Jacobian and Hessian for alignment with a sequence of linear models. Random forests and Gaussian Process have also been widely used for face alignment in recent years [5, 16, 11].

In many real world applications, misaligned face images are often non-frontal. Zhou *et al*. used a Bayesian mixture model for multi-view face alignment [29]. Zhu *et al*. proposed a deformable part model for multi-view facial landmark detection, which achieves state-of-the-art performance on various datasets [31]. Dantone *et al*. proposed the conditional random forests for real-time face alignment [5]. Furthermore, Zhao *et al*. presented an iterative multi-output random forests method to jointly estimate head pose, facial expressions and facial landmarks [28]. To further improve multi-view face alignment accuracy in the wild, this paper integrates several useful techniques into a two-stage multi-task Auto-encoders framework. First, head pose estimation is served as an auxiliary task to improve the generalization performance of the main alignment task. Then we develop a two stage framework with the coarse-to-fine strategy for accurate multi-view face alignment. What's more, the computational cost of our method is much lower than other deep learning based methods. Experimental results have shown that our method achieves state-of-the-art performance on various challenging datasets.

2. Related work

Deep learning based face alignment methods have been well studied in recent years [9, 19, 26, 25, 30]. Sun *et al*. have proposed a three-level Convolutional Neural Networks for face alignment [19]. It has achieved an impressive performance. First, a global network is trained over the entire face to locate each facial landmark. The networks at the following level are trained over the local small regions to refine initial prediction results. The cascade global and local face alignment method has achieved accurate and reliable results. Face alignment is not a stand-alone problem. It is influenced by many other factors, *e.g*., head pose variations, facial expression variations. Zhang *et al*. have further boosted the performance using only one Convolutional Neural Network through multi-task learning [26]. The motivation behind their approach is simple. They jointly estimates the facial landmark locations together with the correlated facial attributes. However, only the low-resolution face images are used for training their algorithm. The detailed facial information has been lost. Besides, the shape constraint has not been fully explored in their models. Zhang *et al*. have cascaded coarse-to-fine Auto-encoders for face alignment [25]. One drawback of their method is that they have not considered other facial attributes which may affect the alignment performance. 3D information is also exploited for face alignment in recent works. A dense 3D face model is fitted to the face image via Convolutional Neural Network for face alignment [30]. Jourabloo and Liu formulate the face alignment as a 3D Morphable Model fitting process [9]. 3D Morphable Model and cascade Convolutional Neural Network regressors are combined to achieve a better localization results.

3. The proposed method

In this section, the traditional cascade regression method for face alignment is reviewed firstly. Then the technical details are presented. Finally, we demonstrate the relationship between previous face alignment methods and our method.

3.1. Cascade regression for face alignment

Face alignment is usually formulated as a classical regression problem. Given a face image, its facial landmarks are often represented as a vector of $2D$ coordinates, *i.e.*, $S = (u_1, v_1, u_2, v_2, ..., u_l, v_l, ..., u_L, v_L)$, where (u_l, v_l) is the coordinate of the $l - th$ landmark location, L is the number of total landmarks. Given N training face images $\{x_1, x_2, ..., x_N\}$, the initial face shape is formulated as $\{S_1^0, S_2^0, ..., S_N^0\}$, and the ground-truth face shape is represented as $\{S_0^g, S_1^g, ..., S_N^g\}$. Regression based face alignment methods aim to learn a regressor f to minimize the following objective function:

$$\sum_{i=1}^{N} \left\| f\left(x_i, S_i^0\right) - S_i^g \right\|_2^2, \tag{1}$$

where f predicts the new shape based on the initial shape for each image. Equation 1 is a complex non-linear problem, and it is very difficult to learn f directly. We usually solve Equation 1 in a cascade fashion. The face shape can be estimated from a initial shape S_i^0, and progressively refines the shape by a cascade of E regressors, *i.e*., we can divide f into a series of simple regressors $\left\{f^1, f^2, ..., f^E\right\}$. Each regressor f^e refines the shape by producing an update ΔS^e from the previous shape, and then updates the previous shape as: $S^e = S^{e-1} + \Delta S^e,\ (e = 1, ..., E)$. The shape update ΔS^e is computed from the regressor f^e, which takes the form as: $\Delta S^e = f^e\left(x, S^{e-1}\right)$.

3.2. The first stage for face alignment

Cascade regression methods have achieved an impressive alignment performance. However, there are still some

challenging factors for cascade regression methods, *e.g.*, large head pose variations, various facial expressions and illumination variations. Among them, head pose variations is one of the most challenging factors. Traditional methods treat head pose estimation and face alignment as two separate problems. A head pose estimator is first trained to divide the face images into several views. Then different face alignment models are trained separately according to different views. The traditional methods rely on the accuracy of the head pose estimator, which is also an opening problem for $2D$ face images in the wild. Besides, both the model complexity and computational cost are increased. We argue that face alignment and head pose estimation is similar to the "chicken-and-egg" problem. They are closely related with each other. Dealing with them together offers significant advantages over treating them separately. Recent works have also shown that head pose estimation can boost the performance of face alignment [26, 28].

In this paper, we use multi-task Auto-encoders for face alignment with related pose information. Similar to [26], our method is also a heterogeneous multi-task learning problem. Face alignment is taken as the main task, while head pose estimation is the related auxiliary task. In the first stage, we take the low-resolution images $\left\{x_i^0\right\}_{i=1}^N$ as input. Suppose that the ground-truth face shape and head pose are represented as $\{S_i^g, y_i^g\}_{i=1}^N$. The stacked Auto-encoders are used to learn a non-linear mapping from the image pixels to the final landmark locations. The original image can be projected into high level image representations gradually by learning a sequence of T non-linear mappings:

$$x^t = \sigma\left(W^t x^{t-1} + b^t\right), \quad t = 1, ..., T-1, \tag{2}$$

where $x^t = \{x_i^t\}_{i=1}^N$, W^t is the projection matrix, b^t is the bias term, and $\sigma(\cdot)$ is the sigmoid function.

The overall objective function of our first stage framework contains two parts. The first part is used to project the high level image representations to the final coordinates of the facial landmark locations. The second part is used to project the high level image representations to the head pose classifications. The proposed multi-task learning framework can be formulated as follows,

$$J = J_r\left(S^g, f\left(x^{T-1}; W^r\right)\right) + J_l\left(y^p - y^g\right), \tag{3}$$

where $S^g = \{S_i^g\}_{i=1}^N$, $y^g = \{y_i^g\}_{i=1}^N$, and $y^p = \{y_i^p\}_{i=1}^N$ denotes the predicted head pose. The first term $J_r\left(S^g, f\left(x^{T-1}; W^r\right)\right)$ is a regression task for face alignment, which can be defined as a least square loss function:

$$\sum_{i=1}^{N} \left\| S_i^g - f\left(x_i^{T-1}; W^r\right)\right\|_2^2, \tag{4}$$

where $f\left(x_i^{T-1}; W^r\right) = W^r x_i^{T-1}$, which means we use a linear function to project the high level image representations to the output coordinates of the facial landmark locations. The second term is a classification task for head pose estimation, which can be represented as a cross-entropy loss function:

$$-\sum_{i=1}^{N} y_i^g \log\left(p\left(y_i^p | x_i^{T-1}; W^l\right)\right) \tag{5}$$

where

$$\begin{aligned} &p\left(y_i^p = m | x_i^{T-1}; W^l\right) \\ &= \frac{\exp\left(W_m^l x_i^{T-1}\right)}{\sum_{j=1}^{K} \exp\left(W_j^l x_i^{T-1}\right)}, (m = 1, ..., K), \end{aligned} \tag{6}$$

K denotes the number of head pose categories, W_j^l represents the $j-th$ column of W^l. The objective function of the first stacked Auto-encoders can be rewritten as:

$$\begin{aligned} &\min_W \sum_{i=1}^{N} \left\| S_i^g - f\left(x_i^{T-1}; W^r\right)\right\|_2^2 \\ &-\lambda_1 \sum_{i=1}^{N} y_i^g \log\left(p\left(y_i^p | x_i^{T-1}; W^l\right)\right) + \lambda_2 \sum_{t=1}^{T} \|W\|_F^2, \end{aligned} \tag{7}$$

where λ_1 is a balance term which denotes the relative importance of the auxiliary task, $\sum_{t=1}^{T} \|W\|_F^2$ is a regularization term which prevents the Auto-encoders from overfitting $\left(W = \left\{W^r, W^l\right\}\right)$.

Equation 7 is a non-convex problem. We adopt the stochastic gradient descend method to solve this problem. The partial derivatives of the objective function with respect to the weight matrix are:

$$\begin{aligned} \frac{\partial J}{\partial W^r} &= \left(W^r x^{T-1} - S^g\right) x^{T-1}, \\ \frac{\partial J}{\partial W^l} &= \left(p\left(y | x^{T-1}; W^l\right) - y^g\right) x^{T-1}. \end{aligned} \tag{8}$$

Based on the partial derivatives, we can update the weight matrix as:

$$\begin{aligned} W^r &= W^r - \eta \frac{\partial J}{\partial W^r}, \\ W^l &= W^l - \eta \frac{\partial J}{\partial W^l}, \end{aligned} \tag{9}$$

where η is the learning rate. Then we can compute the gradients layer by layer, and follow the standard back-propagation to optimize Equation 7.

Note that Equation 7 is actually a heterogeneous multi-task learning problem. Face alignment and head pose estimation have different loss functions and thus may have different convergence rates. In order to guarantee the convergence of the main task, we propose a simple yet effective solution. We decrease λ_1 in Equation 7 gradually during the optimization process. Specifically, we use a relatively large coefficient at the early steps to induce the Auto-encoders to consider the auxiliary task so that it can handle large pose

variations. While at the later steps, we could use a relatively small coefficient to ensure the convergence of the main task. Zhang *et al*. used "early stopping" to solve the heterogeneous multi-task learning problem [26]. Compared with their method, the gradient descend direction of our method is more consistent and smoother. Note that Zhang *et al*. adopted dynamic coefficient strategy in the later work [27]. Compared with their strategy, the gradient descend direction of our method is more stable. Although our method is very simple to deal with heterogeneous multi-task learning problem, it is very effective in practical applications.

3.3. The second stage for face alignment

In order to reduce the computational cost and train a robust model for face alignment, a low resolution image is used to get a rough estimation of the initial facial landmark locations in the first stage. A cascade model is needed to further refine the initial landmark locations because of the following reasons. First, a single model is usually not powerful enough for accurate face alignment. Second, we haven't taken advantage of the local texture information in the high resolution face images. Hence we further refine the facial landmarks on the high resolution face images for the second stage. If we use raw pixels as the input, we need a "deeper" network to learn the features automatically, which needs many human labeled training samples. Thus, we cannot feed the high resolution images to this stage directly because of the limited training samples and redundant global image information. An effective way is to first extract some shape-indexed features in a small neighborhood of the initial landmark locations, and then concatenate them together to form a powerful feature representation [2, 16, 25]. These features can be used as the input to the second stage. The deviation between the initial landmark locations and ground-truth landmark locations is used as the training labels.

Given the high resolution images $\left\{\bar{x}_i^0\right\}_{i=1}^N$, their initial and ground-truth landmark locations are denoted as $\left\{\bar{S}_i^0\right\}_{i=1}^N$ and $\left\{\bar{S}_i^g\right\}_{i=1}^N$, respectively. For the image $\bar{x}_i^0$, we extract the shape-indexed features as $h^0\left(\bar{x}_i^0;\bar{S}_i^0\right)$. Then the shape-indexed features are also projected into high level representation gradually by learning a sequence of T nonlinear mappings,

$$h^t = \sigma\left(W^t h^{t-1} + b^t\right), \quad t = 1, ..., T-1, \tag{10}$$

where $h^t = \left\{h^t\left(\bar{x}_i^t;\bar{S}_i^t\right)\right\}_{i=1}^N$ is the high level representation of the projected shape-indexed features, $\sigma(\cdot)$ is the sigmoid function. Similar to the first stage, the overall objective function for the second stage can be reformulated as:

$$\begin{aligned}&\min_W \sum_{i=1}^N \left\|\Delta S_i - f\left(h_i^{T-1};W^r\right)\right\|_2^2 \\ &-\lambda_1 \sum_{i=1}^N y_i^g \log\left(p\left(y_i^p|x_i^{T-1};W^l\right)\right) + \lambda_2 \sum_{t=1}^T \|W\|_F^2,\end{aligned} \tag{11}$$

where ΔS_i is the shape deviation between the predicted shape $\bar{S}_i^0$ and the ground-truth shape $\bar{S}_i^g$, $f\left(h_i^{T-1};W^r\right) = W^r h_i^{T-1}$ is a linear function to project the high level image representations to the face shape deviation, $\sum_{t=1}^T \|W\|_F^2$ is a regularization term $\left(W = \left\{W^r, W^l\right\}\right)$. The stochastic gradient descend method is also used to solve Equation 11.

Recent deep learning based methods have shown that unsupervised pre-training is an important technique that can improve the performance [8, 22]. Given a $j-th$ layer of the stacked Auto-encoders, we pre-train our method by minimizing the following reconstruction errors,

$$\min_{W_j} \sum_{i=1}^N \left\|\hat{a}_i^{t-1} - a_i^{t-1}\right\|_2^2 + \hat{\lambda}\left(\|W_j\|_F^2 + \left\|(W_j)'\right\|_F^2\right), \tag{12}$$

where $\hat{a}_i^t = f\left(W^t a_i^{t-1} + b^t\right)$, $\hat{a}_i^{t-1} = f\left((W^t)'\hat{a}_i^t + \hat{b}^t\right)$, $f(\cdot) = \sigma(\cdot)$ is the sigmoid function. We pre-train the stacked Auto-encoders in a layer-wise manner. Then the parameters are preserved to be the initialization of the weight matrix. Finally, we fine-tune the weight matrix using back-propagation. This strategy is proved to be better than the pure back-propagation with random parameter initialization [8].

4. Relation to previous work

There are several points when we use deep learning based methods for face alignment. The first one is to improve the generalization capacity of the network. The second one is to preserve shape constraint and local texture information for accurate face alignment. Considering the above points, we have proposed the multi-task Auto-encoders for coarse-to-fine face alignment. Note that compared with [25], head pose information is incorporated into our method. We argue that head pose estimation and facial landmark localization are two closely related problems, which inspires us to solve these two problems together. The objective function (Equation 7 and Equation 11) can influence the weight matrix of the whole network by back propagating the errors. It makes the Auto-encoders more robust for the challenging face alignment problem. Besides, we need less number of Auto-encoders than [25] because of the improved generalization capability. Hence, our method is more efficient than [25].

Although Zhang *et al*. have done similar work to verify the effect of multi-task learning for Convolutional Neu-

ral Network (CNN) [26], they haven't considered the shape constraint and the local texture information. Besides, CNN relies on a large number of training images to learn the features. The cost for acquisition of human labeled face images restricts the wide usage of this method. Our method has the following advantages over [26]. First, it needs less training samples than [26], which means we can still get a satisfied performance even with limited training samples. Besides, the shape constraint, which is very important for face alignment, is naturally encoded into our two-stage framework. Furthermore, the shape-indexed features is used in our method for face alignment. Such high level features are carefully designed so that they are very likely to be mutually uncorrelated and to be complementary with each other. Compared with [26], we only consider the most related head pose information, which has an important impact on the multi-view face alignment results [26]. However, other facial attributes can be easily added to our method for face alignment.

5. Experimental results

5.1. Implementation details

The first stage of our method has four layers which is the same with [25] except for the final output layer. All of the images are resized to a resolution of 50×50. Then the raw pixels are used as the input units. The number of hidden units are 1600, 900 and 400, respectively. The second stage also has four layers. The SIFT features [13] which are extracted from 80×80 face images are used as the input units. The number of hidden units is 400, 200 and 80 respectively. The training dataset is the same with [26, 19], which consists of $10,000$ outdoor face images from the web. Each face image is annotated with five facial landmarks, together with the pose, gender, glass and smiling information. In order to train a robust model, we augment the training samples by small translation, rotation and scaling.

We have tested our method on the AFLW dataset [10] and AFW dataset [31]. The face images in the AFLW and AFW datasets formulate a more challenging scenario than other datasets (*e.g.*, XM2VTS [15]). AFLW dataset contains $24,386$ face images gathered from Flickr. $3,000$ face images are selected to test our algorithm, which is the same as [26]. AFW dataset contains 205 face images. Each face is labeled with 6 landmarks. However, some face images are annotated incompletely due to the challenging viewpoints. The images without 5 common facial points (center of eyes, tip of nose, mouth corners) are simply dropped. Finally we have a total number of 170 testing images on the AFW dataset. The normalized root mean squared error (NRMSE) is adopted to measure the face alignment performance. It is computed by dividing the root mean squared error by the bi-ocular distance. The cumulative error distribution curve is also used to evaluate the alignment results. It is given as a percentage of the face images of which the NRMSE is less than a specific value.

5.2. The effectiveness of multi-task learning

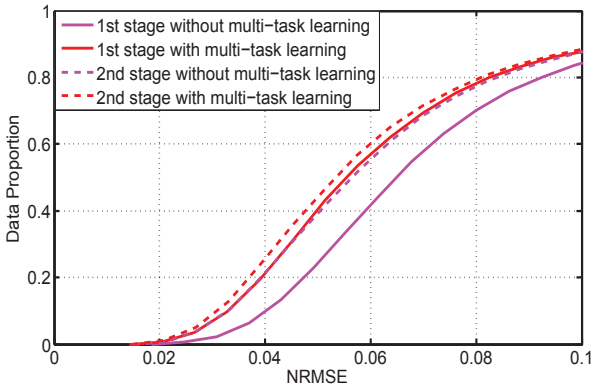

Figure 1. The cumulative error distribution curves of our method with and without multi-task learning on the AFLW database.

In order to verify the effectiveness of multi-task Auto-encoders for face alignment, we train face alignment models with and without multi-task learning for the first stage and second stage, respectively. The cumulative error distribution curves of different models are shown in Figure 1. As shown in Figure 1, we can get a significant improvement for face alignment with the multi-task learning in the first stage. Head pose information provides a strong multi-view face shape prior for learning Auto-encoders in the first stage. We get almost 10 percent improvement for face alignment with the multi-task learning.

Figure 1 also shows the alignment performance of the second stage with and without multi-task learning. As shown in Figure 1, we can also get a better performance with multi-task learning. However, the improvement is marginal compared with the first stage. The possible reason for the marginal improvement is that we already get a multi-view face shape prior after the first stage. Head pose information in this stage has a relatively small influence on the final performance compared with the first stage. However, the shape residue estimation can still benefit from the head pose estimation, which gives us about 4 percent improvement in the second stage.

5.3. Comparison with other deep learning based methods

In this section, we compare our method with other deep learning based methods on the AFLW dataset. The compared methods are TCDCN [26], CFAN [25], Cascaded CNN [19]. The results of different methods are shown in Figure 2. Compared with other deep learning based methods,

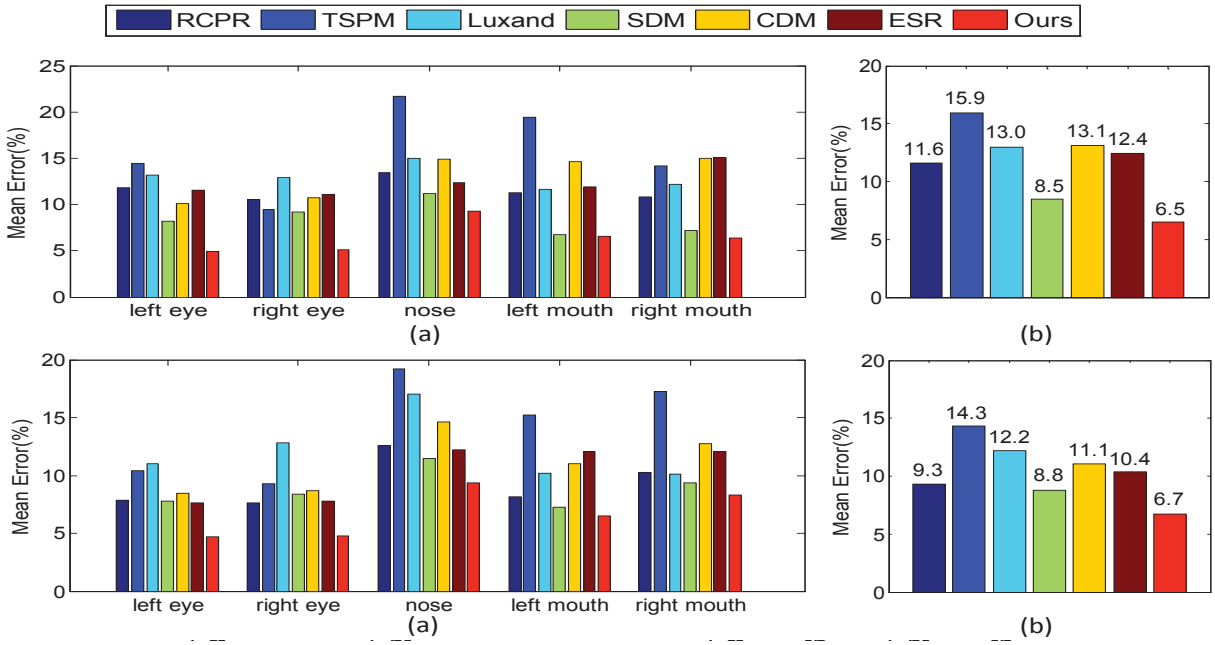

Figure 3. Face alignment results of RCPR, TSPM, Luxand, SDM, CDM, ESR and ours on the AFLW dataset (the first row) and the AFW dataset (the second row). (a) Mean errors of different landmarks. (b) The overall mean errors.

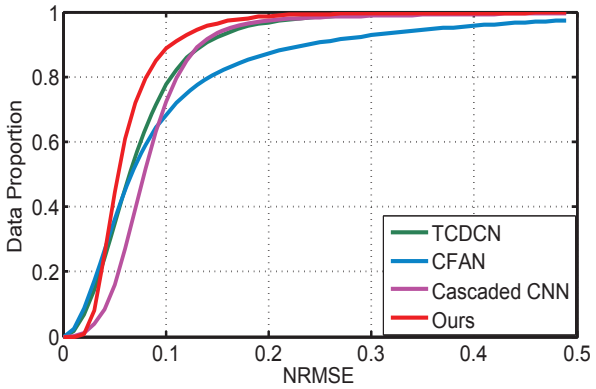

Figure 2. The cumulative error distribution curves of TCDCN, CFAN, Cascaded CNN and ours on the AFLW dataset. The alignment results of TCDCN, CFAN and Cacaded CNN are reported from [27].

CFAN performs not very well when NRMSE is larger than 0.1. The possible reason is that CFAN has not considered head pose information. Thus it is not very robust to large pose variations. TCDCN performs better than CFAN and Cascaded CNN under the multi-task learning framework. However, the shape constraint and local texture information have not been used fully in TCDCN. Our method performs better than CFAN, TCDCN, and Cascaded CNN. The reason is that our method incorporates multi-task information, *e.g.*, head pose information, into face alignment, which increases the generalization capabilities of the network. Besides the two-stage cascaded process preserves the shape constraint, which is very important for face alignment.

An important factor for evaluating different face alignment methods is the computational cost. We evaluate our method on the AFLW dataset and record its average computational time per image. Our method is running using matlab on a PC with 3.4 GHZ CPU. The running time and environment of our method are shown in Table 1. It can be seen from Table 1 that our method is an efficient one even with the unoptimized algorithm and codes. The first stage of our method takes about 0.5ms, while the second stage takes about 2.5ms. Our method has a relatively low computational cost compared with previous deep learning based face alignment methods.

Algorithm	Running time	Environment
Ours	3ms	Intel i7 CPU, matlab

Table 1. Running time and environment of our method.

5.4. Comparison with other face alignment methods

We also compare with other face alignment methods on the AFLW and AFW datasets. The compared methods include Robust Cascade Pose Regression (RCPR) [1], Tree Structured Part Model (TSPM) [31], Luxand face SDK, Explicit Shape Regression (ESR) [2], Cascaded De-

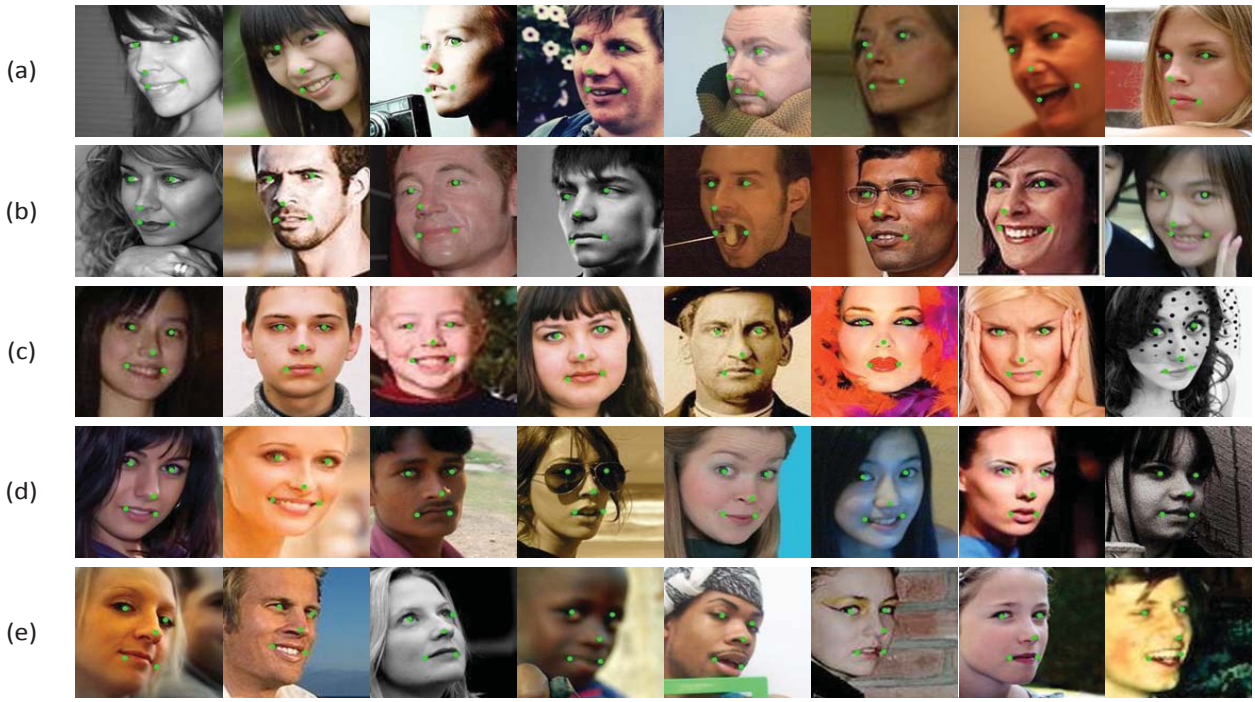

Figure 4. Face alignment results on the AFLW dataset with different head poses. (a) left profile. (b) left. (c) frontal. (d) right. (e) right profile.

formable Model (CDM) [24], Supervised Descend Method (SDM) [23].

Figure 3(a) presents the mean errors of different landmarks (*e.g.*, left eye, right eye) on the AFLW and AFW datasets. The overall mean errors of different methods are shown in Figure 3(b). From Figure 3, we can see that SDM achieves a better alignment performance than most of the other methods. While our algorithm outperforms SDM by a margin of almost 2 percent on both the challenging AFLW and AFW datasets. On the AFLW dataset, our average mean error is 6.5, while SDM is 8.5. On the AFW dataset, our average mean error is 6.7, while SDM is 8.8. We can also see from Figure 3(a) that the alignment error of nose is larger than other facial landmarks for almost all of the face alignment methods. One of the possible reasons is that the texture information around nose is not very informative. The learned or hand-crafted features may not work very well in these areas. Figure 3 have further shown that our method is robust to multi-view face images with large pose variations, which is a very challenging problem for face alignment.

Finally, we present some alignment examples on the AFLW dataset. Figure 4 has shown some examples with different head poses on the AFLW dataset. As shown in Figure 4, our method performs well with multi-view face images. We can locate the facial landmarks accurately with the left profile and right profile face images due to the usage of head pose information under the multi-task learning framework. Besides, our method is also robust to various occlusions and large facial expression variations because of the improved generalization ability.

6. Conclusions

This paper has developed a fast two-stage multi-task Auto-encoders framework for multi-view face alignment by integrating several useful alignment techniques. Head pose information and shape constraint have been naturally encoded into our framework. Deep learned features and hand-crafted features are combined to boost the alignment performance. Experimental results on the challenging AFLW and AFW datasets have shown that the proposed frame achieves state-of-the-art multi-view face alignment results. Future work is to apply the Auto-encoders framework to improve the alignment accuracy of more facial landmarks.

7. Acknowledgements

This work is funded by Beijing Municipal Science and Technology Commission (No.Z161100000216144).

References

[1] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *IEEE International Conference on Computer Vision*, pages 1513–1520, 2013.

[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] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):681–685, 2001.

[4] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *British Machine Vision Conference*, volume 1, page 3, 2006.

[5] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2578–2585, 2012.

[6] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

[7] R. He, S. Li, Z. Lei, and S. Liao. Coarse-to-fine statistical shape model by bayesian inference. In *Asian Conference on Computer Vision*, pages 54–64. 2007.

[8] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[9] A. Jourabloo and X. Liu. Large-pose face alignment via a cnn-based dense 3d model fitting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4188–4196, 2016.

[10] M. Köstinger, P. Wohlhart, P. M. Roth, and H. 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*, pages 2144–2151, 2011.

[11] D. Lee, H. Park, and C. D. Yoo. Face alignment using cascade gaussian process regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4204–4212, 2015.

[12] Q. Li, Z. Sun, Z. Lin, R. He, and T. Tan. Transformation invariant subspace clustering. *Pattern Recognition*, 2016.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[14] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[15] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *International conference on audio and video-based biometric person authentication*, volume 964, pages 965–966, 1999.

[16] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.

[17] J. Saragih and R. Göcke. Learning aam fitting through simulation. *Pattern Recognition*, 42(11):2628–2636, 2009.

[18] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.

[19] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013.

[20] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *International Conference on Computer Vision*, pages 593–600, 2013.

[21] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2014.

[22] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(6):3371–3408, 2010.

[23] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, 2013.

[24] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *IEEE International Conference on Computer Vision*, pages 1944–1951, 2013.

[25] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision*, pages 1–16. 2014.

[26] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. 2014.

[27] 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*, 38(5):918–930, 2016.

[28] X. Zhao, T.-K. Kim, and W. Luo. Unified face analysis by iterative multi-output random forests. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1765–1772, 2014.

[29] Y. Zhou, W. Zhang, X. Tang, and H. Shum. A bayesian mixture model for multi-view face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 741–746, 2005.

[30] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016.

[31] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, 2012.