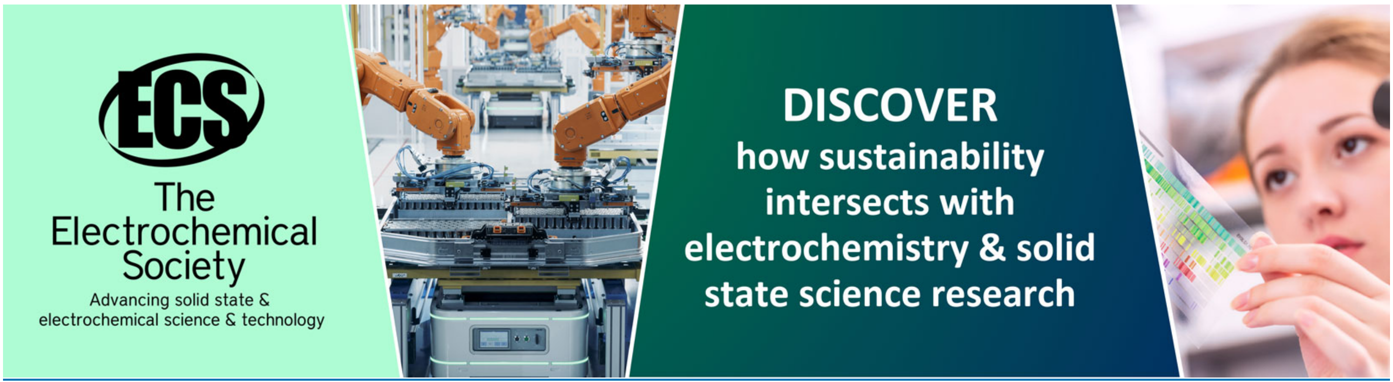
Journal of Physics: Conference Series

The Development and Challenges of Face Alignment Algorithms



The Development and Challenges of Face Alignment Algorithms

Congyi Wang

School of Communication and Information Engineering, Xi’an University of Posts and Telecommunications Xi’an, China

CongyiWang98@outlook.com

Abstract. A comprehensive survey of face alignment using different methods is presented in this paper. Face alignment is the fundamental task of facial applications, e.g., face recognition, 3D face modelling and face expression analysis, etc. State-of-the-art methods can be mainly categorized into the three groups: gradient descent-based, deep learning-based, and 3D model-based. In gradient descent-based methods, landmarks are localized and adjusted by solving a nonlinear regression function. Deep learning-based methods construct one or several cascaded neural networks to improve landmark localization accuracy. Beside the above two categories of methods solving problems on a 2D plane, there is also other category of methods like 3D model-based methods. Despite significant progress that has been made, face alignment faces challenges from real-world conditions: variation across poses, genders and ages, facial expressions, and facial attributes. This paper offers a brief illustration and analysis of several typical methods of face alignment, provides an overall understanding and insight into the field, which will motivate us to explore promising future directions.

1. Introduction

Facial landmarks, also known as facial key points or facial feature points, are mainly located around facial components such as eyes, mouth, nose and chin. When the faces in the images are detected by the face detectors, face alignment or facial landmark detection are then applied to locate these facial landmarks as shown in Figure 1. Based on these landmarks with semantic meaning, we can obtain huge amount of corresponding shape and texture information of original face images for most facial applications, e.g., face verification [1] and recognition [2], expression recognition [3], facial attribution analysis [4], and solution to other computer vision problems. Therefore, face alignment is a fundamental and important task for facial analysis applications.

In Figure 1, the black rectangle is the detected bounding box and the white points are the detected landmarks. The white points can be gathered as a set and be concatenated to represent shape $x = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ (N is the number of landmarks , N equals 68 in Figure 1), where(x_i, y_i) denotes the coordinate of the $\mathrm{i} - th$ point. Many public datasets, like 300-W [5], AFLW [6], are used to train present face alignment algorithms. These datasets are usually split into two subsets, training set and test set. In most occasions, the landmarks of face images in these datasets are manually labelled as the true value or ground truth. In general, the proposed algorithms are first trained on the training subset, then trained models are tested on the test subset. The goal of face alignment algorithms is to achieve performance in these datasets as good as possible.

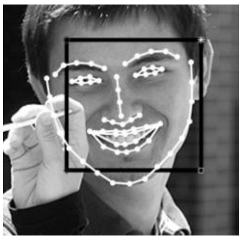


Figure 1. Detected face and 68 facial landmarks

Face alignment generally consists of two phases, training phase and test phase. In the training phase, a model is learned from given labelled training data; in the testing phase, the model is applied to locate facial landmarks of input testing image. Usually, the process starts with a coarse initial shape or the whole image, and simultaneously produce better landmarks output until convergence [7].

Face alignment faces many challenges arising from two categories of deformations: rigid and non-rigid. Rigid deformations are mainly caused by issues of camera, like rotation, scaling, illumination or translation. In contrast, the non-rigid is mainly caused by various facial expression or facial attributes.

According to the method of modelling landmarks and calculating landmark locations, existing face alignment methods can be grouped into three sorts: gradient descent-based methods, deep learning-based methods, and 3D model-based methods.

Gradient descent-based methods are based on the thought of mathematical optimization. Many problems involving face alignment can be treated as nonlinear optimization problems. The goal of gradient descent-based methods is to learn a sequence of descent directions and re-scaling factors of each iteration step. Usually as first, an initial shape is put into the bounding box and then local patches are cropped from the original images and the features like SIFT [8], HoG [9] are extracted to train the models. The models in the current iterations can produce a new shape and new features can be extracted based on this shape. Such sequence produces a series of updates beginning from the initial shape X_0 and converges to final predicted shape X in training data.

Deep learning-based methods utilize neural networks to detect facial landmarks, like convolutional neural networks (CNN) or recurrent neural networks (RNN). The algorithms usually take the whole image as the input and features are extracted from the input by networks. Meanwhile, some models are constructed by multiple levels, called cascaded networks, each containing more than one neural network. Subsequent level takes processed output of prior level as input. The networks get shallower in latter levels, leading to the convergence of landmark prediction. Deep learning-based methods can be subdivided according to network structures and algorithm features, namely, CNN-based methods, cascaded CNN-based methods, MDM/RNN-based methods, multi-task learning-based, etc.

Both gradient descent-based methods and deep learning-based methods focus on a 2D plane. 3D model-based methods are also studied and applied under certain circumstances [10][11]. They can overcome difficulty of face alignment on faces over large poses by converting image into the projection of rotated 3D face model on image plane. With the model normalized and all its vertexes coordinated, we can set up objective function to minimize vertex distances between the face obtained by fitting and the ground truth. Some 3D model-based methods also provide ways of generating training data from public datasets of 2D images, which is crucial for model training.

2. Gradient descent-based methods

Gradient descent-based methods fit an image for the target shape by objective function optimizing, which creates a sequence of updates and finally converge around the ground truth result. Gradient descent-based methods focus on mathematical calculation, in which problems, such as face landmark detection, optical flow, or camera calibration, are treated as continuous nonlinear optimization problems.

2.1. Supervised Descent Method (SDM)

For a given image $d \in R^{m\times 1}$ consisting of m pixels, $\mathrm{d}(x) \in R^{N\times 1}$ are the N landmarks to be extracted in the image. f is a non-linear feature extraction function (e.g., SIFT), and $f(\mathrm{d}(x)) \in R^{128N\times 1}$ if extracts SIFT features. During descent training, the correct N landmarks are referred to as x_*. In the testing condition, the initial shape is first put on the input image and the trained landmark detector is used to detect landmarks.

At first step, the algorithm is not likely to converge, so the SDM produces a sequence of updates along the gradient directions, each update can be viewed as a step of iteration of former landmark situation.

$$x_k = x_{k-1} + R_{k-1}\phi_{k-1} + b_{k-1}, \tag{1}$$

$$\Delta x_{\mathrm{k}} = R_{k-1}\phi_{k-1} + b_{k-1} \tag{2}$$

where $\phi_{k-1} = f(d(x_{k-1}))$ is the feature vector at former landmark position, x_{k-1}. During training process, generic descent directions $\{R_k\}$ and bias terms $\{b_k\}$ are learned for obtaining the shape residual Δx_{k}. In the last iteration, we add Δx_{k} with the predicted shape to produce new shape of current iteration. The initial shape will converge to ground-truth through a sequence of iteration.

R_k and b_k can be learned using regression. Calculation of R_k and b_k can be converted to solving linear least squares problem:

$$\mathrm{argmin}_{R_k,b_k} \sum_{\mathrm{d}^{\mathrm{i}}} \sum_{x_k^i} \left\| \Delta x_*^{ki} - R_k\phi_k^i - b_k \right\|^2. \tag{3}$$

Though the linear regressor and feature extraction of SDM origin from Newton’s method, SDM surpasses original gradient boosting formulation as its feature extraction is not strong. Meanwhile, there are several steps in SDM (usually 4 to 5 steps in practice) instead of one, leading to higher performance than the original method.

2.2. Initialization Optimization Method

An initial shape is always required by traditional gradient-based methods and poor initialization often trap these methods in low level local optima. Meanwhile, traditional methods also suffer from low robustness when coping with large pose variations. Coarse-to-fine shape searching method propose a solution to such problems [12].

Coarse-to-fine shape searching method [13] performs a coarse searching at low levels and provide sub-regions for subsequent finer stages to work on. This method encompasses a number of prospect shapes instead of one at each level and simultaneously discards unpromising results at following level to avoid local optima. Subsequent levels shrink the region and converge it to estimate the final region. In practice, about three levels are required.

3. Deep learning-based methods

Deep learning-based methods mainly use neural networks, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), to detect facial landmarks. Multiple levels of networks, which called cascaded networks, are used for construction of some models. Deep learning-based methods usually take the whole image as the input of first level, and subsequent levels of cascaded networks will take processed output of prior levels as input [14]. Network containing several different networks can reduce the variance by average the prediction. A cascaded network deploys large but coarse networks at low levels to estimate facial landmarks with few large errors, and shallow but fine networks at high levels for restricted adjustment on previous prediction.

In face alignment, deep learning-based methods can be further subdivided into several subclasses: cascaded CNN-based methods, MDM/RNN-based methods, multi-task learning-based, etc.

3.1. Cascaded Convolutional Networks

Cascaded convolutional networks focus on structure design and development of individual networks and their strategies. Several levels of networks with different parameters are contained in a single cascaded convolutional network for coarse-to-fine face alignment [15]. Figure 2. shows how several networks work parallelly in the same level along with the inheritance and convergence of landmark prediction.

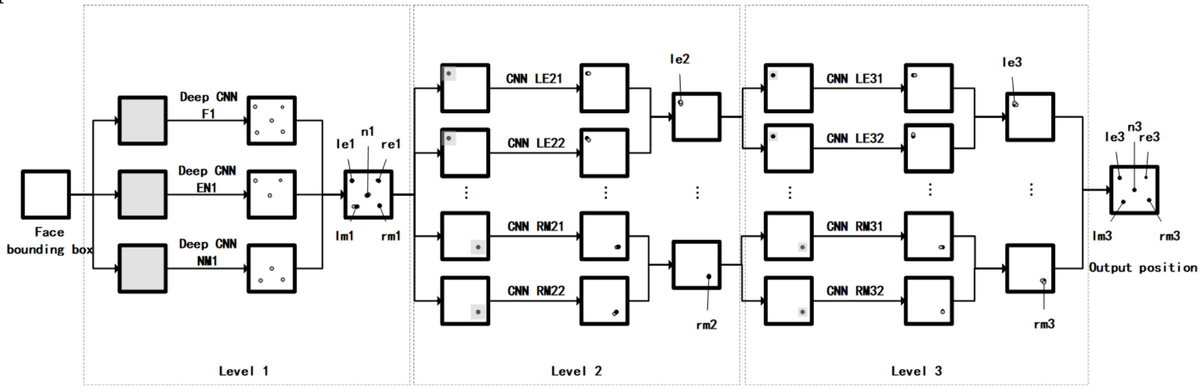


Figure 2. Three-level cascaded convolutional networks. [15]

Cascaded convolutional networks use first level to make initial estimation of facial landmarks from large input regions, this requires first level to be deep and with high non-linearity to form high-level features. Rectified Linear Unit (ReLU) function is crucial to such deep networks as they improve performance by replacing negative values with zero, thus increase non-linear properties. Other functions, like sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ and hyperbolic tangent $tan\,h(x)$ and $|tanh(x)|$ can also increase non-linearity. Performance of convolutional level can be further improved by neurons locally sharing weights on the same map.

One effective way of combining multiple convolutional networks together is multi-level regression. As the only knowledge for first level is the face bounding box, the input regions must be large to cover possible predictions. Subsequent levels are based on a smaller region around the prediction of prior level as to reduce disruption. Consider the inaccuracy of first level and possible drift of latter levels, we express the final prediction for a cascade consisting of n levels at first level with l_i predictions at level i as

$$x = \frac{x_1^{(1)}+\cdots+x_{l_1}^{(1)}}{l_1} + \sum\nolimits_{i=2}^{n} \frac{\Delta x_1^{(i)}+\cdots+\Delta x_{l_i}^{(i)}}{l_i} \tag{4}$$

3.2. Mnemonic Descent Method

For traditional cascaded convolutional networks, parameter and output of each cascade step are learnt independently, so correlations between semantically related image characteristics are not taken into account. Also, cascaded convolutional networks are usually based in non-optimal hand-crafted features and cannot train end-to-end convolutional features. Mnemonic Descent Method (MDM) is proposed to cope with such issues.

Different from CNN, which has all convolutional steps independent and ignores much information, MDM uses a Recurrent Neural Network (RNN) to lay mnemonic constraint on the directions of descent, similar to SDM. MDM is trained in an end-to-end way, starting from the raw image pixels and end at the final predictions.

The memory of MDM is to preserve and facilitate output of the former level, which can be combined as descend gradient into next level. The fundamental equation of RNN is shown as

$$h^{(k+1)} = f_r\left(z^{(k)}, h^{(k)}; \theta_r\right) \tag{5}$$

MDM can provide a sequence of descent directions on a given initial rough estimation, iteratively lead to the optimum. At each step k, internal state is updated by the mnemonic module part of the neural network, according to the energy landscape $z^{(k)}$. During training, the network updates shape displacements at current step through projecting mnemonic element of the algorithm onto the hidden-to-output matrix $W_o \in \mathbb{R}^{u\times d}$. The new time-step can be shown as

$$\Delta x^{(k+1)} = x^{(k)} + W_o h^{(k)} \tag{6}$$

$$x^{(k+1)} = x^{(k)} + \Delta x^{(k+1)} \tag{7}$$

The hidden state of the network, which is often ignored by traditional methods, enables MDM to choose a better descent path by considering the relation of characters. We then repeat the process for a number of times. At last, the objective function of MDM can be presented as

$$\min_\theta \left\| X^* - X^{(0)} + \sum_{k=0}^{K-1} W_o H^{(k)} \right\|_F^2 \tag{8}$$

Where $H(k) = \left[h_1^{(k)}, \cdots, h_n^{(k)}\right] \in \mathbb{R}^{u\times n}$ represents the matrix of all states and θ all the parameters of the model.

3.3. Multi-Task Learning-Based Method

In practice, there are non-negligible correlation among facial landmarks and other facial attributes, like expression and head pose. To exploit this correlation, multi-task learning (MTL) is applied to tackle face alignment task and other facial tasks. Specifically, facial landmark detection is set as the main task and combined with other heterogeneous but related tasks, for example. expression estimation and facial attribute interference. Such models are formulated to facilitate learning converge and deal with different learning difficulties.

Zhanpeng Zhang et al build Tasks-Constrained Deep Convolutional Network (TCDCN) [16] using multi-task learning-based method. There are four tasks related with face alignment in TCDCN, namely, head pose estimation, gender classification, facial expression recognition and facial attribute inference.

A deep convolutional neural network (CNN) is adopted in such method to jointly learn the share feature space x, DCN gradually projects the given face image x_0 to higher level representation by learning a sequence of non-linear mappings. TCDCN uses almost the same structure as DCN during feature extraction, thus, a shared feature vector is provided for multiple tasks in estimation stage.

When it comes to estimation, traditional MTL focus on maximizing the performance of all tasks together, while TCDCN focuses only on main task of face alignment. So, the objective function can be formulated as

$$\operatorname{argmin}_{W^r,\{W^a\}_{a\in A}} \sum_{k=0}^{K-1} l^r\left(y_k^r, f(x_k; W^r)\right) + \sum_{k=0}^{K-1} \sum_{a\in A} \lambda^a l^a\left(y_k^a, f(x_k; W^a)\right) \tag{9}$$

Where λ^a denotes the coefficient of importance of $a - th$ task's error. Different types of loss functions can be optimized together using this function, e.g., regression of landmarks and classification of expressions can be combined.

A task-wise early stop mechanism is introduced to TCDCN in order to prevent being trapped by auxiliary tasks at bad or sub-optimal local optima. Easy tasks that are no longer beneficial after quickly reaching peak performances will be discarded from the iteration process.

4. 3D model-based methods

All methods presented above are based on processing on a 2D plane, assuming that face aligned are in small to medium pose (below 45°) towards the camera. Face alignment in large poses up to 90° faces several hurdles, e.g. occlusion of face and invisibility of landmarks, varies of face appearance and more challenging manual landmark labelling. 3D model-based methods solve such problems by introducing 3D Dense Face Alignment (3DDFA) [17] and using 3D Morphable Models (3DMM) and projection to 'normalize' the face in large pose.

3DMM describes the face in large pose as the rotation, translation and shape appearance of a mean shape face:

$$S = \bar{S} + A_s a_s + A_e a_e, \tag{10}$$

Where S is the 3D face obtained with pose and expression, $\bar{S}$ denotes the mean face shape, A_s and A_e means offsets of face shape and face expression from neutral shape, a_s and a_e mean shape and expression parameters.

The 3DMM is then projected onto image plane with weak prospective projection, model construction and projection function $V(p)$ is shown as:

$$V(p) = f * Op * R * (\bar{S} + A_s a_s + A_e a_e) + t_{2d}, \tag{11}$$

Where f is the scale factor, Op is the orthographic projection matrix, R is the rotation matrix of image containing movement in three directions, namely, pinch, yaw, roll, and t_{2d} is the translation factors. The collection of parameters is presented $p = [f, pitch, yaw, roll, t_{2d}, a_s, a_e]^T$.

3DDFA introduces Projected Normalized Coordinate Code (PNCC), which is a normalized parameter obtained by normalization and projection of input image and provides locations of 3D vertexes on 2D plane. Input image is then stacked with PNCC and transferred to CNN.

CNN is trained to make predictions of parameter update Δp^k:

$$\Delta p^k = Net^k\left(I, PNCC\left(p^k\right)\right) \tag{12}$$

Weighted Parameter Distance Cost (WPDC) is introduced to model, reflecting the importance of each parameter in cost function as influence of each dimension on 3DDFA output is generally different.

3D model-based methods offer a brand-new method of handling large pose face image alignment, which is very difficult for traditional face alignment methods focusing on frontal face. Also, they are proved to be generally as good as common CNN in when dealing with images of small poses. However, the computation cost is relatively high due to 3D transformation.

5. Evaluations

5.1. Databases

There are many public datasets available for face alignment. These datasets have their ground-truth facial landmarks labelled manually or through crowdsourcing. Each face image is individually labelled by several workers or through several methods and weighted average of label results is taken as the ground truth landmarks.

Face databases can be classified into two categories: dataset of controlled conditions, which are arranged and taken under designed experimental settings, and datasets of uncontrolled conditions (i.e. in the wild), which are generally collected from websites or public social medias.

Two popular datasets used in evaluation are 300-W dataset and AFLW dataset.

300-W dataset [5]: This dataset contains multiple alignment databases including AFW, LFPW, HELEN and XM2VTS. The 68 re-annotated landmarks of the “300 Face in-the Wild Challenge” (300-W dataset) were seen as the ground truth for images in LFPW and HELEN. The 300-W dataset provides each image with a prescribed face bounding box, meaning that no external face detectors are needed and no faces are missed. HELEN dataset contains 2000 training and 330 test images, evaluations can be conducted on 194 points and 68/49 points. LFPW dataset contains 1132 training and 300 test images with poses and face expressions, evaluations can be performed on 68/49 points, while some image links are not available, usually only about 800 training and 250 testing images of LFPW can be used.

AFLW dataset [6]: This dataset contains 21080 more challenging in-the-wild faces with large-poses (with a yaw from -90 °to 90 °) and each is annotated with no more than 21 visible landmarks, which is suitable for evaluating alignment performance over large poses.

5.2. Comparisons and Discussions

The distance between estimated landmarks and ground truth landmarks which is normalized according to the number of landmarks and inter-ocular distance (named normalized mean error, NME, in the following test) is commonly used for evaluating a face alignment system as shown in (13):

$$\text{NME} = \frac{\sum_{i=1}^{m} \left\| x_{(i)}^{e} - x_{(i)}^{g} \right\|_2}{N \times \mathrm{d}_{io}} \times 100\%, \tag{13}$$

Where d_{io} denotes the inter-ocular distance, $x^{e}_{(i)}$ is the $i-th$ estimated landmark and $x^{g}_{(i)}$ is the ground truth of the $i-th$ estimated landmark. Meanwhile, failure rate is also used as measurement of face alignment quality due to the variations in error normalization to avoid biases of NME. Calculation of failure rate manually defines a threshold. Point-to-point error exceeding the threshold is considered a failure. Failure rate is thus expressed as the ratio of failure point numbers to total point numbers.

Comparisons is made to illustrate the various characteristics of various methods. Six representative models are chosen for study: SDM-Regression (Xiong and De la Torre, 2013) [18], CCN-DL (Sun et al, 2013) [15], TCDCN (Zhang et al, 2014) [16], MDM (Trigeorgis et al) [19], CFSS (Zhu et al, 2015) [13] and 3DDFA (Zhu et al) [17].

We localize models mostly in two databases, 300-W dataset and AFLW dataset, several other datasets are used for extra difficulty of certain aspects. Some datasets were merged, extracted or expanded from original sets for training and testing for model preference during performance evaluations of different methods.

Normalized mean error (NME) reflects the deviance between detected position and ground truth. Table 1 implies the normalized mean error of SDM under different circumstances.

Table 1. NME of SDM in different situations

Dataset	Training Set	Number of Landmarks	NME (%)
LFPW (300-W)	LFPW (300-W)	17	3.47
LFW	LFW	66	2.70
AFLW	300-W	21	6.10
AFLW	300-W-LP	21	2.45
AFLW2000-3D	300-W	68	7.23
AFLW2000-3D	300-W-LP	68	3.21

SDM continuously shows low NME when trained with and perform face alignment on different datasets. The performance of SDM is dependent on the difficulty of datasets, e.g., the pose of faces and image quality, and number of landmarks. Due to the stability and generality of SDM, it's frequently used as control group in study of other models.

Figure 3 shows the NME performance of CNN-DL (Sun et al, 2013) based on LFPW dataset. Only five characteristic points, namely, the centres of two pupils, the nose tip and two mouth corners, are detected using CNN-DL. As the output shows, CNN-DL displays significantly lower NME in detection of these landmarks.

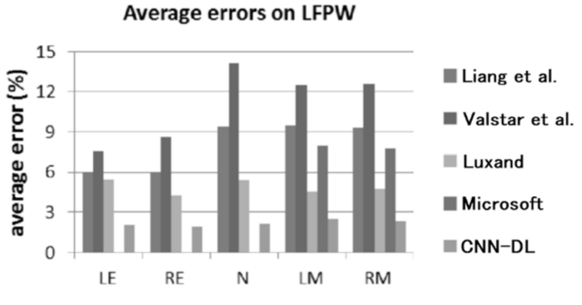


Figure 3. The comparison results on five landmarks respectively [15]

Table 2 presents the failure rate of MDM, CFSS, when training under 300-W dataset. Generally, MDM shows similar or slightly lower failure rate than CFSS. Failure rate goes higher when there are more landmarks need to be localized.

Table 2. Comparison of failure rate of MDM and CFSS

Model	Number of Landmarks	Failure Rate (%)
MDM	51	4.2
MDM	68	6.8
CFSS	51	7.8
CFSS	68	12.3

Table 3 shows the comparison of NME of SDM, 3DDFA and a method combining 3DDFA and SDM. As AFLW is a dataset more challenging than 300-W, containing images of larger pose and other factors interfering face alignment, 3DDFA outperforms SDM in NME. While the combination of both methods achieves better performance measured by mean error, indicating that the combination of two or multiple methods may further boost localization accuracy.

Table 3. Performance of SDM, 3DDFA and collaborating method on large pose images.

Model	Dataset	Training Set	Number of Landmarks	NME (%)
SDM	AFLW	300-W	21	6.10
SDM	AFLW	300-W-LP	21	2.45
SDM	AFLW2000-3D	300-W	68	7.23
SDM	AFLW2000-3D	300-W-LP	68	3.21
3DDFA	AFLW	AFLW	21	0.99
3DDFA	AFLW2000-3D	AFLW2000-3D	68	2.21
3DDFA+SDM	AFLW	AFLW	21	0.92
3DDFA+SDM	AFLW2000-3D	AFLW2000-3D	68	1.97

6. Conclusion

This paper provides a survey of current face alignment methods, including gradient descent-based methods, deep learning-based methods and 3D model-based methods. The above methods are mostly constructed based on deep learning, which achieves a great enhancement in computer vision filed besides face alignment. Although the state-of-the-art methods have achieved comparable performance to humans on some databases, challenges still exist in aligning face under difficult illumination, occlusion or large shape variation conditions. Moreover, most existing datasets are composed of frontal or near frontal face images. The need for high-performance and practical face alignment calls for further breakthrough in the development of new face alignment methods and establishment of new datasets for method training.

References

[1] B Lu, J Chen, C D Castillo and R Chellappa 2019 an experimental evaluation of covariates effects on unconstrained face verification *IEEE Transactions on Biometrics, Behavior, and Identity Science* **1** 42

[2] W Zhang, X Zhao, J Morvan and L Chen 2019 improving shadow suppression for illumination robust face recognition *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41** 611

[3] Y Li, J Zeng, S Shan and X Chen 2019 occlusion aware facial expression recognition using CNN with attention mechanism *IEEE Transactions on Image Processing* **28** 2439

[4] W Wang, Y Yan, Z Cui, J Feng, S Yan and N Sebe 2019 recurrent face aging with hierarchical auto regressive memory IEEE transactions on pattern analysis and machine intelligence **41** 654

[5] Sagonas C, Tzimiropoulos G, Zafeiriou S, & Pantic M 2013 300 faces in-the-wild challenge: the first facial landmark localization challenge *proceedings of the 2013 IEEE International Conference on Computer Vision Workshops. IEEE*

[6] Martin Köstinger, Wohlhart P, Roth P M and Bischof H 2011 annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain, November 6-13, 2011. IEEE*

[7] H Ouanan, M Ouanan and B Aksasse 2016 facial landmark localization: past, present and future *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), Tangier,* 487-493

[8] Lowe D G 2004 distinctive image features from scale-invariant keypoints *International Journal of Computer Vision* **60** 91

[9] Dalal N and Triggs B 2005 histogram of oriented gradients for human detection *2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

[10] M Song, D Tao, S Sun, C Chen and S J Maybank 2014 robust 3D face landmark localization based on local coordinate coding *IEEE Transactions on Image Processing* **23** 5108

[11] A Bulat and G Tzimiropoulos 2017 how far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks) *2017 IEEE International Conference on Computer Vision (ICCV)* 1021

[12] H Mo, L Liu, W Zhu, S Yin and S Wei 2019 face alignment with expression- and pose-based adaptive initialization *IEEE Transactions on Multimedia* **21** 943

[13] Zhu N S, Li N C, Loy C C and Tang X 2015 face alignment by coarse-to-fine shape searching *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

[14] B Shi, X Bai, W Liu and J Wang 2018 face alignment with deep regression *IEEE Transactions on Neural Networks and Learning Systems* **29** 183

[15] Sun Y, Wang X and Tang X 2013 deep convolutional network cascade for facial point detection *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

[16] Zhang Z, Luo P, Loy C C and Tang X 2014 facial landmark detection by deep multi-task learning *European Conference on Computer Vision*

[17] Zhu X, Lei Z, Liu X, Shi H and Li S Z 2016 face alignment across large poses: a 3D solution *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

[18] Xiong X and Fernando D L T 2013 supervised descent method and its applications to face alignment *2013 IEEE Conference on Computer Vision and Pattern Recognition* 532

[19] Trigeorgis G, Snape P, Nicolaou M A, Antonakos E and Zafeiriou S 2016 mnemonic descent method: a recurrent process applied for end-to-end face alignment *IEEE International Conference on Computer Vision & Pattern Recognition (CVPR). IEEE*