A Survey for Conventional Regression- and Deep Learning-based Face Alignment Methods

Tongxuan Gao
University of Wisconsin-Madison, USA
gaotx71@163.com

ABSTRACT

Face alignment, as an important part of facial tasks, will affect the final efficiency and accuracy. Face alignment is to locate the exact shape of a detected face bounding box. There are amount of challenges in face alignment because of large poses, occlusions and illuminations in real-world conditions. The approaches to tackle these challenges can be categorized in methods based on regression, which require operators in feature extraction, and methods based on deep learning, in which the feature extraction is data driven.

Methods applies regression include Supervised Descent Method and Face Alignment by Coarse-to-Fine Shape Searching. Deep Convolutional Neural Networks, Tasks-Constrained Deep Convolutional Network and Multi-task Cascaded Convolutional Networks apply cascaded CNN and they are representational approaches of deep learning method. This article is devoted to the elaboration and summary of these mainstream methods.

CCS CONCEPTS

• **Applied computing** → Computer forensics; Data recovery.

KEYWORDS

Face alignment, Conventional regression, Deep learning

ACM Reference Format:
Tongxuan Gao. 2021. A Survey for Conventional Regression- and Deep Learning-based Face Alignment Methods. In *2021 International Symposium on Electrical, Electronics and Information Engineering (ISEEIE 2021), February 19–21, 2021, Seoul, Republic of Korea.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3459104.3459191

1 INTRODUCTION

Face alignment is to locate the facial key points, such as eyes, nose and mouth corners according to the input face images. Facial key points can be used to align the whole face, detect the head position and detect the facial expression. Therefore, face alignment an essential process and it is widely used in many face-related tasks, such as face recognition [1], facial attributes analysis [2] and automatic synthesis of facial animation.

In face alignment, training and testing are two important processes. During training, the training set, a set of public images, is used to optimize the loss function to get regressors, and as a result discover the potential relationships. In testing process, the test set is used to evaluate the utility of the relationships.

According to feature extraction methods, regression [4, 7] and deep learning-based algorithms [2, 3, 5, 6] are the two most commonly used methods on face alignment.

In regression methods, Supervised Descent Method is a classic optimization algorithm, which can be used to solve the least squares minimization and enhance, leads to the optimization the solution of the objective function in facial feature point detection. The improved method Coarse-to-Fine Shape Searching (CFSS) provides an optimization in face initialization based on SDM, as initialization has a great impact on the final alignment accuracy.

Deep learning methods includes Deep Convolutional Neural Network (DCNN), Tasks-Constrained Deep Convolutional Network and Multi-task Cascaded Convolutional Networks. DCNN is a method that utilizes the context information over the entire face. With the geometric constraints among key points, the face alignment could be more reliable when the images are taken under extreme conditions. Facial Landmark Detection by Deep Multi-task Learning (TCDCN), as a development of DCNN, apply multi-task learning to use the auxiliary information to enhance face alignment. And it also applies early stopping, which prevents the problem caused by different convergence rates of each task. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks (MTCNN) is a joint method of face detection and face alignment that provide better performance. In real world application, face alignment is challenged by different head positions, complex expressions, occlusions, illuminations and many other unsatisfactory conditions. In this paper, an introduction to some classic face alignment algorithms is presented and the corresponding analysis is supplemented for each algorithm.

2 THE FACE ALIGNMENT ALGORITHM

2.1 Supervised Descent Method

The article Supervised Descent Method and its Applications to Face Alignment introduces a method called Supervised Descent Method, which is used to minimize the non-linear Least Squares objective function, in another word, to make the objective function converge to the minimum in a very short time. To achieve this goal, Newton's method is most commonly used among all the methods. Usually, it converges in very fast speed. However, there are problems when applying Newton's method that could cause huge computing cost. Therefore, SDM is proposed to learn the descent direction. Figure 1 (a)(b) can be used to demonstrate the basic principles of Newton's method and SDM method.

ISEEIE 2021, February 19–21, 2021, Seoul, Republic of Korea
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8983-9/21/02…$15.00
https://doi.org/10.1145/3459104.3459191

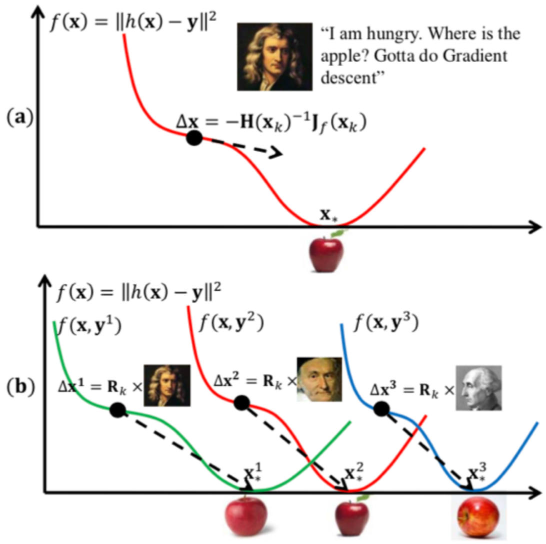
(a)
$f(\mathbf{x}) = \|h(\mathbf{x}) - \mathbf{y}\|^2$
"I am hungry. Where is the apple? Gotta do Gradient descent"
$\Delta\mathbf{x} = -\mathbf{H}(\mathbf{x}_k)^{-1}\mathbf{J}_f(\mathbf{x}_k)$
$\mathbf{x}_*$
(b)
$f(\mathbf{x}) = \|h(\mathbf{x}) - \mathbf{y}\|^2$
$f(\mathbf{x}, \mathbf{y}^1)$
$f(\mathbf{x}, \mathbf{y}^2)$
$f(\mathbf{x}, \mathbf{y}^3)$
$\Delta\mathbf{x}^1 = \mathbf{R}_k \times$
$\Delta\mathbf{x}^2 = \mathbf{R}_k \times$
$\Delta\mathbf{x}^3 = \mathbf{R}_k \times$
$\mathbf{x}_*^1$
$\mathbf{x}_*^2$
$\mathbf{x}_*^3$

Figure 1: (a) Basic principles of Newton's method. (b) Basic principles of SDM method [6].

Given an initial shape X0, X1 can be generated by the function

$$\Delta X_1 = R_0\varphi_0 + b_0 \tag{1}$$

in which R0 represents the descent direction and b0 represents the bias term. Therefore, the Xk would be

$$X_k = X_{K-1} + R_{k-1}\varphi_{k-1} + b_{k-1} \tag{2}$$

The training process is to find the best R and b that ensure the minimum Δx, which now is a primary function of ϕ.

To avoid the using of Jacobian and Hession matrix. The process of learning R and b applies the least squares method is

$$\underset{R_k, b_k}{\arg\min} \sum_{d^i} \sum_{X_k^i} \left\| \Delta X_*^{ki} - R_k\varphi_k^i - b_k \right\|^2 \tag{3}$$

2.2 Coarse-to-Fine Shape Searching

Based on SDM, Face Alignment by Coarse-to-Fine Shape Searching proposes the coarse-to-fine approach that guarantee the performance of initialization. A poor initialization, which is a common problem of cascaded regression approaches, could lead the final solution to be trapped in a local optima.

Different with cascaded regression approach, the coarse-to-fine framework starts by exploring the whole shape space and estimate the sub-region. Each time the whole sub-region will be corrected instead of certain points. Also, the early stage of coarse-to-fine framework put varies of poses into consideration. Once the pose is estimated, the search will focus on the certain category of candidate shapes and get more accurate.

As shown in Figure 2, the approach includes 3 stages. On each stage, the given sub-region is used to estimate a more accurate sub-region.

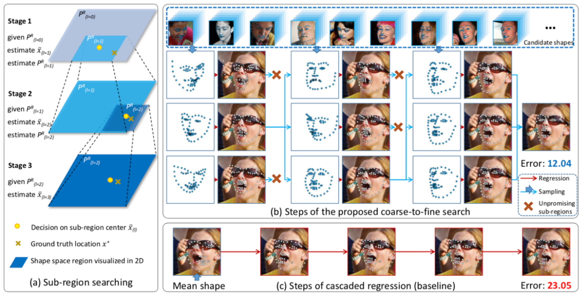

Figure 2: Three stages of Coarse-to-Fine Shape Searching [9].

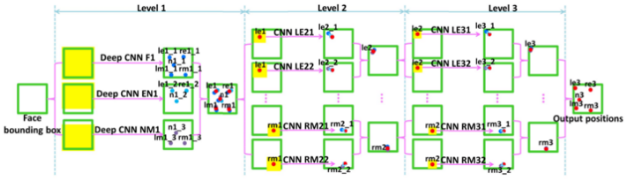

Figure 3: Three-level cascaded convolutional networks [3].

2.3 Deep Convolutional Network Cascade

Apart from the traditional methods such as SDM and CFSS, deep learning is also a very commonly used method in face alignment. Convolutional Neural Networks (CNN) [3] is a neural network specially designed for image recognition, a representative algorithm of deep learning.

Deep Convolutional Network Cascade for Facial Point Detection proposes an approach of facial point detection. In the paper, the three-level cascaded convolution network is introduced. It was mentioned in the paper that many approaches ignore the texture context information.[4]

As shown in the Figure 3, a face bounding box is detected by the face detector. There are three CNN networks in level 1. F1, consists of all 5 key points: left eye, right eye, nose, left mouth corner and right mouth corner; EN1, consists of key points of left eye, right eye and nose; NM1, consist key points of nose, left mouth corner and right mouth corner. In this way, with the geometric constraints, the chance large error of key points' position can be reduced.

Level 1 generates the approximately positions of the 5 points. Level 2 uses a smaller bounding box to bound the 5 regions where these points locate and apply 2 CNN networks for each point to get the average location of two more accurate positions. Similar with level 2, level 3 applies further bounding input image and correcting the positions.

This structure is proved to be very accurate and has better performance than other strategies such as the state-of-the-art method.

2.4 Multi-task Cascaded Convolutional Networks

Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks presents a cascade framework based on CNN that joint face detection and face alignment. In this paper, unified cascaded CNNs by multi-task learning is proposed. The

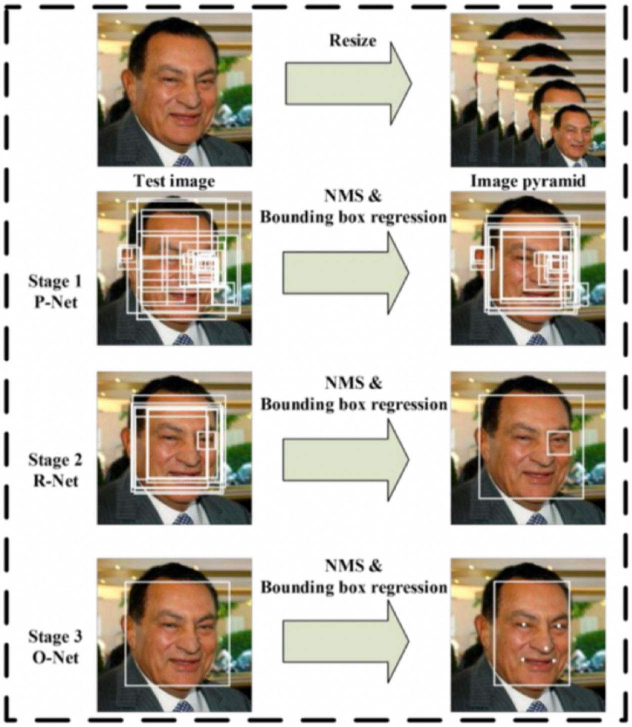

Figure 4: Pipeline of cascaded framework that includes three-stage multi-task deep convolutional networks [7].

MTCNN framework consists of three stages: a shallow CNN that generate candidate window, a complex CNN that eliminates non-face windows and a po werful CNN that output the facial landmarks. As shown in Figure 4, the three stages handle the face goes from coarse to fine.

The CNN regressor is trained using three tasks: face/non-face classification, bounding box regression and facial landmark. In addition, different with traditional hard sample mining, the paper introduces a new online hard sample mining strategy, that is, changing to be adaptive in the training process. The loss of all samples is computed and sorted. And the top 70% are selected as hard sample. In this way, the rest easy samples which are less helpful can be ignored. This self-adaptive strategy can reduce the computational expense and increase the accuracy of detector theoretically and indeed it is proved to perform better without manual sample selection.

2.5 Tasks-Constrained Deep Convolutional Network

Facial Landmark Detection by Deep Multi-task Learning presents an application of multitask learning and CNN, Tasks-Constrained Deep Convolutional Network. As it was mentioned in the article, the solution of facial landmark detection can be influenced by different factors such as gender, if the person is smiling or if the person wears glasses [8]. As a result, the multi-task learning is applied which consists of facial landmarks and four sub-tasks: gender, glasses, smiling and facial posture. However, since different tasks have varied learning difficulties and different convergence rates, the weights of these tasks are not the same. A weighted objective function is presented after optimizing the different learning difficulties in multi-task learning. The first term is the loss function of main task, facial landmark detection and the second term is the loss function of auxiliary tasks, in this paper: gender, glasses, smiling and facial posture.

$$\underset{W^r, \{W^{\ a}\}_{a\in A}}{\arg\min} \sum_{i=1}^{N} \ell^r(y_i^r, f(X_i; W^r)) + \sum_{i=1}^{N} \sum_{a\in A} \lambda^a \ell^a(y_i^a, f(X_i; W^a)) \tag{4}$$

As for the different convergence rates, the paper introduces early stopping. The auxiliary task can be terminated as long as the performance is good enough. The criterion to stop learning a task is

$$\frac{k \cdot med_{j=t-k}^{t} E_{tr}^{a}(j)}{\sum_{j=t-k}^{t} E_{tr}^{a}(j) - k \cdot med_{j=t-k}^{t} E_{tr}^{a}(j)} \cdot \frac{E_{val}^{a}(t) - \min_{j=1..t} E_{tr}^{a}(j)}{\lambda^{a} \cdot \min_{j=1..t} E_{tr}^{a}(j)} > \epsilon \tag{5}$$

where ϵ represents the threshold. The first term represents the tendency of training error and second term is the ratio of generalization error to training error. As the training error dropping, the first term gets smaller, indicates that the performance is still making progress and the task will be continued. Otherwise, the first term is large and the task need to be stopped.

2.6 Mnemonic Descent Method

As a strong improvement, Mnemonic Descent Method: A recurrent process applied for end-to-end face alignment presents a face alignment approach where CNN and RNN are in series. Face alignment often encounters with the non-liner least squares optimization problem. Due to the short coming of Newton's method, a set of "descent directions" are proposed, which are learnt indecently through cascade regression. However, it is lack of consideration of correlations between semantically related image characteristics. To address the issue, Mnemonic Descent Method (MDM) is proposed. MDM can be generally divided into two processes: Convolutional Neural Network (CNN), the extraction of facial features, and Recurrent Neural Network (RNN), the imposition of memory constraint on the descent directions.

MDM is the first approach that is able to be trained in end-to-end manner. It is also proved to be outstanding on the test-set of the 300W competition.

3 EVALUATION

The commonly used measurement of the mean error is the distance between estimated landmarks and the ground truths normalized by interocular distance or the distance between outer eye corners. The algorithms and the used datasets are introduced in the following parts.

3.1 300W Dataset

300W is composed of AFLW, AFW, Helen, IBUG, LFPW, LFW and other datasets. On 300W dataset, the average errors of Face Alignment by Coarse-to-Fine Shape Searching (CFSS) and CFSS practical are 5.76% and 5.99%. Compared with other state-of-art methods including Zhu et.al, DRMF, RCPR, SDM, GN-DPM and CFAN, there is over 16% of error reduction. In addition, CFSS and CFSS-practical outperform cascaded regression on average error. In the 300W competition for both 51-point and 68-point plot, MDM has the lowest failure rate (4.2% and 6.8%) in comparison with ERT, PO-CR, Chehra, Intraface, Balt. et al., Face++, Yan et al. and CFSS.

3.2 LFPW, LFW-A&C and BioID Dataset

LFPW is a dataset in which the images are downloaded from web and have variety in poses, illumination and facial expression. LFW-A&C is a subset of LFW, consisting of the images of people whose names start with 'A' or 'C'. LFW dataset contains over 13,000 face images that are collected from the internet. In this dataset, 1680 people have two or more distinct images. BioID contains 1521 frontal face images with moderate variations on illumination and expression of 23 subjects. On both BioID and LFPW dataset, compared with Component based Discriminative Search, Boosted Regression with Markov Networks, Luxand Face SDK and Microsoft Research Face SDK, Deep Convolutional Network Cascade for Facial Point Detection (DCNC) is able to improve accuracy over 50% and reduce detection error significantly. The average error of DCNC on BioID and LFPW are below 3.0%, which is much lower than the mentioned methods. The evaluation of face feature detection algorithms employ LFPW and LFW-A&C dataset. On LFPW, SDM outperforms the methods trained with only one linear regression and shows comparable average mean error, 3.47%, with that of Belhumeur et al. method, which is 3.43%.

3.3 AFLW and AFW Dataset

AFLW dataset contains over 25,000 face images with various pose, facial expression, illumination, race and other factors. Each face was marked with 21 feature points. AFW dataset is based on a photo sharing website Flickr. It contains 205 images, including 473 faces. For each face, there is a rectangular border box, six landmarks and related postural angles. On AFLW, in comparison against CNN, the proposed Tasks-Constrained Deep Convolutional Network (TCDCN) which consists of facial landmark detection and four auxiliary tasks of recognizing 'pose', 'gender', 'glasses' and 'smiling' performs better in four out of five facial landmarks. In addition, on both AFLW and AFW, TCDCN outperform RCPR, TSPM, CDM, Luxand and SDM. The mean error of TCDCN is 8.0% on AFLW and 8.2% on AFW. The evaluation of face alignment on AFLW indicates that MTCNN outperform the following state-of-art methods: RCPR, TSPM, Luxand face SDK, ESR, CDM, SDM, and TCDCN. The mean error of Multi-task Cascaded Convolutional Networks (MTCNN) is 6.9%.

4 CONCLUSION

This paper presents the comparison of the performance of some recent face alignment methods including SDM, CFSS, DCNN, TCDCN, MTCNN and MDM. The main difference between these methods are that some adopt conventional-regression methods, including SDM and CFSS. The other adopt deep-learning methods in which DCNN, TCDCN, MTCNN and MDM are included. We also take the different training set that are used to compare performance of different methods. In conclusion, deep learning is a face alignment framework with great potential and has made great process recently. However, hardware data becomes more and more prominent. With the increasing complexity and richness of face alignment frameworks and the strict demands of accuracy, data becomes more and more prominent. As far as I am concerned, processing and analysis of massive data will be a crucial research topic.

REFERENCES

[1] Cunjian Chen, Antitza Dantcheva, and Arun Ross. 2013. Automatic facial makeup detection with application in face recognition. *Proc. - 2013 Int. Conf. Biometrics, ICB 2013* June (2013). DOI:https://doi.org/10.1109/ICB.2013.6612994

[2] Huiyu Mo, Leibo Liu, Wenping Zhu, Shouyi Yin, and Shaojun Wei. 2018. Face Alignment With Expression- and Pose-Based Adaptive Initialization. *IEEE Trans. Multimed.* PP, (August 2018), 1. DOI:https://doi.org/10.1109/TMM.2018.2867262

[3] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, and Tsuhan Chen. 2018. Recent advances in convolutional neural networks. *Pattern Recognit.* 77, (2018), 354–377. DOI:https://doi.org/10.1016/j.patcog.2017.10.013

[4] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2013. Deep convolutional network cascade for facial point detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3476–3483. DOI:https://doi.org/10.1109/CVPR.2013.446

[5] George Trigeorgis, Patrick Snape, Mihalis A. Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. 2016. Mnemonic Descent Method: A Recurrent Process Applied for End-to-End Face Alignment. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 4177–4187. DOI:https://doi.org/10.1109/CVPR.2016.453

[6] Xuehan Xiong and Fernando De La Torre. 2013. Supervised descent method and its applications to face alignment. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* (2013), 532–539. DOI:https://doi.org/10.1109/CVPR.2013.75

[7] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment using multi-task cascaded cnn.pdf. *IEEE Signal Process. Lett.* 23, 10 (2016), 1499–1503. DOI:https://doi.org/10.1109/lsp.2016.2603342

[8] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial landmark detection by deep multi-task learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8694 LNCS*, 94–108. DOI:https://doi.org/10.1007/978-3-319-10599-4_7

[9] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. 2015. Face alignment by coarse-to-fine shape searching. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 07-12-June*, 4998–5006. DOI:https://doi.org/10.1109/CVPR.2015.7299134