Neurocomputing 275 (2018) 50–65

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Facial feature point detection: A comprehensive survey

Nannan Wang[a], Xinbo Gao[b,*], Dacheng Tao[c], Heng Yang[d], Xuelong Li[e]

[a] The State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an City, 710071, China
[b] The State Key Laboratory of Integrated Services Networks, School of Electronic Engineering, Xidian University, Xi'an City, 710071, China
[c] UBTech Sydney Artificial Intelligence Institute, The School of Information Technologies, University of Sydney, J12 Cleveland St, Darlington NSW 200, 8, Australia
[d] ULSee Incorporation, Hangzhou City, 310016, China
[e] Center for OPTical IMagery Analysis and Learning (OPTIMAL), Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, China

ARTICLE INFO

Article history:
Received 1 December 2016
Revised 14 March 2017
Accepted 4 May 2017
Available online 1 June 2017

Keywords:
Deep learning
Face alignment
Facial feature point detection
Facial landmark localization

ABSTRACT

This paper presents a comprehensive survey of facial feature point detection with the assistance of abundant manually labeled images. Facial feature point detection favors many applications such as face recognition, animation, tracking, hallucination, expression analysis and 3D face modeling. Existing methods are categorized into two primary categories according to whether there is the need of a parametric shape model: parametric shape model-based methods and nonparametric shape model-based methods. Parametric shape model-based methods are further divided into two secondary classes according to their appearance models: local part model-based methods (*e.g.* constrained local model) and holistic model-based methods (*e.g.* active appearance model). Nonparametric shape model-based methods are divided into several groups according to their model construction process: exemplar-based methods, graphical model-based methods, cascaded regression-based methods, and deep learning based methods. Though significant progress has been made, facial feature point detection is still limited in its success by wild and real-world conditions: large variations across poses, expressions, illuminations, and occlusions. A comparative illustration and analysis of representative methods provides us a holistic understanding and deep insight into facial feature point detection, which also motivates us to further explore more promising future schemes.

1. Introduction

Facial feature points, also known as facial landmarks or facial fiducial points, have semantic meaning. They are mainly located around facial components such as eyes, mouth, nose and chin. Facial feature point detection (FFPD) generally refers to a supervised process using abundant manually labeled images. FFPD usually starts from a rectangular bounding box returned by a face detector [145,169] which locates the face. This bounding box can be employed to initialize the positions of facial feature points.

Facial feature points are different from keypoints for image registration [138] while keypoint detection is usually an unsupervised procedure. However, features on interest points could favor to the facial feature point detection, such as dense interest points for voting facial feature points [167], SIFT (scale-invariant feature transform) for supervised descent method [164] and partial face alignment [168].

Suggested by Cootes et al. [31], facial feature points can be reduced to three types: points labeling parts of faces with application-dependent significance, such as the center of an eye or the sharp corners of a boundary; points labeling application-independent elements, such as the highest point on a face in a particular orientation, or curvature extrema (the highest point along the bridge of the nose); and points interpolated from points of the previous two types, such as points along the chin. According to various application scenarios, different numbers of facial feature points are labeled as, for example, a 17-point model, 29-point model or 68-point model. Whatever the number of points is, these points should cover several frequently-used areas: eyes, nose, and mouth. These areas carry the most important information for both discriminative and generative purposes. Generally speaking, more points indicate richer information, although it is more time-consuming to detect all the points.

The points shown can be concatenated to represent a shape $\mathbf{x} = (x_1, \ldots, x_N, y_1, \ldots, y_N)^T$ where (x_i, y_i) denotes the location of

* Corresponding author.
E-mail address: xbgao.xidian@gmail.com (X. Gao).
URL: https://ulsee.com/en/ (H. Yang)

http://dx.doi.org/10.1016/j.neucom.2017.05.013
0925-2312/

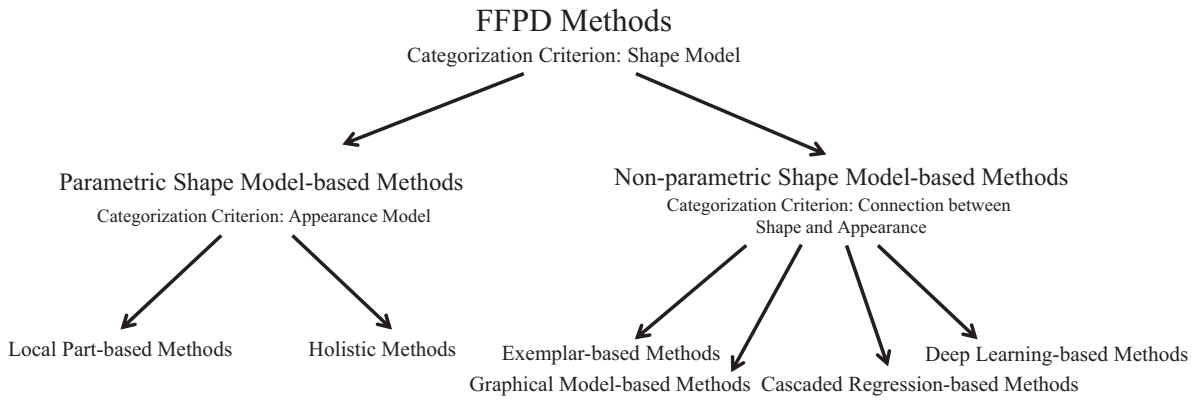

Fig. 1. Tree diagram for FFPD methods.

the *i*th point and *N* is the number of points (*N* is 68 in this figure). Given a sufficiently large number of manually labeled points and corresponding images as the training data, the target of facial feature point detection is to localize the shape of an input testing image according to the facial appearance. Detecting the shape of a facial image is a challenging problem due to both the rigid (scale, rotation, and translation) and non-rigid (such as facial expression variation) face deformation. FFPD generally consists of two phases: in the training phase, a model is learned from the appearance variations to the shape variations; and in the testing phase, the learned model is applied to an input testing image to localize facial feature points (shape). Normally the shape search process starts from a coarse initialization, following that the initial shape is moved to a better position step by step until convergence.

The facial feature point detection problem can be decomposed into three problems, *i.e.* how to construct the face shape model, how to construct the face appearance model and how to model the connection between the shape and the appearance. In order to have a clear understanding of the progress of techniques of FFPD, we classify existing method into different primary categories according to how to model the shape. The secondary criterion for the classification of primary categories is how to model the face appearance or how to construct the model from the shape to appearance. Fig. 1 provides the tree diagram for the categorization of FFPD methods. According to whether there is the need of a parametric shape model, existing FFPD methods are categorized into two primary categories: parametric shape model-based methods and nonparametric shape model-based methods. Parametric model refers to data in the model belonging to some particular distribution, *e.g.* Gaussian, Gaussian mixture model *etc*. Nonparametric model-based methods are distribution free which do not rely on assumptions that the data are drawn from a given probability distribution. The difference between parametric model and nonparametric model is that the former has a fixed number of parameters, while the latter grows the number of parameters with the amount of training data [102]. Parametric shape model-based methods are further divided into two secondary classes according to their appearance models: local part model-based methods, *e.g.* Active Shape Models (ASM), and holistic model-based methods, *e.g.* Active Appearance Models (AAM). Nonparametric shape model-based methods mainly refer to exemplar-based methods, graphical model-based methods, cascaded regression based methods (*e.g.* explicit shape regression method [20], ESR) and deep learning based methods (*e.g.* deep convolution network method [132], DCN).

Local part model-based methods consider the appearance variation around each facial feature point independently. Detected facial feature points are then refined by a global shape model which is generally learned from training shapes. Holistic model-based methods model the appearance variation from a holistic perspective. In addition, both the shape and appearance variation model in holistic methods are usually constructed from a linear combination of some bases learned from training shapes and images. Recently, nonparametric shape model-based methods achieved advanced performance. exemplar-based methods localize facial feature points relying on retrieved top similar exemplars from the training dataset. Graphical model-based methods generally deploy a tree structure or Markov random field to model relationships between facial feature points. Cascaded regression-based methods estimate the shape in a coarse-to-fine manner directly from the appearance without explicitly learning any shape model or appearance model. Deep learning based methods either learn the nonlinear shape and appearance variation or learn the nonlinear mapping from the face appearance to the face shape.

The remainder of this paper is organized as follows: Sections 2 and 3 review parametric and nonparametric shape model-based methods respectively. Section 4 gives some discussions on issues beyond reviewed methods in Sections 2 and 3. Section 5 reviews existing common used databases for FFPD. Section 6 evaluates and analyzes the performance of several representative methods. Finally, Section 6 summarizes the paper, and discusses some promising future directions and tasks regarding FFPD.

2. Parametric shape model-based methods

Local part model-based methods usually detect each facial feature point around some region locally and then these detected points are constrained by a global shape model. Holistic model-based methods usually estimate the location of facial feature points from a holistic texture representation combined with a global shape model. In following subsections we will first detail parametric shape models. Then local part model-based methods and holistic model-based methods are sequentially reviewed.

2.1. Parametric shape model

Fig. 2 illustrates the statistical distribution of facial feature points sampled from 600 facial images. Parametric shape model usually describes each facial point by a distribution, *e.g.* Gaussian or a mixture of Gaussian. In this subsection, we review popular shape models in literature.

2.1.1. Point distribution model

Multivariate Gaussian distribution is the most commonly assumed, otherwise known as the point distribution model (PDM)

Fig. 2. Illustration of statistical distribution of facial feature points. There are 600 shapes (smaller dot points in black) normalized by Procrustes analysis. The larger dot points in red indicate the mean shape of all shapes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

proposed by Cootes and Taylor [28]:

$$\mathbf{s} = \mathbf{s}_0 + \mathbf{P}_s\boldsymbol{\alpha} = \mathbf{s}_0 + \sum_{i=1}^{n} \alpha_i \mathbf{s}_i, \tag{1}$$

where $\mathbf{s}_i (i = 0, \ldots, n)$ can be estimated by the principal component analysis (PCA) on all normalized training shapes and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_i, \ldots, \alpha_N)^T$ is the PDM shape model parameters. Actually, $\mathbf{s}_0$ is the mean of all these normalized training shapes and $\mathbf{s}_1, \ldots, \mathbf{s}_n$ are the eigenvectors corresponding to the n largest eigenvalues $(\lambda_1, \ldots, \lambda_n)$ of the covariance matrix of all normalized training shapes. n is usually determined by preserving 90–98% variance (the ratio between the sum of n largest eigenvalues and sum of all eigenvalues). Mei et al. [98] suggested the above rule to determine whether the value of n is reliable or not and further explored bootstrap stability analysis to improve reliability.

To remove the effect of rigid transformation, all training shapes are aligned by Procrustes analysis before learning the shape model. We call this rigid transformation-free shape $\mathbf{s}$ in a reference frame. We apply the rigid transformation to $\mathbf{s}$ to generate a shape $\mathbf{x}$ in the image frame:

$$\mathbf{x} = s(\mathbf{I}^{(e)} \otimes \mathbf{R})\mathbf{s} + \mathbf{1} \otimes \mathbf{t}, \tag{2}$$

where s, $\mathbf{R}$ and $\mathbf{t}$ are the scale parameter, rotation matrix and translation parameter in the rigid transformation respectively, $\mathbf{x}$ is a shape represented in the image frame, $\mathbf{I}^{(e)}$ is the identity matrix, $\otimes$ denotes the Kronecker product, and $\mathbf{1}$ is a column vector with all ones.

2.1.2. Other parametric shape models

Based on PDM, there are many variants. Considering that PCA can only model the linear structure of shapes, Cootes and coworkers [128,129] generalized the linear PDM to a nonlinear version by exploring polynomial regression and multi-layer perceptron respectively. Zhou et al. [188] extended the PDM to a tangent space. They imposed continuous regularization on shape parameters in contrast to PDM discontinuously truncated shape parameters to constrain the shape variation, which could result in unstable estimation. Gu and Kanade [61] extended the 2D PDM to 3D PDM by means of a weak perspective projection. Vogler et al. [146] proposed to train several ASMs to construct a 3D deformable model, each corresponding to a viewpoint to govern the 2D facial feature variations.

Component configuration and constraint are usually explored to further improve the representation ability of Gaussian shape model assumption (PDM). Liang et al. [85] utilized the component locations as constraints to regularize the configuration of facial feature points belonging to the same component. Le et al. [78] separated the whole face into seven components and constructs a PDM model for each component. To obtain a reasonable configuration of these components, the locations of these components (centroids of these components) are further modeled by Gaussian distribution. In other words, the shape model is decomposed into two modules: component shape fitting and configuration model fitting. Cao et al. [19] proposed a tensor shape model to predict 3D shape points from videos for the purpose of facial animation.

There are several improvements in the prior shape distribution (Gaussian distribution). Since a single Gaussian is inadequate for modeling the distribution over facial feature points, a mixture of Gaussian has been explored [29,50,123]. In PDM (see Eq. (1)), shapes are constrained on the subspace spanned by principal components. Saragih [117] exploited the principal regression analysis to span a constrained subspace.

2.2. Local part model-based methods

Local part model-based methods, also known as constrained local model, fit an input image for the target shape through optimizing an objective function, which is comprised of two terms: shape prior $\mathcal{R}(\mathbf{p})$ and the sum of response maps $\mathcal{D}_i(\mathbf{x}_i; \mathbf{I}),\ (i = 1, \ldots, N)$ obtained from N independent local experts:

$$\min_{\mathbf{p}} \mathcal{R}(\mathbf{p}) + \sum_{i=1}^{N} \mathcal{D}_i(\mathbf{x}_i; \mathbf{I}). \tag{3}$$

where $\mathbf{p}$ is the model parameters (concatenation of the shape model parameters and rigid pose parameters). $\mathbf{x}_i$ is the location of the ith facial feature point on face image $\mathbf{I}$.

Shape models introduced in Section 2.1 are usually taken as the prior refining the configuration of facial feature points. Each local expert is trained from the facial appearance around the corresponding feature point and is utilized to compute the response map which measures detection accuracy. In the offline phase, a shape model and local experts should be learned from training shapes and corresponding images. Then in the online phase, given an input image, the output shape can be solved from the optimization of Eq. (3). Commonly used local experts will be introduced in next subsection. Subsequently, methods are investigated according to the assumption on the distribution of response maps.

A local expert functions to compute a response map on the local region around corresponding facial feature points, *i.e.* we have N local experts in a FFPD model. The region that supports a local expert could be either one-dimensional (*i.e.* a line as in ASM [31] or two-dimensional (such as a rectangular region [101]). A local expert can be a distance metric such as the Mahalanobis distance in ASM [31], a classifier such as linear support vector machine in [92,121,154], or a regressor [37,120].

The fitting of local part model-based methods consists of two main steps: (1) predicting local displacements of shape model points; (2) constraining the configuration of all point to adhere to the shape model. These two steps are iterated until they satisfy a convergence criterion. *Local part model-based methods differs mostly in how to model the response map generated by local experts.*

2.2.1. Isotropic Gaussian for response map

Cootes and Taylor [28,31] proposed to search the "better" candidate point locations along profiles normal to the boundary. Corresponding displacements from current point locations to sought "better" locations should then be refined to adapt the PDM.

2.2.2. Anisotropic Gaussian for response map

Although isotropic Gaussian estimation to the response maps leads to an efficient and simple approximation, it may fail in some cases if the response maps cannot be modeled by isotropic Gaussian distributions. Wang et al. [154] proposed to approximate the response map by anisotropic Gaussian estimators.

2.2.3. Gaussian mixture model for response map

Considering that response maps may be multimodal, a single Gaussian estimator cannot model the density distribution. Gu and Kanade [62] employed a Gaussian mixture model (GMM) to approximate the response maps.

2.2.4. Nonparametric model for response map

Unlike previous methods approximating response maps in parametric forms, Saragih et al. [119,121] proposed a non-parametric estimate in the form of a homoscedastic isotropic Gaussian kernel density estimate.

Regression is another way to nonparametricly approximate response maps. Asthana et al. [7] directly regressed the PDM shape update parameters from the low-dimensional representation of response maps through a series of weak learners. The response maps can be obtained from linear support vector machines and the low-dimensional representation is obtained from the PCA projection. Linear support vector regression plays the role of the weak learner. Amberg and Better [3] utilized decision forest [16] to detect a number of candidate locations for each point. The facial feature localization problem is actually to determine the indexes of points in corresponding candidate points which minimize the distance between the shape model and the image points. Martinez et al. [95] believed that each image patch evaluated by the regressors adds evidence to the target location rather than just taking the last estimate (the last iteration) into account and discarding the rest of these estimates. They aggregated all up-to-date local evidence obtained from support vector regression by an unnormalized mixture of Gaussian distributions.

2.3. Holistic model-based methods

Holistic models refer to approaches which learn a holistic appearance (or texture) representation. Image intensities located at all pixels of a face are used to encode the appearance. Under the framework of parameterized shape model, holistic model-based methods generally consist of three parts: a statistical shape model, a global motion model and an appearance model. Aforementioned PDM is generally utilized as the shape model for holistic model-based methods. The motion model defines how an image should be warped to a reference frame (usually defined by the mean shape), given the shape corresponding to the image. Piece-wise affine [97] and thin plate spline [2,109] are two popular motion models appeared in FFPD literatures.

According to the way for modeling the holistic appearance, holistic model-based methods can be further divided into two groups: parametric holistic model-based methods and nonparametric holistic model-based methods.

2.3.1. Parametric holistic model-based methods

The most well-known parametric strategy to represent the face appearance variation is the linear model in AAM [57], which assumes the shape-free textures distribute as a Gaussian. Since methods in this group (parametric holistic model-based methods) are AAM oriented, this category of methods can also be named after AAM based methods. Subsequently, we would first introduce active appearance models and then review different AAM fitting methods.

Active appearance models: An active appearance model (AAM) [26] can be decoupled into a linear shape model (PDM) and a linear texture model. To construct the texture model, all training faces should be warped to the mean-shape frame by piece-wise affine or thin plate spline; the resultant images should be free of shape variation, called shape-free textures. The texture model can be generated by applying PCA on all normalized textures as follows:

$$\mathbf{a} = \mathbf{a}_0 + \mathbf{P}_a\boldsymbol{\beta} = \mathbf{a}_0 + \sum_{i=1}^{m} \beta_i \mathbf{a}_i, \tag{4}$$

The coupled relationship between the shape model and the texture model is bridged by PCA on concatenated shape parameters $\boldsymbol{\alpha}$ and texture parameters $\boldsymbol{\beta}$. To simplify the parameter representation, here we still utilize $\mathbf{p}$ to incorporate all necessary parameters: $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, pose parameters $\mathbf{q}$, u and v.

The fitting objective of AAM is to find the model parameters $\mathbf{p}$ which best fit a given image $\mathbf{I}$. This is usually achieved by an iterative process which update $\mathbf{p}$ through an update function in a sequential manner:

$$\Delta\mathbf{p} = \mathcal{U}(\mathbf{p}) \circ \mathcal{F}(\mathbf{I}; \mathbf{p}), \tag{5}$$

Where $\mathcal{U}$ is a vector valued update function, with optional dependence on $\mathbf{p}$ and $\mathcal{F}$ is a vector valued feature extractor depending on current parameter settings $\mathbf{p}$.

According to the way the fitting strategy is designed, parametric holistic model-based methods (*i.e.* AAM-based methods) can be further categorized into two classes: discriminative fitting based methods and generative fitting based methods.

Discriminative fitting based Methods: Discriminative fitting based methods directly learn a fixed map from the features $\mathcal{F}(\mathbf{I}; \mathbf{p})$ to the parameter updates $\Delta\mathbf{p}$, given a training set of perturbed model parameters: $\{\mathcal{F}(\mathbf{I}; \mathbf{p}^* - \Delta\mathbf{p}), \Delta\mathbf{p}_i^{N_d}\}$ where $\mathbf{p}^*$ is the ground truth parameter for the ith training sample and N_d represents the capacity of the training set. The advantage of this category of methods is that if the update function samples from a simple function class, then the computation of the updates can be done efficiently [118]. However, this indicates the disadvantage: the function form of $\mathcal{F}$ and $\mathcal{U}$ should be chosen heuristically.

Edwards et al. [48] observed that the relationship between the residual texture $\mathbf{r}(\mathbf{p}) = \mathbf{I}(\mathcal{W}(\mathbf{p})) - \mathbf{a}_m$ and the parameter updates $\Delta\mathbf{p}$ is close to linear around the optimal model parameters. Here $\mathcal{W}(\mathbf{p})$ represents a warping function (*e.g.* piece-wise affine function or thin plate spline function) and $\mathbf{a}_m$ is the texture synthesized by the model (see Eq. (4)). Then the update function $\mathcal{U}$ is resolved from a linear regression problem on aforementioned training set. Some works considered that the linear assumption between residual textures and parameter updates does not hold directly and learned the linear regression from some presentations of residual textures (*e.g.* PCA representation in [66] and canonical correlation analysis representation in [47]).

Although the linear regression strategy achieves some success in obtaining the updated parameters, it is a coarse approximation of the nonlinear relation between texture residuals and warp parameters. When the parameters are initialized far away from the right place, this linear assumption is invalid. To this end, Saragih and Goecke [118] deployed a nonlinear boosting procedure to learn the multivariate regression. Each parameter is updated by a strong regressor consisting of an ensemble of weak learners [55]. This nonlinear modeling results in a more accurate fitting than linear procedures. Tresadern et al. [137] suggested a hybrid AAM which utilizes a nonlinear additive update model at the first several iterations and then a linear additive update model in the last several iterations. Nguyen and Torre [104,105] claimed that AAMs are easily converged to local minima in the fitting process and that the local minima seldom correspond to acceptable solutions. They

proposed a parameterized appearance model which learns a cost function having local minima at and only at desired places.

Generative fitting based methods: Generative fitting based methods minimize some measure of residual textures. The most common utilized is the least squares (or a robust variation thereof [30]):

$$\min_{\mathbf{p}} \|\mathbf{I}(\mathcal{W}(\mathbf{p})) - \mathbf{a}_m\|_2, \tag{6}$$

where $\mathbf{a}_m$ is the texture generated by the model.

The above nonlinear optimization problem could be resolved by Gauss–Newton method, which results in a linear update model [27]. However, a straight forward implementation is time consuming. Therefore most generative fitting methods assume some parts of the method are fixed or rewritten the problem such that they are [118]. The original generative fitting method [26,27] fixed the Jacobian of Eq. (6). Then a fixed linear update model can be precomputed through a pseudo-inverse of the fixed Jacobian. The fixed gradient matrix may lead to poor performance when the texture of a testing image differs dramatically from the mean texture. Batur and Hayes [11] updated the gradient matrix in each iteration by adding a linear combination of basis matrixes to a fixed basic matrix.

In order to improve the efficiency of the fitting process, Matthews and Baker [97] considered the AAM as an image alignment problem and optimized it by inverse compositional method [9] based on independent AAM. Here, independent AAM indicates that the linear shape model and linear texture model are not combined as in the original literature [27]. The inverse compositional method has had many variants since its birth. It has been applied to solve the robust and efficient FFPD objective [142] which aims to detect points under occlusion and illumination changes. Lucey et al. [91] extended the inverse compositional method in the Fourier domain for image alignment and applied this method specifically to the case of AAM fitting [103]. Tzimiropoulos and Pantic [140] proposed a framework for efficiently solving AAM fitting problem in both forward and inverse coordinate frames. They trained and fitted AAM in-the-wild and the trained model could achieve promising performance. Alabort-i-Medina and Zafeiriou [1] proposed a generative Bayesian version of AAM which achieved state-of-the-art performance in-the-wild databases. Due to the removal of the calculation of texture parameters, this method can be implemented more efficiently in comparison to other algorithms.

2.3.2. Nonparametric holistic model-based methods

Unlike Gaussian assumption to the appearance variation in the parametric holistic model, in the nonparametric holistic model, there is no distribution assumption. By contrast, classifiers (*e.g.* RankBoost [54], GentleBoost [56]) or regressors are generally applied to separate the aligned position from the not aligned. The model parameters (shape model parameter and pose parameter) are estimated by maximizing the classification score of the warped test image.

Zheng et al. [185] proposed a rank-based non-rigid shape detection method through RankBoost. RankBoost is utilized to learn a ranking model from Haar-like features extracted from warped training images. One disadvantage of this method is that the detection efficiency is seriously affected by the number of images in the training set.

Liu [88] explored GentleBoost classifier to model the nonlinear relationship between texture and parameter updates. Haar-like rectangular features are fed into each weak classifier. The goal of the fitting procedure is to find the PDM parameter updates which maximize the score of the strong classifier. Zhang et al. [174] utilized granular features to replace the rectangular Haar-like feature to improve computational efficiency, discriminability and a larger search space. Because the weak classifier in [88] is actually utilized to classify the right PDM parameters from the wrong ones, it cannot guarantee that the fitting objective will converge to the optimum solution. Consequently, instead of discriminatively classifying corrected alignment from incorrect alignment, Wu et al. [157] learned GentleBoost classifiers to determine whether to switch from one shape parameters to another parameter corresponding to an improved alignment.

3. Nonparametric shape model-based methods

Nonparametric shape model-based methods do not rely on any shape distribution. However, this does not mean these methods are free from the shape constraint during the prediction process. Actually, the shape constraint may be implicitly embedded in the model (regression based methods, *e.g.* [20]) or explicitly constrained by exemplar shapes in the training set (exemplar-based methods). It can also be in the form of a graphical structure (graphical model-based methods) or a deep structure (deep learning based methods).

3.1. Exemplar-based methods

Exemplar-based FFPD methods generally find some exemplars (shapes) from the training set to constrain the configuration of facial feature points in a data-driven manner. The advantage of exemplar-based methods is that there is no need to learn a parametric shape model (*e.g.* PDM) which often results in a non-convex problem difficult to optimize. However, these methods also have the drawback of time consuming to seek for similar exemplars in a training set, especially when there is a large-scale training set.

Belhumeur et al. [12] proposed a method that combines the output of local experts with a non-parametric global model. The local expert is applied through a support vector machine taking the SIFT feature [90] as the input. Based on the response maps $\mathbf{d}$ of these local experts, the objective is to maximize the posterior probability $p(\mathbf{x}|\mathbf{d})$. Since the location corresponding to the highest response map value is not always the correct location due to occlusions and appearance ambiguities, they explored random sample consensus method (RANSAC) [53] to find some similar exemplar images (with their shapes) from the training set to constrain the configurations of these facial feature points. RANSAC is further used to optimize the objective function.

The method of [12] motivates many researchers to design more effective FFPD methods in the aspects of RANSAC for similar exemplars seeking and shape model learned from exemplar shapes. Smith et al. [127] employed the exemplar seeking strategy RANSAC to perform face parsing. Zhou et al. [186] also adopted the exemplar seeking strategy (RANSAC) as in [12]. An online shape constraint is constructed from exemplars to refine the configuration of point candidates calculated from some local experts as aforementioned. The optimization problem is formulated as a graph matching problem which could be efficiently solved by linear programming.

Shen et al. [122] proposed the detection of each facial feature point through a Hough transform [82] based voting strategy on corresponding points on some exemplar images retrieved from the training dataset. The location corresponding to the peak response in each voting map is the estimated position. Also based on Hough voting, Smith et al. [124] retrieved a subset of top similar exemplar faces from the training set and calculated a voting map for final facial feature point estimation. The global shape regularization technique [12] is employed to distinguish the “correct” position from multiple peak responses in a voting map. Smith and Zhang

[126] applied a similar exemplar seeking, voting, and shape constraint strategy as that in [124] to transferring facial feature point across different datasets which have different number of facial feature points for each face.

3.2. Graphical model-based methods

Graphical model-based FFPD methods usually constrain the configuration of facial feature points through a graphical structure: tree structure or loopy graph structure. Then these methods mainly refer to tree structure based methods and Markov random field (MRF) based methods. Tree structure based methods take each facial feature point as a node and all points as a tree. The locations of facial feature points can be optimally solved by dynamic programming. Unlike the tree structure which has no loop, MRF-based methods model the location of all points with loops.

3.2.1. MRF based methods

Coughlan and Ferreira [33] developed a generative Bayesian graphical model that deployed separate models to describe shape variability (shape prior) and appearance variations (appearance likelihood) to find deformable shapes. The shape prior takes the location of each facial feature point and these points' normal orientation as a node in MRF. An edge map and an orientation map are calculated to model the appearance likelihood. A variant of the belief propagation method is utilized to optimize the problem. MRF has been also explored to constrain the relative position of all facial feature points obtained from the regression procedure in [95,143].

Unlike the method in [33] which models the shape prior only in a local neighborhood, Liang et al. [84] proposed a method that incorporates a global shape prior directly into the Markov network. The local shape prior is enforced by denoting a line segment as a node of the constructed Markov network. Here, line segments draw from one facial point to another neighboring point. Subsequently, Liang et al. [83] claimed that although local part based methods take the global shape prior into account, these methods neglect the neighboring constraint between points since they compute the response map of each point independently. They further incorporated the PDM shape prior into their model.

Another work that considers both the local characteristics and global characteristics of facial shapes is the bi-stage component-based facial feature point detection method [69]. The whole face shape is divided into seven parts. The shape of each part is modeled as a Markov network by taking each point as a node. Belief propagation is explored to find the locations of these components. Then, configurations of these components are constrained by the global shape prior described by the Gaussian process latent variable model.

In consideration of the nonlinear property of the facial feature localization problem and generalization ability, Valstar et al. [143] deployed support vector regression to output the target point location from the input local appearance-based features (Haar-like features). To overcome the overfitting problem due to the high dimensionality of the Haar-like features, Adaboost regression is utilized to perform feature selection.

3.2.2. Tree structure based methods

Kozakaya et al. [76,77] learned an implicit tree structure from training data to constrain the configurations of all facial feature points. For training, some grid sampling points are evenly placed on each face image and an extended feature vector is extracted for each sampling point. The tree structure is implicitly embedded into the feature vector by taking N directional vectors from the sampling point to all N feature points into account. In the detection phase, given an input face image, local descriptors corresponding to each sampling point are extracted. Then a nearest extended feature vector can be found for each sampling point from training set using the approximate nearest neighbor search (ANNS) algorithm [5]. Finally, each feature point is independently estimated from a weighted square distance.

Zhu and Ramanan [193] proposed a unified model for face detection, head pose estimation and point location estimation. Their method is based on a mixture of trees, each of which corresponds to one head pose view. These different trees share a pool of parts. In the training stage, the tree structure is first estimated via Chow-Liu algorithm [24]. Then a model of the tree-structured pictorial structure [52] is constructed for each view. In the testing stage, the input image is scored by all tree structures respectively and the pixel locations corresponding to the tree with maximum score are the final facial feature point locations. Since tree-structure-based methods only consider the local neighboring relation and neglect the global shape configuration, they may easily lead to unreasonable facial shape. Based on the method [193], Ghiasi and Fowlkes [59] incorporated the statistics of occlusion patterns in a hierarchical discriminatively trained model. The hierarchical structure consists of two layers with the face consisting of a collection of parts, each of which is composed of a number of facial feature points.

Dantone et al. [39] proposed a facial feature point detection by extending the concept of regression forests [17,36] to conditional regression forests. They claimed that it is difficult for general regression forests to learn the variations of faces with different head poses. The head pose is evaluated by regression forests. A regression forest is constructed, conditioned on the head pose (*i.e.* there is one regression forest corresponding to each head pose). In the testing phase, the probabilities of the head pose of an input testing image should be first calculated and, according to this distribution, the number of trees selected from each forest can be determined. Finally the position of each facial feature point can be computed through solving a mean-shift problem. Yang and Patras added structural information into the random regression and proposed a structured-output regression forest-based face parts localization method [166]. Then, they [167] proposed to deploy a cascade of sieves to refine the voting map obtained from random regression forest.

3.3. Cascaded regression based methods

Aforementioned methods mostly govern the shape variations through certain parameters, such as PDM coefficient vector $\boldsymbol{\alpha}$ in ASM and AAM. By contrast, cascaded regression-based methods directly learn a regression function from image appearance (feature) to the target output (shape):

$$\mathcal{M} : f(\mathbf{I}) \rightarrow \mathbf{x} \in \mathbb{R}^{2N}, \tag{7}$$

where $\mathcal{M}$ denotes the mapping from image appearance (feature) $f(\mathbf{I})$ to the shape $\mathbf{x}$ and f is the feature extractor. The mapping $\mathcal{M}$ is usually implemented in a coarse-to-fine fashion for cascaded regression-based methods as presented in Algorithm 1. Shape estimation starts from an initial shape $\mathbf{x}^0$ and progressively refines the shape by a cascade of T regressors, $R^{t=1,\ldots,T}$. Each regressor refines

Algorithm 1 Cascaded regression for FFPD.

Require: Image $\mathbf{I}$, initial shape $\mathbf{x}^0$
Ensure: Estimated shape $\mathbf{x}^T$
For $t = 1$ to T **do**
  $f^t = h^t(\mathbf{I}, \mathbf{x}^{t-1})$ // Shape-indexed features
  $\Delta\mathbf{x} = R^t(f^t)$ // Apply regressor R^t
  $\mathbf{x}^t = \mathbf{x}^{t-1} + \Delta\mathbf{x}$ // Update shape
End For

the shape by producing an update $\Delta\mathbf{x}$, which is added up to the current shape estimate.

As shown in Algorithm 1, cascaded regression based methods for FFPD differ from each other mainly in three aspects [165]: strategy of setting initialization, shape-indexed features and regressor. A large variety of image features are utilized as the shape-indexed features, *e.g.* pixel difference [18,20,73], SIFT [90] in [164,191], HOG [38] in [8], and learned features in [115].

There are mainly two initialization schemes appeared in cascaded regression-based methods: random [18,20,191] and mean pose [8,73,115,158,164]. The *random* method usually selects one or several face shapes from the training dataset and then rescale them with respect to the provided face bounding box via similarity transformation. The final shape is the average or median of all estimated shapes from different initializations. The *mean pose* initialization method calculates a mean shape from the training samples within the face bounding box. More detailed analysis on shape initialization are given in Section 4.3.

Cascaded regression for FFPD is inspired by the seminal work: cascaded pose regression [46]. Cao et al. [20] proposed a two-level cascaded regression method (namely ESR, *i.e.* explicit shape regression). Each regressor R^t $(t = 1, \ldots, T)$ in the first level consists of cascaded random fern [108] regressors (r^k, $k = 1, \ldots, K$) in the second level. Shape-indexed features (pixel difference values) are extracted from the whole image and are fed into the regressor. To reduce the complexity of feature selection, the authors proposed a correlation-based feature selection strategy. One remarkable advantage of the proposed learning framework is that it implicitly enforces shape constraint in the regression process if the initialization is a reasonable shape. To cater for this purpose, Shapes randomly selected from the training set are taken as the initialization and the final shape is the media fusion of all shapes resulted from different initializations.

Burgos-Artizzu et al. [18] boosted the robustness of the ESR method [20] against large pose variations and occlusion from two aspects (namely RCPR, *i.e.* robust cascaded pose regression). Firstly, ESR indexes pixel by its local coordinates with respect to its closest facial feature point, which is not robust against large pose variations and shape deformation. RCPR indexes pixels by linear interpolation between two points. Secondly, Burgos-Artizzu et al. presented a strategy to incorporate the occlusion information into the regression which improves the robustness to occlusion. Kazemi and Sullivan [73] substituted the random fern regressor in [20] with regression trees. Gradient tree boosting [63] is employed to train each regressor. A simple exponential prior over the distance between the pixels is used to select features to encourage closer pixel pairs to be chosen.

In view of the boosted regression in [20], which is a greedy method to approximate the function mapping from facial image appearance features to facial shape updates, Xiong and Torre [164] developed the supervised descent method (SDM) to solve a series of linear least squares problems as follows:

$$\operatorname{argmin}_{\mathbf{R}_k,\mathbf{b}_k} \sum_{\mathbf{d}^i} \sum_{\mathbf{x}_k^i} \left\| \Delta\mathbf{x}_*^i - \mathbf{R}_k\boldsymbol{\phi}_k^i - \mathbf{b}_k \right\|^2, \tag{8}$$

where $\Delta\mathbf{x}_*^i = \mathbf{x}_*^i - \mathbf{x}_k^i$ is the ground truth difference between the truth shape $\mathbf{x}_*^i$ of the ith training image $\mathbf{d}^i$ and the shape $\mathbf{x}_k^i$ obtained from the kth iteration, $\boldsymbol{\phi}_k^i$ is the extracted SIFT features around the shape $\mathbf{x}_k^i$ on the training image, $\mathbf{R}_k$ is called the common descent direction in this paper and $\mathbf{b}_k$ is a biased term. This method has a natural derivation process based on the Newton method. A series of $\{\mathbf{R}_k, \mathbf{b}_k\}$ are learned in the training stage, and in the testing stage they are applied to the SIFT features extracted from the testing image to update the shape sequentially. SDM is a local algorithm and it is likely to average conflicting gradient directions. For resolving this problem, authors proposed global SDM (GSDM) in their subsequent work [163], an extension of SDM that divides the search space into regions of similar gradient directions. Tzimiropoulos [139] proposed a project-our cascaded regression method which learns a sequence of averaged Jacobian and Hessian matrices.

Since the major computation of SDM method is the feature extraction and simple linear multiplication, it has attracted great attentions. Ren et al. [115] considered that the handcrafted general purpose features (SIFT) are not optimal for specific FFPD. They proposed a learning (random forest regression [17]) based approach to encode the whole face by some sparse local binary features. Later, they applied the framework to jointly perform face detection and alignment [21]. Asthana et al. [8] extended the cascaded linear regression to incremental learning for face alignment both in a sequential and a parallel formulation. Zhu et al. [191] proposed to learn the linear regressor from some shapes and faces sampled from training data according to some learned distribution as in [12]. Within the SDM framework, Wu and Ji [158] extended it to be robust against significant head poses and occlusion by sequentially predicting the landmark visibility probabilities and they also extended the SDM framework for simultaneous facial action unit recognition and facial landmark localization [159]. Liu et al. [87] jointly modeled the face alignment problem with face reconstruction problem together by cascaded linear regressors. Considering the hardness of localization is unbalanced across different facial landmarks, Ge et al. [58] conduct local cascaded regression for each landmark after a global cascaded regression for all landmarks. Wang et al. [155] proposed to employ multiple cascaded linear regressors which take different partial features as the input to reduce the problem of occlusion. Deng et al. [42] introduced the sparse shape constraint into the cascaded linear regression framework and they claimed that the sparse shape constraint could suppress the ambiguity in local features and outlier caused by occlusion.

3.4. Deep learning-based methods

Deep networks (*e.g.* convolutional neural network (CNN) [80], deep autoencoder (DAE) [65] and restricted Boltzmann machine (RBM) [64]) have achieved great success in many computer vision applications such as scene classification [51], image segmentation [106], objection recognition [25] and tracking [177]. For FFPD, deep networks are used for either learning the nonlinear shape and appearance variation or learning the nonlinear mapping from the face appearance to the face shape.

3.4.1. Deep learning for nonlinear shape/appearance variation

Wu et al. [160] explored deep belief networks to capture face shape variations due to facial expression variations. In order to localize landmarks for faces on different poses, they further utilized a 3-way RBM to capture the relationship between frontal face shapes and non-frontal face shapes. It is reported that it performs better than the linear model AAM which learns the face shape and appearance variation in a linear fashion. However, it has limited performance on faces with large pose variations and exaggerated expression variations. Subsequently, they proposed a hierarchical probabilistic model for the sake of expressions and poses changing facial feature point location dramatically [161].

3.4.2. Deep learning for nonlinear mapping from appearance to shape

Luo et al. [93] proposed a hierarchical face parsing method based on deep learning. They recast the FFPD problem as the process of finding the label maps (segmentation) which clearly indicate the pixels belong to a certain component. The feature point can then be easily obtained from the boundary of the label maps.

The proposed hierarchical framework consists of four layers: face detector (the first layer), facial parts detectors (the second layer), facial component detectors (the third layer), and facial component segmentation. The structure of this model is somewhat like a pictorial structure [52]: the face detector can be seen as the root node and other detectors (part detectors and component detectors) as the child nodes. The objective function can be formulated in a Bayesian (maximum a posterior) form. The prior term denotes the spatial consistency between detectors of different layers and is modeled as the Gaussian distribution. The likelihood term represents the detectors and segmentation. All detectors can be learned by RBM and segmentation can be learned by a deep autoencoder-like method.

Sun et al. [132] proposed a three-level cascaded deep convolutional network framework (DCNN) for point detection in a coarse-to-fine manner similar as cascaded regression-based methods. Each level is composed of several numbers of convolutional networks. The first level gives an initial estimate to the point position and the following two levels then refine the obtained initial estimate to a more accurate one. Though great accuracy can be achieved, this method needs to model each point by a convolutional network which improves the complexity of the whole model. Moreover, with the increase in the number of facial feature points, the time consumption to detect all points is high.

Motivated by the successful application of DCNN to FFPD [132], Zhang et al. [179] proposed a tasks-constrained DCNN to jointly optimize FFPD with a set of related tasks: head pose estimation, gender classification, age estimation, facial expression recognition or facial attribute inference. Multiple tasks share and learn common deep layers and result in a shared representation which facilitates the learning of the main task: FFPD. Inspired by the work [132] and cascaded regression learning framework [164], Zhang et al. [176] proposed a coarse-to-fine autoencoder network (CFAN) based method (*i.e.* cascaded autoencoder network). The CFAN framework is composed of four successive stacked auto-encoder networks (SANs). Each SAN attempts to characterize the nonlinear mappings from the face image to the face shape in different resolutions. The first SAN are used for initialize the face shape by taking global image features as the input. The subsequent three SANs are utilized to fine tune the face shape taking local shape-indexed features as the input.

In addition to convolutional neural networks (CNN), recurrent neural networks (RNN) are another technique utilized for the task of FFPD. Xiao et al. [162] proposed a recurrent attentive-refinement (RAR) model which refines the landmark locations sequentially at each recurrent stage. Peng et al. [112] employed recurrent encoder-decoder network to model the spatial and temporal variations for video-based face alignment. Chen et al. [23] employed long short term memory architecture of RNN to obtain the initial landmark estimation and then deep neural network are applied to perform fine local search.

4. Discussions

In this section, we give some discussions on issues beyond above reviewed methods. These issues incorporate some special methods or used in special scenarios not included in above categories, FFPD applications and some important issues affecting the performance of a FFPD approach.

4.1. Joint face alignment

In literatures, jointly aligning an ensemble of face images undergoing a variety of geometric and appearance variations are called joint face alignment. Learned-Miller [79] was the first to propose the idea of congealing which is to start with a set of images and make them appear as similar as possible in an unsupervised manner. Actually these congealing-based unsupervised joint face alignment methods [34,35,67,189] are different from aforementioned FFPD methods. FFPD methods are used to localize some specific facial features or identifying parts of faces while these congealing-based methods are used to place the faces into the same canonical pose for subsequent processing such as face identification task.

Motivated by the congealing-style joint alignment method [79] and sparse and low-rank decomposition method [113], Zhao et al. [181] designed a joint AAM by assuming that the images of the same face should lie in the same linear subspace and the person-specific space should be proximate the generic appearance space. Smith and Zhang [125] stated that the method of [181] breaks down under several common conditions, such as significant occlusion or shadow, image degradation, and outliers. Then they introduced the global shape model in [12] combined with a local appearance model into the joint alignment framework. Zhao et al. [183] proposed to distinguish the “good” alignments from the “bad” ones among all these initial estimations obtained from [62]. A discriminative face alignment evaluation metric is designed by virtue of cascaded AdaBoost framework [145] and Real AdaBoost [56]. Tong et al. [134,135] proposed a semi-supervised facial landmark localization approach which utilizes a small number of manually labeled images. To obtain a reasonable shape, an online learned PDM shape model is imposed as a constraint.

Though above methods has achieved some success for batch face alignment, they generally have a high computation complexity. In order to achieve real-time experience for users in real-world applications, *e.g.* internet photo sharing sites such as Facebook and Flickr, personal photo library software like iPhoto and Picasa, and video conferencing, more efficient joint face alignment methods should be designed in a much more efficient manner.

4.2. Independent facial feature point detectors

The aforementioned methods predict the locations of all facial feature points or a group of points simultaneously. There are other methods which detect each point independently. Here, methods which do not rely on manually labeled images, such as approach [6], are not included.

Vukadinovic and Pantic [147] detected each point by a local expert as utilized in local part based methods. Here, Gabor feature-based boosted classifier is utilized to classify the positive image patch from the negative image patch. The position with the peak response among the response map of each point is the sought location.

Considering the fact that there is great variability among faces and facial features, such as eye centers and eye corners, Ding and Martinez [44,45] employed subclass discriminant analysis [190] to divide vectors (features or context) of the same class into subclasses. Vectors centered on the facial feature point are called features and vectors centered on points surrounding the facial feature point are called context. The K-means clustering method is explored to divide each class into a number of subclasses. Given the detected face box, facial feature points can be exhaustively searched in some windows located relative to the bounding box by comparison with the learned subclasses at different scales. The final facial feature point is achieved by a voting strategy on different detected positions at different scales.

The advantage of independent facial feature point detectors is the initialization free character. One major disadvantage is the ambiguity problem. This means there exist more than one positions looking like the target point, especially under complex environ-

ment like deliberately disguise, occlusion or pose variation. To address this problem, Zhao et al. [184] proposed to jointly estimate correct positions of all points from some candidates obtained by independent facial feature point detectors.

4.3. Initialization

Initialization is important for a FFPD method to accurately localize facial feature points. Most existing methods are initialized by a mean shape and a bounding box returned by a face detector [173].

The most simple strategy is to put the mean shape at the center of the bounding box and then scale the mean shape by the size of the bounding box. The drawback of this strategy is that it is heavily depending on the performance of face detectors, *i.e.* if the face returned bounding box has some translation displacement and scale variation in comparison to ground truth, it would be very difficult to accurately localize point positions. An advanced strategy is to model the relative position of a mean shape to the detected face by a distribution such as Gaussian [164] relying on the training set. Then in the testing phase, the mean shape is positioned to some location according to the scale and translation displacement sampled from the learned Gaussian distribution. This strategy can be applied to each point independently as in [143].

Another strategy is to randomly select some shapes from the training set to initialize the FFPD method [20]. The final detection points is the average or median of all results from different initial shapes. To take advantage of the correlation among results from different initial shapes, Burgos-Artizzu et al. [18] proposed a smart strategy by dropping out some initial shapes whose corresponding detection results are greatly different from others. Zhu et al. [191] employed the dominant set approach [111] to linearly combine the multiple estimations.

Above initialization strategies could work well for frontal faces but for extreme poses and some expressions, these initialization strategy may fail. An effective strategy to overcome this is to initialize the shape by similar exemplar shapes from the training set. This strategy has been applied in the exemplar-based FFPD approaches such as [124,126].

Zhang et al. [179] proposed a deep learning based FFPD method which could achieve best up to date performance on detecting five points: two eye centers, two mouth corners and the nose trip. Benefited from the efficient and accurate detection performance, this method can be used to initialize existing FFPD methods and they reported a great improvement in comparison to a state-of-the method [18] without using this method for initialization.

Face detection also affects the initialization since most recent FFPD methods are built on top of face detection but from different face detectors, *e.g.* Viola-Jones [145], dlib [74] and HeadHunter [96]. Yang et al. [165] gave a detailed and comprehensive analysis to the influence of face center shifts and scale variations (*i.e.* to FFPD accuracies.

4.4. Applications

Many related research topics and real-world applications could benefit from the accurate detection of facial feature points. FFPD for face alignment is an essential preprocessing step in face hallucination [153], face recognition [43,182], and facial swapping [14]. Facial animation [156] generally detects facial feature points to control the variation of facial appearance. Chen et al. [22] employed ASM [28] to separate the shape from the texture to favor the sketch generation process. Similar ideas appear in the face sketch portrait synthesis works [148-152,178]. Zhou et al. [187] proposed a fusion strategy to incorporate subspace model constraints into active shape model for robust shape tracking. Stegmann et al. [130] applied AAM [26] to medical image analysis. Lee and Kim [81] explored the fitted shape and shape-normalized appearance of the proposed tensor-based AAM to transform the input image into a normalized image (frontal pose, neural expression, and normal illumination) to conduct variation-robust face recognition. Tresadern et al. [136] explored Haar-like features to provide a computationally inexpensive linear projection in AAM for efficiency to facilitate facial feature point tracking on a mobile device. Anderson et al. [4] applied AAM to track robustly and quickly over a very large corpus of expressive facial data and to synthesize video realistic renderings in the visual text-to-speech system. The correspondence of facial feature points plays an important role in 3D face modeling [15]. The 3D information is usually explored to improve the robustness to pose variation in FFPD. For example, the combination of 2D and 3D view-based AAM is utilized to robustly describe the variation of facial expression across different poses [133].

5. Databases

There are many face databases publically available due to the easy acquisition of images and the fast development of social networks such as Facebook, Flickr, and Google+. The ground truth facial feature points are usually labeled manually by employing workers or through crowdsourcing, *e.g.* the Amazon mechanical turk (MTurk). Each face image is generally labeled by several workers and the average of these labeled results is taken as the final ground truth. These face databases can be classified into two categories: databases captured in controlled conditions and databases captured in uncontrolled conditions (*i.e.* in the wild). Controlled databases are taken under the framework of predefined experimental settings such as the variation of illumination, occlusions, head pose and facial expressions. Databases in the wild are generally collected from websites such as Facebook and Flickr. Following are some representation collections which are popularly used in empirical studies.

5.1. Databases collected under well-controlled conditions

CMU Multi-PIE [60] face database was collected in four sessions between October 2004 and March 2005. It aims to support the development of algorithms for recognition of faces across pose, illumination and expression conditions. This database contains 337 subjects and more than 750,000 images for 305GB of data. A total of six different expressions are recorded: neutral, smile, surprise, squint, disgust and scream. Subjects were recorded across 15 views and under 19 different illumination conditions. A subset of this database has been labeled either 68 points or 39 points depending on their view but facial feature points are not published online.

Extended M2VTS database (XM2VTS) database [99] collected 2360 color images, sound files and 3D face models of 295 people. The database contains four recordings of these 295 subjects taken over a period of four months. Each recording was captured when the subject was speaking or rotating his/her head. These 2360 color images are labeled with 68 points.

AR database [94] contains over 4000 color images corresponding to the faces of 126 people (70 men and 56 women). Images were taken under strictly controlled conditions and with different facial expressions, illumination conditions, and occlusions (sunglasses and scarf). Each person appeared in two sessions, separated by two weeks. Ding and Martinez [45] manually annotated 130 points on each face image.

IMM database [107] contains 240 color images of 40 persons (7 females and 33 males). Each image is labeled with 58 points around the eyebrows, eyes, nose, mouth and jaw.

MUCT database [100] consists of 3755 face images of 276 subjects and each image is marked with 76 manual points. Faces in this database are captured under different lighting conditions, at various ages, and are of several different ethnicities.

PUT database [72] collected 9,971 high resolution images (2048 × 1536) of 100 people taken in partially controlled illumination conditions with rotations along the pitch and yaw angle. Each image is labeled with 30 points. A subset of 2193 near-frontal images is provided with 194 control points.

Databases in the wild

BioID database [70] was recorded in an indoor lab environment, but "real world" conditions were used. This database contains 1521 grey level face images of 23 subjects and each image is labeled with 20 points.

LFW database [68] contains 13,233 face images of 5749 subjects collected from the web. Each face in the database has been labeled with the name of the person pictured. 1680 of the people pictured have two or more distinct photos in the data set. The constructors of this database did not provide manually labeled points but there are other available sites.

Annotated Facial Landmarks in the Wild (AFLW) database [75] is a large-scale, multi-view, real-world face database with annotated facial feature points. Images were collected from Flickr using a wide range of face relevant key words such as face, mugshot, and profile face. This database includes 25,993 images in total and each image is labeled with 21 landmarks.

Labeled Face Parts in the Wild (LFPW) database [12] is composed of 1400 face images (1100 as the training set and the other 300 images are taken as the testing set) downloaded from the web using simple text queries on websites such as Google.com, Flickr.com, and Yahoo.com. Due to copyright issues, the authors did not distribute image files but provided a list of image URLs. However, some image links are no longer available. 35 points are labeled in total; 29 of them are usually utilized in literatures.

Annotated Faces in the Wild (AFW) database [193] contains 205 images with a highly cluttered background and large variations both in face scale and pose. Each image is labeled with 6 points and the bounding box of the corresponding face.

Helen database [78] contains 2300 high resolution face images collected from Flickr.com. Each face image is labeled with 194 points.

300 Faces in-the-Wild Challenge (300-W) [116] is a mixed database consisting of face images from several published databases (LFPW, Helen, AFW, and XM2VTS) and a new collected database IBUG. All these images are re-annotated with 68 points. This database is published for the first Automatic Facial Landmark Detection in-the-Wild Challenge (300-W 2013) held in conjunction with the International Conference on Computer Vision 2013.

Caltech Occluded Faces in the Wild (COFW) database [18] is composed of 1007 face images showing large variations in shape and occlusions due to differences in pose, expression, use of accessories such as sunglasses and hats and interactions with objects (*e.g.* food, hands, microphones, *etc.*). 29 points are marked for each image. The major difference between this database and other ones is that each point is explicitly labeled whether it is occluded. This database presents a great challenging task for facial feature point detection due to the large amount and variety of occlusions and large shape variations.

6. Facial feature point detection evaluations

6.1. Evaluation metric

The distance from the estimated points to the ground truth normalized by the interocular distance is a common informative metric for evaluating a facial feature point detection system (named mean normalized error, MNE, in the following text):

$$e_i = \frac{\left\| \mathbf{x}_{(i)}^e - \mathbf{x}_{(i)}^g \right\|_2}{d_{io}}, \tag{9}$$

where $\mathbf{x}_{(i)}^e$ is the ith estimated point and $\mathbf{x}_{(i)}^g$ is its corresponding ground truth. In some works (*e.g.* [1,193]), the interocular distance is substitute by the face size. The error in Eq. (9) is landmark-wise. In order to evaluate the performance of a set of images, there are mainly two methods. One is the mean error, sample-wise, landmark-wise or overall. The other one is the Cumulative Error Distribution (CED) curve, which is the cumulative distribution function of the normalized error.

The single value form of mean error is very straightforward and intuitive. However, this measure is heavily impacted by the presence of some big failures, *i.e.* outliers, in particular when the average error level is very low. Though CED is a better way to handle outliers, it is not intuitive given its curve representation. Yang et al. [165] proposed a novel intuitive evaluation metric based on CED:

$$AUC_\alpha = \int_0^\alpha f(e)de, \tag{10}$$

where e is the normalized error, $f(e)$ is the CED function and α is the upper bound that is used to calculate the definite integration. The value of AUC_α will not be influenced by points with error bigger than α.

6.2. Comparisons and discussions

In order to show the performance of FFPD methods, we categorize evaluation datasets into different scenarios according to images are captured under controlled environment or uncontrolled environment. Table 1 presents the published performance on datasets under controlled environment. It can be seen that the mean error is generally under 5% and it is relatively an easy task for FFPD on datasets under controlled environment.

For the datasets under uncontrolled environment, Zhang et al. [180] compared the FFPD performance of some state-of-the-art methods (STASM [101], CompASM [78], DRMF [7], ESR [20], RCPR [18], SDM [164], LBF [115], CFAN [176], CDM [171], GN-DPM [141], TREES [73], CFSS [191], TCDCN [179]) on the Helen database and 300W database respectively as shown in Tables 2 and 3. Compared with the performance on datasets under controlled environment, though FFPD on datasets in-the-wild achieves great success (state-of-the-art methods achieve comparable performance against human being on some datasets in-the-wild, *e.g.* LFPW), it is promising to further improve FFPD performance due to many factors such as occlusion, large shape exaggeration and so on. From these tables, we found that cascaded regression-based methods (et al. ESR [20], SDM [164], RCPR [18], LBF [115] and CFSS [191]) achieved the best performance. We should note that some state-of-the-art methods did not published the exact mean error (see Section 6.1) but the CED curve. It is reported that the deep learning based method (CFAN) [176] outperforms the SDM method on the LFPW database and the Helen database. The TCDCN method [179] outperforms the ESR method [20] and the SDM method [164] on the AFW database and the AFLW database. Due to the limited representation ability of parametric shape models, parametric shape model-based methods obtain less satisfying performance than the nonparametric shape model-based methods.

Yang et al. [165] has compared some parametric shape model-based methods and nonparametric shape model-based methods on the 300-W database. The training dataset includes images of AFW, the training images of LFPW and the training images of HELEN, totally with 3148 samples. The testing dataset is composed of the test images of LFPW, the test images of HELEN and the images in

Table 1
Mean errors (percent) on databases under controlled environment.

Method	Database	**Mean error**	**Number of landmarks**
Sukno et al. [131]	XM2VTS	2.03	64
Sukno et al. [131]	AR	1.63	98
Le et al. [78]	MUCT+BioID	4.5	17
Valstar et al. [143]	FERET [114]+MMI [144]	5.11	22
Martinez et al. [95]	MMI [144]+FERET [114]+XM2VTS+BioID	3.575	20
Wu et al. [160]	MMI [144]	5.5275	26

Table 2
Mean errors (percent) on Helen database.

Method	STASM	CompASM	DRMF	ESR	RCPR	SDM	LBF	CFAN	CDM	GN-DPM	TREES	CFSS	TCDCN
194 landmarks	11.10	9.10	–	5.70	6.50	5.85	5.41	–	–	–	4.90	4.74	4.63
68 landmarks	–	–	6.70	–	5.93	5.50	–	5.53	9.90	5.69	–	4.63	4.60

Table 3
Mean errors (percent) on 300W database (68 landmarks).

Method	CDM	DRMF	RCPR	GN-DPM	CFAN	ESR	SDM	TREES	LBF	CFSS	TCDCN
Common subset	10.10	6.65	6.18	5.78	5.50	5.28	5.57	–	4.95	4.73	4.80
Challenging subset	19.54	19.79	17.26	–	16.78	17.00	15.40	–	11.98	9.98	8.60
Fullset	11.94	9.22	8.35	–	7.69	7.58	7.50	6.40	6.32	5.76	5.54

Table 4
Evaluations and properties of FFPD methods (ME: Mean Error).

Methods	CCNF	GNDPM	DRMF	LBF	SDM	TCDCN	CFSS	IFA	RCPR	CFAN	TREES
Best BB	IBUG	IBUG	V&J	IBUG	V&J	IBUG	IBUG	HOG+SVM	V&J	IBUG	HOG+SVM
Landmarks#	68	49	66	68	49	68	68	49	68	68	68
Training set	300-W+MPIE	300-W	–	300-W	–	CelebA+300-W	300-W	300-W	300-W	–	–
Run-time (FPS)	30	70	0.5	10	40	50	10	20	80	20	300
Language	MATLAB	MATLAB	MATLAB	MATLAB	MATLAB	MATLAB	MATLAB	MATLAB	C++	MATLAB	C++
ME ($\times 10^{-2}$)	12.05	16.50	11.40	8.75	9.04	7.03	6.58	7.90	7.11	7.66	6.19
$AUC_{0.2}$($\times 10^{-2}$)	12.23	11.59	11.09	12.94	13.27	13.33	13.95	14.01	14.02	14.52	14.92

the IBUG database, with 689 samples in total. The 300-W database provides the ground truth locations of 68 facial landmarks. Eleven state-of-the-art off-the-shelf methods are used to compare their performance: the CCNF method [10], the GNDPM method [141], the DRMF method [7], the LBF method [115], the SDM method [164], the TCDCN method [179], the CFSS method [191], the IFA method [8], the RCPR method [18], the CFAN method [176], and the TREES method [73]. Table 4 gives normalized mean error and the properties of these evaluated face alignment methods as shown in [165]. "Best BB" denotes "best bounding box" for face detection used in respective FFPD methods. We should note that the larger the $AUC_{0.2}$ value is, the better performance the corresponding method achieves.

Since these released software are either training on different databases or conducting face detection by different methods, it is difficult to give a fair comparison to these FFPD algorithms. However, it does illustrate the performance of these released software. From the table, we found that cascaded regression-based software [8,18,73,115,164,191] and deep learning-based software [179,176] outperform parametric shape model-based software [7,10,141]. In addition, most of these compared software methods could reach real-time detection even under the MATLAB implementations. The TREES software [73] achieved the best performance in terms of both mean error and $AUC_{0.2}$. Fig. 3 gives the cumulative error distribution of these evaluated 11 FFPD methods.

By using the off-the-shelf model, we are unable to make a fair comparison cross different methods due to the difference in experimental setting, training data, and face detection. Yang et al. [165] select four representative methods for a fairer comparison by re-training their models using their default setting on the same

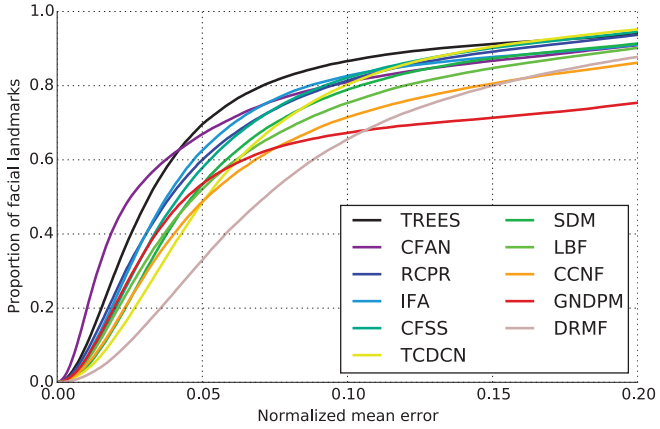

Fig. 3. Cumulative error distribution of 11 evaluated off-the-shelf software.

training data and the same data setting: CFSS [191], TREES [73], SDM [164] (re-implemented by [191]), and ESR [20]. They re-train the models on 300-W training set (3148 samples) using the HOG+SVM face detection and test the models on 300-W test set (689 samples). As all of them are cascaded regression-based methods, the random initialization number is set to 20.

Table 5 gives the normalized mean error and the $AUC_{0.2}$ values of these four different methods and Fig. 4 gives the cumulative error distribution. From Table 5 and Fig. 4 it can be seen the CFSS method achieves the best performance among the four compared cascaded regression-based methods under the same experimental settings.

Table 5
Evaluations of four re-trained models (ME: Mean Error).

Methods	ESR	SDM	TREES	CFSS
ME ($\times 10^{-2}$)	10.02	7.83	8.11	7.01
$AUC_{0.2}$ ($\times 10^{-2}$)	12.01	13.27	13.43	13.98

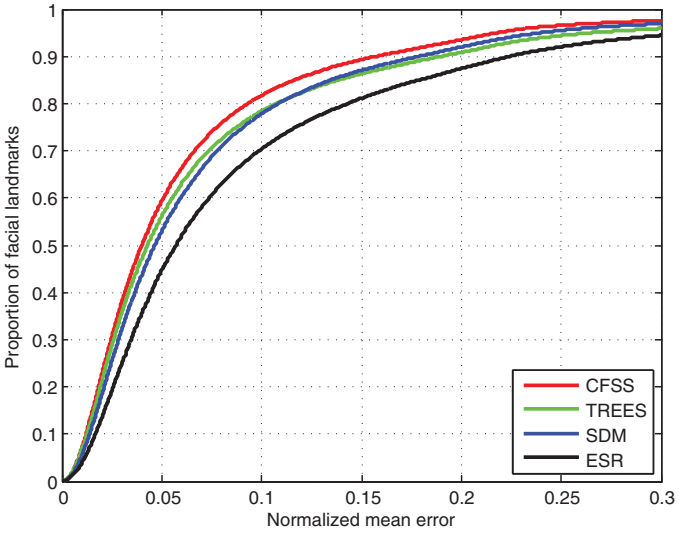

Fig. 4. Cumulative error distribution of four re-trained FFPD methods.

7. Conclusion and promising future directions

This paper reviews FFPD methods, which can be grouped into two major categories, parametric shape model-based methods and nonparametric shape model-based methods, according to a parametric or nonparametric shape model utilized in the method. Parametric shape model-based methods are further divided into two classes according to a part appearance or holistic appearance model used: local part model-based methods and holistic model-based methods. Nonparametric shape model-based methods are further divided into four categories depending on their model construction process: exemplar-based methods, graphical model-based methods, cascaded regression based methods and deep learning based methods. Through comprehensive analyses and comparisons of these methods, some remarks and promising future directions are summarized as follows:

Parametric shape model vs. nonparametric shape model: Parametric shape model-based methods achieve early successes in facial feature point localization. However these approaches generally tend to break down under extreme in-the-wild conditions [172], such as large pose variation, lighting variation and shape variation due to facial expressions. This is because parametric shape model generally has limited flexibility in the representation. Nonparametric shape model-based methods could achieve high accuracy owing to their greater representation ability and draw great attention in recent several years. Yet some parametric shape model-based methods (*e.g.* [1,7]) achieved superior performance in comparison to state-of-the-art nonparametric shape model-based methods. In reality, different kinds of methods can be selected depending on the problem. For example, person-specific AAM could be effectively utilized to perform personalized facial feature localization or tracking; and customized occlusion-robust or other robust methods can be applied in corresponding scenarios.

Handcrafted features vs. learned features: Most existing methods improve the efficiency and robustness depending on the carefully designed features such as pixel difference features [18,20] and SIFT features [164]. Though these features achieve some success, they still cannot adaptively deal with various shape variations and appearance variations. Recently, Ren et al. [115] presented an effective way to learn a set of local binary features to represent the facial image and achieved better performance in comparison to handcrafted features. Another promising way to adaptively learn features is by virtue of deep learning [13] which achieves state-of-the-art performance on many computer vision tasks including FFPD. The recently proposed deep autoencoder based FFPD method [176] is one of the representatives.

Although some state-of-the-art methods are ostensibly comparable to humans on some databases, FFPD is limited in its success by some wild and real-world conditions: intrinsic variations from people themselves (*e.g.* aging and expression) and extrinsic variations from environmental settings (*e.g.* occlusions, pose, and distance to the camera (image resolution)).

FFPD at a distance: Due to the image formation model of a typical charge-coupled device (CCD) camera, face images have a low resolution when they are captured at a distance [110]. Classic FFPD methods achieved poorly performance in the case of low-resolution images [40,41]. Several methods have been proposed to detect facial feature points on low resolution images. [40,41] proposed a resolution-aware algorithm to adapt to low-resolution images which substitutes the classic fitting criterion of *L2* norm error with a new formulation, taking the image formation model into account. Liu et al. [89] trained several AAMs, each of which corresponds to a specific resolution, to model the compactness at a lower resolution.

Though above methods could detect facial feature points on low resolution face images, they generally highly depend on the training data. However, it is a difficult and boring task for human beings to manually label points on low resolution face images. As shown in [110], face images captured at a distance of 5 meters are not easily recognized by human beings. To improve FFPD accuracy on face images at a distance, besides designing effective FFPD methods for low resolution images, another strategy is enhance the resolution of face images. There are generally two patterns to enhance the resolution. One way is to acquire high resolution face image through a special camera system such as the coaxial and concentric pan-tilt-zoom camera system [110]. Another way is to reconstruct a high resolution image from the low resolution one by deploying super-resolution techniques [153].

FFPD on face images with occlusions: Occlusions appeared on faces are often caused by *e.g.* hairs, hand, sunglasses, shadows and people. In spite of many works devoting to design a occlusion-robust algorithm [59,172,186], there is still a long way to go since many existing methods cannot perform as well as people on faces with occlusions. Recovering the genuine appearance for the occluded parts provides a promising solution to this problem to some extent. Zhang et al. [175] designed a de-corrupt autoencoder networks to automatically recover the appearance and then deep regression networks are utilized to predict the facial feature points by leveraging the recovered parts together with the non-occluded parts.

Besides the above occlusions, partial faces (*e.g.* captured due to limited field of view or sensor saturation) introduce another challenge problem [86]. FFPD on partial faces plays a key role on face recognition performance [49]. Generic FFPD methods perform poorly on partial faces since they are designed for holistic faces. It is even difficult to detect partial faces not to speak of initializing a shape. A feasible way is to utilize keypoint detection techniques. In this area, [168] has proposed a partial face alignment method, which overcomes the difficulties brought by face incompletion through a SIFT based generative face model. The SIFT based generative face is constructed from clustering all detected keypoints into some anchor points. This partial face alignment method

has also been taken as a preprocessing technique for heterogeneous face recognition between near infrared and visual images [170]. Further effective strategies to detect facial feature points on partial visible faces are still on the way.

FFPD on face images with pose variations: Most existing methods perform well on frontal or near-frontal faces (rotation angle less than 30°). For face images with a large rotation angle (*e.g.* larger than 60°), several models, each of which governs a special extent of views (*e.g.* Multiview AAM [32], tree structure based methods [59,193]), need to be learned. Recently, Zhu et al. [192] proposed a 3D solution to the large pose face alignment problem. Jourabloo and Liu [71] combined CNN model with 3D model to overcome the large pose variations. These methods generally have large computational complexity and cannot achieve real-time performance in many scenarios. Then, designing efficient and effective pose-robust FFPD methods is a challenging yet promising research area.

FFPD on face images with large shape variations: Due to the facial expression variations and aging problem, faces in the wild are often with large shape variation. Among existing methods, designing effective features is the common idea to overcome this challenge. For example, Cao et al. [20] utilized shape index features (pixel difference, which is indexed in a local coordinate system) and achieved some robustness to shape variation. Burgos-Artizzu et al. [18] subsequently improve the way of indexing features which further improve the robustness to large shape variation. However, it is generally difficult to manually designing an effective feature for various types of shape variation. Deep learning [179] may provide some cues for developing new effective methods robust to large shape variation since recent proposed deep learning based methods achieves state-of-the-art performance.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (under Grants 61501339, 61671339, 61432014, U1605252, and 61601158), in part by Young Elite Scientists Sponsorship Program by CAST (under Grant YESS20160026), in part by the Fundamental Research Funds for the Central Universities under Grant JB160104, in part by the Program for Changjiang Scholars, in part by the Leading Talent of Technological Innovation of Ten-Thousands Talents Program under Grant CS31117200001, in part by the China Post-Doctoral Science Foundation under Grants 2015M580818 and 2016T90893, and in part by the Shaanxi Province Post-Doctoral Science Foundation.

References

[1] J. Alabort-i-Medina, S. Zafeiriou, Bayesian active appearance models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3438–3445.
[2] B. Amberg, B. Andrew, V. Thomas, On compositional image alignment, with an application to active appearance models, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2009, pp. 1714–1721.
[3] B. Amberg, T. Better, Optimal landmark detection using shape models and branch and bound, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 455–462.
[4] R. Anderson, B. Stenger, R. Cipolla, V. Wan, Expressive visual text-to-speech using active appearance models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3382–3389.
[5] S. Arya, D. Mount, R. Silverman, A. Wu, An optimal algorithm for approximate nearest neighbor searching, J. ACM 45 (6) (1998) 891–923.
[6] S. Asteriadis, N. Nikolaidis, I. Pitas, Facial feature detection using distance vector fields, Pattern Recognit. 42 (7) (2009) 1388–1398.
[7] A. Asthana, S. Cheng, S. Zafeiriou, M. Pantic, Robust discriminative response map fitting with constrained local models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3444–3451.
[8] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Incremental face alignment in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1859–1866.
[9] S. Baker, R. Gross, I. Matthews, Lucas–Kanade 20 Years on: A Unifying Framework: Part 3, Technical Report, Carnegie Mellon University, 2003.
[10] T. Baltrusatis, P. Robinson, L. Morency, Continuous conditional neural fields for structured regression, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 593–608.
[11] A. Batur, M. Hayes, Adaptive active appearance models, IEEE Trans. Image Process. 14 (11) (2005) 1707–1721.
[12] P. Belhumeur, D. Jacobs, D. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 545–552.
[13] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[14] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, S. Nayar, Face swapping: automatically replacing faces in photographs, in: Proceedings of the 2008 SIGGRAPH, 2008, pp. 39.1–39.8.
[15] V. Blanz, T. Vetter, A morphable model for the synthesis of 3d faces, in: Proceedings of the 1999 SIGGRAPH, 1999, pp. 187–194.
[16] L. Breiman, Classification and Regression Trees, Chapman & Hall/CRC, Boca Raton, 1984.
[17] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[18] X. Burgos-Artizzu, P. Perona, P. Dollar, Robust face landmark estimation under occlusion, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1513–1520.
[19] C. Cao, Y. Weng, S. Lin, K. Zhou, 3D shape regression for real-time facial animation, in: Proceedings of the 2013 SIGGRAPH, 2013, pp. 1–10.
[20] X. Cao, Y. Wei, F. Wen, J. Sun, Face alignment by explicit shape regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2887–2894.
[21] D. Chen, S. Ren, Y. Wei, X. Cao, J. Sun, Joint cascaded face detection and alignment, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 109–122.
[22] H. Chen, Y. Xu, H. Shum, S. Zhu, N. Zheng, Example-based facial sketch generation with non-parametric sampling, in: Proceedings of the IEEE International Conference on Computer Vision, 2001, pp. 433–438.
[23] Y. Chen, J. Yang, J. Qian, Recurrent neural network for facial landmark detection, Neurocomputing 219 (2017) 26–38.
[24] C. Chow, C. Liu, Approximating discrete probability distributions with dependence trees, IEEE Trans. Inf. Theory 14 (3) (1968) 462–467.
[25] S. Christian, T. Alexander, E. Dumitru, Deep neural networks for object detection, in: Proceedings of the Advances in Neural Information Processing Systems, 2013, pp. 2553–2561.
[26] T. Cootes, G. Edwards, C. Taylor, Active appearance models, in: Proceedings of the European Conference on Computer Vision, 1998, pp. 484–498.
[27] T. Cootes, G. Edwards, C. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.
[28] T. Cootes, C. Taylor, Active shape models – 'smart snakes', in: Proceedings of the British Machine Vision Conference, 1992, pp. 266–275.
[29] T. Cootes, C. Taylor, A mixture model for representing shape variation, Image Vis. Comput. 17 (8) (1999) 567–573.
[30] T. Cootes, C. Taylor, Constrained active appearance models, in: Proceedings othe IEEE International Conference on Computer Vision, 2001, pp. 748–754.
[31] T. Cootes, C. Taylor, D. Cooper, J. Graham, Active shape models-their training and application, Comput. Vis. Image Underst. 61 (1) (1995) 38–59.
[32] T. Cootes, G. Wheeler, K. Walker, C. Taylor, View-based active appearance models, Image Vis. Comput. 20 (9–10) (2002) 657–664.
[33] J. Coughlan, S. Ferreira, Finding deformable shapes using loopy belief propagation, in: Proceedings of the European Conference on Computer Vision, 2002, pp. 453–468.
[34] M. Cox, S. Sridharan, S. Lucey, J. Cohn, Least squares congealing for unsupervised alignment of images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[35] M. Cox, S. Sridharan, S. Lucey, J. Cohn, Least-squares congealing for large numbers of images, in: Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 1–8.
[36] A. Criminisi, J. Shotton, E. Konukoglu, Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning, Found. Trends Comput. Graph. Vis. 7 (2–3) (2012) 81–227.
[37] D. Cristinacce, T. Cootes, Boosted regression active shape models, in: Proceedings of the British Machine Vision Conference, 2007, pp. 1–10.
[38] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.
[39] M. Dantone, J. Gall, G. Fanelli, L. van Gool, Real-time facial feature detection using conditional regression forests, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2578–2585.
[40] G. Dedeoglu, S. Baker, T. Kanade, Resolution-aware fitting of active appearance models to low resolution images, in: Proceedings of the European Conference on Computer Vision, 2006, pp. 83–97.
[41] G. Dedeoglu, T. Kanade, S. Baker, The asymmetry of image registration and its application to face tracking, IEEE Trans. Pattern Anal. Mach. Intell. 29 (5) (2007) 807–823.
[42] J. Deng, Y. Sun, Q. Liu., H. Lu, Low rank driven robust facial landmark regression, Neurocomputing 151 (2015) 196–206.
[43] C. Ding, J. Choi, D. Tao, L. Davis, Multi-Directional Multi-Level Dual-CrossPatterns for Robust Face Recognition, Technical Report, arxiv preprint, 2014. https://arxiv.org/abs/1401.5311.

[44] L. Ding, A. Martinez, Precise detailed detection of faces and facial features, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–7.
[45] L. Ding, A. Martinez, Features versus context: an approach for precise and detailed detection and delineation of faces and facial features, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 2022–2038.
[46] P. Dollar, P. Welinder, P. Perona, Cascaded pose regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 1078–1085.
[47] R. Donner, M. Reiter, L. Georg, P. Peloschek, H. Bischof, Fast active appearance model search using canonical correlation analysis, IEEE Trans. Pattern Anal. Mach. Intell. 28 (10) (2006) 1690–1694.
[48] G. Edwards, C. Taylor, T. Cootes, Interpreting face images using active appearance models, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition, 1998, pp. 300–305.
[49] H. Ekenel, R. Stiefelhagen, Why is facial occlusion a challenging problem, in: Proceedings of the IAPR/IEEE International Conference on Biometrics, 2009, pp. 299–308.
[50] M. Everingham, J. Sivic, A. Zisserman, "Hello! My name is... Buffy" – automatic naming of characters in tv video, in: Proceedings of the British Machine Vision Conference, 2006, pp. 899–908.
[51] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1915–1929.
[52] P. Felzenszwalb, D. Huttenlocher, Pictorial structures for object recognition, Int. J. Comput. Vis. 61 (1) (2005) 55–79.
[53] M. Fischler, R. Bolles, Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography, Commun. ACM. 24 (6) (1981) 381–395.
[54] Y. Freund, R. Iyer, R. Schapire, Y. Singer, An efficient boosting algorithm for combining preferences, J. Mach. Learn. Res. 4 (6) (2003) 933–969.
[55] J. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (5) (2001) 1189–1232.
[56] J. Friedman, T. Hastie, R. Tibshiani, Additive logistic regression: a statistical view of boosting, Ann. Stat. 38 (2) (2000) 337–374.
[57] X. Gao, Y. Su, X. Li, D. Tao, A review of active appearance models, IEEE Trans. Syst. Man. Cybern. Part C Appl. Rev. 40 (2) (2010) 145–158.
[58] Y. Ge, C. Peng, M. Hong, Joint local regressors learning for face alignment, Neurocomputing 208 (2016) 262–268.
[59] G. Ghiasi, C. Fowlkes, Occlusion coherence: localizing occluded faces with a hierarchical deformable part model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1899–1906.
[60] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, Image Vis. Comput. 28 (5) (2010) 807–813.
[61] L. Gu, T. Kanade, 3D alignment of face in a single image, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2006, pp. 1305–1312.
[62] L. Gu, T. Kanade, A generative shape regularization model for robust face alignment, in: Proceedings of the European Conference on Computer Vision, 2008, pp. 413–426.
[63] T. Hastie, R. Tibshirani, Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, New York, 2001.
[64] G. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[65] G. Hinton, R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[66] X. Hou, S. Li, H. Zhang, Q. Cheng, Direct appearance models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2001, pp. 828–833.
[67] G. Huang, V. Jain, Learned-MilerE., Unsupervised joint alignment of complex images, in: Proceedings of the IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[68] G. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, Technical Report 07-49, University of Massachusetts, 2007.
[69] Y. Huang, Q. Liu, D. Metaxas, A component based deformable model for generalized face alignment, in: Proceedings of the IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[70] O. Jesorsky, K. Kirchberg, R. Frischholz, Robust face detection using the Hausdoff distance, in: Proceedings of the International Conference on Audio- and Video-based Biometric Person Authentication, 2001, pp. 90–95.
[71] A. Jourabloo, X. Liu, Large-pose face alignment via CNN-based dense 3D model fitting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1–9.
[72] A. Kasinski, A. Florek, A. Schmidt, The PUT face database, Image Process. Commun. 13 (3) (2008) 59–64.
[73] V. Kazemi, J. Sullivan, One millisecond face alignment with ensemble of regression trees, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1867–1874.
[74] D. King, Dlib-ml: a machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.
[75] M. Kostinger, P. Wohlhart, P. Roth, H. Bischof, Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization, in: Proceedings of the International Conference on Computer Vision WorkShops, 2011, pp. 2144–2151.
[76] T. Kozakaya, T. Shibata, M. Yuasa, O. Yamaguchi, Facial feature localization using weighted vector concentration approach, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition, 2008, pp. 1–6.
[77] T. Kozakaya, T. Shibata, M. Yuasa, O. Yamaguchi, Facial feature localization using weighted vector concentration approach, Image Vis. Comput. 28 (5) (2010) 772–780.
[78] V. Le, J. Brandt, Z. Lin, L. Bourdev, T. Huang, Interactive facial feature localization, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 679–692.
[79] E. Learned-Miller, Data driven image models through continuous joint alignment, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2) (2006) 236–250.
[80] Y. LeCun, Y. Bengio, Convolutional networks for images, speech, and time series, in: The Handbook of Brain Science and Neural Networks, MIT Press, Cambridge, MA, USA, 1995, pp. 255–258.
[81] H. Lee, D. Kim, Tensor-based AAM with continuous variation estimation: application to variation-robust face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (6) (2009) 1102–1116.
[82] B. Leibe, A. Leonardis, B. Schiele, An implicit shape model for combined object categorization and segmentation, in: Proceedings of the European Conference on Computer Vision Workshop on Statistical Learning in Computer Vision, 2004, pp. 508–524.
[83] L. Liang, F. Wen, X. Tang, Y. Xu, An integrated model for accurate shape alignment, in: Proceedings of the European Conference on Computer Vision, 2006, pp. 333–346.
[84] L. Liang, F. Wen, Y. Xu, X. Tang, H. Shum, Accurate face alignment using shape constrained Markov network, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2006, pp. 1313–1319.
[85] L. Liang, R. Xiao, F. Wen, J. Sun, Face alignment via component-based discriminative search, in: Proceedings of the European Conference on Computer Vision, 2008, pp. 72–85.
[86] S. Liao, A. Jain, S. Li, Partial face recognition alignment-free approach, IEEE Trans. Pattern Anal. Mach. Intell. 35 (5) (2013) 1193–1205.
[87] F. Liu, D. Zeng, Q. Zhao, M. Liu, Joint face alignment and 3D face reconstruction, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 545–560.
[88] X. Liu, Discriminative face alignment, IEEE Trans. Pattern Anal. Mach. Intell. 31 (11) (2009) 1941–1954.
[89] X. Liu, P. Tu, F. Wheeler, Face model fitting on low resolution images, in: Proceedings of the British Machine Vision Conference, 2006, pp. 1079–1088.
[90] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[91] S. Lucey, R. Navarathna, A. Ashraf, S. Sridharan, Fourier Lucas–Kanade algorithm, IEEE Trans. Pattern Anal. Mach. Intell. 35 (6) (2013) 1383–1396.
[92] S. Lucey, Y. Wang, M. Cox, S. Sridharan, J. Cohn, Efficient constrained local model fitting for non-rigid face alignment, Image Vis. Comput. 27 (12) (2009) 1804–1813.
[93] P. Luo, X. Wang, X. Tang, Hierarchical face parsing via deep learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2480–2487.
[94] A. Martinez, R. Benavente, The AR Face Database, Technical Report, University of Barcelona, 1998.
[95] B. Martinez, M. Valstar, X. Binefa, M. Pantic, Local evidence aggregation for regression-based facial point detection, IEEE Trans. Pattern Anal. Mach. Intell. 35 (5) (2013) 1149–1163.
[96] M. Mathias, R. Benenson, M. Pedersoli, L. Van Gool, Face detection without bells and whistles, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 720–735.
[97] I. Matthews, S. Baker, Active appearance models revisited, Int. J. Comput. Vis. 60 (2) (2004) 135–164.
[98] L. Mei, M. Figl, A. Darzi, D. Rueckert, P. Edwards, Sample sufficiency and PCA dimension for statistical shape models, in: Proceedings of the European Conference on Computer Vision, 2008, pp. 492–503.
[99] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, XM2VTSDB:the extended M2VTS database, in: Proceedings of the International Conference on Audio- and Video-based Biometric Person Authentication, 1999, pp. 72–77.
[100] S. Miborrow, J. Morkel, F. Nicolls, The MUCT landmarked face database, in: Proceedings of the Pattern Recognition Association of South Africa, 2010, pp. 1–6.
[101] S. Miborrow, F. Nicolls, Locating facial features with an extended active shape model, in: Proceedings of the European Conference on Computer Vision, 2008, pp. 504–513.
[102] K. Murphy, Machine Learning: A Probabilistic Perspective, The MIT Press, 2012.
[103] R. Navarathna, S. Sridharan, S. Lucey, Fourier active appearance models, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 1919–1926.
[104] M. Nguyen, F. Torre, Local minima free parameterized appearance models, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[105] M. Nguyen, F. Torre, Metric learning for image alignment, Int. J. Comput. Vis. 88 (1) (2010) 69–84.
[106] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1520–1528.

[107] M. Nordstrom, M. Larsen, J. Sierakowski, M. Stegmann, The IMM Face Database – An Annotated Dataset of 240 Face Images, Technical Report, Technical University of Denmark, 2004.
[108] M. Ozuysal, M. Calonder, V. Lepetit, P. Fua, Fast keypoint recognition using random ferns, IEEE Trans. Pattern Anal. Mach. Intell. 32 (3) (2010) 448–461.
[109] G. Papandreou, P. Maragos, Adaptive and constrained algorithms for inverse compositional active appearance model fitting, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[110] U. Park, H. Choi, A. Jain, S. Lee, Face tracking and recognition at a distance: a coaxial and concentric PTZ camera system, IEEE Trans. Inf. Forens. Secur. 8 (10) (2013) 1665–1677.
[111] M. Pavan, M. Pelillo, Dominant sets and pairwise clustering, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 167–172.
[112] X. Peng, R. Feris, X. Wang, N. Metaxas, A recurrent encoder–decoder network for sequential face alignment, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 38–56.
[113] Y. Peng, A. Ganesh, J. Wright, W. Xu, Y. Ma, RASL: robust alignment by sparse and low-rank decomposition for linearly correlated images, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2233–2246.
[114] P. Phillips, H. Moon, P. Rauss, S. Rizvi, The FERET evaluation methodology for face recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (2000) 1090–1104.
[115] S. Ren, X. Cao, Y. Wei, J. Sun, Face alignment at 3000 fps via regressing local binary features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1685–1692.
[116] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 faces in-the-wild challenge: the first facial landmark localization challenge, in: Proceedings of the International Conference on Computer Vision, 2013, pp. 397–403.
[117] J. Saragih, Principal regression analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 2881–2888.
[118] J. Saragih, R. Goecke, A nonlinear discriminative approach to AAM fitting, in: Proceedings of the IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[119] J. Saragih, S. Lucey, J. Cohn, Face alignment through subspace constrained mean-shifts, in: Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 1034–1041.
[120] J. Saragih, S. Lucey, J. Cohn, Probabilistic constrained adaptive local displacement experts, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2009, pp. 288–295.
[121] J. Saragih, S. Lucey, J. Cohn, Deformable model fitting by regularized landmark mean-shift, Int. J. Comput. Vis. 91 (2) (2011) 200–215.
[122] X. Shen, Z. Lin, J. Brandt, Y. Wu, Detecting and aligning faces by image retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3460–3467.
[123] J. Sivic, M. Everingham, A. Zisserman, "Who are you?" – learning person specific classifiers from video, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2009, pp. 1145–1152.
[124] B. Smith, J. Brandt, Z. Lin, L. Zhang, Nonparametric context modeling of local appearance for pose- and expression-robust facial landmark localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1741–1748.
[125] B. Smith, L. Zhang, Joint face alignment with non-parametric shape models, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 43–56.
[126] B. Smith, L. Zhang, Collaborative facial landmark localization for transferring annotations across datasets, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 78–93.
[127] B. Smith, L. Zhang, J. Brandt, Z. Lin, J. Yang, Exemplar-based face parsing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3484–3491.
[128] P. Sozou, T. Cootes, C. Taylor, E. Mauro, A non-linear generalization of point distribution models using polynomial regression, Image Vis. Comput. 13 (5) (1995) 451–457.
[129] P. Sozou, T. Cootes, C. Taylor, E. Mauro, Non-linear point distribution modeling using a multi-layer perceptron, Image Vis. Comput. 15 (6) (1997) 457–463.
[130] M. Stegmann, B. Ersboll, R. Larsen, FAME-a flexible appearance modeling environment, IEEE Trans. Med. Imaging 22 (10) (2003) 1319–1331.
[131] F. Sukno, S. Ordas, C. Butakoff, S. Cruz, A. Frangi, Active shape models with invariant optimal features: application to facial analysis, IEEE Trans. Pattern Anal. Mach. Intell. 29 (7) (2007) 1105–1117.
[132] Y. Sun, X. Wang, X. Tang, Deep convolutional network cascade for facial point detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3476–3483.
[133] J. Sung, D. Kim, Pose-robust facial expression recognition using view-based 2D+3D AAM, IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. 38 (4) (2008) 852–866.
[134] Y. Tong, X. Liu, F. Wheeler, P. Tu, Automatic facial landmark labeling with minimal supervision, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2009, pp. 2097–2104.
[135] Y. Tong, X. Liu, F. Wheeler, P. Tu, Semi-supervised facial landmark annotation, Comput. Vis. Image Underst. 116 (8) (2012) 922–935.
[136] P. Tresadern, M. Ionita, T. Cootes, Real-time facial feature tracking on a mobile device, Int. J. Comput. Vis. 96 (3) (2012) 280–289.
[137] P. Tresadern, P. Sauer, T. Cootes, Additive update predictors in active appearance models, in: Proceedings of the British Machine Vision Conference, 2010, pp. 1–12.
[138] T. Tuytelaars, K. Mikolajczyk, Local invariant feature detectors: a survey, Found. Trends Comput. Graph. Vis. 3 (3) (2007) 177–280.
[139] G. Tzimiropoulos, Project-out cascaded regression with an application to face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3659–3667.
[140] G. Tzimiropoulos, M. Pantic, Optimization problems for fast AAM fitting in-the-wild, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 593–600.
[141] G. Tzimiropoulos, M. Pantic, Gauss-newton deformable part models for face alignment in-the-wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1851–1858.
[142] G. Tzimiropoulos, S. Zafeiriou, M. Pantic, Robust and efficient parametric face alignment, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 1847–1854.
[143] M. Valstar, B. Martinez, X. Binefa, M. Pantic, Facial point detection using boosted regression and graph models, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2010, pp. 2729–2736.
[144] M. Valstar, M. Pantic, Induced disgust, happiness and surprise: an addition to the MMI facial expression database, in: Proceedings of the International Conference on Language Resources and Evaluation, Workshop EMOTION, 2010, pp. 65–70.
[145] P. Viola, M. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[146] C. Vogler, Z. Li, A. Kanaujia, The best of both worlds: combining 3D deformable models with active shape models, in: Proceedings of the IEEE International Conference on Computer Vision, 2007, pp. 1–7.
[147] D. Vukadinovic, M. Pantic, Fullyautomatic facial feature point detection using Gabor feature based boosted classifiers, in: Proceedings of the International Conference on Systems, Man, and Cybernetics, 2005, pp. 1692–1698.
[148] N. Wang, X. Gao, J. Li, Random sampling and locality constraint for face sketch, arXiv Preprint, https://arxiv.org/abs/1701.01911.
[149] N. Wang, X. Gao, J. Li, Bayesian face sketch synthesis, IEEE Trans. Image Process. 26 (3) (2017) 1264–1274.
[150] N. Wang, X. Gao, J. Li, B. Song, Z. Li, Evaluation on synthesized face sketches, Neurocomuting 214 (2016) 991–1000.
[151] N. Wang, J. Li, D. Tao, X. Li, X. Gao, Heterogeneous image transformation, Pattern Recognit. Lett. 34 (1) (2013) 77–84.
[152] N. Wang, D. Tao, X. Gao, X. Li, A comprehensive survey to face hallucination, Int. J. Comput. Vis. 31 (1) (2014) 9–30.
[153] N. Wang, D. Tao, X. Gao, X. Li, J. Li, A comprehensive survey to face hallucination, Int. J. Comput. Vis. 106 (1) (2014) 9–30.
[154] Y. Wang, S. Lucey, J. Cohn, Enforcing convexity for improved alignment with constrained local models, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[155] Y. Wang, J. Yue, Y. Dong, Z. Hu, Robust discriminative regression for facial landmark localization under occlusion, Neurocomputing 214 (2016) 881–893.
[156] T. Weise, S. Bouaziz, H. Li, M. Pauly, Realtime performance-based facial animation, in: Proceedings of the 2011 SIGGRAPH, 2011, pp. 77.1–77.9.
[157] H. Wu, X. Liu, G. Doretto, Face alignment via boosted ranking model, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[158] Y. Wu, Q. Ji, Robust facial landmark detection under significant head poses and occlusion, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3658–3666.
[159] Y. Wu, Q. Ji, Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1–9.
[160] Y. Wu, Z. Wang, Q. Ji, Facial feature tracking under varying facial expressions and face poses based on restricted Boltzmann machine, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3452–3459.
[161] Y. Wu, Z. Wang, Q. Ji, A hierarchical probabilistic model for facial feature detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1781–1788.
[162] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, A. Kassim, Robust facial landmark detection via recurrent attentive-refinement networks, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 57–72.
[163] X. Xiong, D. la Torre F, Global supervised descent method, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2664–2673.
[164] X. Xiong, F. Torre, Supervised descent method and its application to face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 532–539.
[165] H. Yang, X. Jia, C. Chen, P. Robinson, An Empirical Study of Recent Face Alignment Methods, Technical Report, arxiv preprint, 2015. https://arxiv.org/abs/1511.05049.
[166] H. Yang, I. Patras, Face parts localization using structured output regression forests, in: Proceedings of the Asian Conference on Computer Vision, 2012, pp. 667–679.

[167] H. Yang, I. Patras, Sieving regression forest votes for facial feature detection in the wild, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1936–1943.
[168] J. Yang, S. Liao, S. Li, Automatic partial face alignment in NIR video sequences, in: Proceedings of the IAPR/IEEE International Conference on Biometrics, 2009, pp. 249–258.
[169] M. Yang, D. Kriegman, N. Ahuja, Detecting faces in images: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 24 (1) (2002) 34–58.
[170] D. Yi, S. Liao, Z. Lei, J. Sang, S. Li, Partial face matching between near infrared and visual images in MBGC portal challenge, in: Proceedings of the IAPR/IEEE International Conference on Biometrics, 2009, pp. 733–742.
[171] X. Yu, J. Huang, S. Zhuang, W. Yan, D. Metaxas, Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1944–1951.
[172] X. Yu, Z. Lin, J. Brandt, D. Metaxas, Consensus of regression for occlusion-robust facial feature localization, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 105–118.
[173] C. Zhang, Z. Zhang, A Survey of Recent Advances in Face Detection, Technical Report, Microsoft Research, 2010.
[174] H. Zhang, D. Liu, M. Poel, A. Nijholt, Face alignment using boosting and evolutionary search, in: Proceedings of the Asian Conference on Computer Vision, 2009, pp. 110–119.
[175] J. Zhang, M. Kan, S. Shan, X. Chen, Occlusion-free face alignment: deep regression networks coupled with de-corrupt autoencoders, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3428–3437.
[176] J. Zhang, S. Shan, M. Kan, X. Chen, Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 1–16.
[177] K. Zhang, Q. Liu, Y. Wu, M. Yang, Robust visual tracking via convolutional networks without training, IEEE Trans. Image Process. 25 (4) (2016) 1779–1792.
[178] M. Zhang, J. Li, N. Wang, X. Gao, Compositional model-based sketch generator in facial entertainment. (2017), DOI: 10.1109/TCYB.2017.2664499.
[179] Z. Zhang, P. Luo, C. Chen, X. Tang, Facial landmark detection by deep multi--task learning, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 94–108.
[180] Z. Zhang, P. Luo, C. Loy, X. Tang, Learning deep representation for face alignment with auxiliary attributes, IEEE Trans. Pattern Anal. Mach. Intell. 38 (5) (2016) 918–930.
[181] C. Zhao, W. Cham, X. Wang, Joint face alignment with a generic deformable face model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 561–568.
[182] W. Zhao, R. Chellappa, P. Phillips, A. Rosenfeld, Face recognition: a literature survey, ACM Comput. Surv. 35 (4) (2003) 399–458.
[183] X. Zhao, X. Chai, S. Shan, Joint face alignment: rescue bad alignments with good ones by regularized re-fitting, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 616–630.
[184] X. Zhao, S. Shan, X. Chai, X. Chen, Cascaded shape space pruning for robust facial landmark detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1033–1040.
[185] Y. Zheng, X. Zhou, B. Georgescu, S. Zhou, D. Comaniciu, Example based non–rigid shape detection, in: Proceedings of the European Conference on Computer Vision, 2006, pp. 423–436.
[186] F. Zhou, J. Brandt, Z. Lin, Exemplar-based graph matching for robust facial landmark localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1025–1032.
[187] X. Zhou, D. Comaniciu, A. Gupta, An information fusion framework for robust shape tracking, IEEE Trans. Pattern Anal. Mach. Intell. 27 (1) (2005) 115–129.
[188] Y. Zhou, L. Gu, H. Zhang, Bayesian tangent shape model: estimating shape and pose parameters via Bayesian inference, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2003, pp. 109–116.
[189] J. Zhu, L. Gool, S. Hoi, Unsupervised face alignment by robust nonrigid mapping, in: Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 1265–1272.
[190] M. Zhu, A. Martinez, Subclass discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 28 (8) (2006) 1274–1286.
[191] S. Zhu, C. Li, C. Loy, X. Tang, Face alignment by coarse-to-fine shape searching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4998–5006.
[192] X. Zhu, Z. Lei, X. Liu, H. Shi, S. Li, Face alignment across large poses: a 3D solution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1–11.
[193] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in: Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2879–2886.

Nannan Wang received the B.Sc. degree in information and computation science from Xi'an University of Posts and Telecommunications in 2009. He received his Ph.D. degree in information and telecommunications engineering in 2015. Now, he works with the state key laboratory of integrated services networks at Xidian University. From September 2011 to September 2013, he has been a visiting Ph.D. student with the University of Technology, Sydney, NSW, Australia. His current research interests include computer vision, pattern recognition, and machine learning. He has published more than 30 papers in refereed journals and proceedings including IJCV, IEEE T-PAMI, T-NNLS, T-Cyber, IEEE T-IP,T-CSVT etc.

Xinbo Gao (M'02-SM'07) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a Research Fellow at the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Post-Doctoral Research Fellow at the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong. Since 2001, he has been at the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of Ministry of Education, a Professor of Pattern Recognition and Intelligent System, and the Director of the State Key Laboratory of Integrated Services Networks, Xi'an, China. His current research interests include multimedia analysis, computer vision, pattern recognition, machine learning, and wireless communications. He has published five books and around 200 technical articles in refereed journals and proceedings, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, International Journal of Computer Vision, and Pattern Recognition in the above areas. He is on the Editorial Boards of several journals, including Signal Processing (Elsevier), and Neurocomputing (Elsevier). He served as the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences. He is currently a fellow of the Institution of Engineering and Technology.

Dacheng Tao is Professor of Computer Science and ARC Future Fellow in the School of Information Technologies and the Faculty of Engineering and Information Technologies, and the Founding Director of the UBTech Sydney Artificial Intelligence Institute at the University of Sydney. He was Professor of Computer Science and Director of Centre for Artificial Intelligence in the University of Technology Sydney. He mainly applies statistics and mathematics to Artificial Intelligence and Data Science. His research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and 500+ publications at prestigious journals and prominent conferences, such as IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, CIKM, ICML, CVPR, ICCV, ECCV, AISTATS, ICDM; and ACM SIGKDD, with several best paper awards, such as the best theory/algorithm paper runner up award in IEEE ICDM'07, the best student paper award in IEEE ICDM'13, and the 2014 ICDM 10-year highest-impact paper award. He received the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award and the 2015 UTS Vice-Chancellor's Medal for Exceptional Research. He is a Fellow of the IEEE, OSA, IAPR and SPIE.

Heng Yang is now a research scientist with ULSee Incorporation, Hangzhou. He received his B.Eng. degree in National University of Defense Technology. He received his Ph.D. degree in Queen Mary University of London in 2015. His current research interests include computer vision, pattern recognition, and machine learning. He has published more than 20 papers in refereed journals and proceedings including IEEE Trans. IP, CSVT, CVPR, etc.

Xuelong Li is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, China.