International Journal of Computer Vision (2019) 127:115–142
https://doi.org/10.1007/s11263-018-1097-z

Facial Landmark Detection: A Literature Survey

Yue Wu[1] · Qiang Ji[1]

Received: 3 November 2016 / Accepted: 26 April 2018 / Published online: 8 May 2018

Abstract

The locations of the fiducial facial landmark points around facial components and facial contour capture the rigid and non-rigid facial deformations due to head movements and facial expressions. They are hence important for various facial analysis tasks. Many facial landmark detection algorithms have been developed to automatically detect those key points over the years, and in this paper, we perform an extensive review of them. We classify the facial landmark detection algorithms into three major categories: holistic methods, Constrained Local Model (CLM) methods, and the regression-based methods. They differ in the ways to utilize the facial appearance and shape information. The holistic methods explicitly build models to represent the global facial appearance and shape information. The CLMs explicitly leverage the global shape model but build the local appearance models. The regression based methods implicitly capture facial shape and appearance information. For algorithms within each category, we discuss their underlying theories as well as their differences. We also compare their performances on both controlled and in the wild benchmark datasets, under varying facial expressions, head poses, and occlusion. Based on the evaluations, we point out their respective strengths and weaknesses. There is also a separate section to review the latest deep learning based algorithms. The survey also includes a listing of the benchmark databases and existing software. Finally, we identify future research directions, including combining methods in different categories to leverage their respective strengths to solve landmark detection "in-the-wild".

Keywords Facial landmark detection · Face alignment · Survey

1 Introduction

The face plays an important role in visual communication. By looking at the face, human can automatically extract many nonverbal messages, such as humans' identity, intent, and emotion. In computer vision, to automatically extract those facial information, the localization of the fiducial facial key points (Fig. 1) is usually a key step and many facial analysis methods are built up on the accurate detection of those landmark points. For example, facial expression recognition (Pantic and Rothkrantz 2000) and head pose estimation algorithms (Murphy-Chutorian and Trivedi 2009) may heavily rely on the facial shape information provided by the landmark locations. The facial landmark points around eyes can provide the initial guess of the pupil center positions for eye detection and eye gaze tracking (Hansen and Ji 2010). For facial recognition, the landmark locations on 2D image are usually combined with 3D head model to "frontalize" the face and help reduce the significant within-subject variations to improve recognition accuracy (Taigman et al. 2014). The facial information gained through the facial landmark locations can provide important information for human and computer interaction, entertainment, security surveillance, and medical applications.

Facial landmark detection algorithms aim to automatically identify the locations of the facial key landmark points on facial images or videos. Those key points are either the dominant points describing the unique location of a facial component (e.g., eye corner) or an interpolated point connecting those dominant points around the facial components and facial contour. Formally, given a facial image denoted as $\mathcal{I}$, a landmark detection algorithm predicts the locations of

Communicated by T.E. Boult.

✉ Qiang Ji
jiq@rpi.edu

Yue Wu
wuyuesophia@gmail.com

[1] Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180-3590, USA

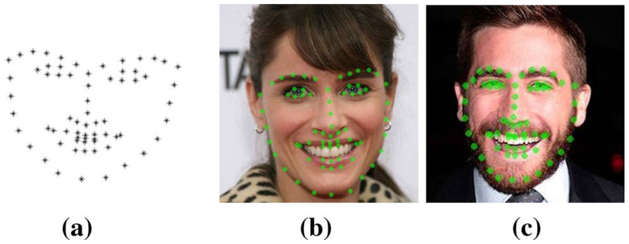

Fig. 1 **a** Facial landmarks defining the face shape. **b** Sample images (Belhumeur et al. 2013) with annotated facial landmarks

D landmarks $\mathbf{x} = \{x_1, y_1, x_2, y_2, ..., x_D, y_D\}$, where $x_.$ and $y_.$ resentment the image coordinates of the facial landmarks.

Facial landmark detection is challenging for several reasons. First, facial appearance changes significantly across subjects under different facial expressions and head poses. Second, the environmental conditions such as the illumination would affect the appearance of the faces on the facial images. Third, facial occlusion by other objects or self-occlusion due to extreme head poses would lead to incomplete facial appearance information.

Over the past few decades, there have been significant developments of the facial landmark detection algorithms. The early works focus on the less challenging facial images without the aforementioned facial variations. Later, the facial landmark detection algorithms aim to handle several variations within certain categories, and the facial images are usually collected with "controlled" conditions. For example, in "controlled" conditions, the facial poses and facial expressions can only be in certain categories. More recently, the research focuses on the challenging "in-the-wild" conditions, in which facial images can undergo arbitrary facial expressions, head poses, illumination, facial occlusions, etc. In general, there is still a lack of a robust method that can handle all those variations.

Facial landmark detection algorithms can be classified into three major categories: the *holistic methods*, the *Constrained Local Model (CLM) methods*, and *regression-based methods*, depending on how they model the facial appearance and facial shape patterns. The facial appearance refers to the distinctive pixel intensity patterns around the facial landmarks or in the whole face region, while face shape patterns refer to the patterns of the face shapes as defined by the landmark locations and their spatial relationships. As summarized in Table 1, the holistic methods explicitly model the holistic facial appearance and global facial shape patterns. CLMs rely on the explicit local facial appearance and explicit global facial shape patterns. Thes regression-based methods use holistic or local appearance information and they may embed the global facial shape patterns implicitly for joint landmark detection. In general, the regression-based methods show better performances recently (details will be discussed later). Note that, some recent methods combine the deep learning models and global 3D shape models for detection and they are outside the scope of the three major categories. They will be discussed in detail in Sect. 4.3.

The remaining parts of the paper are organized as follows. In Sects. 2, 3, and 4, we discuss methods in the three major categories: the holistic methods, the Constrained Local Model methods, and the regression-based methods. Section 4.3 is dedicated to the review of the recent deep learning based methods. In Sect. 5, we discus the relationships among methods in the three major categories. In Sect. 6, we discuss the limitations of the existing algorithms in "in-the-wild" conditions and some advanced algorithms that are specifically designed to handle those challenges. In Sect. 7, we discuss related topics, such as face detection, facial landmark tracking, and 3D facial landmark detection. In Sect. 8, we discuss facial landmark annotations, the popular facial landmark detection databases, software, and the evaluation of the leading algorithms. Finally, we summarize the paper in Sect. 9, where we point out future directions.

2 Holistic Methods

Holistic methods explicitly leverage the holistic facial appearance information as well as the global facial shape patterns for facial landmark detection (Fig. 2). In the following, we first introduce the classic holistic method: the Active Appearance Model (AAM) (Cootes et al. 2001). Then, we introduce its several extensions.

2.1 Active Appearance Model

The Active Appearance Model (AAM)[1] was introduced by Taylor and Cootes Edwards et al. (1998) and Cootes et al. (2001). It is a statistical model that fits the facial images with a small number of coefficients, controlling both the facial appearance and shape variations. During model construction, AAM builds the global facial shape model and the holistic facial appearance model sequentially based on Principal Component Analysis (PCA). During detection, it identifies the landmark locations by fitting the learned appearance and shape models to the testing images.

There are a few steps for AAM to construct the appearance and shape models, given the training facial images with landmark annotations, denoted as $\{\mathcal{I}_i, x_i\}_{i=1}^{N}$, where N is the number of training images. First, Procrustes Analysis (Gower 1975) is applied to register all the training facial shapes. It removes the affine transformation (c, θ, t_c, t_r denote the scale, rotation and translation parameters) of each face shape $\mathbf{x}_i$ and generates the normalized training facial shapes $\mathbf{x}'_i$.

[1] In this paper, we refer Active Appearance Model to the model, independent of the fitting algorithms.

Table 1 Comparison of the major facial landmark detection algorithms

Algorithms	Appearance	Shape	Performance	Speed
Holistic method	Whole face	Explicit	Poor generalization/good	Slow/fast
Constrained Local Method (CLM)	Local patch	Explicit	Good	Slow/fast
Regression-based method	Local patch/whole face	Implicit	Good/very good	Fast/very fast

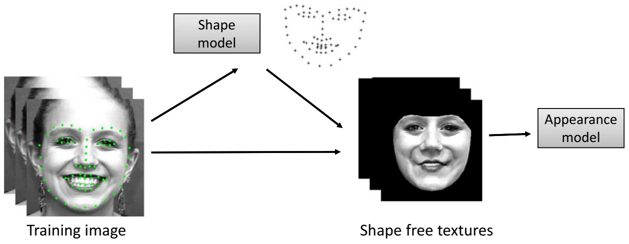

Fig. 2 Holistic model

Second, given the normalized training facial shapes $\{\mathbf{x}'_i\}_{i=1}^N$, PCA is applied to learn the mean shape $\mathbf{s}_0$ and orthonormal bases $\{\mathbf{s}_n\}_{n=1}^{K_s}$ that capture the shape variations, where K_s is the number of bases (Fig. 3). Given the learned facial shape bases $\{\mathbf{s}_n\}_{n=0}^{K_s}$, a normalized facial shape $\mathbf{x}'$ can be represented using the shape coefficients $\mathbf{p} = \{p_n\}_{n=1}^{K_s}$:

$$\mathbf{x}' = \mathbf{s}_0 + \sum_{n=1}^{K_s} p_n * \mathbf{s}_n. \tag{1}$$

Third, to learn the appearance model, image wrapping is applied to register the image to the mean shape and generate the shape normalized facial image, denoted as $\mathcal{I}_i(\mathcal{W}(\mathbf{x}'_i))$, where $\mathcal{W}(.)$ indicates the wrapping operation. Then, PCA is applied again on the shape normalized facial images $\{\mathcal{I}_i(\mathcal{W}(\mathbf{x}'_i))\}_{i=1}^N$ to generate a mean appearance A_0 and K_a appearance bases $\mathcal{A} = \{A_m\}_{m=1}^{K_a}$, as shown in Fig. 4. Given the appearance model $\mathcal{A} = \{A_m\}_{m=0}^{K_a}$, each shape normalized facial image can be represented using the appearance coefficients $\boldsymbol{\lambda} = \{\lambda_m\}_{m=1}^{K_a}$.

$$\mathcal{I}(\mathcal{W}(\mathbf{x}')) = \mathcal{A}_0 + \sum_{m=1}^{K_a} \lambda_m \mathcal{A}_m \tag{2}$$

An optional third model may be applied to learn the correlations among the shape coefficients $\mathbf{p}$ and appearance coefficients $\boldsymbol{\lambda}$.

In landmark detection, AAM finds the shape and appearance coefficients $\mathbf{p}$ and $\boldsymbol{\lambda}$, as well as the affine transformation parameters $\{c, \theta, t_c, t_r\}$ that best fit the testing image, which determine the landmark locations:

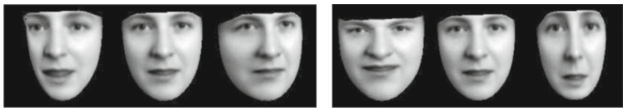

Fig. 3 Learned shape variations using AAM model, adapted from Cootes et al. (2001)

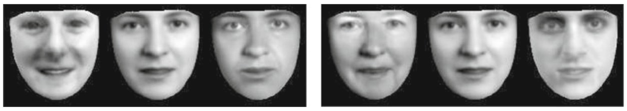

Fig. 4 Learned appearance variations using AAM, adapted from Cootes et al. (2001)

$$\mathbf{x} = cR_{2d}(\theta)\left(\mathbf{s}_0 + \sum_{n=1}^{K_s} p_n * \mathbf{s}_n\right) + t. \tag{3}$$

Here, $R_{2d}(\theta)$ denotes the rotation matrix, and $t = \{t_c, t_r\}$. To simplify the notation, in the following the shape coefficients would include both PCA coefficients and affine transformation parameters.

In general, the fitting procedure can be formulated by minimizing the distance between the reconstructed images $\mathcal{A}_0 + \sum_{m=1}^{K_a} \lambda_m \mathcal{A}_m$ and the shape normalized testing image $\mathcal{I}(\mathcal{W}(\mathbf{p}))$. The difference is usually referred to as the error image, denoted as $\Delta\mathcal{A}$:

$$\Delta\mathcal{A}(\boldsymbol{\lambda}, \mathbf{p}) = \mathrm{Diff}\left(\mathcal{A}_0 + \sum_{m=1}^{K_a} \lambda_m \mathcal{A}_m, \mathcal{I}(\mathcal{W}(\mathbf{p}))\right) \tag{4}$$

$$\boldsymbol{\lambda}^*, \mathbf{p}^* = \arg\min_{\boldsymbol{\lambda}, \mathbf{p}} \Delta\mathcal{A}(\boldsymbol{\lambda}, \mathbf{p}) \tag{5}$$

In the conventional AAM (Edwards et al. 1998; Cootes et al. 2001), model coefficients are estimated by iterative calculation of the error image based on the current model coefficients and model coefficient update prediction based on the error image.

2.2 Fitting Algorithms

Most of the holistic methods focus on improving the fitting algorithms, which involve solving Eq. (5). They can be classified into the *analytic fitting methods* and *learning-based fitting methods*.

2.2.1 Analytic Fitting Methods

The analytic fitting methods formulate AAM fitting problem as a nonlinear optimization problem and solve it analytically. In particular, the algorithm searches the best set of shape and appearance coefficients $\mathbf{p}, \lambda$ that minimize the difference between reconstructed image and the testing image with a nonlinear least squares formulation:

$$\tilde{\mathbf{p}}, \tilde{\boldsymbol{\lambda}} = \arg\min_{\mathbf{p},\boldsymbol{\lambda}} \left\| \mathcal{A}_0 + \sum_{m}^{M} \lambda_m \mathcal{A}_m - \mathcal{I}(\mathcal{W}(\mathbf{p})) \right\|_2^2 . \tag{6}$$

Here, $\mathcal{A}_0 + \sum_m^M \lambda_m \mathcal{A}_m$ represents the reconstructed face in the shape normalized frame depending on the shape and appearance coefficients, and the whole objective function represents the reconstruction error.

One natural way to solve the optimization problem is to use the Gaussian-Newton methods. However, since the Jacobin and Hessian matrix for both $\mathbf{p}$ and $\boldsymbol{\lambda}$ need to be calculated for each iteration (Jones and Poggio 1998), the fitting procedure is usually very slow. To address this problem, Baker and Matthews proposed a series of algorithms, among which the Project Out Inverse Compositional algorithm (POIC) (Matthews and Baker 2004) and the Simultaneous Inverse Compositional (SIC) algorithm (Baker et al. 2002) are two popular works. In POIC (Matthews and Baker 2004), the errors are projected into space spanned by the appearance eigenvectors $\{\mathcal{A}_m\}_{m=1}^{K_a}$, and its orthogonal complement space. The shape coefficients are firstly searched in the appearance space, and the appearance coefficients are then searched in the orthogonal space, given the shape coefficients. In SIC (Baker et al. 2002), the appearance and shape coefficients are estimated jointly with gradient descent algorithm. Compared to POIC, SIC is more computationally intensive, but generalizes better than POIC (Gross et al. 2005).

Recently, more advanced analytic fitting methods only estimate the shape coefficients, which fully determine the landmark locations as in Eq. (3). For example, in Alabort-I-Medina and Zafeiriou (2014), the Bayesian Active Appearance Model formulates AAM as a probabilistic PCA problem. It treats the texture coefficients as hidden variables and marginalizes them out to solve for the shape coefficients:

$$\tilde{\mathbf{p}} = \arg\max_{\mathbf{p}} ln p(\mathbf{p}) = \arg\max_{p} ln \int_{\boldsymbol{\lambda}} p(\mathbf{p}|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}) d\boldsymbol{\lambda}. \tag{7}$$

But exactly integrating out $\boldsymbol{\lambda}$ can be computationally expensive. To alleviate this problem, in Tzimiropoulos and Pantic (2013), Tzimiropoulos and Pantic proposed the fast-SIC algorithm and the fast-forward algorithm. In both algorithms, the appearance coefficient updates $\Delta\lambda$ are represented in terms of the shape coefficient updates $\Delta\mathbf{p}$, and they are plugged into the nonlinear least square formulation, which is then directly minimized to solve for $\Delta\mathbf{p}$. Different from Alabort-I-Medina and Zafeiriou (2014), Tzimiropoulos and Pantic (2013) follows a deterministic approach.

2.2.2 Learning-Based Fitting Methods

Instead of directly solving the fitting problem analytically, the learning-based fitting methods learn to predict the shape and appearance coefficients from the image appearances. They can be further classified into *linear regression fitting methods*, *nonlinear regression fitting methods*, and *other learning-based fitting methods*.

Linear regression fitting methods The linear regression fitting methods assume that there is linear relationship between model coefficient updates and the error image $\Delta\mathcal{A}(\boldsymbol{\lambda}, \mathbf{p})$ or image features $\mathcal{I}(\boldsymbol{\lambda}, \mathbf{p})$. They learn linear regression function for the prediction, which follows the conventional AAM as illustrated in the previous section.

$$\Delta\mathcal{A}(\boldsymbol{\lambda}, \mathbf{p})\ or\ \mathcal{I}(\boldsymbol{\lambda}, \mathbf{p}) \xrightarrow{Linear\ Regression} \Delta\boldsymbol{\lambda}, \Delta\mathbf{p} \tag{8}$$

They therefore estimate the model coefficients by iteratively estimating the model coefficient updates, and add them to the currently estimated coefficients for the prediction in the next iteration. For example, in Donner et al. (2006), Canonical Correlation Analysis (CCA) is applied to model the correlation between the error image and the model coefficient updates. It then learns the linear regression function to map the canonical projections of the error image to the coefficient updates. Similarly, in Direct Appearance Model (Hou et al. 2001), linear model is applied to directly predict the shape coefficient updates from the principal components of the error images. The linear regression fitting methods usually differ in the used image features, linear regression models, and whether to go to a different feature space to learn the mapping (Donner et al. 2006; Hou et al. 2001).

Nonlinear regression fitting methods The linear regression fitting methods assume that the relationship between features and the error image as shown in Eq. (4) around the true solution of model coefficients are close to quadratic, which ensures that an iterative procedure with linear updates and adaptive step sizes would lead to convergence. However, this linear assumption is only true when the initialization is around the true solution which makes the linear regression fitting methods sensitive to the initialization. To tackle this problem, the nonlinear regression fitting methods use nonlinear models to learn the relationship among the image features and the model coefficient updates:

$$\Delta\mathcal{A}(\boldsymbol{\lambda}, \mathbf{p})\ or\ \mathcal{I}(\boldsymbol{\lambda}, \mathbf{p}) \xrightarrow{Nonlinear\ Regression} \Delta\boldsymbol{\lambda}, \Delta\mathbf{p} \tag{9}$$

For example, in Saragih and Goecke (2007), boosting algorithm is proposed to predict the coefficient updates from the appearance. It combines a set of weak learners to form a strong regressor, and the weak learners are developed based on the Haar-like features and the one-dimensional decision stump. The strong nonlinear regressor can be considered as an additive piecewise function. In Tresadern et al. (2010), Tresadern et al. compared the linear and nonlinear regression algorithms. The used nonlinear regression algorithms include the additive piecewise function developed with boosting algorithm in Saragih and Goecke (2007) and the Relevance Vector Machine (Williams et al. 2005). They empirically showed that nonlinear regression method is better at the first few iterations to avoid the local minima, while linear regression is better when the estimation is close to the true solution.

2.2.3 Discussion: Analytic Fitting Methods Versus Learning-Based Fitting Methods

Compared to the analytic fitting methods solved with gradient descent algorithm with explicit calculation of the Hessian and Jacobian matrices, the learning-based fitting methods use constant linear or nonlinear regression functions to approximate the steepest descent direction. As a result, the learning-based fitting methods are generally fast but they may not be accurate. The analytic methods do not need training images, while the fitting methods do. The learning-based fitting methods usually use a third PCA to learn the joint correlations among the shape and appearance coefficients and further reduce the number of unknown coefficients, while the analytic fitting methods usually do not. But, for the analytic fitting methods, the interaction among appearance and shape coefficients can be embedded in the joint fitting objective function as in Eq. (6). The learned correlation between shape and appearance coefficients can reduce the number of parameters. Such learned correlation may not generalize well to different images. The joint estimation of shape and appearance coefficients using Eq. (6) can be more accurate. But they are more difficult.

2.3 Other Extensions

2.3.1 Feature Representation

There are other extensions of the conventional AAM methods. One particular direction is to improve the feature representations. It is well known that the AAM model has limited generalization ability and it has difficulty fitting the unseen face variations (e.g. across subjects, illumination, partial occlusion, etc.) (Gross et al. 2004, 2005). This limitation is partially due to the usage of raw pixel intensity as features. To tackle this problem, some algorithms use more robust image features. For example, in Hu et al. (2003), instead of using the raw pixel intensity, the wavelet features are used to model the facial appearance. In addition, only the local appearance information is used to improve the robustness to partial occlusion and illumination. In Jiao et al. (2003), Gabor wavelet with Gaussian Mixture model is used to model the local image appearance, which enables fast local point search. Both methods improve the performances of the conventional AAM method.

2.3.2 Ensemble of AAM Models

A single AAM model inherently assumes linearity in face shape and appearance variation. Realizing this limitation, some methods utilize ensemble models to improve the performance. For example, in Patrick Sauer and Taylor (2011), sequential regression AAM model is proposed, which trains a serials of AAMs for sequential model fitting in a cascaded manner. AAM in the early stage takes into account of the large variations (e.g. pose), while those in the later stage fit to the small variations. In this work, both independent ensemble AAMs and coupled sequential AAMs are used. The independent ensemble AAMs use independently perturbed model coefficients with different settings, while the coupled ensemble AAMs apply the learned prediction model in the first few levels to generate the perturbed training data in the later level. Both boosting regression and random forest regressions are utilized to predict the model updates. Similar to the coupled ensemble AAMs in Patrick Sauer and Taylor (2011) and Saragih and Gocke (2009), AAM fitting problem is formulated as an optimization problem with stochastic gradient descent solution. It leads to an approximated algorithm that iteratively learns the linear prediction model from the training data with perturbed model coefficients at the first iteration, and then updates the model coefficients for training data to be used in the next iteration following a cascaded manner. Different from Patrick Sauer and Taylor (2011), Saragih and Gocke (2009) uses different subsets of training data in different stages to escape from the local minima.

3 Constrained Local Methods

As shown in Fig. 5, the CLM methods (Cristinacce and Cootes 2006; Saragih et al. 2011) infer the landmark locations $\mathbf{x}$ based on the global facial shape patterns as well as the independent local appearance information around each landmark, which is easier to capture and more robust to illumination and occlusion, comparing to the holistic appearance.

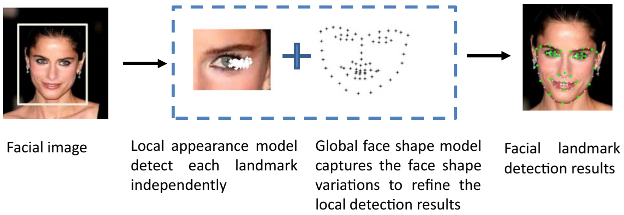
Facial image
Local appearance model detect each landmark independently
Global face shape model captures the face shape variations to refine the local detection results
Facial landmark detection results

Fig. 5 Constrained local method

3.1 Problem Formulation

In general, CLM can be formulated either as a deterministic or probabilistic method. In the deterministic view point, CLMs (Cristinacce and Cootes 2006; Saragih et al. 2011) find the landmarks by minimizing the misalignment error subject to the shape patterns:

$$\tilde{\mathbf{x}} = \arg\min_{\mathbf{x}} \mathbf{Q}(\mathbf{x}) + \sum_{d=1}^{D} \mathbf{D}_d(x_d, \mathcal{I}) \tag{10}$$

Here, x_d represents the positions of different landmarks in $\mathbf{x}$. $\mathbf{D}_d(x_d, \mathcal{I})$ represents the local confidence score around x_d . $\mathbf{Q}(\mathbf{x})$ represents a regularization term to penalize the infeasible or anti-anthropology face shapes in a global sense. The intuition is that we want to find the best set of landmark locations that have strong independent local support for each landmark and satisfy the global shape constraint.

The shape regularization can be applied to the shape coefficients $\mathbf{p}$. If we denote the regularization term as $\mathbf{Q}_p(\mathbf{p})$, Eq. (10) becomes:

$$\tilde{\mathbf{p}} = \arg\min_{\mathbf{p}} \mathbf{Q}_p(\mathbf{p}) + \sum_{d=1}^{D} \mathbf{D}_d(x_d(\mathbf{p}), \mathcal{I}) \tag{11}$$

Here, each landmark location x_d is determined by $\mathbf{p}$ as in Eq. (3).

In the probabilistic view point, CLM can be interpreted as maximizing the product of the prior probability of the facial shape patterns $p(\mathbf{x}; \eta)$ of all points and the local appearance likelihoods $p(x_d|\mathcal{I}; \theta_d)$ of each point:

$$\tilde{\mathbf{x}} = \arg\max_{\mathbf{x}} p(\mathbf{x}; \eta) \prod_{d=1}^{D} p(x_d|\mathcal{I}; \theta_d) \tag{12}$$

Similar to the deterministic formulation, the prior can also be applied to the shape coefficients $\mathbf{p}$, and Eq. (12) becomes:

$$\tilde{\mathbf{p}} = \arg\max_{\mathbf{p}} p(\mathbf{p}; \eta_p) \prod_{d=1}^{D} p(x_d(\mathbf{p})|\mathcal{I}; \theta_d). \tag{13}$$

For both the deterministic and the probabilistic CLMs, there are two major components. The first component is the local appearance model embedded in $\mathbf{D}_d(x_d, \mathcal{I})$ or $p(x_d|\mathcal{I}; \theta_d)$ in Eqs. (10–13). The second component refers to the facial shape pattern constraints either applied to the shape model coefficients $\mathbf{p}$ or the shape $\mathbf{x}$ itself, as penalty terms or probabilistic prior distributions. The two components are usually learned separately during training and they are combined to infer landmark locations during landmark detection. In the following, we will discuss each component, and how to combine them for landmark detection.

3.2 Local Appearance Model

The local appearance model assigns confidence score $\mathbf{D}_d(x_d, \mathcal{I})$ or probability $p(x_d|\mathcal{I}; \theta_d)$ that the landmark with index d is located at a specific pixel location x_d based on the local appearance information around x_d of image $\mathcal{I}$. The local appearance models can be categorized into *classifier-based local appearance models* and the *regression-based local appearance models*.

3.2.1 Classifier-Based Local Appearance Model

The classifier-based local appearance model trains binary classifier to distinguish the positive patches that are centered at the ground truth locations and the negative patches that are far away from the ground truth locations. During detection, the classifier can be applied to different pixel locations to generate the confidence scores $\mathbf{D}_d(x_d, \mathcal{I})$ or probabilities $p(x_d|\mathcal{I}; \theta_d)$ through voting. Different image features and classifiers are used. For example, in the original CLM work (Cristinacce and Cootes 2006) and FPLL (Zhu and Ramanan 2012), template based method is used to construct the classifier. The original CLM uses the raw image patch, while FPLL uses the HOG feature descriptor. In Belhumeur et al. (2011, 2013), SIFT feature descriptor and SVM classifier are used to learn the appearance model. In Cristinacce and Cootes (2007), Gentle Boost classifier is used. One issue with the classifier-based local appearance model is that it is unclear which feature representation and classifier to use. Even though SIFT and HOG features and SVM classifier are the popular choice, there is some work (Wu and Ji 2015) that learns the features using the deep learning methods, and it is shown that the learned features are comparable to HOG and SIFT.

3.2.2 Regression-Based Local Appearance Model

During training, the goal of the regression-based local appearance model is to predict the displacement vector $\Delta x_d^* = x_d^* - x$, which is the difference between any pixel location x and the ground truth landmark location x_d^* from

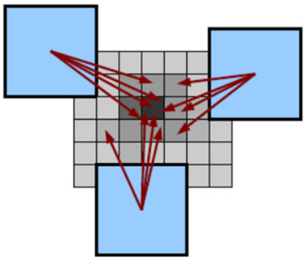

Fig. 6 Regression-based local appearance model (Cootes et al. 2012)

the local appearance information around x using the regression models. During detection, the regression model can be applied to patches at different locations x in a region of interest to predict Δx_d, which can be added to the current location to calculate x_d.

$$Regression : \mathcal{I}(x) \rightarrow \Delta x_d \tag{14}$$

$$x_d = x + \Delta x_d; \tag{15}$$

Predictions from multiple patches can be merged to calculate the final prediction of the confidence score or probability through voting (Fig. 6). Different image features and regression models are used. In Cristinacce and Cootes (2007), Cristinacce and Cootes proposed to use the Gentleboost as the regression function. In Cootes's later work (2012), he extended the method and used the random forests as the regressor, which shows better performance. In Valstar et al. (2010), Adaboost feature selection method is combined with SVM regressor to learn the regression function. In Smithk et al. (2014), nonparametric appearance model is used for location voting based on the contextual features around the landmark. Similar to the classifier-based methods, it's unclear which feature and regression function to use. It is empirically shown in Martinez et al. (2013) that the LBP features are better than the Haar features and LPQ descriptor. Another issue is that since the regression-based local appearance model performs one-step prediction. The prediction may not be accurate if the current positions are far away from the true target.

3.2.3 Discussion: Local Appearance Model

There are several issues related to the local appearance model. First, there exists accuracy-robustness tradeoffs. For example, a large local patch is more robust, while it is less accurate for precise landmark localization. A small patch with more distinctive appearance information would lead to more accurate detection results. To tackle this problem, some algorithms (Ren et al. 2014) combine the large patch and small patch for estimation and adapt the sizes of the patches or searching regions across iterations.

Second, it is unclear which approach to follow, among the classifier-based methods and the regression-based methods. One advantage of the regression-based approach is that it only needs to calculate the features and predict the displacement vectors for a few sample patches in testing. It is more efficient than the classifier-based approach that scans all the pixel locations in the region of interest. It is empirically shown in Cristinacce and Cootes (2007) that the Gentleboost regressor as a regression-based appearance model is better than the Gentleboost classifier as a classifier-based local appearance model.

3.3 Face Shape Models

The face shape model captures the spatial relationships among facial landmarks, which constrain and refine the landmark location search. In general, they can be classified into *deterministic face shape models* and *probabilistic face shape models*.

3.3.1 Deterministic Face Shape Models

The deterministic face shape models utilize deterministic models to capture the face shape patterns. They assign low fitting errors to the feasible face shapes and penalize infeasible face shapes. For example, Active shape model (ASM) (Cootes et al. 1995) is the most popular and conventional face shape model. It learns the linear subspaces of the training face shapes using Principal Component Analysis as in Eq. (1). The face can be evaluated by the fitness to the subspaces. It has been used both in the holistic AAM method and the CLM methods. Since one linear ASM may not be effective to model the global face shape variations, in Le et al. (2012), two levels of ASMs are constructed. One level of ASMs are used to capture the shape patterns of each facial component independently, and the other level of ASM is used for modeling the joint spatial relationships among facial components. Zhu and Ramanan (2012) built pose-dependent tree structure face shape model to capture the local nonlinear shape patterns, in which each landmark is represented as a tree node. Improving upon (Zhu and Ramanan 2012), the method in Hsu et al. (2015) builds two levels of tree structured models focusing on different numbers of landmark points on images with different resolutions. In Baltrušaitis et al. (2012), instead of using the 2D facial shape models, Baltrusaitis et al. proposed to embed the facial shape patterns into a 3D facial deformable model to handle pose variations. During detection, both the 3D model coefficients and the head pose parameters are jointly estimated for landmark detection. This method, however, requires to learn 3D deformable model, and to estimate head pose. 3D landmark detection will be further discussed in Sect. 7.3.

3.3.2 Probabilistic Face Shape Models

The probabilistic face shape models capture the facial shape patterns in a probabilistic manner. They assign high probabilities to face shapes that satisfy the anthropological constraint learned from training data and low probabilities to other infeasible face shapes. The early probabilistic face shape model is the switching model proposed in Tong et al. (2007). It can automatically switch the states of facial components (e.g. mouth open and close) to handle different facial expressions. In Valstar et al. (2010) and Martinez et al. (2013), a generative Boosted Regression and graph Models based method (BoRMaN) constructed based on the Markov Random Field is proposed. Each node in the MRF corresponds to the relative positions of three points, and the MRF as a whole can model the joint relationships among all landmarks. In Belhumeur et al. (2011, 2013), the authors proposed a non-parametric probabilistic face shape model with optimization strategy to fit the facial images. In Wu et al. (2013) and Wu and Ji (2015), the authors proposed a discriminative deep face shape model based on the Restricted Boltzmann Machine model. It explicitly handles face pose and expression variation, by decoupling the face shapes into head pose related part and expression related part. Compared to the other probabilistic face shape models, it can better handle facial expressions and poses within a unified model.

3.3.3 Discussion: Face Shape Model

Despite of the recent developments of the face shape model, its construction is still an unsolved problem. There is a lack of a unified shape model that can capture all natural facial shape variations (some methods will be discussed in Sect. 6). In addition, it's time-consuming to generate the facial landmark annotations under different facial shape variations, and more complex model requires more data to train. For example, due to facial occlusion, complete landmark annotation is infeasible on the self-occluded profile faces.

3.4 Landmark Point Detection via Optimization

Given the local appearance models and the face shape models described above, CLMs combine them for detection using Eqs. (10), (11), (12) or (12). This is a non-trivial task, since the analytic representation of the local appearance model is usually not directly computable from the local appearance model and the whole objective function is usually non-convex. To solve this problem, there are two sets of approaches, including the *iterative methods* and the *joint optimization methods*.

```
Algorithm 1: Iterative methods
Data: The initial searching regions
      $\boldsymbol{\Omega}^{(1)} = \{\Omega_1^{(1)}, \Omega_2^{(1)}, ..., \Omega_D^{(1)}\}$ for all $D$ landmarks.
Result: The detected landmark locations $\mathbf{x}$ or the shape
        coefficients $\mathbf{p}$ that determine the landmark locations.
for $t$=1 until convergence do
    Within the searching region $\Omega_d^{(t)}$, detect each facial landmark
    point independently using the local appearance models, and
    treat them as the measurements $\mathbf{m}^t$.
    Refine the measurements $\mathbf{m}$ jointly with the face shape model
    constraint, and output the estimated locations $\mathbf{x}^t$ or shape
    coefficients $\mathbf{p}^t$.
    Modify each searching region $\Omega_d^{(t+1)}$ to be around currently
    estimated landmark location.
```

3.4.1 Iterative Methods

The iterative methods decompose the optimization problem into two steps: landmark detection by local appearance model and location refinement by the shape model. They occur alternately until the estimation converges. Specifically, it estimates the optimal landmark positions that best fit the local appearance models for each landmark in the local region independently, and then refines them jointly with the face shape model. The detailed algorithm is shown in Algorithm 1. There are a few algorithms (Cristinacce and Cootes 2006, 2007; Cootes et al. 2012; Valstar et al. 2010; Martinez et al. 2013; Wu et al. 2013, 2014) that follow the iterative framework. Those methods differ in the used local appearance models and facial shape models, and we have discussed their particular techniques in the above sections.

3.4.2 Joint Optimization Methods

Different from the iterative methods, the joint optimization methods aim to perform joint inference. The challenge is that they have to find a way to represent the detection results from independent local point detectors and combine them with the face shape model for joint inference. To tackle this problem, in Saragih et al. (2011), the independent detection results are represented with Kernel Dense Estimation (KDE) and the optimization problem is solved with EM algorithm subject to the shape constraint, which treats the true landmark location as hidden variables. It also discusses some other methods to represent the local detection results, such as using the Isotropic Gaussian Model (Cootes et al. 1995), the Anisotropic Gaussian Model (Nickels and Hutchinson 2002), and Gaussian Mixture Model (Gu and Kanade 2008). In the Consensus of exemplars work (Belhumeur et al. 2011, 2013), in a Bayesian formulation, the local detector is combined with the nonparametric global model. Because of the special probabilistic formulation, the objective function can be optimized in a "brutal-force" way with RANSAC strategy.

There are also algorithms that simplify the shape model to use efficient inference methods. For example, in Zhu and Ramanan (2012), due to the usage of simple tree structure face shape model, dynamic programming can be applied to solve the optimization problem efficiently. In Cristinacce and Cootes (2004), by converting the objective function into a linear programming problem, Nelder-Meade simplex method is applied to solve the optimization problem.

3.4.3 Discussion: Optimization

The iterative methods and joint optimization methods have their own benefits and disadvantages. On one hand, the iterative methods are generally more efficient, but it may fail into the local minima due to the iterative procedure. On the other hand, the joint optimization methods are usually more difficult to solve and are computationally expensive.

Note that, as shown in Eqs. (10), (11), (12) or (12), in CLM, we would either infer the exact landmark locations **x** or the shape coefficients that can fully determine the landmark locations (e.g. shape coefficients **p** when using ASM as the face shape model as in Eq. 3). There is a dilemma. On one hand, since small shape model errors may lead to large landmark detection errors, directly predicting the landmark locations may be a better choice. On the other hand, it may be easier to design the cost function for the model coefficients than the shapes. For example, in ASM (Cootes et al. 1995), it is relatively easier to set up the range for the model coefficients while it is difficult to directly set the constraint for the face shapes.

4 Regression-Based Methods

The regression-based methods directly learn the mapping from image appearance to the landmark locations. Different from the Holistic Methods and Constrained Local Model methods, they usually do not explicitly build any global face shape model. Instead, the face shape constraints may be implicitly embedded. In general, the regression-based methods can be classified into *direct regression methods*, *cascaded regression methods*, and *deep-learning based regression methods*. Direct regression methods predict the landmarks in one iteration without any initialization, while the cascaded regression methods perform cascaded prediction and they usually require initial landmark locations. The deep-learning based methods follow either the direct regression or the cascaded regression. Since they use unique deep learning methods, we discuss them separately.

4.1 Direct Regression Methods

The direct regression methods learn the direct mapping from the image appearance to the facial landmark locations without any initialization of landmark locations. They are typically carried out in one step. They can be further classified into *local approaches* and *global approaches*. The local approaches use image patches, while the global approaches use the holistic facial appearance.

Local approaches The local approaches sample different patches from the face region, and build structured regression models to predict the displacement vectors (target face shape to the locations of the extracted patches), which can be added to the current patch location to produce all landmark locations jointly. The final facial landmark locations can be calculated by merging the prediction results from multiple sampled patches. Note that, this is different from the regression-based local appearance model that predicts each point independently (Sect. 3.2.2), while the local approaches here predict the updates for all points simultaneously. For example, in Dantone et al. (2012), conditional regression forests are used to learn the mapping from randomly sampled patches in the face region to the face shape updates. In addition, several head pose dependent models are built and they are combined together for detection. Similarly, privileged information-based conditional random forests model (Yang and Patras 2013) uses additional facial attributes (e.g. head pose, gender, etc.) to train regression forests to predict the face shape updates. Different from Dantone et al. (2012) which merges the prediction from different pose-dependent models, in testing, it predicts the attributes first and then performs attribute dependent landmark location estimation. In this case, the landmark prediction accuracy will be affected by attribute prediction. One issue with the local regression methods is that the independent local patches may not convey enough information for global shape estimation. In addition, for images with occlusion, the randomly sampled patches may lead to bad estimations.

Global approaches The global approaches learn the mapping from the global facial image to landmark locations directly. Different from the local approaches, the holistic face conveys more information for landmark detection. But, the mapping from the global facial appearance to the landmark locations is more difficult to learn, since the global facial appearance has significant variations, and they are more susceptible to facial occlusion. The leading approaches (Sun et al. 2013; Zhang et al. 2014) all use the deep learning methods to learn the mapping, which we will discuss in details in Sect. 4.3. Note that, since the global approaches directly predict landmark locations, they are different from the holistic methods in Sect. 2 that construct the shape and appearance models and predict the model coefficients.

Algorithm 2: Cascaded regression detection

```
Initialize the landmark locations $\mathbf{x}^0$ (e.g. mean face).
for t=1, 2, ..., T or convergence do
    Update the landmark locations, given the image and the current landmark location.
    $f_t : \mathcal{I}, \mathbf{x}^{t-1} \rightarrow \Delta\mathbf{x}^t$
    $\mathbf{x}^t = \mathbf{x}^{t-1} + \Delta\mathbf{x}^t$
Output the estimated landmark locations $\mathbf{x}^T$ .
```

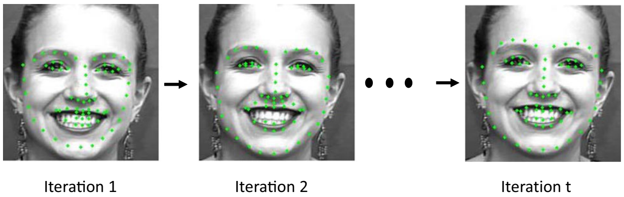

Fig. 7 Cascaded regression methods

4.2 Cascaded Regression Methods

In contrast to the direct regression methods that perform one-step prediction, the cascaded regression methods start from an initial guess of the facial landmark locations (e.g. mean face), and they gradually update the landmark locations across stages with different regression functions learned for different stages (Fig. 7). Specifically, in training, in each stage, regression models are applied to learn the mapping between shape-indexed image appearances (e.g., local appearance extracted based on the currently estimated landmark locations) to the shape updates. The learned model from the early stage will be used to update the training data for the training in the next stage. During testing, the learned regression models are sequentially applied to update the shapes across iterations. Algorithm 2 summarizes the detection process.

Different shape-indexed image appearance and regression models are used. For example, in Cao et al. (2014), the author proposed the shape-indexed pixel intensity features which are the pixel intensity differences between pairs of pixels whose locations are defined by their relative positions to the current shape. In Ren et al. (2014), the author proposed to learn discriminative binary features by the regression forests for each landmark independently. Then, the binary features from all landmarks are concatenated and a linear regression function is used to learn the joint mapping from appearance to the global shape updates. In Kazemi and Sullivan (2014), ensemble of regression trees are used as regression models for face alignment. By modifying the objective function, the algorithm can use training images with partially labeled facial landmark locations.

Among different cascaded regression methods, the Supervised Descent Method (SDM) in Xiong and De la Torre Frade (2013) achieves promising performances. It formulates face alignment as a nonlinear least squares problem. In particular, assuming the appearance features (e.g. SIFT) of the local patches around the true landmark locations $\mathbf{x}^*$ are denoted as $\Phi(\mathcal{I}(\mathbf{x}^*))$, the goal of landmark detection is to estimate the location updates $\delta\mathbf{x}$ starting from an initial shape $\mathbf{x}_0$, so that the feature distance is minimized:

$$\begin{aligned}\delta\tilde{\mathbf{x}} &= \arg\min_{\delta\mathbf{x}} f(\mathbf{x}_0 + \delta\mathbf{x}) \\ &= \arg\min_{\delta\mathbf{x}} \|\Phi(\mathcal{I}(\mathbf{x}^*)) - \Phi(\mathcal{I}(\mathbf{x}_0 + \delta\mathbf{x}))\|_2^2\end{aligned} \tag{16}$$

By applying the second order Taylor expansion with Newton-type method, the shape updates are calculated:

$$f(\mathbf{x}_0 + \delta\mathbf{x}) \approx f(\mathbf{x}_0) + \mathbf{J}_f(\mathbf{x}_0)^T\delta\mathbf{x} - \frac{1}{2}\delta\mathbf{x}^T\mathbf{H}_f(\mathbf{x}_0)\delta\mathbf{x} \tag{17}$$

$$\begin{aligned}\delta\mathbf{x} &= -\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_f(\mathbf{x}_0) \\ &= -2\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_\Phi^T(\Phi(\mathcal{I}(\mathbf{x}_0)) - \Phi(\mathcal{I}(\mathbf{x}^*)))\end{aligned} \tag{18}$$

To directly calculate $\delta\mathbf{x}$ analytically is difficult, since it requires the calculation of the Jacobin and Hessian matrix for different $\mathbf{x}_0$. Therefore, supervised descent method is proposed to learn the descent direction with regression method. It is then simplified as the cascaded regression method with linear regression function, which can predict the landmark location updates from shape-indexed local appearance.

$$R = -2\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_\Phi^T \tag{19}$$

$$b \approx 2\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_\Phi^T\Phi(\mathcal{I}(\mathbf{x}^*)) \tag{20}$$

$$\delta\mathbf{x} = R\Phi(\mathcal{I}(\mathbf{x}_0)) + b \tag{21}$$

The regression functions are different for different iterations, but they ignore different possible starting shape $\mathbf{x}_0$ within one iteration.

There are some other variations of the cascaded regression methods. For example, in Asthana et al. (2014), instead of learning the regression functions in a cascaded manner (the later level depends on the former level), a parallel learning method is proposed, so that the later level only needs the statistic information from the previous level. Based on the parallel learning framework, it's possible to incrementally update the model parameters in each level by adding a few more training samples, which achieves fast training.

The cascaded regression methods are more effective than the direct regression since they follow the coarse-to-fine strategy. The regression functions in the early stage can focus on the large variations while the regression functions in the later stage may focus on the fine search.

However, for cascaded regression methods, it is unclear how to generate the initial landmark locations. The popular choice is to use the mean face, which may be sub-optimal for images with large head poses. To tackle this problem, there are some hybrid methods that use the direct regression methods to generate the initial estimation for cascaded regression methods. For example, in Zhang et al. (2014), a model based on auto-encoder is proposed. It first performs direct regression on down-sampled lower-resolution images, and then refines the prediction in a cascaded manner with higher resolution images. In Zhu et al. (2015), a coarse to fine searching strategy is employed and the initial face shape is continuously updated based on the estimation from last stage. Therefore, a more close-to-solution initialization will be generated and fine facial landmark detection results are easier to get.

Another issue about the cascaded regression method is that the algorithms apply a fixed number of cascaded prediction, and there is no way to judge the quality of landmark prediction and adapt the necessary cascaded stages for different testing images. In this case, it is possible that the prediction is already trapped in a local minima while the iteration continues. It is also possible that the prediction is already close to the optimum after a few stages. In the existing methods (Ren et al. 2014), it is only shown that the cascaded regression methods can improve the performance over different cascaded stages, but it doesn't know when to stop.

4.3 Deep Learning Based Methods

Recently, deep learning methods become popular tools for computer vision problems. For facial landmark detection and tracking, there is a trend to shift from traditional methods to deep learning based methods. In the early work (Wu et al. 2013), the deep boltzmann machine model, which is a probabilistic deep model, was used to capture the facial shape variations due to poses and expressions for facial landmark detection and tracking. More recently, the Convolutional Neural Network (CNN) models become the dominate deep learning models for facial landmark detections, and most of them follow the global direct regression framework or cascaded regression framework. Those methods can be broadly classified into *pure-learning methods* and *hybrid methods*. The pure-learning methods directly predict the facial landmark locations, while the hybrid methods combine deep learning methods with computer vision projection model for prediction.

Pure-learning methods Methods in this category use the powerful CNN models to directly predict the landmark locations from facial images. Sun et al. (2013) is the early work and it predicts five facial key points in a cascaded manner. In the first level, it applies a CNN model with four convolution layers (Fig. 8) to predict the landmark locations given the facial image determined by the face bounding box. Then, several shallow networks refine each individual point locally.

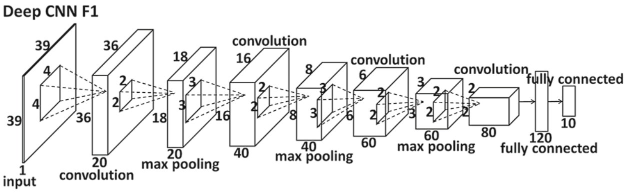

Fig. 8 CNN model structure, adapted from Sun et al. (2013)

Ever since then, there are several improvements over (Sun et al. 2013) in two directions. In the first direction, Zhang et al. (2014, 2016) and Ranjan et al. (2016) leverage multi-task learning idea to improve the performance. The intuition is that multiple tasks could share the same representation and their joint relationships would improve the performances of individual tasks. For example, in Zhang et al. (2014, 2016), multi-task learning is combined with CNN model to jointly predict facial landmarks, facial head pose, facial attributes etc. A similar multi-task CNN framework is proposed in Ranjan et al. (2016) to jointly perform face detection, landmark localization, pose estimation, and gender recognition. Different from Zhang et al. (2014, 2016), it combines features from multiple convolutional layers to leverage both the coarse and fine feature representations.

In the second direction, some works improve the cascaded procedure of method (Sun et al. 2013). For example, in Zhou et al. (2013), similar cascaded CNN model is constructed to predict many more points (68 landmarks instead of 5). It starts from the prediction of all 68 points and gradually decouples the prediction into local facial components. In Zhang et al. (2014), the deep auto-encoder model is used to perform the same cascaded landmark search. In Trigeorgis et al. (2016), instead of training multiple networks in a cascaded manner, Trigeorgis et. al trained a deep convolutional Recurrent Neural Network (RNN) for end-to-end facial landmark detection to mimic the cascaded behavior. The cascaded stage is embedded into the different time slices of RNN.

Hybrid deep methods The hybrid deep methods combine the CNN with 3D vision, such as the projection model and 3D deformable shape model (Fig. 9). Instead of directly predicting the 2D facial landmark locations, they predict 3D shape deformable model coefficients and the head poses. Then, the 2D landmark locations can be determined through the computer vision projection model. For example, in Zhu et al. (2016), a dense 3D face shape model is construct. Then, an iterative cascaded regression framework and deep CNN models are used to update the coefficients of 3D face shape and pose parameters. In each iteration, to incorporate the currently estimated 3D parameters, the 3D shape is projected to 2D using the vision projection model and the 2D shape is used as additional input of the CNN model for regression

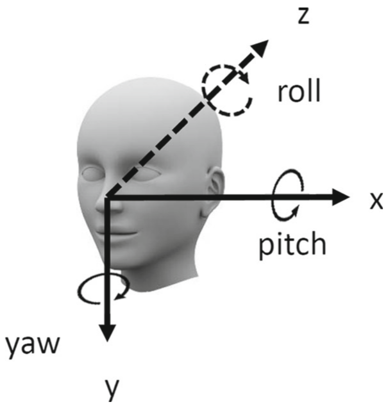

Fig. 9 3D face model and its projection based on the head pose parameters (i.e. pitch, yaw, roll angles)

prediction. Similarly, in Jourabloo and Liu (2016), in a cascaded manner, the whole facial appearance is used in the first cascaded CNN model to predict the updates of 3D shape parameters and pose, while the local patches are used in the later cascaded CNN models to refine the landmarks.

Compared to the pure-learning methods, the 3D shape deformable model and pose parameters of the hybrid methods are more compact ways to represent the 2D landmark locations. Therefore, there are fewer parameters to estimate in CNN and shape constraint can be explicitly embedded in the prediction. Furthermore, due to the introduction of 3D pose parameters, they can better handle pose variations.

Table 2[2] summarizes the CNN structures of the leading methods. We list their numbers of convolutional layers, the numbers of fully connected layer, whether a 3D model is used, and whether the cascaded method is used. For the cascaded methods, if different model structures are used for different layers, we only list the model in the first level. As can be seen, the models proposed for facial landmark detection usually contain around four convolutional layers and one fully connected layer. The model complexity is on par with the deep models used for other related face analysis tasks, such as head pose estimation (Patacchiola and Cangelosi 2017), age and gender estimation (Levi and Hassncer 2015), and facial expression recognition (Lopes et al. 2017), which usually have similar or smaller numbers of convolutional layers. For the face recognition problem, the CNN models are usually more complex with more convolutional layers (e.g. eight layers) and fully connected layers (Sun et al. 2015; Schroff et al. 2015; Taigman et al. 2014). It is in part due to the fact that there is much more training data (e.g., 10M+, 100M+ images) for face recognition comparing to the data set used for facial landmark detections (e.g., 20K+ images) (Ranjan et al. 2016). It is still an open question whether adding more data would improve the performances of facial landmark detection. Another promising direction is to leverage the multi-task learning idea to jointly predict related tasks (e.g., landmark detection, pose, age and gender) with a deeper model to boost the performances for all tasks (Ranjan et al. 2016).

4.4 Discussion: Regression-Based Methods

Among different regression methods, cascaded regression method achieves better results than direct regression. Cascaded regression with deep learning can further improve the performance. One issue for the regression-based methods is that since they learn the mapping from the facial appearance within the face bounding box region to the landmarks, they may be sensitive to the used face detector and the quality of the face bounding box. Because the size and location of the initial face is determined by the face bounding box, algorithms trained with one face detector may not work well if a different biased face detector is used in testing. This issue has been studied in Sagonas et al. (2016).

Even though we mentioned that the regression-based methods do not explicitly build the facial shape model, the facial shape patterns are usually implicitly embedded. In particular, since the regression-based methods predict all the facial landmark locations jointly, the structured information as well as the shape constraint are implicitly learned through the process.

5 Discussions: Relationships Among Methods in Three Major Categories

In the previous three sections, we discussed the facial landmark detection methods in three major categories: the holistic methods, the Constrained Local Model (CLM) methods, and the regression-based methods as summarized in Fig. 10. There exist similarities and relationships among the three major approaches.

First, both the holistic methods and CLMs would capture the global facial shape patterns using the explicitly constructed facial shape models, which are usually shared between them. CLMs improve over the holistic methods in that they use the local appearance around landmarks instead of the holistic facial appearance. The motivation is that it's more difficult to model the holistic facial appearances, and the local image patches are more robust to illumination changes and facial occlusion compared to the holistic appearance models.

Second, the regression-based methods, especially for the cascaded regression methods (Xiong and De la Torre Frade 2013) share similar intuitions as the holistic AAM (Baker et al. 2002; Saragih and Gocke 2009). For example, both of

[2] For Ranjan et al. (2016), we list the landmark prediction model instead of the multi-task prediction model for fair comparison.

Table 2 CNN model structures of the leading methods

Methods	# convolutional layer	# fully connected layer, # features	3D model	Cascaded method
Sun et al. (2013)	4	1 (120)	N	Y
Zhang et al. (2014)	4	1 (100)	N	N
Ranjan et al. (2016)	5 + Dim. reduction	1 (3072)	N	N
Zhou et al. 2013	4	1 (120)	N	Y
Zhu et al. (2016)	4	2 (256, 234)	Y	Y
Jourabloo and Liu (2016)	3	1 (150)	Y	Y

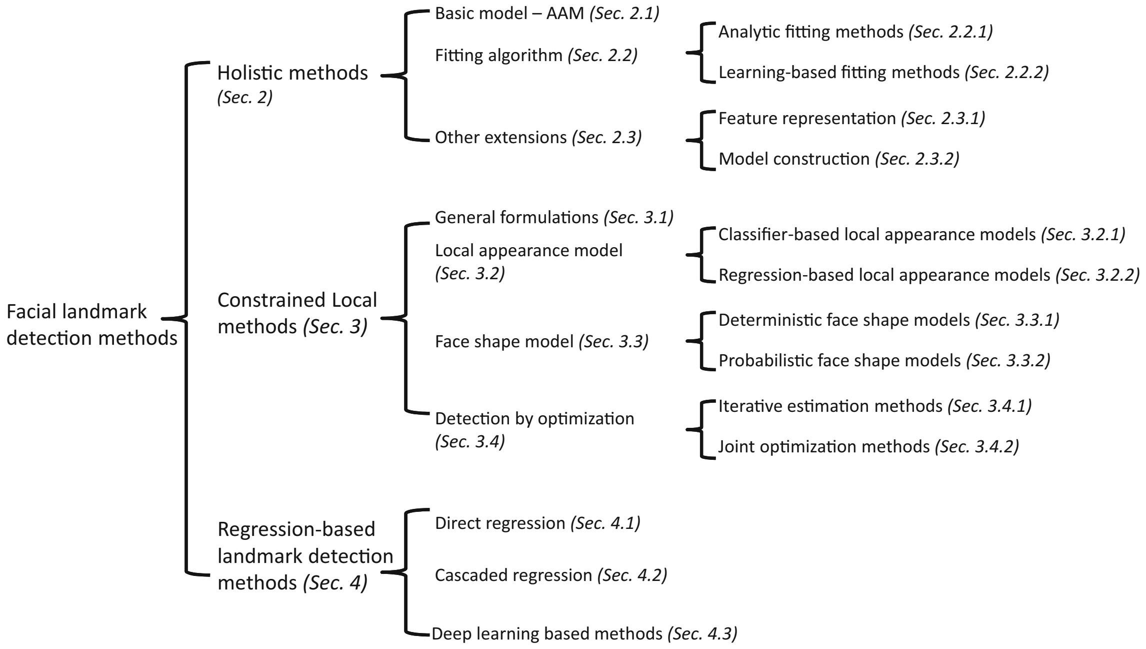
Facial landmark detection methods
Holistic methods (Sec. 2)
Basic model – AAM (Sec. 2.1)
Fitting algorithm (Sec. 2.2)
Analytic fitting methods (Sec. 2.2.1)
Learning-based fitting methods (Sec. 2.2.2)
Other extensions (Sec. 2.3)
Feature representation (Sec. 2.3.1)
Model construction (Sec. 2.3.2)
Constrained Local methods (Sec. 3)
General formulations (Sec. 3.1)
Local appearance model (Sec. 3.2)
Classifier-based local appearance models (Sec. 3.2.1)
Regression-based local appearance models (Sec. 3.2.2)
Face shape model (Sec. 3.3)
Deterministic face shape models (Sec. 3.3.1)
Probabilistic face shape models (Sec. 3.3.2)
Detection by optimization (Sec. 3.4)
Iterative estimation methods (Sec. 3.4.1)
Joint optimization methods (Sec. 3.4.2)
Regression-based landmark detection methods (Sec. 4)
Direct regression (Sec. 4.1)
Cascaded regression (Sec. 4.2)
Deep learning based methods (Sec. 4.3)

Fig. 10 Major categories of facial landmark detection algorithms

them estimate the landmarks by fitting the appearance and they all can be formulated as a nonlinear least squares problem as shown in Eqs. (6) and (16). However, the holistic methods predict the 2D shape and appearance model coefficients by fitting the holistic appearance model, while the cascaded regression methods predict the landmarks directly by fitting the local appearances without explicit 2D shape model. The fitting problem of holistic methods can be solved with learning-based approaches or analytically as discussed in Sect. 2.2, while all the cascaded regression methods perform estimation by learning. While the learning-based fitting methods for holistic models usually use the same model for coefficient updates in an iterative manner, the cascaded regression methods learn different regression models in a cascaded manner. The AAM model (Saragih and Gocke 2009) discussed in Sect. 2.3.2 as one particular type of holistic method is very similar to the Supervised Descent Methods (SDM) (Xiong and De la Torre Frade 2013) as one particular type of the cascaded regression method. Both train cascaded models to learn the mapping from shape-indexed features to shape (coefficient) updates. The trained model in the current cascaded stage will modify the training data to train the regression model in the next state. While the former holistic method fits the holistic appearance and predicts the model coefficients, SDM fits the local appearance and predicts the landmark locations.

Third, there are similarities among the regression-based local appearance model used in CLM in Sect. 3.2.2 and the regression-based methods in Sect. 4. Both of them predict the location updates from an initial guess of the landmark locations. The former approach predicts each landmark location independently, while the later approach predicts them jointly, so that shape constraint can be embedded implicitly. The former approach usually performs one-step prediction with the same regression model, while the later approach can apply different regression functions in a cascaded manner.

Fourth, compared to the holistic methods and constrained local methods, the regression-based methods may be more promising. The regression-based methods bypass the explicit face shape modeling and embed the face shape pattern constraint implicitly. The regression-based methods directly predict the landmarks, instead of the model coefficients as in the holistic methods and some CLMs. Directly predicting the shape usually can achieve better accuracy since small model coefficient errors may lead to large landmark errors.

6 Facial Landmark Detection "in-the-wild"

Most of the aforementioned algorithms focus on facial images in "controlled conditions" without significant variations. However, in real world, the facial images would undergo varying facial expressions, head poses, illuminations, facial occlusion etc., which are generally referred to as "in-the-wild" conditions. Some of the aforementioned algorithms may be able to handle those variations implicitly (e.g., deep learning based methods), while some others may fail. In this section, we focus on algorithms that are explicitly designed to handle those challenging conditions.

6.1 Head Poses

Significant head pose (e.g. profile face) is one of the major cause of the failure of the facial landmark detection algorithms (Fig. 11). There are a few difficulties. First, the 3D rigid head movement affects the 2D facial appearance and face shape. There would be significant facial appearance and shape variations caused by different head poses. Traditional shape models such as the PCA-based shape model used in AAM and ASM can no longer model the large facial shape variations since they are linear in nature and large facial pose shape variation is non-linear. Second, large head pose may lead to self-occlusion. Due to the missing of the facial landmark points, some facial landmark detection algorithms may not be directly applicable. Third, there is limited training data with large head poses, and it may need extra efforts to annotate the head pose labels to train the algorithms.

To handle large head poses, one direction is to train pose dependent models, and these methods differ in the detection procedures (Yan et al. 2003; Cootes et al. 2000; Zhu and Ramanan 2012; Dantone et al. 2012). They either select the best model or merge the results from all models. There are two ways to select the model. The first way is to estimate the head poses using existing head pose estimation methods. For example, in the early work (Yan et al. 2003), multiple pose dependent AAM models are built in training and the model is selected from the multi-view face detector during testing. In Yu et al. (2013), the head pose is first estimated based on the detection of a few facial key points. Then, head pose dependent fitting algorithm is applied to further refine the landmark detection results using the selected pose model.

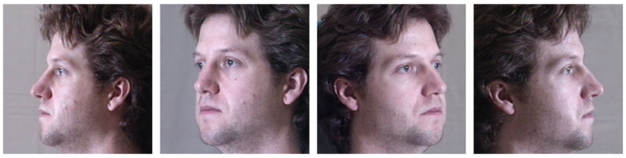

Fig. 11 Facial images (Gross et al. 2010) with different head poses

The head pose can also be selected based on the confidence scores using different pose dependent models. For example, in the early work (Cootes et al. 2000), three AAM models are built for faces in different head poses (e.g. left profile, frontal, and right profile) during training. During detection, the result with the smallest fitting error is considered as the final output. In Zhu and Ramanan (2012), multiple models are built for each discrete head pose and the best fit during testing is outputted as the final result.

The algorithms that select the best head pose dependent model would fail, if the model is not selected correctly. Therefore, it may be better to merge the results from different pose dependent models. For example, in Dantone et al. (2012), a probabilistic head pose estimator is trained and the facial landmark detection results from different pose dependent models are merged through Bayesian integration.

More recently, there are a few algorithms that build one unified model to handle all head poses. For example, in Wu and Ji (2015), self-occlusion caused by large head poses is considered as the general facial occlusion, and a unified model is proposed to handle facial occlusion, which explicitly predicts the landmark occlusion along with the landmark locations. In Xiong and la Torre (2015), landmark detection follows the cascaded iterative procedure and the pose dependent model is automatically selected based on the estimation from the last iteration. In Zhu et al. (2016) and Jourabloo and Liu (2016), Convolutional Neural Network (CNN) is combined with 3D deformable facial shape model to jointly estimate the head pose and facial landmarks on images with large head poses, following the cascaded regression framework. In summary, methods handling head poses include pose dependent shape models, unified pose models, and pose invariant features. They all have their strengths and weaknesses. It depends on the applications to choose which one to follow. Also, for some applications, it may be best to combine different types of methods.

6.2 Occlusion

Facial occlusion is another cause of the failure of the facial landmark detection algorithms. Facial occlusion could be caused by objects or self-occlusion due to large head poses. Figure 12 shows facial images with object occlusion. Some images in Fig. 11 also contain facial occlusion. There are a few difficulties to handle facial occlusions. First, the algo-

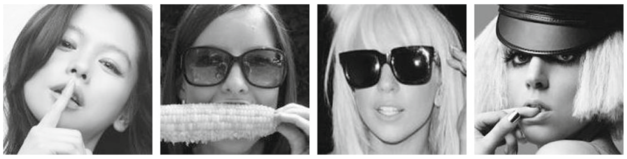

Fig. 12 Facial images (Burgos-Artizzu et al. 2013) with different occlusions

rithm should rely more on the facial parts without occlusion than the part with occlusion. However, it's difficult to predict which facial part or which facial landmarks are occluded. Second, since arbitrary facial part could be occluded by objects with arbitrary appearances and shapes, the facial landmark detection algorithms should be flexible enough to handle different cases (e.g. mouth may be occluded by mask or nose may be occluded by hand). Third, the occlusion region is usually locally consistent (e.g. it is unlikely that every other point is occluded), but it is difficult to embed this property as a constraint for occlusion prediction and landmark detection.

Due to these difficulties, there is limited work that can handle occlusion. Most of the current algorithms build occlusion dependent models by assuming that some parts of the face are occluded, and merge those models for detection. For example, in Burgos-Artizzu et al. (2013), face is divided into nine parts and it is assumed that only one part is not occluded. Therefore, the facial appearance information from one particular part is utilized to predict the facial landmark locations and facial occlusions for all parts. The predictions from all nine parts are merged together based on the predicted facial occlusion probability for each part. Similarly, it is assumed in Yu et al. (2014) that a few manually pre-designed regions (e.g. one facial component) are occluded, and different models are trained for prediction based on the non-occluded parts.

The aforementioned occlusion dependent models may be sub-optimal, since they assume that a few pre-defined regions are occluded, while the facial occlusion could be arbitrary. Therefore, those algorithms may not cover all rich and complex occlusion cases in real world scenario. Another issue related to the aforementioned methods is that the limited facial appearance information from one part of the face (e.g. from the mouth region) maybe insufficient for the prediction of the facial landmarks in the whole face region. To alleviate this problem, some algorithms handle facial occlusion in a unified framework. For example, in Ghiasi and Fowlkes (2014), a probabilistic model is proposed to predict the facial landmark locations by jointly modeling the local facial appearance around each facial landmark, the landmark visibility for each landmark, occlusion pattern, and hidden states of facial components. It jointly predicts the landmark locations and landmark occlusions through inference in the probabilistic model. In Wu and Ji (2015), a constrained cascaded regression model is proposed to iteratively predict the facial landmark locations and landmark visibility probabilities, based on local appearance around currently predicted landmark points iteratively. For landmark visibility prediction, it gradually updates the landmark visibility probabilities and it explicitly adds occlusion patterns as a constraint in the prediction. For landmark location prediction, it assigns weights to facial appearance around different facial landmarks based on their landmark visibility probabilities, so that the algorithm relies more on the local appearance from visible landmarks than that from occluded landmarks. Different from Ghiasi and Fowlkes (2014), the model can handle facial occlusion caused by both object occlusion and large head poses.

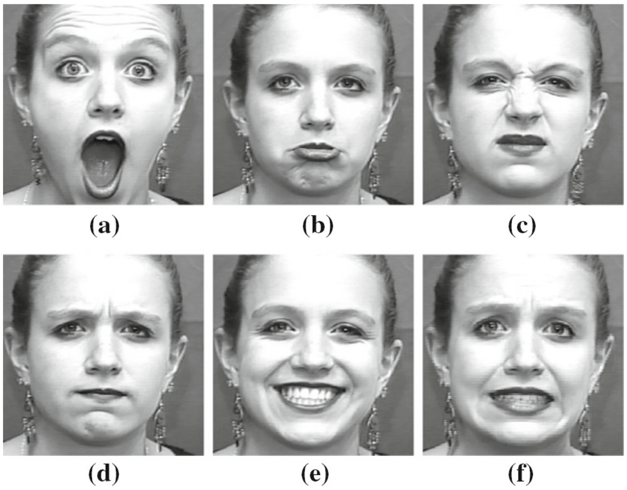

Fig. 13 Facial images (Kanade et al. 2000; Lucey et al. 2010) with different facial expressions. **a** surprised, **b** sadness, **c** disgust, **d** anger, **e** happy and **f** fear

6.3 Facial Expression

Facial expression would lead to non-rigid facial motions which would affect facial landmark detection and tracking. For example, as shown in Fig. 13, the six basic facial expressions, including happy, surprise, sadness, angry, fear and disgust, would cause the changes of facial appearance and shape. In more natural conditions, facial images would undergo more spontaneous facial expressions other than the six basic expressions. Generally speaking, recent facial landmark detection algorithms have been able to handle facial expressions to some extent.

Even though most algorithms handle facial expressions implicitly, there are some algorithms that are explicitly designed to handle significant facial expression variations. For example, in Tong et al. (2007), a hierarchical dynamic probabilistic model is proposed, which can automatically switch between specific states of the facial components caused by different facial expressions. Due to the correlation between facial expression and facial shape, some algorithms also perform joint facial expression and facial landmark detection. For example, in Li et al. (2013), a dynamic bayesian network model is proposed to model the dependencies among facial action units, facial expressions, and face shapes for joint facial behavior analysis and facial landmark tracking. It shows that exploiting the joint relationships and interactions improves the performances of both facial expression recognition and facial landmark detection. Similarly, in Wu and Ji (2016), a constrained joint cascaded regression framework is proposed for simultaneous facial action unit recognition and facial landmark detection. It first learns the joint relationships among facial action units and facial shapes, and then uses the relationship as a constraint to iteratively update the action unit activation probabilities and landmark locations in a cascaded iterative manner.

There are also some works that handle both facial expression and head pose variations simultaneously for facial landmark detection. For example, in Perakis et al. (2013) and Baltrušaitis et al. (2012), 3D face models are used to handle facial expression and pose variations. Baltrušaitis et al. (2012) can predict both the landmarks and the head poses. In Wu et al. (2013) and Wu and Ji (2015), the face shape model is constructed to handle the variations due to facial expression changes. The model decouples the face shape into expression related parts and head pose related parts. In Zhao et al. (2014), not only the facial landmark locations, but also the expression and pose are estimated jointly in a cascaded manner with Random Forests model. In addition, in each cascade level, the poses and expressions are firstly updated and they are used further for the estimation of facial landmarks. In Wu et al. (2014), a hierarchical probabilistic model is proposed to automatically exploit the relationships among facial expression, head pose, and facial shape changes of different facial components for facial landmark detection. Some multi-task deep learning methods discussed in Sect. 4.3 also can be included in this category.

7 Related Topics

7.1 Face Detection for Facial Landmark Detection

It is usually assumed that face is already detected and given for most of the existing facial landmark detection algorithms. The detected face would provide the initial guess of the face location and face scale. However, there are some issues. First, face detection is still an unsolved problem and it would fail especially on images with large variations. The failure of the face detection would directly lead to the failure of most facial landmark detection algorithms. Second, facial landmark detection accuracy may be significantly affected by the face detectors accuracy. For example, in the regression-based methods, the initial shape is generated by placing the mean shape in the center of the face bounding box, where the scale is also estimated from the bounding box. In CLM, the initial regions of interest for each independent local point detector are determined by the face bounding box. Third, to ensure real-time facial landmark detection, fast face detector is usually preferred.

The most popular face detector has been the Viola-Jones face detector (VJ) (Viola and Jones 2001). The usage of the integral image and the adaboost learning ensures both fast computation and effective face detection. The part-based approaches (Heisele et al. 2007; Felzenszwalb et al. 2010; Mathias et al. 2014) use a slightly different framework. They consider the face as a object consisting of several parts with spatial constraints. More recently, the region-based convolutional neural networks (RCNN) methods (Girshick et al. 2014; Girshick 2015; Ren et al. 2015) have been used for face detection. They are based on a region proposal component which identifies the possible face regions and a detection component, which further refines the proposal regions for face detection. Generally, the RCNN based methods are more accurate especially for images with large pose, illumination, occlusion variations. However, their computational costs (about 5 frame/second with GPU) are much higher than the traditional face detectors which provide real-time detection. A more detailed study and survey of the face detection algorithms can be found in Zhang and Zhang (2010).

There are some algorithms that perform joint face detection and landmark localization. For example, in Zhu and Ramanan (2012), deformable part model is applied to jointly perform face detection, face alignment and head pose estimation. In Shen et al. (2013), face detection and face alignment are formulated as an image retrieval problem. Similarly, In Chen et al. (2014), face detection and face alignment are performed jointly in a cascaded regression framework. The face shape is iteratively updated and the bounding box would be rejected if the confidence based on the current face shape is less than a threshold.

7.2 Facial Landmark Tracking

Facial landmark detection algorithms are generally designed to handle individual facial images, and they can be extended to handle facial image sequences or video. The simplest way is tracking by detection, where facial landmark detection is applied. However, methods in this category ignore the dependency and temporal smoothness among consecutive frames, which are sub-optimal.

There are three types of works that perform facial landmark tracking by leveraging the temporal relationship: tracker based independent tracking methods, joint tracking methods, and probabilistic graphical model based methods.

The tracker based independent tracking methods (Bourel et al. 2000; Tong et al. 2007; Wu et al. 2013) perform facial landmark tracking on the individual points based on the general object trackers, such as the Kalman Filter and Kanade–Lucas–Tomasti tracker (KLT). In each frame, the face shape model is applied to restrict the independently tracked points, so that the face shape in each frame satisfies the face shape pattern constraint (Bourel et al. 2000; Tong et al. 2007; Wu et al. 2013). The joint tracking methods perform facial landmark points update jointly, and they initialize the model parameters or landmark locations based on the information from the last frame. For example, in Ahlberg (2002), AAM model coefficients estimated in the last frame are used to initialize the model coefficients in the current frame. In the cascaded regression framework (Xiong and De la Torre Frade 2013), the tracked facial landmark locations in the last frame are used to determine the location and size of the initial facial shape for the cascaded regression in the current frame. Since detection and tracking are initialized differently (detection uses face bounding box to determine the face size and location, while it uses landmark locations in the last frame in tracking), the method needs train different sets of regression functions for detection and tracking. The probabilistic graphical model based methods build dynamic models to jointly embed the spatial and temporal relationships among facial landmark points for facial landmark tracking. For example, Markov Random Field (MRF) model is used in Cosar and Cetin (2011), and Dynamic Bayesian Network is used in Li et al. (2013). The dynamic probabilistic models capture both the temporal coherence as well as shape dependencies among landmark points. More evaluation and review about facial landmark tracking can be found here (Chrysos et al. 2017).

7.3 3D Facial Landmark Detection

3D facial landmark detection algorithms detect the 3D facial landmark locations. The existing works can be classified into *2D-data* based methods using 2D images and *3D-data* based methods using 3D face scan data.

Given the limited information, the detection of the 3D facial landmark locations from 2D image is a ill-posed problem. To solve this issue, existing methods either leverage 3D training data or a pre-trained 3D facial shape model and combine them with machine learning. For example, in Tulyakov and Sebe (2015), Tulyakov and Sebe extended the cascaded regression method from 2D to 3D by adding the prediction of the depth information. The method directly learns the regressors to predict the 3D facial landmark locations from 2D image given 3D training data. Since 3D data is difficult to generate, some algorithms learn a 3D facial shape model instead with limited training data. For example, in Gou et al. (2016), 2D facial landmarks are firstly detected using the cascaded regression method, and they are combined with a 3D deformable model to determine the face pose and coefficients of the deformable model, based on which they can then recover the positions of the 3D landmark points. Similarly, in Jeni et al. (2015), cascaded regression method is used to predict both dense 2D facial landmarks and their visibilities. An iterative method is then applied to fit the 2D shape to a pre-trained dense 3D model to estimate the 3D model parameters. There are some methods that use both 3D deformable model and 3D training data. For example, in Jourabloo and Liu (2015), cascaded regression method is used to estimate the 3D model coefficients and pose parameters from 2D images, which determine both the 2D and the 3D facial landmark locations.

There are a few algorithms that perform 3D facial landmark detection on 3D face scan. For example, in Papazov et al. (2015), eight 3D facial landmark locations are estimated. The algorithm first uses two 3D local shape descriptors, including the shape index feature and the spin image to generate landmark location candidates for each landmark. Then the final landmark locations are selected by fitting the 3D face model. Similarly, in Liang et al. (2013), 17 dominate landmarks are firstly detected on 3D face scan using their particular geometric properties. Then, a 3D template is matched to the testing face to estimate the 3D locations of 20 landmarks. In Papazov et al. (2015), dense 3D features are estimated from the 3D point cloud generated with depth sensor. In particular, a triangular surface patch descriptor is designed to select and match the training patches to the randomly generated patches from the testing image. Then, the associated 3D face shapes of the training patches are used to vote the 3D shape of the testing image.

Compared to the 2D facial landmark detection, 3D facial landmark detection is still new. There is lack of a large 3D face database with abundant 3D annotations. Compared to largely available 2D images, 3D face scans are difficult to obtain. Labeling 3D facial landmarks on 2D images or 3D face scan are usually more difficult than 2D landmarks.

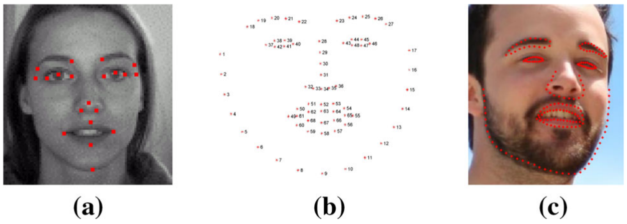

Fig. 14 Facial landmark annotations. Images are adapted from https://www.bioid.com/About/BioID-Face-Database, Sagonas et al. (2013) and Le et al. (2012). **a** 20 points, **b** 68 points and **c** 194 points

8 Databases and Evaluations

8.1 Landmark Annotations

Facial landmark annotations refer to the manual annotations of the groundtruth facial landmark locations on facial images. There are usually two types of facial landmarks: the facial key points and interpolated landmarks. The facial key points are the dominant landmarks on face, such as the eye corners, nose tip, mouth corners, etc. They possess unique local appearance/shape patterns. The interpolated landmark points either describe the facial contour or connect the key points (Fig. 14). In the early research, only sparse key landmark points are annotated and detected (Fig. 14a). Recently, more points are annotated in the new databases (Fig. 14b, c). For example, in BioID, 20 landmarks are annotated, while there are 68 and 194 landmarks annotated in ibug and Helen databases.

There are some issues with the existing landmark annotations. First, the landmark annotations are inherently bias and they are inconsistent across databases. As a result, it's difficult to combine multiple databases for evaluations. The annotation inconsistency also exists for individual landmark annotation. For example, for the annotation of eye corners, some databases tend to provide annotation within the eye region, while the others may annotate the point outside the eye region. To solve this issue, in Smith and Zhang (2014), a method is proposed to combine databases with different facial landmark annotations. It generates a union of landmark annotations by transferring the landmark annotations from source database to target database.

The second issue is that manual annotation is a time-consuming process. There are some works improving the annotation process. In Feng et al. (2015), 3D face scan and projection models are used to generate the synthetic 2D facial landmarks and corresponding landmark annotations. The synthetic images are then combined with real images to train the facial landmark detector. In Sagonas et al. (2013a), an iterative semi-automatic landmark annotation method is proposed. A facial landmark detector is initially trained with a small number of training data, and it is used to fit new testing images, which are selected by user to retrain the detector. Similarly, in Tong et al. (2012), a semi-supervised facial landmark annotation method is proposed. Even though the aforementioned methods improve the facial landmark annotation process, the annotation is still time-consuming and expensive. Overall, the existing training images and databases may still not be adequate for some landmark detection algorithms, such as the deep learning based methods. Finally, to scale up annotation to large datasets, online crowd-sourcing such as Amazon Mechanical Turk may be a potential method for facial landmark annotation.

8.2 Databases

There are two types of databases: databases collected under the "controlled" conditions or databases with "in-the-wild" images. See Table 3 for the summary.

8.2.1 Databases Under "Controlled" Conditions

Databases under "controlled" conditions refer to databases with videos/images collected indoor with certain restrictions (e.g. pre-defined expressions, head poses etc.).

- BioID (https://www.bioid.com/About/BioID-Face-Database): The data set contains 1521 gray scale indoor images with a resolution of 384 × 286 from 23 subjects. Images are taken under different illuminations and backgrounds. Subjects may show moderate expression variations. It contains landmark annotations of 20 points.
- AR (Martínez and Benavente 1998): The set contains 4000 frontal color images of 126 people with expressions, illumination, and facial occlusions (e.g. sun glasses and scarf). 22 landmark annotations are provided (http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/tarfd_markup/tarfd_markup.html).
- Extended YaleB (Georghiades et al. 2001): The extended Yale Face database B contains 16,128 images of 28 subjects under 9 poses and 64 illumination conditions. The database provides original images, the cropped facial images, and three annotated landmarks.
- FERET (Phillips et al. 1997): The Facial Recognition Technology (FERET) database contains 14,051 gray scale facial images, covering about 20 discrete head poses that differ in yaw angles. Frontal faces also have illumination and facial expression variations. 11 landmarks on selected profile faces are provided by Wu and Ji (2015).
- CK/CK+ (Kanade et al. 2000; Lucey et al. 2010): The Cohn-Kanade AU-coded expression database (CK) contains 486 (593 in CK+) video sequences of frontal faces from 97+ subjects with 6 basic expressions, including happy, surprised, sadness, disgust, fear and anger. The videos start from the neural expression and go to apex. CK+ is an extended version of CK database. It includes both posed and spontaneous expressions. AAM landmark tracking results are provided by the database.
- Multi-PIE database (Gross et al. 2010): The Multi-PIE face database contains more than 750,000 images of 337 subjects. The facial images are taken under 15 view points and 19 illumination conditions. A few facial expressions are included, such as neutral, smile, surprise, squint, disgust, and scream. 68 or 39 facial landmarks are annotated, depending on the head poses.
- XM2VTSDB (Messer et al. 1999): The Extended M2VTS database contains videos of 295 subjects with speech and

Table 3 Summary of the databases

Databases	Video (v)/image (i)	Gray (g)/color (c)	Amount of data	Variations	Number of landmark points
BioID (https://www.bioid.com/About/BioID-Face-Database)	i	g	1521	e, i	20
AR (Martínez and Benavente 1998)	i	c	4000+	e, i, o	22 (http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/tarfd_markup/tarfd_markup.html)
Extended YaleB (Georghiades et al. 2001)	i	g	16128	i, p	3
FERET (Phillips et al. 1997)	i	g	14501	e, i, p	11 (Wu and Ji 2015)
CK/CK+ (Kanade et al. 2000; Lucey et al. 2010)	v	g&c	486 (593)	e	68
MultiPIE (Gross et al. 2010)	i	c	750,000	e, i, p	68 or 39
XM2VTSDB (Messer et al. 1999)	i	c	1180	p	68 (Sagonas et al. 2013)
FRGC v2 (Phillips et al. 2005)	i	c	50,000	e, i	68 (Sagonas et al. 2013)
BU-4DFE (Yin et al. 2008)	i	c	3000	e	68 (Tulyakov and Sebe 2015)
AFLW (Koestinger et al. 2011)	i	c	25,000	(e, i, o, p)	21
LFPW (Belhumeur et al. 2013)	i	c	1432	(e, i, o, p)	29 (68) (Sagonas et al. 2013)
Helen (Le et al. 2012)	i	c	2330	(e, i, o, p)	168 (68) (Sagonas et al. 2013)
AFW (Zhu and Ramanan 2012)	i	c	205	(e, i, o), p	6 (68) (Sagonas et al. 2013)
ibug300 (Sagonas et al. 2013)	i	c	135	e, i, o, p	68
ibug300-VW (Shen et al. 2015)	v	c	114	(e, i, o, p)	68
COFW (Burgos-Artizzu et al. 2013)	i	c	1852	(e, i, p), o	29

We use the following notations to represent different variations. e: expression, i: illumination, o: occlusion, p: pose. "(.)" represents moderate/spontaneous variations

rotation head movements. 3D head model of each subject is also provided. 68 facial landmark annotations are provided by Sagonas et al. (2013).

- FRGC v2 (Phillips et al. 2005): The Face Recognition Grand Challenge (FRGC) database contains 50,000 facial images from 4,003 subject sessions with different lighting conditions and two facial expressions (smile and neutral). 3D images acquired by special sensor (Minolta Vivid 900/910) consisting of both range and texture images are also provided. 68 facial landmark annotations on selected images are provided by Sagonas et al. (2013).
- BU-4DFE (Yin et al. 2008): The Binghamton University 4D Facial Expression database (BU-4DFE) contains 2D and 3D videos for six prototypic facial expressions (e.g., anger, disgust, happiness, fear, sadness, and surprise) from 101 subjects (58 female and 43 male). There are approximately 60k+ images. 68 2D and 3D facial landmark annotations on selected images are provided by Tulyakov and Sebe (2015).

8.2.2 "In-the-wild" Databases

Recently, researchers focus on developing more robust and effective algorithms to handle facial landmark detection in real day-life situations. To evaluate the algorithms in those conditions, a few "in-the-wild" databases are collected from the webs, such as Flicks, facebook etc. They contain all sorts of variations, including head pose, facial expression, illumination, ethnicity, occlusion, etc. They are much more difficult than images with "controlled" conditions. Those databases are listed as follows:

- AFLW (Koestinger et al. 2011): The Annotated Facial Landmark in the Wild (AFLW) database contains about 25K images. The annotations include up to 21 landmarks based on their visibility.
- LFPW (Belhumeur et al. 2013): The Labeled Face Parts in the Wild (LFPW) database contains 1432 facial images. Since only the URLs are provided, some images are no longer available. 29 landmark annotations are provided by the original database. Re-annotations of 68 facial landmarks for 1132 training images and 300 testing images are provided by Sagonas et al. (2013).
- Helen database (Le et al. 2012): The Helen database contains 2330 high resolution images with dense 194 facial landmark annotations. Re-annotations of 68 landmarks are also provided by Sagonas et al. (2013).
- AFW (Zhu and Ramanan 2012): The Annotated Faces in the Wild (AFW) database contains about 205 images with relatively larger pose variations than the other "in-the-wild" databases. 6 facial landmark annotations are provided by the database, and re-annotations of 68 landmarks are provided by Sagonas et al. (2013).
- Ibug 300-W (Sagonas et al. 2013, b): The ibug dataset from 300 faces in the Wild (300-W) database [3] is the most challenging database so far with significant variations. It only contains 135 images with annotations of 68 landmarks.
- Ibug 300-VW (Shen et al. 2015): The 300 Video in the Wild (300) database contains 114 video sequences for three different scenarios from easy to difficult. 68 facial landmark annotations are provided.
- COFW (Burgos-Artizzu et al. 2013): The Caltech Occluded Faces in the Wild (COFW) database contains images with significant occlusions. There are 1345 training images and 507 testing images. There are annotations of 29 landmark locations and landmark occlusions.

8.3 Evaluation and Discussion

8.3.1 Evaluation Criteria

Facial landmark detection and tracking algorithms output the facial landmark locations in the facial images or videos. The accuracy is evaluated by comparing the detected landmark locations to the groundtruth facial landmark locations. In particular, if we denote the detected and groundtruth landmark locations for landmark i as $d_i = \{d_{x,i}, d_{y,i}\}$ and $g_i = \{g_{x,i}, g_{y,i}\}$, the detection error for the ith point is:

$$error_i = \|d_i - g_i\|_2 \tag{22}$$

One issue with the above criteria is that the error could change significantly for faces with different sizes. To handle this issue, there are several ways to normalize the error. The *inter-ocular distance* is the most popular criteria. If we denote the left and right pupil centers as g_{le} and g_{re}, we can calculate the normalized error as follows:

$$norm\_error_i = \frac{\|d_i - g_i\|_2}{\|g_{le} - g_{re}\|_2} \tag{23}$$

Besides the inter-ocular distance, some works (Sagonas et al. 2013a) may choose the distance between outer eye corners as a normalization constant. For particular images, such as images with extreme head poses ($\geq$ 60°) or occlusion (e.g., Figs. 11, 12), the eyes may not be visible. Therefore, some other normalization constants, such as the face size from the face bounding box (Zhu and Ramanan 2012) or the distance between outer eye corner and outer mouth corner (same side of the face) (Wu and Ji 2015) can be used as the normalization constants.

To accumulate the errors of multiple landmarks for one image, the average normalized errors are used:

$$norm\_error\_image = \frac{1}{N}\sum_{i}^{N}\frac{\|d_i - g_i\|_2}{\|g_{le} - g_{re}\|_2} \tag{24}$$

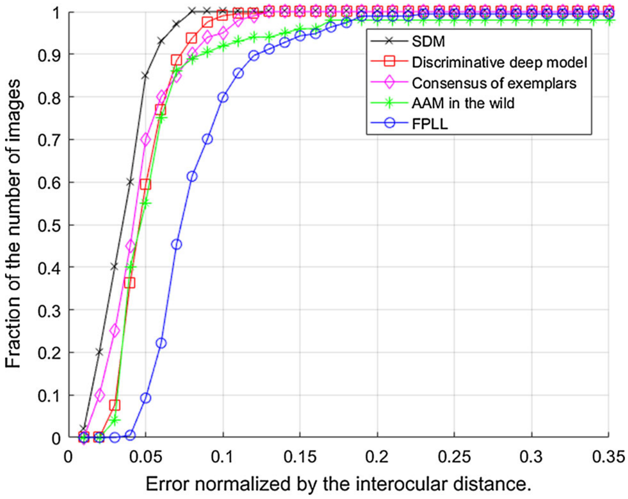

Fig. 15 Comparison of the cumulative distribution curves of several algorithms: SDM (Xiong and De la Torre Frade 2013), Discriminative deep model (Wu and Ji 2015), Consensus of exemplars (Belhumeur et al. 2013), AAM in the wild (Tzimiropoulos and Pantic 2013), and FPLL (Zhu and Ramanan 2012) on LFPW databases (Belhumeur et al. 2013). Figure is adapted from Wu and Ji (2015)

[3] Ibug 300-W database contains public available training images and private testing images. The training images include the annotations of public available databases and several newly collected images. Here, we name the newly collected images as Ibug 300-W database.

Table 4 Accuracy comparison of leading algorithms

Databases	# points	Methods	Type	Normalized error
BU-4DFE (Yin et al. 2008)	68 (Tulyakov and Sebe 2015)	One millisecond face alignment (Kazemi and Sullivan 2014)	R	5.22 (Tulyakov and Sebe 2015)
		3D regression (Tulyakov and Sebe 2015)	R	5.15*
AFLW (Koestinger et al. 2011)	21	Explicit shape regression (Cao et al. 2014)	R	8.24 (Zhu et al. 2016) (fs)
		RCPR (Burgos-Artizzu et al. 2013)	R	7.85 (Zhu et al. 2016) (fs)
		SDM (Xiong and De la Torre Frade 2013)	R	6.55 (Zhu et al. 2016) (fs)
		3DDFA (Zhu et al. 2016)	DR	5.32* (fs)
		HyperFace (Ranjan et al. 2016)	DR	4.26* (fs)
AFW (Zhu and Ramanan 2012)	5	Explicit shape regression (Cao et al. 2014)	R	10.4 (Zhang et al. 2014)
		RCPR (Burgos-Artizzu et al. 2013)	R	9.3 (Zhang et al. 2014)
		SDM (Xiong and De la Torre Frade 2013)	R	8.8 (Zhang et al. 2014)
		TCDCN (Zhang et al. 2014)	DR	8.2*
LFPW (Belhumeur et al. 2013)	29	Consensus of exemplars (Belhumeur et al. 2013)	C	3.99 (Cao et al. 2014)
		Robust facial landmark detection (Wu and Ji 2015)	R	3.93*
		One millisecond face alignment (Kazemi and Sullivan 2014)	R	3.8*
		RCPR (Burgos-Artizzu et al. 2013)	R	3.50*
		SDM (Xiong and De la Torre Frade 2013)	R	3.47*
		Explicit shape regression (Cao et al. 2014)	R	3.43*
		Face alignment 3000 fps (Ren et al. 2014)	R	3.35*
	68	FPLL (Zhu and Ramanan 2012)	C	8.29 (Zhu et al. 2015)
		DRMF (Asthana et al. 2013)	C	6.57 (Zhu et al. 2015)
		RCPR (Burgos-Artizzu et al. 2013)	R	6.56 (Zhu et al. 2015)
		Gaussian-Newton DPM (Tzimiropoulos and Pantic 2014)	C	5.92 (Zhu et al. 2015)
		SDM (Xiong and De la Torre Frade 2013)	R	5.67 (Zhu et al. 2015)
		CFAN (Zhang et al. 2014)	DR	5.44 (Zhu et al. 2015)
		CFSS (Zhu et al. 2015)	R	4.87*
Helen (Le et al. 2012)	68	FPLL (Zhu and Ramanan 2012)	C	8.166 (Zhu et al. 2015)
		DRMF (Asthana et al. 2013)	C	6.70 (Zhu et al. 2015)
		RCPR (Burgos-Artizzu et al. 2013)	R	5.93 (Zhu et al. 2015)
		Gaussian-Newton DPM (Tzimiropoulos and Pantic 2014)	C	5.69 (Zhu et al. 2015)
		CFAN (Zhang et al. 2014)	DR	5.53 (Zhu et al. 2015)
		SDM (Xiong and De la Torre Frade 2013)	R	5.50 (Zhu et al. 2015)
		CFSS (Zhu et al. 2015)	R	4.63*
	194	Stasm(ASM) (Milborrow and Nicolls 2008)	C	11.1 (Le et al. 2012)
		Component-based ASM (Le et al. 2012)	C	9.1*
		RCPR (Burgos-Artizzu et al. 2013)	R	6.50*
		SDM (Xiong and De la Torre Frade 2013)	R	5.85 (Ren et al. 2014)
		Explicit shape regression (Cao et al. 2014)	R	5.7*
		Robust facial landmark detection (Wu and Ji 2015)	R	5.49*
		Face alignment 3000 fps (Ren et al. 2014)	R	5.41*

Table 4 continued

Databases	# points	Methods	Type	Normalized error
		One millisecond face alignment (Kazemi and Sullivan 2014)	R	4.9*
		CFSS (Zhu et al. 2015)	R	4.74*
Ibug 300-W (Sagonas et al. 2013, b)	68	FPLL (Zhu and Ramanan 2012)	C	10.20 (Zhu et al. 2015)
		DRMF (Asthana et al. 2013)	C	9.22 (Zhu et al. 2015)
		RCPR (Burgos-Artizzu et al. 2013)	R	8.35 (Zhu et al. 2015)
		CFAN (Zhang et al. 2014)	DR	7.69 (Fan and Zhou 2016)
		Explicit shape regression (Cao et al. 2014)	R	7.58 (Ren et al. 2014)
		SDM (Xiong and De la Torre Frade 2013)	R	7.52 (Ren et al. 2014)
		One millisecond face alignment (Kazemi and Sullivan 2014)	R	6.4 (Zhu et al. 2015)
		Face alignment 3000 fps (Ren et al. 2014)	R	6.32*
		3DDFA (Zhu et al. 2016)	DR	6.31*
		CFSS (Zhu et al. 2015)	R	5.76*
		TCDCN (Zhang et al. 2014)	DR	5.54 (Fan and Zhou 2016)

Types: H: holistic methods, C: constrained local methods, R: regression based methods, DR: deep learning based regression methods. The number provided by the original paper is marked as “*”. The error normalized by face size is indicated as “fs”

Table 5 Efficiency comparison of leading algorithms

Methods	Type	# points	fps
TCDCN (Zhang et al. 2014)	DR	5	58
HyperFace (Ranjan et al. 2016)	DR	21	5
Consensus of exemplars (Belhumeur et al. 2013)	C	29	1 (Ren et al. 2014)
3DDFA (Zhu et al. 2016)	DR	68	13
CFAN (Zhang et al. 2014)	DR	68	40
CFSS (Zhu et al. 2015)	R	68	25
SDM (Xiong and De la Torre Frade 2013)	R	68	30
3D Regression (Tulyakov and Sebe 2015)	R	68	111
Explicit shape regression (Cao et al. 2014)	R	87	345
RCPR (Burgos-Artizzu et al. 2013)	R	194	6
One millisecond face alignment (Kazemi and Sullivan 2014)	R	194	1000
Face alignment 3000 fps (Ren et al. 2014)	R	194	200/1500

Types: H: holistic methods, C: constrained local methods, R: regression based methods, DR: deep learning based regression methods

To calculate the performances on multiple images, the *mean error* or the *cumulative distribution error* are used. The mean error calculates the mean of the normalized errors of multiple images. The cumulative distribution error calculates the percentages of images that lie under certain thresholds (see Fig. 15).

To evaluate the efficiency, the number of processed frames is used. Normally, facial landmark detection algorithms are evaluated on regular PC (e.g., laptop) without powerful GPU or parallel computing implementation etc.

8.3.2 Evaluation of Existing Algorithms

In Table 4, we list the performances of leading algorithms on the benchmark databases, their categories and the landmark detection errors. In Fig. 15, we show the cumulative distribution curves of some algorithms on LFPW dtabase. Note that, in this paper, we focus on the reported results from the existing literatures. There are additional detailed references Chrysos et al. (2017), Sagonas et al. (2016) and Shen et al. (2015) that provide original evaluations by running the software and implementations of known algorithms on different databases.

Table 6 Summary of the academic software

Methods	Detection (d) or tracking (t)	Realtime (y) or not (n)	Source code (sc) or binary code (bc)	Number of points	Links
Stasm (ASM) (Milborrow and Nicolls 2008)	d	y	sc	77	http://www.milbo.users.sonic.net/stasm/
DeMoLib (AAM, ASM etc.)	d		sc		http://staff.estem-uc.edu.au/roland/research/demolib-home/
Generic AAM (Tzimiropoulos et al. 2012)	d	n	sc	68	http://ibug.doc.ic.ac.uk/resources/aoms-generic-face-alignment/
AAM in the wild (Tzimiropoulos and Pantic 2013)	d	n	sc	68	http://ibug.doc.ic.ac.uk/resources/fitting-aams-wild-iccv-2013/
FPLL (Zhu and Ramanan 2012)	d	n	sc	68 or 39	http://www.ics.uci.edu/~xzhu/face/
Pose-free (Yu et al. 2013)	d	n	bc	66	http://www.research.rutgers.edu/~xiangyu/face_align.html
Flandmark (Uřičář et al. 2012)	d	y	sc	8	http://cmp.felk.cvut.cz/~uricamic/flandmark/
BoRMaN (Valstar et al. 2010), LEAR (Martinez et al. 2013)	d	n	bc	20	http://ibug.doc.ic.ac.uk/resources/facial-point-detector-2010/
DRMF (Asthana et al. 2013)	d	n	bc	66	http://ibug.doc.ic.ac.uk/resources/drmf-matlab-code-cvpr-2013/;
SDM (Xiong and De la Torre Frade 2013)	d & t	y	bc	49	http://www.humansensing.cs.cmu.edu/intraface/index.php
Face alignment 3000 fps (Ren et al. 2014)	d	y	sc	68	https://github.com/jwyang/face-alignment
RCPR (Burgos-Artizzu et al. 2013)	d&t	y	sc	29	http://www.vision.caltech.edu/xpburgos/ICCV13/
One millisecond face alignment (Kazemi and Sullivan 2014)	d	y	sc	194	http://www.csc.kth.se/~vahidk/face_ert.html
CNN (Sun et al. 2013)	d	y	bc	5	http://mmlab.ie.cuhk.edu.hk/archive/CNN_FacePoint.htm
Incremental face alignment (Asthana et al. 2014)	d&t	y	bc	49	http://ibug.doc.ic.ac.uk/resources/chehra-tracker-cvpr-2014/
Conditional regression forest (Dantone et al. 2012)	d		sc	10	http://www.dantone.me/projects-2/facial-feature-detection/
CLMZ (OpenFace) (Baltrušaitis et al. 2012)	d&t	y	sc	49	https://github.com/TadasBaltrusaitis/OpenFace/
CCNF (Baltrusaitis et al. 2014)	d	y	sc	30	https://www.cl.cam.ac.uk/~tb346/res/ccnf.html

Table 7 Summary of the commercial software

Methods	Detection (d) or tracking (t)	Realtime (y) or not (n)	Trial version	Number of points	Links
Face++	d	y	y	83/25/5	http://www.faceplusplus.com/demo-landmark/
Betaface	d	n	y	101	http://betaface.com/wpa/index.php/demo-gallery
Lambda Labs	d	y	y	6	https://lambdal.com/face-recognition-api#src
Visage	d&t	y	y	51	http://visagetechnologies.com/products-and-services/visagesdk/
LUXAND	d&t	y	y	66	https://www.luxand.com/facesdk/

The trial version refers to the free evaluation/download without full access/functionalities

There are several observations. First, generally, the regression based methods achieve much better performances than the holistic methods and the constrained local model methods, especially on images with significant variations (e.g., ibug 300-w). Second, deep learning based regression methods (e.g., Zhang et al. 2014) are the leading techniques and they achieve the state-of-the-art performances on several databases. Third, the performances of the same algorithm are different across database, but the rank of multiple algorithms is generally consistent.

The efficiencies of leading algorithms are shown in Table 5. Note that, the computational speeds of different algorithms are reported from their original papers and their evaluation methods may vary. For example, they have different implementation choices (matlab vs. C++), and they run on different computers. Some algorithms may only report the processing time by excluding the image loading time etc. Generally, we can see that the traditional cascaded regression methods (Kazemi and Sullivan 2014; Ren et al. 2014) are faster than the other methods.

The results shown here are generally consistent with the findings in Chrysos et al. (2017) and Sagonas et al. (2016). In Chrysos et al. (2017), it shows that the one millisecond face alignment (Kazemi and Sullivan 2014), the Supervised Descent method (Xiong and De la Torre Frade 2013), and CFSS (Zhu et al. 2015) are good options considering both the speed and accuracy.

8.4 Software

In Tables 6 and 7, we list a few academic software and commercial software. The academic software refers to the implementations of the existing methods with paper publications. The commercial software is usually only available in a limited sense. For commercial software, visage SDK covers many applications, including facial landmark detection, head pose estimation, and facial expression recognition, which is a good option.

9 Conclusion

In this paper, we reviewed the facial landmark detection algorithms in three major categories: the holistic methods, the constrained local methods, and the regression-based methods. In addition, we specifically discussed a few recent algorithms that try to handle facial landmark detection “in-the-wild” under different variations caused by head poses, facial expressions, facial occlusion etc. Furthermore, we discussed the popular benchmark databases, performances of leading algorithms and a few existing software.

There are still a few open questions about facial landmark detection. First, the current facial landmark detection and tracking algorithms still have problems on facial images under challenging “in-the-wild” conditions, including extreme head poses, facial occlusion, strong illumination, etc. The existing algorithms focus on solving one or a few conditions. There is still lack of a facial landmark detection and tracking algorithm that can handle all those cases. Second, there is a lack of a large facial image database that covers all different conditions with facial landmark annotations, which may significantly speed up the development of the algorithms. The existing databases only cover a few conditions (e.g. head poses and expressions). Third, facial landmark detection still heavily relies on the face detection accuracy, which may still fail in certain conditions. Fourth, the computational cost for some landmark detection and tracking algorithms is still high. The facial landmark detection and tracking algorithms should meet the real-time processing requirement.

There are a few future research directions. First, since there are similarities as well as unique properties about the methods in three major approaches, it would be beneficial to have a hybrid approach that combines all three approaches. For example, it would be interesting to see how and whether the appearance and shape models used in holistic methods and CLM can help the regression based methods. It is also interesting to study whether the analytic solutions used for the holistic methods can be applied to solve the cascaded regression at each stage, since they share similar object functions as discussed in Sect. 5. Vice versa, the cascaded regression idea may be applied to the holistic methods to predict the model coefficients in a cascaded manner. Second, currently, the dynamic information is utilized in a limited sense. The facial motion information should be combined with the facial appearance and facial shape for facial landmark tracking. For example, it would be interesting to see how and whether the dynamic features would help facial landmark tracking. Landmark tracking with facial structure information is also an interesting direction. Third, since there are relationships among facial landmark detection and other facial behavior analysis tasks, including head pose estimation and facial expression recognition, their interactions should be utilized for joint analysis. By leveraging their dependencies, we can incorporate the computer vision projection models and improve the performances for all tasks. Finally, to fully exploit the power of deep learning, a large annotated database of millions of images under different conditions is needed. Annotation of such a large image requires a hybrid annotation methods, including human annotation, online crowd sourcing, and automatic annotation algorithms.

References

Ahlberg, J. (2002). An active model for facial feature tracking. *EURASIP Journal on Advances in Signal Processing*, *2002*(6), 569,028.

Alabort-I-Medina, J., & Zafeiriou, S. (2014). Bayesian active appearance models. In *IEEE conference on computer vision and pattern recognition*.

Asthana, A., Zafeiriou, S., Cheng, S., & Pantic, M. (2013). Robust discriminative response map fitting with constrained local models. In *IEEE conference on computer vision and pattern recognition, CVPR '13*, pp. 3444–3451.

Asthana, A., Zafeiriou, S., Cheng, S., & Pantic, M. (2014). Incremental face alignment in the wild. In *IEEE conference on computer vision and pattern recognition*, pp. 1859–1866.

Baker, S., Gross, R., & Matthews, I. (2002). Lucas-kanade 20 years on: A unifying framework: Part 3. *International Journal of Computer Vision*, *56*, 221–255.

Baltrusaitis, T., Robinson, P., & Morency, L. P. (2014). Continuous conditional neural fields for structured regression. In *European conference on computer vision* (pp. 593–608). Springer.

Baltrušaitis, T., Robinson, P., & Morency, L. P. (2012). 3D constrained local model for rigid and non-rigid facial tracking. In *IEEE conference on computer vision and pattern recognition*.

Belhumeur, P., Jacobs, D., Kriegman, D., & Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2930–2940.

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., & Kumar, N. (2011). Localizing parts of faces using a consensus of exemplars. In *IEEE conference on computer vision and pattern recognition*.

BioID. https://www.bioid.com/About/BioID-Face-Database. Accessed 30 August 2015.

Bourel, F., Chibelushi, C., & Low, A. (2000). Robust facial feature tracking. In *British Machine Vision Conference*, pp. 24.1–24.10.

Burgos-Artizzu, X. P., Perona, P., & Dollar, P. (2013). Robust face landmark estimation under occlusion. In *IEEE international conference on computer vision*, pp. 1513–1520.

Cao, X., Wei, Y., Wen, F., & Sun, J. (2014). Face alignment by explicit shape regression. *International Journal of Computer Vision*, *107*, 177–190.

Chen, D., Ren, S., Wei, Y., Cao, X., & Sun, J. (2014). Joint cascade face detection and alignment. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *European Conference on Computer Vision, Lecture Notes in Computer Science* (Vol. 8694, pp. 109–122). Berlin: Springer.

Chrysos, G. G., Antonakos, E., Snape, P., Asthana, A., & Zafeiriou, S. (2017). A comprehensive performance evaluation of deformable face tracking "in-the-wild". *International Journal of Computer Vision*, *126*, 198–232.

Cootes, T., Walker, K., & Taylor, C. (2000). View-based active appearance models. In *IEEE international conference on automatic face and gesture recognition*, pp. 227–232.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(6), 681–685.

Cootes, T. F., Ionita, M. C., Lindner, C., & Sauer, P. (2012). Robust and accurate shape model fitting using random forest regression voting. In *European Conference on Computer Vision—Volume Part VII*, pp. 278–291.

Cootes, T. F., Taylor, C. J., Cooper, D. H., & Graham, J. (1995). Active shape models their training and application. *Computer Vision and Image Understanding*, *61*(1), 38–59.

Cosar, S., & Cetin, M. (2011). A graphical model based solution to the facial feature point tracking problem. *Image and Vision Computing*, *29*(5), 335–350.

Cristinacce, D., & Cootes, T. (2007). Boosted regression active shape models. In *British Machine Vision Conference*, pp. 880–889.

Cristinacce, D., & Cootes, T. F. (2004). A comparison of shape constrained facial feature detectors. In *International conference on automatic face and gesture recognition*, pp. 375–380.

Cristinacce, D., & Cootes, T. F. (2006). Feature detection and tracking with constrained local models. In *British Machine Vision Conference*.

Dantone, M., Gall, J., Fanelli, G., & Gool, L. V. (2012). Real-time facial feature detection using conditional regression forests. In *IEEE conference on computer vision and pattern recognition*.

Donner, R., Reiter, M., Langs, G., Peloschek, P., & Bischof, H. (2006). Fast active appearance model search using canonical correlation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(10), 1690–1694.

Edwards, G. J., Taylor, C. J., & Cootes, T. F. (1998). Interpreting face images using active appearance models. In *IEEE international conference on face and gesture recognition* (pp. 300–305). IEEE Computer Society.

Fan, H., & Zhou, E. (2016). Approaching human level facial landmark localization by deep learning. *Image and Vision Computing*, *47*(C), 27–35.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intellgence*, *32*(9), 1627–1645.

Feng, Z. H., Huber, P., Kittler, J., Christmas, W., & Wu, X. J. (2015). Random cascaded-regression copse for robust facial landmark detection. *IEEE Signal Processing Letters*, *22*(1), 76–80.

Georghiades, A., Belhumeur, P., & Kriegman, D. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(6), 643–660.

Ghiasi, G., & Fowlkes, C. (2014). Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *IEEE conference on computer vision and pattern recognition*, pp. 1899–1906.

Girshick, R. (2015). Fast r-cnn. In *The IEEE international conference on computer vision (ICCV)*.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Gou, C., Wu, Y., Wang, F. Y., & Ji, Q. (2016). Shape augmented regression for 3D face alignment, pp. 604–615. Cham.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, *40*(1), 33–51.

Gross, R., Matthews, I., & Baker, S. (2004). Appearance-based face recognition and light-fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(4), 449–465.

Gross, R., Matthews, I., & Baker, S. (2005). Generic vs. person specific active appearance models. *Image Vision and Computing*, *23*(12), 1080–1093.

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-pie. *Image Vision and Computing*, *28*(5), 807–813.

Gu, L., & Kanade, T. (2008). A generative shape regularization model for robust face alignment. In *European Conference on Computer Vision: Part I* (pp. 413–426). Berlin, Heidelberg: Springer.

Hansen, D. W., & Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(3), 478–500.

Heisele, B., Serre, T., & Poggio, T. (2007). A component-based framework for face detection and identification. *International Journal of Computer Vision*, *74*(2), 167–181.

Hou, X., Li, S., Zhang, H., & Cheng, Q. (2001). Direct appearance models. In *IEEE conference on computer vision and pattern recognition*, Vol. 1.

Hsu, G. S., Chang, K. H., & Huang, S. C. (2015). Regressive tree structured model for facial landmark localization. In *IEEE International conference on computer vision*, pp. 3855–3861.

Hu, C., Feris, R., & Turk, M. (2003). Real-time view-based face alignment using active wavelet networks. In *IEEE international workshop on analysis and modeling of faces and gestures*, pp. 215–221.

Jeni, L. A., Cohn, J. F., & Kanade, T. (2015). Dense 3D face alignment from 2D videos in real-time. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. articles/Jeni15FG_ZFace.pdf.

Jiao, F., Li, S., Shum, H., & Schuurmans, D. (2003). Face alignment using statistical models and wavelet features. In *IEEE conference on computer vision and pattern recognition*.

Jones, M., & Poggio, T. (1998). Multidimensional morphable models: A framework for representing and matching object classes. *International Journal of Computer Vision*, *29*(2), 107–131.

Jourabloo, A., & Liu, X. (2015). Pose-invariant 3D face alignment. In *2015 IEEE international conference on computer vision (ICCV)*, pp. 3694–3702.

Jourabloo, A., & Liu, X. (2016). Large-pose face alignment via CNN-based dense 3D model fitting. In *IEEE conference on computer vision and pattern recognition*. Las Vegas, NV.

Kanade, T., Cohn, J. F., & Tian, Y. Comprehensive database for facial expression analysis. In *IEEE international conference on automatic face and gesture recognition*, pp. 46–53.

Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1867–1874.

Koestinger, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2011). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE international workshop on benchmarking facial image analysis technologies*.

Landmark annotation for AR database. http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/tarfd_markup/tarfd_markup.html.

Le, V., Brandt, J., Lin, Z., Bourdev, L., & Huang, T. S. (2012). Interactive facial feature localization. In *European Conference on Computer Vision—Volume Part III*, pp. 679–692.

Levi, G., & Hassncer, T. (2015). Age and gender classification using convolutional neural networks. In *2015 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 34–42.

Li, Y., Wang, S., Zhao, Y., & Ji, Q. (2013). Simultaneous facial feature tracking and facial expression recognition. *IEEE Transactions on Image Processing*, *22*(7), 2559–2573.

Liang S Wu J, Liang, S., Wu, J., Weinberg, S. M., & Shapiro, L. G. (2013). Improved detection of landmarks on 3D human face data. In *Annual international conference of the IEEE Engineering in Medicine and Biology Society*.

Lopes, A. T., de Aguiar, E., Souza, A. F. D., & Oliveira-Santos, T. (2017). Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition*, *61*, 610–628.

Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE conference on computer vision and pattern recognition workshops*, pp. 94–101.

Martínez, A., & Benavente, R. (1998). The AR face database.

Martinez, B., Valstar, M. F., Binefa, X., & Pantic, M. (2013). Local evidence aggregation for regression-based facial point detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(5), 1149–1163.

Mathias, M., Benenson, R., Pedersoli, M., & Van Gool, L. (2014). Face detection without bells and whistles. In *European Conference on Computer Vision*.

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

Messer, K., Matas, J., Kittler, J., & Jonsson, K. (1999). XM2VTSDB: The extended M2VTS database. In *International conference on audio and video-based biometric person authentication*, pp. 72–77.

Milborrow, S., & Nicolls, F. (2008). Locating facial features with an extended active shape model. In *European Conference on Computer Vision: Part IV* (pp. 504–513). Berlin, Heidelberg: Springer.

Murphy-Chutorian, E., & Trivedi, M. (2009). Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(4), 607–626.

Nickels, K., & Hutchinson, S. (2002). Estimating uncertainty in SSD-based feature tracking. *Image and Vision Computing*, *20*, 47–58.

Pantic, M., & Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: The state of the art. *IEEE Transanctions on Pattern Analysis and Machine Intellgence*, *22*(12), 1424–1445.

Papazov, C., Marks, T., & Jones, M. (2015). Real-time head pose and facial landmark estimation from depth images using triangular surface patch features. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 4722–4730). IEEE.

Patacchiola, M., & Cangelosi, A. (2017). Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, *71*, 132–143.

Patrick Sauer, T. C., & Taylor, C. (2011). Accurate regression procedures for active appearance models. In *British Machine Vision Conference*.

Perakis, P., Passalis, G., Theoharis, T., & Kakadiaris, I. A. (2013). 3D facial landmark detection under large yaw and expression variations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(7), 1552–1564.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., et al. (2005). Overview of the face recognition grand challenge. In *IEEE conference on computer vision and pattern recognition, CVPR '05* (pp. 947–954). Washington, DC: IEEE Computer Society.

Phillips, P. J., Moon, H., Rauss, P., & Rizvi, S. A. (1997). The FERET evaluation methodology for face-recognition algorithms. In *IEEE conference on computer vision and pattern recognition, CVPR '97* (pp. 137–143). Washington, DC: IEEE Computer Society.

Ranjan, R., Patel, V. M., & Chellappa, R. (2016). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. CoRR arXiv:1603.01249.

Ren, S., Cao, X., Wei, Y., & Sun, J. (2014). Face alignment at 3000 FPS via regressing local binary features. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1685–1692.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, *47*, 3–18. 300-W, the First Automatic Facial Landmark Detection in-the-Wild Challenge.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE international conference on computer vision, 300 Faces in-the-Wild Challenge (300-W)*. Sydney, Australia.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013a). A semi-automatic methodology for facial landmark annotation. In *2013 IEEE conference on computer vision and pattern recognition workshops*, pp. 896–903.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013b). A semi-automatic methodology for facial landmark annotation. In *IEEE conference on computer vision and pattern recognition workshop*. Portland Oregon, USA.

Saragih, J., & Gocke, R. (2009). Learning AAM fitting through simulation. *Pattern Recognition*, *42*(11), 2628–2636.

Saragih, J., & Goecke, R. (2007). A nonlinear discriminative approach to AAM fitting. In *International conference on computer vision*, pp. 1–8.

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

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering.

Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., & Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *The IEEE international conference on computer vision (ICCV) workshops*.

Shen, X., Lin, Z., Brandt, J., & Wu, Y. (2013). Detecting and aligning faces by image retrieval. In *IEEE conference on computer vision and pattern recognition*.

Smith, B., Brandt, J., Lin, Z., & Zhang, L. (2014). Nonparametric context modeling of local appearance for pose- and expression-robust facial landmark localization. In *IEEE conference on computer vision and pattern recognition*, pp. 1741–1748.

Smith, B. M., & Zhang, L. (2014). *Collaborative facial landmark localization for transferring annotations across datasets* (pp. 78–93). Cham: Springer.

Sun, Y., Liang, D., Wang, X., & Tang, X. (2015). Deepid3: Face recognition with very deep neural networks. CoRR arXiv:1502.00873.

Sun, Y., Wang, X., & Tang, X. (2013). Deep convolutional network cascade for facial point detection. In *IEEE conference on computer vision and pattern recognition*, pp. 3476–3483.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification.

Tong, Y., Liu, X., Wheeler, F. W., & Tu, P. H. (2012). Semi-supervised facial landmark annotation. *Computer Vision and Image Understanding*, *116*(8), 922–935.

Tong, Y., Wang, Y., Zhu, Z., & Ji, Q. (2007). Robust facial feature tracking under varying face pose and facial expression. *Pattern Recognition*, *40*(11), 3195–3208.

Tresadern, P., Sauer, P., & Cootes, T. (2010). Additive update predictors in active appearance models. In *British Machine Vision Conference* (pp. 91.1–91.12). BMVA Press.

Trigeorgis, G., Snape, P., Nicolaou, M. A., Antonakos, E., & Zafeiriou, S. (2016). Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4177–4187. Las Vegas, NV, USA.

Tulyakov, S., & Sebe, N. (2015). Regressing a 3D face shape from a single image. In *IEEE international conference on computer vision*, pp. 3748–3755.

Tzimiropoulos, G., i medina, J. A., Zafeiriou, S., Pantic, M. (2012). Generic active appearance models revisited. In *Asian Conference on Computer Vision*, pp. 650–663. Daejeon, Korea.

Tzimiropoulos, G., & Pantic, M. Optimization problems for fast aam fitting in-the-wild. In *IEEE international conference on computer vision*, pp. 593–600.

Tzimiropoulos, G., & Pantic, M. (2014). Gauss-Newton deformable part models for face alignment in-the-wild. In *IEEE conference on computer vision and pattern recognition*, pp. 1851–1858.

Uřičář, M., Franc, V., & Hlaváč, V. (2012). Detector of facial landmarks learned by the structured output SVM. In *International conference on computer vision theory and applications* (Vol. 1, pp. 547–556). Portugal.

Valstar, M., Martinez, B., Binefa, V., & Pantic, M. (2010). Facial point detection using boosted regression and graph models. In *IEEE conference on computer vision and pattern recognition*, pp. 13–18.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *IEEE conference on computer vision and pattern recognition*, Vol. 1, pp. I-511–I-518.

Williams, O., Blake, A., & Cipolla, R. (2005). Sparse Bayesian learning for efficient visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1292–1304.

Wu, Y., & Ji, Q. (2015). Discriminative deep face shape model for facial point detection. *International Journal of Computer Vision*, *113*(1), 37–53.

Wu, Y., & Ji, Q. (2015). Robust facial landmark detection under significant head poses and occlusion. In *International conference on computer vision*.

Wu, Y., & Ji, Q. (2016). Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *IEEE conference on computer vision and pattern recognition*.

Wu, Y., Wang, Z., & Ji, Q. (2013). Facial feature tracking under varying facial expressions and face poses based on restricted Boltzmann machines. In *IEEE conference on computer vision and pattern recognition*, pp. 3452–3459.

Wu, Y., Wang, Z., & Ji, Q. (2014). A hierarchical probabilistic model for facial feature detection. In *IEEE conference on computer vision and pattern recognition*, pp. 1781–1788.

Xiong, X., & De la Torre Frade, F. (2013). Supervised descent method and its applications to face alignment. In *IEEE international conference on computer vision and pattern recognition*.

Xiong, X., & la Torre, F. D. (2015). Global supervised descent method. In *IEEE conference on computer vision and pattern recognition*, pp. 2664–2673.

Yan, S., Hou, X., Li, S. Z., Zhang, H., & Cheng, Q. (2003). Face alignment using view-based direct appearance models. Special issue on facial image processing, analysis and synthesis. *International Journal of Imaging Systems and Technology*, *13*, 106–112.

Yang, H., & Patras, I. (2013). Privileged information-based conditional regression forest for facial feature detection. In *IEEE international conference and workshops on automatic face and gesture recognition*, pp. 1–6.

Yin, L., Chen, X., Sun, Y., Worm, T., & Reale, M. (2008). A high-resolution 3D dynamic facial expression database. FG 2,3,5.

Yu, X., Huang, J., Zhang, S., Yan, W., & Metaxas, D. (2013). Pose free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *IEEE international conference on computer vision*.

Yu, X., Lin, Z., Brandt, J., & Metaxas, D. N. (2014). Consensus of regression for occlusion-robust facial feature localization. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *European Conference on Computer Vision, Lecture Notes in Computer Science* (Vol. 8692, pp. 105–118). Berlin: Springer.

Zhang, C., & Zhang, Z. (2010). *A survey of recent advances in face detection*. Tech. Rep. MSR-TR-2010-66.

Zhang, J., Shan, S., Kan, M., & Chen, X. (2014). Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In *European Conference on Computer Vision, Part II*, pp. 1–16.

Zhang, Z., Luo, P., Loy, C., & Tang, X. (2014). Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision, Part II*, pp. 94–108.

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2016). Learning deep representation for face alignment with auxiliary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(5), 918–930.

Zhao, X., Kim, T. K., & Luo, W. (2014). Unified face analysis by iterative multi-output random forests. In *IEEE conference on computer vision and pattern recognition*, pp. 1765–1772.

Zhou, E., Fan, H., Cao, Z., Jiang, Y., & Yin, Q. (2013). Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *IEEE international conference on computer vision workshops*, pp. 386–391.

Zhu, S., Li, C., Change Loy, C., & Tang, X. (2015). Face alignment by coarse-to-fine shape searching. In *IEEE conference on computer vision and pattern recognition*.

Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S. (2016). Face alignment across large poses: A 3D solution. In *IEEE conference on computer vision and pattern recognition*. Las Vegas, NV.

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *IEEE conference on computer vision and pattern recognition*, pp. 2879–2886.