Facial Feature Point Detection: A Comprehensive Survey

Nannan Wang · Xinbo Gao · Dacheng Tao · Xuelong Li

arXiv:1410.1037v1 [cs.CV] 4 Oct 2014

Abstract This paper presents a comprehensive survey of facial feature point detection with the assistance of abundant manually labeled images. Facial feature point detection favors many applications such as face recognition, animation, tracking, hallucination, expression analysis and 3D face modeling. Existing methods can be categorized into the following four groups: constrained local model (CLM)-based, active appearance model (AAM)-based, regression-based, and other methods. CLM-based methods consist of a shape model and a number of local experts, each of which is utilized to detect a facial feature point. AAM-based methods fit a shape model to an image by minimizing texture synthesis errors. Regression-based methods directly learn a mapping function from facial image appearance to facial feature points. Besides the above three major categories of methods, there are also minor categories of methods which we classify into *other* methods: graph-ical model-based methods, joint face alignment methods, independent facial feature point detectors, and deep learning-based methods. Though significant progress has been made, facial feature point detection is limited in its success by wild and real-world conditions: variations across poses, expressions, illuminations, and occlusions. A comparative illustration and analysis of representative methods provide us a holistic understanding and deep insight into facial feature point detection, which also motivates us to explore promising future directions.

N. Wang
VIPS Lab, School of Electronic Engineering, Xidian University, 710071, Xi'an, P. R. China
E-mail: nannanwang.xidian@gamil.com

X. Gao
VIPS Lab, School of Electronic Engineering, Xidian University, 710071, Xi'an, P. R. China
E-mail: xbgao@mail.xidian.edu.cn

D. Tao
Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering & Information Technology, University of Technology, Sydney, 235 Jones Street, Ultimo, NSW 2007, Australia
E-mail: Dacheng.Tao@uts.edu.au

X. Li
School of Computer Science and Information Systems, Birkbeck College, University of London, Malet Street, London, WC1E 7HX, U.K.
E-mail: xuelong@dcs.bbk.ac.uk

1 Introduction

Facial feature points, also known as facial landmarks or facial fiducial points, have semantic meaning. Facial feature points are mainly located around facial components such as eyes, mouth, nose and chin (see Fig. 1). Facial feature point detection (FFPD) refers to a supervised or semi-supervised process using abundant manually labeled images. FFPD usually starts from a rectangular bounding box returned by face detectors (Viola and Jones, 2004; Yang et al, 2002) which implies the location of a face. This bounding box can be employed to initialize the positions of facial feature points. Facial feature points are different from keypoints for image registration (Ozuysal et al, 2010) and keypoint detection is usually an unsupervised procedure.

Suggested by (Cootes et al, 1995), facial feature points can be reduced to three types: points labeling parts of faces with application-dependent significance, such as the center of an eye or the sharp corners of a
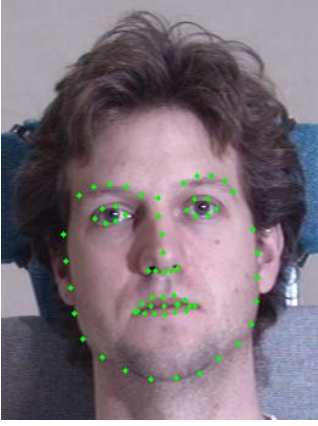
Fig. 1 Illustration of an example image with 68 manually labeled points from the Multi-PIE database (Gross et al, 2010).

boundary; points labeling application-independent elements, such as the highest point on a face in a particular orientation, or curvature extrema (the highest point along the bridge of the nose); and points interpolated from points of the previous two types, such as points along the chin. According to various application scenarios, different numbers of facial feature points are labeled as, for example, a 17-point model, 29-point model or 68-point model. Whatever the number of points is, these points should cover several frequently-used areas: eyes, nose, and mouth. These areas carry the most important information for both discriminative and generative purposes. Generally speaking, more points indicate richer information, although it is more time-consuming to detect all the points.

The points shown in Fig. 1 can be concatenated to represent a shape $\mathbf{x} = (x_1, \cdots, x_N, y_1, \cdots, y_N)^T$ where (x_i, y_i) denotes the location of the i-th point and N is the number of points (N is 68 in this figure). Given a sufficiently large number of manually labeled points and corresponding images as the training data, the target of facial feature point detection is to localize the shape of an input testing image according to the facial appearance. Detecting the shape of a facial image is a challenging problem due to both the rigid (scale, rotation, and translation) and non-rigid (such as facial expression variation) face deformation. FFPD generally consists of two phases: in the training phase, a model is learned from the appearance variations to the shape variations; and in the testing phase, the learned model is applied to an input testing image to localize facial feature points (shape). Normally the shape search process starts from a coarse initialization, following which the initial shape is moved to a better position step by step until convergence. According to the method of modeling the shape variation and the appearance variation, existing FFPD methods can be grouped into four cate-

gories: constrained local model (CLM)-based methods (here, the term CLM should be not confused with that in Cristinacce and Cootes (2006b) which is a special case of CLM in our nomenclature), active appearance model (AAM)-based methods, regression-based methods and other methods.

CLM-based methods consider the appearance variation around each facial feature point independently. One response map can therefore be calculated from the appearance variation around each facial feature point with the assistance of a corresponding local expert. Facial feature points are then predicted from these response maps refined by a shape prior which is generally learned from training shapes. AAM-based methods model the appearance variation from a holistic perspective. In addition, both the shape and appearance variation model are usually constructed from a linear combination of some bases learned from training shapes and images. Regression-based methods estimate the shape directly from the appearance without learning any shape model or appearance model. There are also other FFPD methods which do not fall into any of the aforementioned categories and are classified into the category of 'other methods'. This category can be further divided into four sub-categories: graphical model-based methods, joint face alignment methods, independent facial feature point detectors, and deep learning-based methods. Table 1 and Fig.2 present the development timeline of the four categories of methods. As shown in the table and figure, the topic has attracted growing interest.

Many related research topics and real-world applications could benefit from the accurate detection of facial feature points. Lee and Kim (Lee and Kim, 2009) explored the fitted shape and shape-normalized appearance of the proposed tensor-based active appearance model (AAM) (Cootes et al, 1998a) to transform the input image into a normalized image (frontal pose, neural expression, and normal illumination) to conduct variation-robust face recognition. Stegmann et al (2003) applied AAM to medical image analysis. (Zhou et al, 2005) proposed a fusion strategy to incorporate subspace model constraints for robust shape tracking. Chen et al (2001) applied active shape model (ASM) (Cootes and Taylor, 1992) to separate the shape from the texture to favor the sketch generation process. FFPD for face alignment is an essential preprocessing step in face hallucination (Wang et al, 2014) and facial swapping (Bitouk et al, 2008). Facial animation (Weise et al, 2011) generally detects facial feature points to control the variation of facial appearance. The combination of 2D and 3D view-based AAM is utilized to robustly describe the variation of facial expression across different poses (Sung and Kim, 2008). The correspondence of facial feature points

Table 1 Development timeline of four categories of methods.

Year	CLM	AAM	Regression	*Other* Methods			
				GM	Joint	Independent	DL
1992	[1] Cootes and Taylor (1992)						
1993	[2] Cootes and Taylor (1993)						
1994	[3] Cootes et al (1994)						
1995	[4] Cootes et al (1995); [5] Sozou et al (1995)						
1997	[6] Sozou et al (1997)						
1998		[7] Cootes et al (1998a); [8] Cootes et al (1998b)					
2001		[9] Cootes et al (2001); [10] Cootes and Taylor (2001); [11] Hou et al (2001)					
2002		[12] Cootes et al (2002)		[13] Coughlan and Ferreira (2002)			
2003	[14] Cristinacce and Cootes (2003); [15] Zhou et al (2003)	[16] Batur and Hayes (2003)					
2004	[17] Cristinacce and Cootes (2004); [18] Cristinacce et al (2004)	[19] Matthews and Baker (2004)					
2005		[20] Batur and Hayes (2005); [21] Gross et al (2005)				[22] Vukadinovic and Pantic (2005)	
2006	[23] Cristinacce and Cootes (2006a); [24] Cristinacce and Cootes (2006b)	[25] Cootes and Taylor (2006); [26] Dedeoglu et al (2006); [27] Donner et al (2006); [28] Liu et al (2006)		[29] Liang et al (2006a); [30] Liang et al (2006b)			
2007	[31] Cristinacce and Cootes (2007); [32] Sukno et al (2007); [33] Vogler et al (2007)	[34] Gonzalez-Mora et al (2007); [35] Kahraman et al (2007); [36] Matthews et al (2007); [37] Peyras et al (2007); [38] Roberts et al (2007); [39] Saragih and Goecke (2007); [40] Sung et al (2007)	[41] Zhou and Comaniciu (2007)	[42] Huang et al (2007b)			
2008	[43] Cristinacce and Cootes (2008); [44] Gu and Kanade (2008); [45] Liang et al (2008); [46] Miborrow and F. (2008); [47] Wang et al (2008a); [48] Wang et al (2008b); [49] Wimmer et al (2008)	[50] Nguyen and De la Torre (2008); [51] Nguyen and Torre (2008); [52] Papandreou and Maragos (2008); [53] Saragih et al (2008); [54] Sung et al (2008)	[55] Kozakaya et al (2008a); [56] Kozakaya et al (2008b)			[57] Ding and Martinez (2008)	
2009	[58] Li et al (2009); [59] Lucey et al (2009); [60] Paquet (2009); [61] Saragih et al (2009a); [62] Saragih et al (2009b); [63] Saragih et al (2009c); [64] Tresadern et al (2009)	[65] Amberg et al (2009); [66] Asthana et al (2009); [67] Hamsici and Martinez (2009); [68] Lee and Kim (2009); [69] Liu (2009); [70] Saragih and Gocke (2009)			[71] Tong et al (2009)	[72] Asteriadis et al (2009)	
2010		[73] Ashraf et al (2010); [74] Martins et al (2010); [75] Nguyen and Torre (2010); [76] Tresadern et al (2010)	[77] Kozakaya et al (2008b); [78] Valstar et al (2010)			[79] Ding and Martinez (2010)	
2011	[80] Belhumeur et al (2011); [81] Chew et al (2011); [82] Li et al (2011); [83] Roh et al (2011); [84] Saragih (2011); [85] Saragih et al (2011)	[86] Asthana et al (2011); [87] Hansen et al (2011); [88] Navarathna et al (2011); [89] Sauer et al (2011)	[90] Kazemi and Cullivan (2011)		[91] Zhao et al (2011)		
2012	[92] Baltrusaitis et al (2012); [93] Cootes et al (2012); [94] Le et al (2012); [95] Martins et al (2012a); [96] Martins et al (2012b)	[97] Huang et al (2012); [98] Kinoshita et al (2012); [99] Tresadern et al (2012); [100] Tzimiropoulos et al (2012)	[101] Cao et al (2012); [102] Dantone et al (2012); [103] Rivera and Martinez (2012); [104] Sanchez-Lozano et al (2012); [105] Yang and Patras (2012)	[106] Michal et al (2012); [107] Zhu and Ramanan (2012)	[108] Smith and Zhang (2012); [109] Tong et al (2012); [110] Zhao et al (2012)		[111] Luo et al (2012)
2013	[112] Asthana et al (2013); [113] Belhumeur et al (2013); [114] Yu et al (2013)	[115] Anderson et al (2013); [116] Fanelli et al (2013); [117] Martins et al (2013); [118] Tzimiropoulos and Pantic (2013); [119] Lucey et al (2013)	[120] Burgos-Artizzu et al (2013); [121] Martinez et al (2013); [122] Xiong and De la Torre (2013); [123] Yang and Patras (2013)	[124] Zhao et al (2013)		[125] Shen et al (2013)	[126] Smith et al (2013); [127] Sun et al (2013); [128] Wu et al (2013)

Note: Representative methods after 2006 are surveyed very carefully. Important works before 2006 are also included. In the table, "GM" denotes the sub-category of graphical model-based methods, "Joint" represents the sub-category of joint face alignment, "Independent" means the sub-category of independent facial feature point detectors, and "DL" is the abbreviation of deep learning.
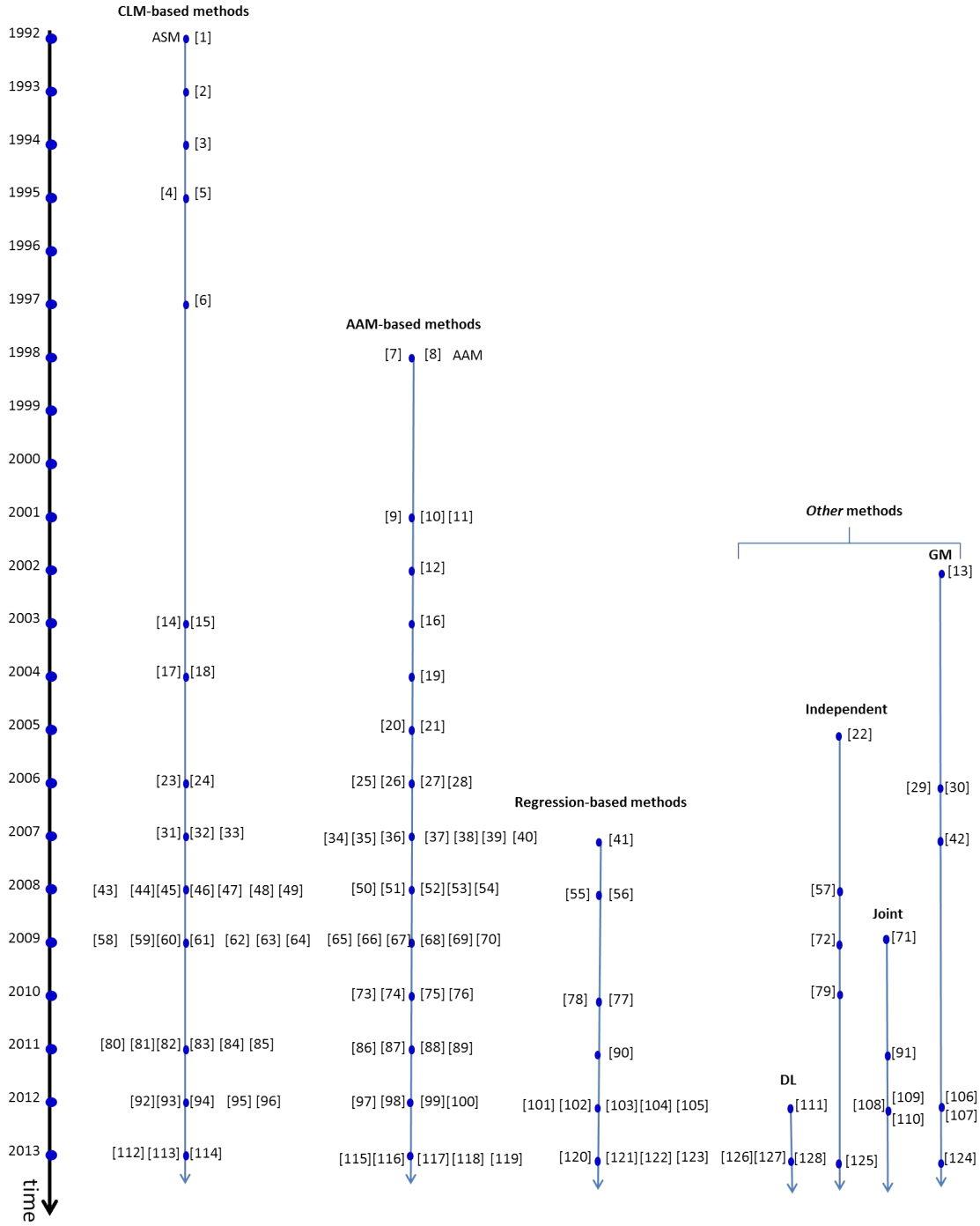
Fig. 2 Development timeline of four categories of methods. The reference number in this figure is indexed according to Table 1.

plays an important role in 3D face modeling (Blanz and Vetter, 1999). Anderson et al (2013) applied AAM to track robustly and quickly over a very large corpus of expressive facial data and to synthesize video realistic renderings in the visual text-to-speech system.

Table 2 shows general notations commonly appearing in this paper. The remainder of this paper is orga-

nized as follows: Sections 2 to 5 investigate the aforementioned four categories of methods, respectively. Section 6 evaluates and analyzes the performance of several representative methods. Finally, Section 7 summarizes the paper, and discusses some promising future directions and tasks regarding FFPD.

Table 2 Notations

Symbols	Descriptions
N	The number of landmarks labeled in each image
m	The number of principal modes in texture models of AAM
n	The number of principal modes in the point distribution model
$\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_i, \cdots, \alpha_N)^T$	Shape parameters in the point distribution model (PDM)
$\boldsymbol{\beta} = (\beta_1, \cdots, \beta_i, \cdots, \beta_N)$	Texture parameters in the texture model of AAM
λ_i	The i-th eigenvalue in the PDM
$\mathbf{P}_s$	Shape projection matrix in PDM
$\mathbf{P}_a$	Texture projection matrix in the texture model of AAM
$\mathbf{I}$	An input testing image
$\mathbf{I}^{(e)}$	Identity matrix whose order is determined in the context
$\mathbf{x}$	A shape represented in the image frame
$(x_i, y_i)^T$ or $\mathbf{x}_{(i)}$	Coordinate of the i-th point in the image frame
$\mathbf{s}$	A shape represented in the reference (mean-shape) frame
$\mathbf{s}_i$	The i-th shape basis in the PDM
$\mathbf{a}$	A texture representation in the reference frame
$\mathbf{a}_i$	The i-th texture basis in the texture model of AAM
$\mathbf{s}_0$	The mean shape
$\mathbf{a}_0$	The mean texture in the reference frame
$\mathbf{c}$	Appearance parameters in AAM
s	Sclae in the rigid transformation
$\mathbf{R}$	2×2 rotation matrix with orientation θ
$\mathbf{t}$	2×1 translation vector in rigid transformation
$\mathbf{q}$	Pose parameters (4×1 vector: s, θ, and $\mathbf{t}$)
$\mathbf{p}$	Rigid and non-rigid shape parameters ($\mathbf{p} = (\mathbf{q}^T; \boldsymbol{\alpha}^T)^T$)

2 Constrained Local Model-Based Methods

CLM-based methods fit an input image for the target shape through optimizing an objective function, which is comprised of two terms: shape prior $\mathcal{R}(\mathbf{p})$ and the sum of response maps $\mathcal{D}_i(\mathbf{x}_i; \mathbf{I})$, $(i = 1, \cdots, N)$ obtained from N independent local experts (Saragih et al, 2011):

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

A shape model is usually learned from training facial shapes and is taken as the prior refining the configuration of facial feature points. Each local expert is trained from the facial appearance around the corresponding feature point and is utilized to compute the response map which measures detection accuracy. The CLM objective function in the equation (1) can be interpreted from a probabilistic perspective:

$$\max_{\mathbf{p}} p(\mathbf{p}) \prod_{i=1}^{N} p(l_i = 1 | \mathbf{x}_i, \mathbf{I}). \tag{2}$$

where $l_i \in \{1, -1\}$ indicates whether the i-th point is aligned or misaligned, $\mathcal{R}(\mathbf{p}) = -ln\{p(\mathbf{p})\}$ and $\mathcal{D}_i(\mathbf{x}_i; \mathbf{I}) = -lnp(l_i = 1 | \mathbf{x}_i, \mathbf{I})$. The CLM objective function (either (1) or (2)) implicitly assumes that N response maps are calculated independently.

In the offline phase, a shape model and local experts should be learned from training shapes and corresponding images. Then in the online phase, given an input image, the output shape can be solved from the optimization of equation (1). We will investigate commonly used shape models and local experts sequentially. Finally, methods on how to combine the shape model and local experts for optimization are investigated.

2.1 Shape Model

Fig. 3 illustrates the statistical distribution of facial feature points sampled from 600 facial images. Regarding the shape prior, multivariate Gaussian distribution is commonly assumed, otherwise known as the point distribution model (PDM) proposed by Cootes and Taylor (Cootes and Taylor, 1992):
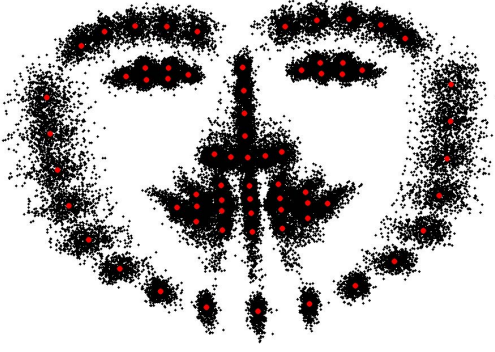
Fig. 3 Illustration of statistical distribution of facial feature points. There are 600 shapes (smaller dot points in black) normalized by Procrustes analysis. The larger dot points in red indicate the mean shape of all shapes.

$$\mathbf{s} = \mathbf{s}_0 + \mathbf{P}_s\boldsymbol{\alpha} = \mathbf{s}_0 + \sum_{i=1}^{n} \alpha_i \mathbf{s}_i, \qquad (3)$$

where $\mathbf{s}_i(i = 0, \cdots, n)$ can be estimated by the principal component analysis (PCA) on all aligned training shapes. Actually, $\mathbf{s}_0$ is the mean of all these shapes and $\mathbf{s}_1, \cdots, \mathbf{s}_n$ are the eigenvectors corresponding to the n largest eigenvalues of the covariance matrix of all aligned training shapes. n is usually determined by preserving $90\% \sim 98\%$ variance (the ratio between the sum of n largest eigenvalues and sum of all eigenvalues). Mei et al (2008) suggested the above rule to determine whether the value of n is reliable or not and further explored bootstrap stability analysis to improve reliability. To remove the effect of rigid transformation, all training shapes are aligned by Procrustes analysis before learning the shape model. We call this rigid transformation-free shape $\mathbf{s}$ in a reference frame. We apply rigid transformation to $\mathbf{s}$ to generate a shape $\mathbf{x}$ in the image frame:

$$\mathbf{x}_{point} = s\mathbf{R}\mathbf{s}_{point} + \mathbf{t}_{point}, \qquad (4)$$

where $\mathbf{t}_{point}$ consists of N replications of $\mathbf{t}$ and $\mathbf{s}_{point}$ denotes a rearranged $2 \times N$ matrix with each column corresponding to one point in $\mathbf{s}$. Similarly, $\mathbf{x}$ is the rearrangement of $\mathbf{x}_{point}$.

An eigenspace shown in equation (3) can be represented by a quadruple: mean vector, matrix of eigenvectors, eigenvalues, and the number of observations to construct the eigenspace. Eigenspace fusion (Hall et al, 2000) merges two eigenspaces into one eigenspace, which has great significance for online updating. Butakoff and Frangi (Butakoff and Frangi, 2006) generalized the eigenspace fusion model (Hall et al, 2000) to a weighted version and applied it to merge multiple ASMs

(or AAMs). Their experimental results show that fused ASMs have similar performance to full ASMs (model constructed from full set of observations) in terms of both segmentation error and time cost. They also applied the above fusion model to multi-view face segmentation (Butakoff and Frangi, 2010), which can be casted as a two-model fusion problem: the fusion of a frontal view model and a left profile model; and the fusion of a frontal view model and a right profile model. Faces in intermediate view can be interpolated through fusion weight estimation.

In addition to PDM, there are several improvements on the prior shape distribution. Considering that PCA can only model the linear structure of shapes, Cootes et al. (Sozou et al, 1995, 1997) generalized the linear PDM to a nonlinear version by exploring polynomial regression and multi-layer perceptron respectively. Gu and Kanade (2006) proposed a 3D face alignment method in a single testing image based on a 3D PDM. To project the 3D shapes to the 2D plane, a weak perspective projection is assumed between the 3D space and the 2D plane. De la Torre and Nguyen (2008) proposed a kernel PCA-based nonlinear shape model. Since a single Gaussian is inadequate for modeling the distribution over facial feature points, a mixture of Gaussian has been explored (Cootes and Taylor, 1999; Everingham et al, 2006; Sivic et al, 2009). In PDM (see equation (3)), shapes are constrained on the subspace spanned by principal components. Saragih (2011) exploited the principal regression analysis to span a constrained subspace. Since PDM assumes Gaussian observation noise and learns a shape model using all the training data, it is vulnerable to gross feature detection errors due to partial occlusions or spurious background features. Li et al. (Li et al, 2009, 2011) thus presented a robust shape model exploring random sample consensus (Fischler and Bolles, 1981). This method is in a hypothesis-and-test form, i.e. given a set of hypotheses, the one that satisfies certain optimal conditions should be chosen. Object shape and pose hypotheses (parameters) are first generated from randomly sampled partial shape-subsets of feature points. Subsequently, the hypotheses are tested to find the one that minimizes the shape prediction error.

Besides the above explicit shape models, Cristinacce et al (2004) proposed an implicit shape model known as pairwise reinforcement of feature responses, which models a shape by learning the pairwise distribution of all ground truth feature point locations relative to the optimal match of each corresponding individual feature detector.
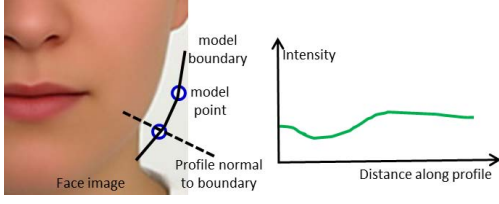
Fig. 4 ASM search profile.

2.2 Local Expert

A local expert functions to compute a response map on the local region around corresponding facial feature points, i.e. we have N local experts in a FFPD model. The region that supports a local expert could be either one-dimensional (i.e. a line) or two-dimensional (such as a rectangular region). A local expert can be a distance metric such as the Mahalanobis distance (Cootes et al, 1995), a classifier such as linear support vector machine (Wang et al, 2008a), or a regressor (Cristinacce and Cootes, 2007; Saragih et al, 2009c).

Regarding ASM, Cootes et al (1995) defined the support region as the profile normal to the model boundary through each shape model point (see Fig. 4). Along the profile, k pixels are sampled from both sides of the model point in the i-th training image. Then $2k + 1$ samples (actually gradients of these pixels) can be concatenated into a column vector $\mathbf{g}_i$. After being normalized by the sum of the absolute value of elements in the vector, the mean $\bar{\mathbf{g}}$ and the covariance $\mathbf{S}_g$ can be estimated from all training vectors $\{\mathbf{g}_i\}$. They adopted the multivariate Gaussian distribution assumption for the vectors. The fitting response for a new sample vector $\mathbf{g}_s$ is given by

$$(\mathbf{g}_s - \bar{\mathbf{g}})^T \mathbf{S}_g^{-1} (\mathbf{g}_s - \bar{\mathbf{g}}), \tag{5}$$

which is also known as the Mahalanobis distance of the sample vector from the model mean. The authors then provided a quantitative evaluation of the active shape model search using these local grey-level models (Cootes and Taylor, 1993).

The aforementioned Mahalanobis distance-based methods assume that the local appearance is Gaussian distributed. This Gaussian distribution assumption does not always hold and thus may result in inferior performance. Classifier-based local experts separate aligned from misaligned locations, and so they ignore the local appearance variations. These experts are trained from positive image patches (centered at corresponding facial feature points) and negative image patches (with their centers displaced from the correct facial feature point positions). The linear support vector machine is
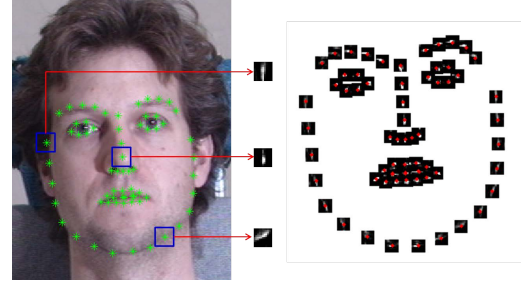

Fig. 5 Illustration of response maps. The area within the blue box is the region used to calculate the response maps. The red dots indicate the ground truth facial feature point locations.

frequently chosen due to its efficiency (Saragih et al, 2011; Wang et al, 2008a). Taking the local expert corresponding to the i-th facial feature point as an example:

$$\mathcal{C}_i(\mathbf{x}; \mathbf{I}) = \mathbf{w}_i \mathbf{f} + \gamma_i, \tag{6}$$

where $\{\mathbf{w}_i, \gamma_i\}$ denote the gain and the bias, respectively and $\mathbf{f}$ represents the normalized patch vector with zero mean and unit variance. To reformulate the output of the classifier in a probability form, the logistic regression is employed to refine the equation (6):

$$\frac{1}{1 + e^{a_i \mathcal{C}_i(\mathbf{x}; \mathbf{I}) + b_i}}. \tag{7}$$

Given an estimated shape, we can calculate the response map within the region around each facial feature point according to equation (7). Fig. 5 shows the response maps of 66 classifiers (specifically, linear support vector machines).

An alternative way to model the local expert is to exploit regressors instead of classifiers. Cristinacce and Cootes (2007) explored GentleBoost (Friedman et al, 2000) to learn a regressor from the local neighborhood appearance to the displacement between the center of the local neighborhood and the true facial feature point location. Saragih et al (2009c) claimed that a fixed mapping function (regressor) would take a complex form to incorporate the issues of generalizability and computational complexity. Considering a fixed mapping function cannot adapt to face variations in identity, pose, illumination and expression, they developed a bilinear model. Cootes et al (2012) introduced random forest (Breiman, 2001) to the CLM framework. Random forest learns response maps taking Haar-like features as the regressor input. PDM statistically models the shape models and regularizes the global shape configuration. The motivation behind the regressor rather than the classifier is that the regressor can potentially provide

more useful information, such as the distance of negative patches from a positive patch, while classifiers only determine whether an image patch is positive or negative. However, learning a regressor is more difficult than constructing a classifier.

2.3 Improvements and Extensions

The fitting of CLM-based methods consists of two main steps: (1) predicting local displacements of shape model points; (2) constraining the configuration of all point to adhere to the shape model. These two steps are iterated until they satisfy a convergence criterion.

Cootes and Taylor (Cootes and Taylor, 1992; Cootes et al, 1995) proposed to search the "better" candidate point locations along profiles normal to the boundary. Corresponding displacements from current point locations to sought "better" locations should then be refined to adapt the PDM. The fitting objective function can be written in a similar form as equation (1) (Saragih et al, 2011):

$$\min_{\mathbf{p}} \|\boldsymbol{\alpha}\|^2_{\boldsymbol{\Lambda}^{-1}} + \sum_{i=1}^{N} \omega_i \left\| \mathbf{x}_{(i)} - \boldsymbol{\mu}_{(i)} \right\|^2, \qquad (8)$$

where $\boldsymbol{\Lambda} = diag\{[\lambda_1; \cdots ; \lambda_n]\}$, $\boldsymbol{\mu}_{(i)}$ is the sought location of the i-th facial feature point corresponding to the peak response (the maximum value of the equation (4)), the weights $\{\omega_i\}_{i=1}^{N}$ measure the significance of the peak response coordinates. In the above optimization problem, the first term neglects regularization on rigid transformation parameters $\mathbf{q}$ by assuming a non-informative prior. To minimize the problem (8), the first order Taylor expansion of the PDM's points is applied:

$$\mathbf{x}_{(i)} \approx \mathbf{x}_{(i)}^c + \mathbf{J}_i \triangle \mathbf{p}, \qquad (9)$$

where $\mathbf{x}^c = [\mathbf{x}_{(1)}^c; \cdots ; \mathbf{x}_{(N)}^c]$ denotes the current approximated PDM shape, and $\mathbf{J} = [\mathbf{J}_1; \cdots ; \mathbf{J}_N]$ is the PDM's Jacobian matrix. Substituting the equation (9) into (8) we can then obtain the increment for updating the parameters:

$$\triangle \mathbf{p} = -\mathbf{H}^{-1} \left(\tilde{\boldsymbol{\Lambda}}^{-1} \mathbf{p} + \sum_{i=1}^{N} \omega_i \mathbf{J}_i^T (\mathbf{x}_i^c - \boldsymbol{\mu}_i) \right), \qquad (10)$$

where $\mathbf{H} = \tilde{\boldsymbol{\Lambda}}^{-1} + \sum_{i=1}^{N} \omega_i \mathbf{J}_i^T \mathbf{J}_i$ is the Gauss-Newton Hessian and $\tilde{\boldsymbol{\Lambda}} = diag\{[\mathbf{0}; \lambda_1; \cdots ; \lambda_n]\}$. The parameters can be updated in an additive manner: $\mathbf{p} \leftarrow \mathbf{p} + \triangle \mathbf{p}$. Indeed, from a probabilistic perspective (Saragih et al,

2011), the ASM's fitting procedure is equivalent to modeling the response maps by the isotropic Gaussian estimators $\{\mathcal{N}(\mathbf{x}_{(i)}; \boldsymbol{\mu}_i, \omega_i^{-1} \mathbf{I}^{(e)})\}_{i=1}^{N}$.

Since the emergence of the seminal work ASM (Cootes et al, 1995), quantity variants have been proposed. Cootes et al (1994) proposed a multi-resolution strategy to improve the fitting performance from coarse to fine. The optimized solution on a low resolution image is taken as the initialization of the next higher resolution image. This strategy overcomes the sensitivity to initialization to some extent. Roh et al (2011) found that the least squares in the equation (7) determines whether the problem will have optimal results only when the assumption of Gaussian noise is satisfied. However, since non-Gaussian noise is regularly encountered, they proposed to employ two strategies (M-estimator and random sampling) to robustly estimate these parameters. Zhou et al (2003) formulated FFPD into a maximum a posterior (MAP) problem in the tangent space and designed an expectation-maximization (EM)-based fitting algorithm to solve the MAP optimization. The tangent shape is iteratively updated by a weighted average of model shapes and tangent projection of the observed shape while the shape is reconstructed from model shapes in ASM. Furthermore, continuous regularization of shape parameters was applied while traditional ASM discontinuously truncated shape parameters to constrain the shape variation, which could result in unstable estimation. Vogler et al (2007) proposed to combine a 3D deformable model based on ASM for reliable real-time tracking. The 3D deformable model mainly governs the overall variation of a face (such as shape, orientation, and location). Several ASMs are trained, each corresponding to a viewpoint to govern the 2D facial feature variations. Miborrow et al (2010) extended the 1D profile to the 2D profile (actually a squared area) which outperformed the traditional ASM. Wimmer et al (2008) investigated how to learn local objective functions for face model fitting. They claimed that a best local objective function should have the following two properties: (1) a global minimum corresponding to the best model fit; (2) no local extrema or saddle points. They then learned objective functions under the framework of ASM. Rather than using the Mahalanobis distance, they explored tree-based regression to learn an objective function mapping from the extracted Haar-like and edge features.

To further facilitate the localization of facial feature points, some component-based methods have also been proposed. Liang et al (2008) utilized the component locations as constraints to regularize the configuration of facial features. This method first detects 11 components by cascaded boosting classifier (Viola and Jones,

2004). By solving a fitting objective similar to that of ASM except for an additional constraint term of component locations, a fitted shape can be resolved. To further improve the detection accuracy of components, the authors proposed to utilize direction classifiers to determine the search direction for component locations. These direction classifiers (3 classifiers for left/right profile, brows, and upper/lower lips and 9 classifiers for other components) are trained from positive and negative samples with respect to corresponding components. To determine the appropriate position along the above detected direction, a customized searching strategy is designed. Since new positions of components are found, an updated shape can be achieved by solving the aforementioned fitting objective function. Through several such iterations, a reasonable shape can ultimately be reached. Le et al (2012) presented a component-based ASM model and an interactive refinement algorithm. According to the aforementioned descriptions, ASM consists of two models: a profile model (local expert) and a shape model. Unlike the ASM method, which models all points by a single multivariate Gaussian distribution, this approach separates the whole face into seven components and constructs a Gaussian model for each component. To obtain a reasonable configuration of these components, the locations of these components (centroids of these components) are further modeled by Gaussian distribution. In other words, the shape model is decomposed into two modules: component shape fitting and configuration model fitting. For the profile model, besides the unary scores (owing to the fact that N local detectors are independent), binary constraint (tree-structure) is introduced to refine each pair of neighboring landmarks. Similar binary constraints have been imposed by tree-structure (Zheng et al, 2006) and MRF (graph structure with loops) in Tresadern et al (2009). Lastly, dynamic programming is explored to solve the fitting problem. Moreover, the authors introduced a user-assisted facial feature point localization strategy to further decrease localization error.

van Ginneken et al (2002) substituted the fixed normalized first derivative profile (Cootes et al, 1995) with a distinct set of optimal features for each facial feature point. A nonlinear k-nearest neighbor (kNN) classifier instead of linear Mahalanobis distance has also been explored to search the optimal displacement for points. Sukno et al (2007) proposed a generalization of optimal features ASM (van Ginneken et al, 2002). A reduced set of differential invariant features is taken as the local appearance descriptors, which are invariant to rigid transformation. In the fitting phase, a sequential features selection method is adopted to choose a subset of features for each point. To further speed the procedure, multivalued neurons (MVN) (Aizenberg et al, 2000) are adopted to replace the kNN classifier in van Ginneken et al (2002).

Inspired by the cascaded face detection method (Viola and Jones, 2004), Cristinacce and Cootes (2003) proposed to detect each local facial feature point by trained an Adaboost classifier. To constrain the global configurations of these points and reliably locate each point, multivariate Gaussian was assumed for the shape point distribution. They then further (Cristinacce and Cootes, 2004) extended this model by utilizing three templates to compute the response map for each individual feature point: normalized correlation template, orientation maps and boosted classifier. The fitting objective is to maximize the sum of all these response maps under the constraint that each PDM shape parameter should be under the threshold of three standard deviations. The Nelder-Mead simplex method (Nelder and Mead, 1965) was explored to optimize this problem. They then (Cristinacce and Cootes, 2006a) proposed an adaptive strategy to update the template to compute the response maps. They (Cristinacce and Cootes, 2006b, 2008) first proposed the term "constrained local model" consisting of two steps: the first step is to calculate the response maps for each facial feature point; the second step is to maximize the sum of response scores under the Gaussian prior constraint, as in equation (7). Given an input testing image, templates are generated through an appearance model (AAM (Cootes et al, 2001)) constructed from vectors, each of which is the concatenation of image patches extracted from each facial feature point in a training image. These templates are iteratively updated in the fitting process. The Nelder-Mead simplex method (Nelder and Mead, 1965) is utilized to optimize the problem.

Lucey et al (2009) proposed an improved version of the method (Cristinacce and Cootes, 2008), named exhaustive local search, in the following aspects: (1) substitute the original generative patch experts with a discriminative expertt trained by linear support vector machine; (2) decompose the complex fitting function into N independent fitting problems, which greatly favors real-time performance; (3) exploit a composite rather than additive warp update step. However, this method only utilizes the maximum of a response map for each facial feature point, neglecting the distribution of response maps. Furthermore, constraints to the shape configuration are not taken into account, which may lead to an invalid shape.

Although isotropic Gaussian estimation to the response maps leads to an efficient and simple approxi-

mation, it may fail in some cases if the response maps cannot be modeled by isotropic Gaussian distributions.

Wang et al (2008a) proposed to approximate the response map by: $\{\mathcal{N}(\mathbf{x}_{(i)}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}_{i=1}^{N}$, anisotropic Gaussian estimators. Here $\boldsymbol{\Sigma}_i$ is the full covariance matrix. $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ can be inferred from a convex quadratic function fitted to the negative log of the response maps (obtained from a linear support vector machine). The fitting problem can be written as (Saragih et al, 2011):

$$\min_{\mathbf{p}} \|\boldsymbol{\alpha}\|_{\boldsymbol{\Lambda}^{-1}}^2 + \sum_{i=1}^{N} \left\| \mathbf{x}_{(i)} - \boldsymbol{\mu}_{(i)} \right\|_{\boldsymbol{\Sigma}_i^{-1}}^2, \qquad (11)$$

Then the Gauss-Newton update is:

$$\triangle \mathbf{p} = -\mathbf{H}_{ani}^{-1} \left(\tilde{\boldsymbol{\Lambda}}^{-1} \mathbf{p} + \sum_{i=1}^{N} \mathbf{J}_i^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}_i^c - \boldsymbol{\mu}_i) \right), \qquad (12)$$

where $\mathbf{H}_{ani} = \tilde{\boldsymbol{\Lambda}}^{-1} + \sum_{i=1}^{N} \mathbf{J}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{J}_i$. They subsequently applied this strategy to non-rigid face tracking (Wang et al, 2008b). Paquet proposed a Bayesian version of the method (Wang et al, 2008a) which can be seen as the maximum likelihood solution of the proposed Bayesian method.

Considering that response maps may be multimodal, a single Gaussian estimator cannot model the density distribution. Gu and Kanade (2008) employed a Gaussian mixture model (GMM) to approximate the response maps: $\{\sum_{k=1}^{K_i} \pi_{ik} \mathcal{N}(\mathbf{x}_{(i)}; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik})\}_{i=1}^{N}$ where K_i is the number of Gaussian components to model the response map corresponding to the i-th point and π_{ik} are the mixing coefficients. GMM parameters are estimated from the GMM fitting process to the response maps. Finally the optimization problem is (Saragih et al, 2011):

$$\min_{\mathbf{p}} \|\boldsymbol{\alpha}\|_{\boldsymbol{\Lambda}^{-1}}^2 + \sum_{i=1}^{N} \sum_{k=1}^{K_i} \omega_{ik} \left\| \mathbf{x}_{(i)} - \boldsymbol{\mu}_{(ik)} \right\|_{\boldsymbol{\Sigma}_{ik}^{-1}}^2, \qquad (13)$$

where $\omega_{ik} = \frac{\pi_{ik} \mathcal{N}(\mathbf{x}_{(i)}; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik})}{\sum_{j=1}^{K_i} \pi_{jk} \mathcal{N}(\mathbf{x}_{(i)}; \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij})}$. Through Gaussian-Newton optimization, the update can be calculated:

$$\triangle \mathbf{p} = -\mathbf{H}_{GMM}^{-1} \left(\tilde{\boldsymbol{\Lambda}}^{-1} \mathbf{p} + \sum_{i=1}^{N} \sum_{k=1}^{K_i} \omega_{ik} \mathbf{J}_i^T \boldsymbol{\Sigma}_{ik}^{-1} (\boldsymbol{\mu}_{ik} - \mathbf{x}_i^c) \right), \qquad (14)$$

where $\mathbf{H}_{GMM} = \tilde{\boldsymbol{\Lambda}}^{-1} + \sum_{i=1}^{N} \sum_{k=1}^{K_i} \omega_{ik} \mathbf{J}_i^T \boldsymbol{\Sigma}_{ik}^{-1} \mathbf{J}_i$. Saragih et al (2009a) utilized GMM to approximate response maps of the proposed mixture of local experts. They have a similar fitting objective as in equation (13) except without the shape prior regularization.

Unlike previous methods approximating response maps in parametric forms, Saragih et al. (Saragih et al, 2009b, 2011) proposed a non-parametric estimate in the form of a homoscedastic isotropic Gaussian kernel density estimate: $\{\sum_{\mathbf{y}_{(j)} \in \boldsymbol{\Psi}_i} \pi_{\mathbf{y}_{(j)}} \mathcal{N}(\mathbf{x}_{(i)}; \mathbf{y}_{(j)}, \rho \mathbf{I}^{(e)})\}_{i=1}^{N}$. Here $\boldsymbol{\Psi}_i$ represents all integer pixel locations within the rectangular region around i-th facial feature point, $\pi_{\mathbf{y}_{(i)}}$ denotes the likelihood that the i-th point is aligned at location $\mathbf{y}_{(i)}$ which can be estimated from the equation (7), and ρ denotes the variance of the noise on point locations which can be determined from training data as $\rho = \frac{1}{N-n} \sum_{i=n+1}^{N} \lambda_i$. The fitting objective function is as follows:

$$\min_{\mathbf{p}} \|\boldsymbol{\alpha}\|_{\boldsymbol{\Lambda}^{-1}}^2 + \sum_{i=1}^{N} \sum_{\mathbf{y}_{(j)} \in \boldsymbol{\Psi}_i} \frac{\omega_{\mathbf{y}_{(j)}}}{\rho} \left\| \mathbf{x}_{(i)} - \mathbf{y}_{(j)} \right\|^2, \qquad (15)$$

where $\omega_{\mathbf{y}_{(j)}} = \frac{\pi_{\mathbf{y}_{(i)}} \mathcal{N}(\mathbf{x}_{(i)}; \mathbf{y}_{(i)}, \rho \mathbf{I}^{(e)})}{\sum_{\mathbf{z}_{(j)}} \pi_{\mathbf{z}_{(j)}} \mathcal{N}(\mathbf{x}_{(i)}; \mathbf{z}_{(j)}, \rho \mathbf{I}^{(e)})}$. The update is:

$$\triangle \mathbf{p} = -(\rho \tilde{\boldsymbol{\Lambda}}^{-1} + \mathbf{J}^T \mathbf{J})^{-1} (\rho \tilde{\boldsymbol{\Lambda}}^{-1} \mathbf{p} - \mathbf{J}^T \mathbf{v}), \qquad (16)$$

where $\mathbf{v} = [\mathbf{v}_1; \cdots; \mathbf{v}_N]$ and
$\mathbf{v}_i = \left(\sum_{\mathbf{y}_{(j)} \in \boldsymbol{\Psi}_i} \frac{\pi_{\mathbf{y}_{(j)}} \mathcal{N}(\mathbf{x}_{(i)}; \mathbf{y}_{(j)}, \rho \mathbf{I}^{(e)})}{\sum_{\mathbf{z}_{(j)}} \pi_{\mathbf{z}_{(j)}} \mathcal{N}(\mathbf{x}_{(i)}; \mathbf{z}_{(j)}, \rho \mathbf{I}^{(e)})} \right) - \mathbf{x}_i^c$
which is in a similar form with mean-shift. To further handle partial occlusions, they used an M-estimator to substitute the least square in the equation (15).

The above method (Saragih et al, 2011) has been extensively investigated due to its effectiveness and efficiency. Chew et al (2011) have applied this method to facial expression detection. Excluding the influence of occluded points through random sample consensus (Fischler and Bolles, 1981) hypothesis-and-test strategy, Roh et al (2011) proposed an algorithm robust to occlusion. Response maps achieved from linear SVM are represented in a multi-model fashion resulting from the mean-shift segmentation (each segmented region is modeled by a 2D-Gaussian distribution). Baltrusaitis et al (2012) extended the above method (Saragih et al, 2011) to a 3D version. In addition to general face images, they explored the information of depth images. The mean of response maps estimated from the general image and corresponding depth image is taken as the final response map. Yu et al (2013) explored the mean-shift method (Saragih et al, 2011) to rapidly approach the global optimum in their proposed two-stage cascaded deformable shape model and then utilized component-wise active contours to discriminatively refine the subtle shape variation.

Unlike the aforementioned non-parametric and parametric approximations to response maps, Asthana et al

(2013) directly regressed the PDM shape update parameters from the low-dimensional representation of response maps through a series of weak learners. The response maps can be obtained from linear support vector machines and the low-dimensional representation is obtained from the PCA projection. Linear support vector regression plays the role of the weak learner.

Martins et al (2012a) claimed that the above subspace-constrained mean-shift method (Saragih et al, 2009b, 2011) is vulnerable to outliers owing to a least squares projection. They formulated the objective as maximum a posterior (MAP) of PDM shape parameters $p(\boldsymbol{\alpha}|\mathbf{s})$ and pose parameters $p(\mathbf{q}|\mathbf{s})$ respectively conditioned on the observed shape. The observed shape here is obtained according to response maps. The MAP $p(\boldsymbol{\alpha}|\mathbf{s})$ can be decomposed into $p(\mathbf{s}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})$. According to PDM (see equation (3)), $p(\mathbf{s}|\boldsymbol{\alpha})$ can be modeled as a Gaussian distribution:

$$p(\mathbf{s}|\boldsymbol{\alpha}) \propto$$
$$exp\Big(-\frac{1}{2}\big(\mathbf{s}-(\mathbf{s}_0+\sum_{i=1}^{n}\alpha_i\mathbf{s}_i)\big)^T\boldsymbol{\Sigma}_{\mathbf{s}}^{-1}\big(\mathbf{s}-(\mathbf{s}_0+\sum_{i=1}^{n}\alpha_i\mathbf{s}_i)\big)\Big),$$
$$(17)$$

where $\boldsymbol{\Sigma}_{\mathbf{s}}$ indicates the uncertainty of the spatial localization of all points and can be estimated from response maps. To simplify the optimization procedure, they adopted the conjugate prior (Martins et al, 2012b), i.e. $p(\boldsymbol{\alpha})$ distributes as a Gaussian. The MAP problem $p(\mathbf{q}|\mathbf{s})$ was processed in the same way. Finally the MAP problem was optimized by linear dynamical systems.

Belhumeur et al (2011) proposed a method that combines the output of local experts with a non-parametric global model. The local expert is applied through a support vector machine taking the SIFT feature (Lowe, 2004) as the input. Based on the response maps of these local experts, the objective is to maximize the posterior probability $p(\mathbf{x}|\mathbf{d})$ where $\mathbf{d}$ represents the response maps of all local experts. Since the location corresponding to the highest response map value is not always the correct location due to occlusions and appearance ambiguities, they further designed a non-parametric set of global models from the similarity transformation of training exemplar images to constrain the configurations of these facial feature points. The random sample consensus method (Fischler and Bolles, 1981) is explored to optimize the global model.

Amberg and Better (2011) casted the FFPD detection problem as a discrete programming problem given a number of candidate positions for each point. They utilized decision forest (Breiman, 1984) to detect a number of candidate locations for each point. The

facial feature localization problem is actually to determine the indexes of points in corresponding candidate points which minimize the distance between the shape model and the image points. A fixed 3D shape projected according to a weak perspective camera is taken as the shape model. The objective is globally minimized by the branch and bound method.

3 Active Appearance Model-Based Methods

3.1 Active Appearance Model

An active appearance model (AAM) (Gao et al, 2010) can be decoupled into a linear shape model and a linear texture model. The linear shape is obtained in the same way as in the CLM framework (see equation (3)). To construct the texture model, all training faces should be warped to the mean-shape frame by triangulation or thin plate spline method; the resultant images should be free of shape variation, called shape-free textures. Each shape-free texture is raster scanned into a grey-level vector $\mathbf{z}_i$. To eliminate the effect of global lighting variation, $\mathbf{z}_i$ is normalized by a scaling u and offset v:

$$\mathbf{a} = \frac{\mathbf{z}_i - v \cdot \mathbf{1}}{u}, \qquad (18)$$

where u and v represent the variance and the mean of the texture $\mathbf{z}_i$ respectively and $\mathbf{1}$ is a vector of all 1s with the same length as $\mathbf{z}_i$. The texture model can be generated by applying PCA on all normalized textures as follows:

$$\mathbf{a} = \mathbf{a}_0 + \mathbf{P}_a\boldsymbol{\beta} = \mathbf{a}_0 + \sum_{i=1}^{m}\beta_i\mathbf{a}_i, \qquad (19)$$

The coupled relationship between the shape model and the texture model is bridged by PCA on shape and texture parameters:

$$\begin{pmatrix}\mathbf{W}_s\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{pmatrix} = \begin{pmatrix}\mathbf{W}_s\mathbf{P}_s^T(\mathbf{s}-\mathbf{s}_0) \\ \mathbf{P}_a^T(\mathbf{a}-\mathbf{a}_0)\end{pmatrix} = \begin{pmatrix}\mathbf{Q}_s \\ \mathbf{Q}_a\end{pmatrix}\mathbf{c} \qquad (20)$$

where $\mathbf{W}_s$ is a diagonal weighting matrix measuring the difference between the shape and texture parameters. The appearance parameter vector $\mathbf{c}$ governs both the shape and texture variation. To simplify the parameter representation, here we still utilize $\mathbf{p}$ to incorporate all necessary parameters: appearance parameters $\mathbf{c}$, pose parameters $\mathbf{q}$ and texture transformation parameters u and v.

The fitting objective of AAM is to minimize the difference between the texture $\mathbf{a}_s$ sampled from the testing

image and the texture $\mathbf{a}_m$ synthesized by the model. Let $\mathbf{r}(\mathbf{p}) = \mathbf{a}_s - \mathbf{a}_m$. Cootes et al (1998a) first proposed to model the relationship between $\mathbf{r}(\mathbf{p})$ (it was warped to the image frame in Cootes et al (1998a)) and parameter update $\delta\mathbf{p}$ by linear regression:

$$\delta\mathbf{p} = \mathbf{A} \cdot \mathbf{r}(\mathbf{p}), \qquad (21)$$

where $\mathbf{A}$ was solved by multiple multivariate linear regression on a sample of known model displacements $\delta\mathbf{p}$ and the corresponding difference texture $\mathbf{r}(\mathbf{p})$. Cootes et al (2001) later developed a Gaussian-Newton optimization method. Applying a first order Taylor expansion to $\mathbf{r}(\mathbf{p})$:

$$\mathbf{r}(\mathbf{p} + \delta\mathbf{p}) = \mathbf{r}(\mathbf{p}) + \frac{\partial\mathbf{r}}{\partial\mathbf{p}}\delta\mathbf{p}, \qquad (22)$$

Through solving the optimization problem: $\min_{\delta\mathbf{p}}\|\mathbf{r}(\mathbf{p}+\delta\mathbf{p})\|^2$, we receive the optimal solution:

$$\delta\mathbf{p} = -\mathbf{Br}(\mathbf{p}), \mathbf{B} = \left(\frac{\partial\mathbf{r}^T}{\partial\mathbf{p}}\frac{\partial\mathbf{r}}{\partial\mathbf{p}}\right)^{(-1)}\frac{\partial\mathbf{r}^T}{\partial\mathbf{p}}. \qquad (23)$$

Considering the fact that updating $\frac{\partial\mathbf{r}^T}{\partial\mathbf{p}}$ at every iteration is expensive, the authors fixed it in a constant matrix which can be estimated from training images by numeric differentiation.

3.2 Improvements and Extensions

Due to the flexible and simple framework of AAM, it has been extensively investigated and improved. However, many difficulties are encountered when AAM is applied to real applications. These difficulties are generally encountered from the following three aspects: the low efficiency for real-time applications, the less discrimination for classification, and the lack of robustness under inconstant circumstances. As our previous work (Gao et al, 2010) does, we review the developments of AAM from these three aspects.

3.2.1 Efficiency

Due to the high-dimensional texture representation and the unconstrained optimization, the original AAM suffers from low efficiency in real-time systems. We investigate improvements from these two aspects respectively.

1) Texture representation

To reduce the redundancy information contained in the texture, Cootes et al (1998b) only subsampled a number of pixels. Pixels corresponding to a number of

the largest elements in the regression matrix are assumed to be helpful and are preserved. This procedure decreases the dimension of texture representation. However, since the assumption is not always tenable, it cannot be guaranteed to obtain reasonable results.

Since learning the regression matrix in (21) or (23) is time- and memory-consuming, Hou et al (2001) learned the regression from the low-dimensional representation (PCA projection) of texture difference to the position displacement. Moreover, considering that the mapping from the texture to the shape is many-to-one, they proposed to linearly model the relationship between the texture and the shape to cater for this.

Tresadern et al (2012) explored Haar-like features to provide a computationally inexpensive linear projection for efficiency to facilitate facial feature point tracking on a mobile device. To provide high accuracy, a hierarchical model that utilizes tailored training data is designed.

2) Optimization

In order to improve the efficiency of the fitting process, Matthews and Baker (2004) considered the AAM as an image alignment problem and optimized it by inverse compositional method (Baker et al, 2003) based on independent AAM. Here, independent AAM indicates that the linear shape model and linear texture model are not combined, as in the original literature. The method aims to minimize the following objective function:

$$\sum_{\mathbf{s}_{(k)} \in \mathbf{s}_0} \left[\mathbf{a}_0(\mathbf{s}_{(k)}) + \sum_{i=1}^{m}\beta_i\mathbf{a}(\mathbf{s}_{(k)}) - \mathbf{I}\Big(\mathcal{Q}(\mathcal{W}(\mathbf{s}_{(k)};\mathbf{p});\mathbf{q})\Big)\right]^2, \qquad (24)$$

where $\mathbf{s}_{(k)}$ denotes any pixel location with the area enclosed by the mean shape $\mathbf{s}_0$, $\mathcal{W}(\mathbf{s}_{(k)};\mathbf{p})$ represents the pixel location after warping $\mathbf{s}_{(k)}$ with a warp $\mathcal{W}(\cdot;\mathbf{p})$, $\mathcal{Q}(\mathbf{s}_{(k)};\mathbf{q})$ has a similar meaning and a composition relation exists: $\mathcal{Q} \circ \mathcal{W}(\mathbf{s}_{(k)};\mathbf{p},\mathbf{q})$. The advantage of the inverse compositional method is that in the fitting process, many variants such as the Jacobian matrix and the Hessian matrix can be precomputed.

The inverse compositional method has had many variants since its birth. It has been applied to solve the robust and efficient FFPD objective (Tzimiropoulos et al, 2011) which aims to detect points under occlusion and illumination changes. Gross et al (2005) proposed a simultaneous inverse compositional algorithm which simultaneously updates the warp parameters $\mathbf{p}$ and the texture parameters β. Moreover, they also claimed that: (1) the person specific AAM is much easier to build and fit than the generic AAM (person-independent AAM)

and can also achieve better performance; (2) the generic shape model is far easier to build than the generic texture model; (3) the origin of the idea that fitting the generic AAM is far harder than the person specific AAM lies in fact that the effective dimensionality of the generic shape model is far higher than that of the person-specific shape models. Papandreou and Maragos (2008) presented two improvements to the inverse compositional AAM fitting method to overcome significant appearance variation: fitting algorithm adaptation through fitting matrix adjustment and AAM mean template update by incorporating the prior information to constrain the fitting process. Saragih et al (2008) applied a mixed inverse-compositional-forward-additive parameter update scheme to optimize the objective subject to soft correspondence constraints between the image and the model. Amberg et al (2009) claimed that the inverse compositional method has a small convergence radius and proposed two improvements to enlarge the radius at the expense of it being four times slower and preserving the same time consumption respectively. Lucey et al. (Ashraf et al, 2010; Lucey et al, 2013) extended the inverse compositional method in the Fourier domain for image alignment and applied this method specifically to the case of AAM fitting (Navarathna et al, 2011). Tzimiropoulos et al (2012) proposed a generative model called active orientation model which is as computationally efficient as the standard project-out inverse compositional algorithm. Subsequently, Tzimiropoulos and Pantic (2013) proposed a framework for efficiently solving AAM fitting problem in both forward and inverse coordinate frames. Benefiting from the efficiency of proposed framework, they trained and fitted AAM in-the-wild and the trained model could achieve promising performance.

Donner et al (2006) claimed that the multivariate regression technology explored in conventional AAM neglects the correlations between the response variables. This results in slow convergence (more iterations) in the fitting procedure. Since canonical correlation analysis models the correlations between response variables, it is employed to calculate a more accurate gradient matrix.

Tresadern et al (2010) utilized an additive update model (boosting model, both linear and nonlinear) to substitute for the original linear predictors in AAM taking Haar-like features as the regression input. They found that the linear additive model is faster than original linear regression (Cootes et al, 1998a) but it preserves comparable accuracy and is also as effective as nonlinear models when close to the true solution. Therefore, they suggested a hybrid AAM which utilizes a nonlinear additive update model at the first several itera-

tions and then a linear additive update model in the last several iterations.

Although the linear regression strategy achieves some success in obtaining the updated parameters, it is a coarse approximation of the nonlinear relation between texture residuals and warp parameters. When the parameters are initialized far away from the right place, this linear assumption is invalid. To this end, Saragih and Goecke (2007) deployed a nonlinear boosting procedure to learn the multivariate regression. Each parameter is updated by a strong regressor consisting of an ensemble of weak learners (Friedman, 2001). Each weak learner is fed with Haar-like features to output the parameter. This nonlinear modeling results in a more accurate fitting than linear procedures.

Liu (2007, 2009) explored GentleBoost classifier (Friedman et al, 2000) to model the nonlinear relationship between texture and parameter updates. A strong classifier consists of an ensemble of weak classifiers (arctangent functions). Haar-like rectangular features are fed into each weak classifier. The goal of the fitting procedure is to find the PDM parameter updates which maximize the score of the strong classifier. Zhang et al (2009) utilized granular features to replace the rectangular Haar-like feature to improve computational efficiency, discriminability and a larger search space. In addition, they explored the evolutionary search process to overcome the deficiency searching problem in the large feature space. Because the weak classifier in Liu (2007, 2009) is actually utilized to classify the right PDM parameters from the wrong ones, it cannot guarantee that the fitting objective will converge to the optimum solution. Consequently, instead of discriminatively classifying corrected alignment from incorrect alignment, Wu et al (2008) learned classifiers (GentleBoost) to determine whether to switch from one shape parameters to another parameter corresponding to an improved alignment. Based on the ranking appearance model (Wu et al, 2008), Gao et al (2012) preferred to use gradient boosted regression trees (Friedman, 2001) instead of GentleBoost classifiers. Modified census transform features and pseudo census transform features (Gao et al, 2011) are fed to the regression trees.

Sauer et al (2011) compared the performance of linear predictor, boosting-based predictor and random regression-based predictor. Their experimental results illustrate the random regression-based method achieves the best generalization ability. Furthermore, it can achieve performance that is as efficient as boosting procedures without significant reduction in accuracy.

3.2.2 Discrimination

Regarding discrimination, here we mainly refer to the ability to accurately fit a model to an image (Gao et al, 2010). Many aspects may affect this, such as prior knowledge, texture representation and nonlinear modeling the relation between texture residuals and parameters.

Instead of simply minimizing a sum of square measure, Cootes and Taylor (2001) reformulated the AAM problem in a MAP form $p(\mathbf{p}|\mathbf{I}) = p(\mathbf{I}|\mathbf{p})p(\mathbf{p})$ is a zero-mean Gaussian with covraiance matrix $\mathbf{S}_p^{-1}$, the MAP problem can be simplified in a log-probability form to minimize the following problem:

$$E(\mathbf{p}) = \sigma_r^2 \mathbf{r}^T \mathbf{r} + \mathbf{p}^T \mathbf{S}_p^{-1} \mathbf{p}. \tag{25}$$

Following a similar procedure to that described in Section3.1, the above problem can be resolved. To deploy the prior knowledge such as the position of certain points, a further prior can be added to equation (25) and the optimization is still a similar procedure.

Traditional AAM (Cootes et al, 2001) fixed the gradient matrix (Jacobian of the residual to the parameters) which may lead to poor performance when the texture of a testing image differs dramatically from the mean texture. Batur and Hayes (2003, 2005) update the gradient matrix in each iteration by adding a linear combination of basis matrixes to a fixed basic matrix. Cootes and Taylor (2006) presented a strategy-like quasi-Newton method to update the gradient matrix. The update of the gradient matrix will increase the fitting accuracy to some extent but will also reduce efficiency. Saragih and Gocke (2009) pointed out that a fixed linear update model, as in traditional AAM, has limited ability to account for the various error terrains about the optimum in different images and an adaptive update model according to the image at hand requires a time consuming process of re-calculating it for every iteration. They adopted a compromise manner: learning a set of fixed linear update models to be applied to the image sequentially in the fitting process.

Zheng et al (2006) proposed a rank-based non-rigid shape detection method through RankBoost (Freund et al, 2003). RankBoost is utilized to learn a ranking model from Haar-like features extracted from warped training images. The ranking model is then applied to Haar-like features extracted from images warped from the testing image to calculate the response scores. The final shape is achieved from a linear combination of the K training shapes corresponding to previously calculated top K response scores. One disadvantage of this method is that the detection efficiency is seriously affected by the number of images in the training set.

Standard AAM achieves limited accuracy in fitting a face image for an individual unseen in the training set. This is mainly because the appearance model of AAM (created in a generative manner) has limited generalization ability. Peyras et al (2007) proposed a multi-level segmented method which constructs multiple AAMs, each corresponding to different parts of a face, e.g. eye, mouth, and nose. The whole fitting strategy is in a coarse-to-fine fashion (multi-resolution) and a different number of AAMs are correspondingly constructed.

Nguyen et al. (Nguyen and De la Torre, 2008; Nguyen and Torre, 2008, 2010) claimed that AAMs are easily converged to local minima in the fitting process and that the local minima seldom correspond to acceptable solutions. They proposed a parameterized appearance model which learns a cost function having local minima at and only at desired places. This is guaranteed by a quadratic error function with a symmetric positive semidefinite coefficient matrix corresponding to the quadratic term. The objective function is optimized by a variant of the Newton iteration method.

3.2.3 Robustness

The robustness of AAM is generally influenced by the inconstant circumstances, e.g. pose variations, resolutions, illumination changes, occlusion, and any other wild conditions.

Pose: Cootes et al (2002) demonstrated that five AAMs corresponding to five views ($-90°$, $-45°$, $0°$, $45°$, $90°$) can capture the appearance variation across a wide range of rotations. Each AAM can capture faces rotated in a range of view angles. The rotation angle θ is dependent on the appearance parameters $\mathbf{c}$ through a model:

$$\mathbf{c} = \mathbf{c}_0 + \cos\theta \mathbf{c}_x + \sin\theta \mathbf{c}_y. \tag{26}$$

Huang et al (2012) combined view-based AAM (Cootes et al, 2002) with Kalman filter to perform pose robust face tracking. Instead of model parameters controlling the shape and appearance variations, this paper only utilized the shape parameters to construct the view space. Gonzalez-Mora et al (2007) decoupled variations on both the shape and texture into pose and expression/identity parts. The shape and texture are modeled by a bilinear model respectively.

An alternative way to model the pose rotation is by exploring 3D information (Xiao et al, 2004), and studies show that 3D linear face models generally have better qualifications than 2D linear models in three aspects: representational power, construction and real-time fitting (Matthews et al, 2007). Sung et al (2008) deployed

cylinder head models (La Cascia et al, 2000) to predict the global head pose parameters, which are fed to the subsequent AAM procedures. Their experimental results illustrate that face training by the combined method is more pose robust than that of AAM, having a 170% higher tracking rate and 115% wider pose coverage. Asthana et al (2009) applied the 3D facial pose estimator (Grujic et al, 2008) to obtain the pose range given a testing image. Facial feature points are then detected by the AAM trained from images with the corresponding view. They also exploited regression relationship from annotated frontal facial images to non-frontal facial images to handle pose and expression variation (Asthana et al, 2011). Grujic et al (2008) presented a 3D AAM that does not require any pre-alignment of shapes, thanks to the inherent properties of complex spherical distributions: invariance to scale, translation and rotation. Martins et al (2010, 2013) combined 3D PDM and a 2D appearance model through a full perspective projection. The fitting objective can be optimized by two methods based on the Lucas-Kanade framework (Baker and Matthews, 2004): the simultaneous forwards additive algorithm and the normalization forwards additive algorithm. Hansen et al (2011) presented a nonlinear shape model-based on Riemannian elasticity framework instead of linear subspace model (PDM) in a conventional AAM framework to handle the poor pose initialization problem. However, due to the complexity of the nonlinear shape formulation, the efficiency is reduced. Fanelli et al (2013) proposed a 3D AAM-based on intensity and depth images. Random forest (Breiman, 2001) is explored to model the relationship between textures (from both intensity and depth images) and model parameters.

Resolutions: Dedeoglu et al (2006, 2007) observed that classic AAM performed poorly in the case of low-resolution images. This is due to the image formation model of a typical charge-coupled device (CCD) camera. Consequently, they proposed a resolution-aware algorithm to adapt to low-resolution images which substitutes the classic fitting criterion of L2 norm error with a new formulation, taking the image formation model into account. Liu et al (2006) trained several AAMs, each of which corresponds to a special resolution, to model compactness at lower resolution.

Illumination: Classic AAM models the texture variation with the Gaussian. Sometimes this assumption may result in errors when the illumination changes considerably. Kahraman et al (2007) decomposed the original texture space into two orthogonal subspaces: identity and illumination subspace. Any texture can then be described by two projection vectors $\boldsymbol{\beta}_{id}$ and $\boldsymbol{\beta}_{illu}$:

$$\mathbf{a} = \mathbf{a}_0 + \mathbf{P}_{id}\boldsymbol{\beta}_{id} + \mathbf{P}_{illu}\boldsymbol{\beta}_{illu}. \tag{27}$$

where $\mathbf{P}_{id}$ and $\mathbf{P}_{illu}$ consist of basis vectors which span the identity and illumination subspace respectively. Kozakaya et al (2010) explored multilinear analysis (tensor) to model the variations across face identity, pose, expression, and illumination. The tensor consists of an image tensor and a model tensor. The image tensor is utilized to estimate image variations which can be solved in a discretion or continuous manner. The model tensor is applied to construct a variation-specific AAM from a tensor representation.

Outliers: Roberts et al (2007) observed that AAMs are not robust to a large set of gross outliers; they explored the Geman-McClure kernel (M-estimator) with two sets of learned scaling parameters to alleviate this problem.

Occlusions: AAM learns the texture model from a holistic view and faces challenges in achieving good performance, such as sensible to partial occlusion, while ASM opts for local texture descriptors. Sung et al (2007) therefore combined ASM and AAM to give a united objective function:

$$E = (1 - \omega)(E_{aam} + E_{reg}) + \omega E_{asm}, \tag{28}$$

where E_{aam} and E_{asm} are the residual errors of the AAM and ASM appearance model respectively, E_{reg} is the regularization error term constrained on shape parameters, and ω is a trade-off parameter to balance the ASM residual error with other errors. Martins et al (2013) proposed two robust fitting methods based on the Lucas-Kanade forwards additive method (Baker et al, 2003) to handle partial and self-occlusions.

4 Regression-Based Methods

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

$$\mathcal{M} : \mathcal{F}(\mathbf{I}) \rightarrow \mathbf{x} \in \mathbb{R}^{2N}, \tag{29}$$

where $\mathcal{M}$ denotes the mapping from image appearance (feature) $\mathcal{F}(\mathbf{I})$ to the shape $\mathbf{x}$ and $\mathcal{F}$ is the feature extractor. Haar-like features (Viola and Jones, 2004),

SIFT (Lowe, 2004), local binary patterns (LBP) (Ojala et al, 1996) and other gradient-based features are generally used feature types.

Zhou and Comaniciu (2007) proposed a shape regression method based on boosting (Freund and Schapire, 1997; Friedman et al, 2000). Their method proceeds in two stages: first, the rigid parameters are found by casting the problem as an object detection problem which is solved by a boosting-based regression method; secondly, a regularized regression function is learned from perturbed training examples to predict the non-rigid shape. Haar-like features are fed to the non-rigid shape regressors.

Kozakaya et al. (Kozakaya et al, 2008a,b, 2010) proposed a weighted vector concentration approach to localize facial features without any specific prior assumption on facial shape or facial feature point configuration. In the training phase, grid sampling points are evenly placed on each face image and an extended feature vector is extracted for each sampling point of each training image. The extended feature vector is composed of histograms of oriented gradients (HOG descriptor (Dalal and Triggs, 2005)), N directional vectors from the sampling point to all N feature points, and local likelihood patterns at the feature points. In the detection phase, given an input face image, local descriptors corresponding to each sampling point are extracted. Then a nearest local pattern descriptor can be found for each sampling point of the input image among the descriptors located at the same position of training images using the approximate nearest neighbor search (ANNS) algorithm (Arya et al, 1998). Simultaneously, a group of directional vectors and local likelihood patterns can also be obtained. Finally, feature points are computed from a weighted square distance from the point to the line through sampling points and the directional vector. Each facial feature point can be detected independently after all nearest neighbors are found by the ANNS method. This paper does not take faces with different expressions into consideration in their experiments.

In consideration of the nonlinear property of the facial feature localization problem and generalization ability, Valstar et al (2010) deployed support vector regression to output the target point location from the input local appearance-based features (Haar-like features). To overcome the overfitting problem due to the high dimensionality of the Haar-like features, Adaboost regression is utilized to perform feature selection.

Kazemi and Cullivan (2011) divided a face into four parts: eyes (left and right), nose and mouth. Several regression strategies such as ridge regression, ordinary least squares regression, principal component regression
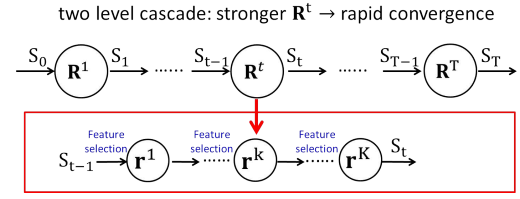

Fig. 6 Two-level cascaded regressor of (Cao et al, 2012).

were then explored to regress the local appearance (a variant of HOG descriptor) of each part to the target landmark location. Their experimental results illustrate that these several regression methods, ridge regression achieves the best performance. Moreover, their method has comparable performance as to AAM methods but is more robust.

Cao et al (2012) proposed a two-level cascaded learning framework (see Fig. 6) based on boosted regression (Duffy, 2002). Unlike the above method which learns the regression map of each landmark of those landmarks that correspond to the same component, this method directly learns a vectorial output for all landmarks. Shape-indexed features such as that in (Dollar et al, 2010) are extracted from the whole image and are fed into the regressor. To reduce the complexity of feature selection but still achieve reasonable performance, the authors further proposed a correlation-based feature selection strategy. Each regressor ($\mathbf{R}^t$ in Fig. 6, $t = 1, \cdots, T$) in the first level consists of cascaded random fern regressors ($\mathbf{r}^k$ in Fig. 6, $k = 1, \cdots, K$) (Ozuysal et al, 2010) in the second level. This method achieves state-of-the-art performance in a very efficient manner. In particular, it achieves the highest accuracy on the LFPW database: labeled face parts in the wild database (Belhumeur et al, 2011), images of which are taken under uncontrolled conditions.

Considering the method (Cao et al, 2012) is not robust to occlusions and large shape variations, Burgos-Artizzu et al (2013) improved it from three aspects. First, Cao et al (2012) references pixel by its local coordinates with respect to its closest landmark, which is not enough against large pose variations and shape deformation. Burgos-Artizzu et al (2013) proposed to reference pixels by linear interpolation between two landmarks. Secondly, Burgos-Artizzu et al. presented a strategy to incorporate the occlusion information into the regression which improves the robustness to occlusion. Thirdly, they designed a smart initialization restart scheme to deploying the similarity between different predictions resulted from different initializations. Experimental results on several existing databases in the wild and their newly constructed database illustrate the proposed method achieves state-of-the-art performance.

In view of the boosted regression in Cao et al (2012), which is a greedy method to approximate the function mapping from facial image appearance features to facial shape updates, Xiong and De la Torre (2013) developed the supervised descent method (SDM) to solve a series of linear least squares problems as follows:

$$\mathrm{argmin}_{\mathbf{R}_k, \mathbf{b}_k} \sum_{\mathbf{d}^i} \sum_{\mathbf{x}_k^i} \left\| \triangle \mathbf{x}_*^i - \mathbf{R}_k \boldsymbol{\phi}_k^i - \mathbf{b}_k \right\|^2, \qquad (30)$$

where $\triangle \mathbf{x}_*^i = \mathbf{x}_*^i - \mathbf{x}_k^i$ is the ground truth difference between the truth shape $\mathbf{x}_*^i$ of the i-th training image $\mathbf{d}^i$ and the shape $\mathbf{x}_k^i$ obtained from the k-th iteration, $\boldsymbol{\phi}_k^i$ is the extracted SIFT features around the shape $\mathbf{x}_k^i$ on the training image, $\mathbf{R}_k$ is called the common descent direction in this paper and $\mathbf{b}_k$ is a biased term. This method has a natural derivation process based on the Newton method. A series of $\{\mathbf{R}_k, \mathbf{b}_k\}$ are learned in the training stage, and in the testing stage they are applied to the SIFT features extracted from the testing image to update the shape sequentially. SDM efficiently achieves comparable performance to (Cao et al, 2012) on database LFPW (Belhumeur et al, 2011).

Martinez et al (2013) believed that each image patch evaluated by the regressors adds evidence to the target location rather than just taking the last estimate (the last iteration) into account and discarding the rest of these estimates. They aggregated all up-to-date local evidence obtained from support vector regression by an unnormalized mixture of Gaussian distributions. LBP is deployed as the local texture descriptor and a correlation-based feature selection method is introduced to reduce the dimensionality of LBP features.

Dantone et al (2012) proposed a facial feature point detection by extending the concept of regression forests (Breiman, 2001; Criminisi et al, 2012) to conditional regression forests. They claimed that it is difficult for general regression forests to learn the variations of faces with different head poses. The head pose is evaluated by regression forests. A regression forest is constructed, conditioned on the head pose (i.e. there is one regression forest corresponding to each head pose). In the testing phase, the probabilities of the head pose of an input testing image should be first calculated and, according to this distribution, the number of trees selected from each forest can be determined. Finally the position of each facial feature point can be computed through solving a mean-shift problem. Yang and Patras added structural information into the random regression and proposed a structured-output regression forest-based face parts localization method (Yang and Patras, 2012). Then, they (Yang and Patras, 2013) pro-

posed to deploy a cascade of sieves to refine the voting map obtained from random regression forest.

Rivera and Martinez (2012) casted the facial feature point detection problem as a regression problem. The input of a regressor consists of features extracted from input images, either pixel intensities or C1 features (Serre et al, 2007). The output of a regressor is PDM coefficients (shape parameters). The regressor is either a kernel ridge regression or ϵ-support vector regression. Their experimental results show that kernel ridge regression with pixel intensities achieves the best performance when images have a low resolution.

Considering the fact existing parameterized appearance models do not sample parameter space uniformly, which may result in a biased model, Sanchez-Lozano et al (2012) proposed a continuous regression method to solve this biased learning problem. Instead of discretely sampling the parameter space, this method directly integrates on the parameter space. A closed-form solution can be achieved. To alleviate the small sample size problem, the closed-form solution is further projected onto the principal components.

5 Other Methods

In addition to the aforementioned three categories of methods, there are also some methods that do not belong to any of them. Some methods deploy graphical model to describe the relation between facial feature points, which are assigned to the sub-category of graphical model-based methods in the following text. Some methods align a set of facial images simultaneously, which is known as joint face alignment. Other methods may detect facial features points independently from the image texture and ignore the correlation between points, and we call this sub-category of methods independent detectors.

5.1 Graphical Model-based Methods

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

Coughlan and Ferreira (2002) developed a generative Bayesian graphical model that deployed separate models to describe shape variability (shape prior) and appearance variations (appearance likelihood) to find

deformable shapes. The shape prior takes the location of each facial feature point and these points' normal orientation as a node in MRF. An edge map and an orientation map are calculated to model the appearance likelihood. A variant of the belief propagation method is utilized to optimize the problem. MRF has been also explored to constrain the relative position of all facial feature points obtained from the regression procedure in Valstar et al (2010); Martinez et al (2013). Gu et al (2007) learned a sparse Gaussian MRF structure to regularize the spatial configuration of face parts by lasso regression.

Unlike the method in Coughlan and Ferreira (2002) which models the shape prior only in a local neighborhood, Liang et al (2006a) proposed a method that incorporates a global shape prior directly into the Markov network. The local shape prior is enforced by denoting a line segment as a node of the constructed Markov network. Here, line segments draw from one facial point to another neighboring point. Subsequently, Liang et al (2006b) claimed that although CLM-based methods take the global shape prior into account, these methods neglect the neighboring constraint between points since they compute the response map of each point independently. Based on the thought in Liang et al (2006a), Liang et al. further incorporated the PDM shape prior into their model.

Another work that considers both the local characteristics and global characteristics of facial shapes is the bi-stage component-based facial feature point detection method (Huang et al, 2007b). The whole face shape is divided into seven parts. The shape of each part is modeled as a Markov network by taking each point as a node. Belief propagation is explored to find the locations of these components. Then, configurations of these components are constrained by the global shape prior described by the Gaussian process latent variable model.

Zhu and Ramanan (2012) proposed a unified model for face detection, head pose estimation and landmark estimation. Their method is based on a mixture of trees, each of which corresponds to one head pose view. These different trees share a pool of parts. In the training stage, the tree structure is first estimated via Chow-Liu algorithm (Chow and Liu, 1968). Then a model of a tree-structured pictorial structure (Felzenszwalb and Huttenlocher, 2005) is constructed for each view. In the testing stage, the input image is scored by all tree structures respectively and the pixel locations corresponding to the tree with maximum score are the final landmark locations. Michal et al (2012) also modeled the relative position of facial feature points as a tree-structure. Since tree-structure-based methods only consider the

local neighboring relation and neglect the global shape configuration, they may easily lead to unreasonable facial shape.

5.2 Joint Face Alignment Methods

Joint face alignment jointly aligns a batch of images undergoing a variety of geometric and appearance variations (Zhao et al, 2011), motivated by the congealing-style joint alignment method (Learned-Miller, 2006) and sparse and low-rank decomposition method (Peng et al, 2012). Zhao et al (2011) designed a joint AAM by assuming that the images of the same face should lie in the same linear subspace and the person-specific space should be proximate the generic appearance space. The problem is formulated as a nonlinear problem constrained by a rank term which can be transformed to a nuclear norm. An augmented Lagrangian method is explored to optimize the nonlinear problem.

Smith and Zhang (2012) stated that the method (Zhao et al, 2011) breaks down under several common conditions, such as significant occlusion or shadow, image degradation, and outliers. Considering the fact that a non-parametric set of global shape models (Belhumeur et al, 2011) results in excellent facial feature point localization accuracy on facial images undergoing significant occlusions, shadows, and pose and expression variation, they introduced the same shape model combined with a local appearance model into the joint alignment framework.

Different from the aforementioned two joint face alignment methods, which both incorporate the rank term into the objective, Zhao et al (2012) proposed a novel two-stage approach to align a set of images of the same person. The initial facial feature point estimation is first computed by an off-the-shelf approach (Gu and Kanade, 2008). To distinguish the "good" alignments from the "bad" ones among all these initial estimations, a discriminative face alignment evaluation metric is designed by virtue of cascaded AdaBoost framework (Viola and Jones, 2004) and Real AdaBoost (Friedman et al, 2000). Selected "good" alignments are utilized to improve the accuracy of "bad" ones through appearance consistency between the "bad" estimate and its selected K neighboring "good" estimates.

Tong et al (2009, 2012) proposed a semi-supervised facial landmark localization approach which utilizes a small number of manually labeled images. Their objective function is to minimize the sum of squared error of two distances: the distance between the labeled and unlabeled images, and the distance between the unlabeled images. To obtain a reasonable shape, an on-line learned PDM shape model is imposed as a constraint.

To further improve the preciseness of the above model, they perform the above procedures in a coarse-to-fine manner, which proceeds by dividing the whole face into patches with different sizes at different levels.

5.3 Independent Facial Feature Point Detectors

The aforementioned methods predict the locations of all facial feature points or a group of points simultaneously. There are other methods which detect each point independently. Here, methods which do not rely on manually labeled images, such as approach (Asteriadis et al, 2009), are not included.

Vukadinovic and Pantic (2005) detected each point by a local expert as utilized in CLM-based methods. Here, Gabor feature-based boosted classifier is utilized to classify the positive image patch from the negative image patch. The position with the peak response among the response map of each point is the sought location. Shen et al (2013) proposed the detection of each facial feature point through a voting strategy on corresponding points on some exemplar images retrieved from the training dataset. The location corresponding to the peak response in each voting map is the estimated position.

Considering the fact that there is great variability among faces and facial features, such as eye centers and eye corners, Ding and Martinez (2008, 2010) employed subclass discriminant analysis (Zhu and Martinez, 2006) to divide vectors (features or context) of the same class into subclasses. Vectors centered on the facial feature point are called features and vectors centered on points surrounding the facial feature point are called context. The K-means clustering method is explored to divide each class into a number of subclasses. Given the detected face box, facial feature points can be exhaustively searched in some windows located relative to the bounding box by comparison with the learned subclasses at different scales. The final facial feature point is achieved by a voting strategy on different detected positions at different scales.

The advantage of independent facial feature point detectors is the initialization free character. One major disadvantage is the ambiguity problem. This means there exist more than one positions looking like the target landmark, especially under complex environment like deliberately disguise, occlusion or pose variation. To address this problem, Zhao et al (2013) proposed to jointly estimate correct positions of all landmarks from some candidates obtained by independent facial feature point detectors.

5.4 Deep Learning-Based Methods

Luo et al (2012) proposed a hierarchical face parsing method based on deep learning (Hinton et al, 2006; Hinton and Salakhutdinov, 2006). They recast the facial feature point localization problem as the process of finding the label maps (segmentation) which clearly indicate the pixels belong to a certain component. The feature can then be easily obtained from the boundary of the label maps. The proposed hierarchical framework consists of four layers: face detector (the first layer), facial parts detectors (the second layer), facial component detectors (the third layer), and facial component segmentation. The structure of this model is somewhat like a pictorial structure (Felzenszwalb and Huttenlocher, 2005): the face detector can be seen as the root node and other detectors (part detectors and component detectors) as the child nodes. The objective function can be formulated in a Bayesian (maximum a posterior) form. The prior term denotes the spatial consistency between detectors of different layers and is modeled as the Gaussian distribution. The likelihood term represents the detectors and segmentation. All detectors can be learned by restricted Boltzmann machine (Hinton et al, 2006) and segmentation can be learned by a deep autoencoder-like (Hinton and Salakhutdinov, 2006) method. Inspired by Luo et al. (Luo et al, 2012), Smith et al (2013) deployed exemplar-based strategy as in Belhumeur et al (2011) to parse a face image.

Sun et al (2013) proposed a three-level cascaded deep convolutional network framework for point detection in a coarse-to-fine manner. Each level is composed of several numbers of convolutional networks. The first level gives an initial estimate to the point position and the following two levels then refine the obtained initial estimate to a more accurate one. Though great accuracy can be achieved, this method needs to model each point by a convolutional network which improves the complexity of the whole model. Moreover, with the increase in the number of facial feature points, the time consumption to detect all points is high.

Wu et al (2013) explored deep belief networks to capture face shape variation due to facial expression variations and utilized a 3-way restricted Boltzmann machine to capture the relationship between frontal face shapes and non-frontal face shapes. They applied the proposed model to facial feature tracking.

6 Evaluations

6.1 Databases

There are many face databases publically available due to the easy acquisition of images and the fast development of social networks such as Facebook, Flickr, and Google+. The ground truth facial feature points are usually labeled manually by employing workers or through crowdsourcing, e.g. the Amazon mechanical turk (MTurk). Each face image is generally labeled by several workers and the average of these labeled results is taken as the final ground truth. These face databases can be classified into two categories: databases captured in controlled conditions and databases captured in uncontrolled conditions (i.e. in the wild). Controlled databases are taken under the framework of predefined experimental settings such as the variation of illumination, occlusions, head pose and facial expressions. Databases in the wild are generally collected from websites such as Facebook and Flickr. Table 3 describes representation collections which are popularly used in empirical studies.

6.2 Comparisons and Discussions

The distance from the estimated points to the ground truth normalized by the inter-ocular distance and the number of points is a common informative metric for evaluating a facial feature point detection system (named mean normalized error, MNE, in the following text). Sometimes the figure of the proportion of testing images with the increase of MNE is plotted as a comparison metric among different approaches. The performance of facial feature point detection methods cannot be verified by experimenting on each database listed as Table 3 shown since there are too many databases. Table 4 shows the published performance of representative methods of aforementioned categories on several different databases.

To further illustrate the characteristics of various categories of methods, we have collected some software published online and listed them as shown in Table 5. Eight representative methods were chosen for study: DRMF-CLM (Asthana et al, 2013), OPM-CLM (Yu et al, 2013), FF-AAM (Tzimiropoulos and Pantic, 2013), CNN-DL (Sun et al, 2013), the graphical model (GM) method (Zhu and Ramanan, 2012), BorMan-Regression (Valstar et al, 2010), SDM-Regression (Xiong and De la Torre, 2013), and RCPR-Regression (Burgos-Artizzu et al, 2013). We localized FFPs in three databases, COFW (Burgos-Artizzu et al, 2013), LFPW (Belhumeur

et al, 2011), and Helen (Le et al, 2012), using the published software. The 68 re-annotated landmarks of the "300 Faces in-the-Wild Challenge" were used as the ground truth for images in LFPW and Helen. For COFW, we used the augmented version presented in (Burgos-Artizzu et al, 2013), which contains 1345 training images and 507 test images. Examples from these three databases are shown in Fig. 7.

Since only the trained models, and not the source code, were published in some cases, it was difficult to make equitable comparisons (for example, some software contained different face detectors). In addition, different methods labeled different numbers of facial landmarks (see Table 6). In Table 6, "any" denotes that the authors published the training code, and thus the models could be trained for different numbers of facial landmarks. Face detection rates were quantified according to the percentage of detected faces being labeled in the corresponding database. "GM-99" and "GM-1050" indicate graphical models composed of 99 and 1050 parts, respectively. Errors were measured as the percentage of the interocular distance d_{io}, as shown in equation (31), i.e., the mean normalized error (MNE), where $\mathbf{x}_{(i)}^{e}$ is the i-th estimated point and $\mathbf{x}_{(i)}^{g}$ is its corresponding ground truth:

$$e = \frac{\sum_{i=1}^{N} \left\| \mathbf{x}_{(i)}^{e} - \mathbf{x}_{(i)}^{g} \right\|_{2}}{N \times d_{io}} \times 100\%, \tag{31}$$

Fig. 8, Fig. 9, and Fig. 10 show the cumulative error curves for the above three databases. It can be seen that CNN (Sun et al, 2013) achieves promising performance on all three databases. There are two main reasons for this: first, deep learning is highly capable of performing feature learning followed by classification or detection, especially when there are many training samples (CNN utilizes approximately ten thousand training samples); secondly, CNN detects five characteristic points: the center of the two pupils, the nose tip, and the two eye corners, which are relatively easy to detect. The cascaded regression method, SDM (Xiong and De la Torre, 2013), also achieves good performance for detecting 49 facial points distributed around the eyebrows, eyes, nose, and mouth, and without points around the outline of the face. RCPR, another cascaded regression method, also appears promising, although inferior to SDM; this is likely to be because SDM fails to detect several difficult test images and detects 49 points without the facial outline. Table 7 shows a comparison of the normalized error of RCPR retrained on 49 points on the same faces detected by SDM. The model could not be retrained on COFW, since 29 points label the

Table 3 Datasets for Facial Feature Point Detection

Databases Collected under Well-Controlled Conditions:

CMU Multi-PIE2008 - CMU Multi-PIE (Gross et al, 2010) face database was collected in four sessions between October 2004 and March 2005. It aims to support the development of algorithms for recognition of faces across pose, illumination and expression conditions. This database contains 337 subjects and more than 750,000 images for 305 GB of data. A total of six different expressions are recorded: neutral, smile, surprise, squint, disgust and scream. Subjects were recorded across 15 views and under 19 different illumination conditions. A subset of this database has been labeled either 68 points or 39 points depending on their view but landmarks are not published online. Details on obtaining this dataset can be found at: `http://www.multipie.org`.

Extended M2VTS database1999 (XM2VTS) - XM2VTS database (Messer et al, 1999) collected 2,360 color images, sound files and 3D face models of 295 people. The database contains four recordings of these 295 subjects taken over a period of four months. Each recording was captured when the subject was speaking or rotating his/her head. This database is available on request: `www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/`. These 2,360 color images are labeled with 68 landmarks and are published online: `http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/xm2vts/xm2vts_markup.html`.

AR1998 - AR database (Martinez and Benavente, 1998) contains over 4,000 color images corresponding to the faces of 126 people (70 men and 56 women). Images were taken under strictly controlled conditions and with different facial expressions, illumination conditions, and occlusions (sunglasses and scarf). Each person appeared in two sessions, separated by two weeks. Ding and Martinez (Ding and Martinez, 2010) manually annotated 130 landmarks on each face image which have been published online with the database: `www2.ece.ohio-state.edu/~aleix/ARdatabase.html`.

IMM2004 - IMM database (Nordstrom et al, 2004) contains 240 color images of 40 persons (7 females and 33 males). Each image is labeled with 58 landmarks around the eyebrows, eyes, nose, mouth and jaw. Face images and landmarks can be downloaded at: `http://www2.imm.dtu.dk/~aam/datasets/datasets.html`.

MUCT2010 - MUCT2010 database (Miborrow et al, 2010) consists of 3,755 face images of 276 subjects and each image is marked with 76 manual landmarks. Faces in this database are captured under different lighting conditions, at various ages, and are of several different ethnicities. The database is available at: `www.milbo.org/muct/`.

PUT2008 - PUT database (Kasinski et al, 2008) collected 9,971 high resolution images (2048 × 1536) of 100 people taken in partially controlled illumination conditions with rotations along the pitch and yaw angle. Each image is labeled with 30 landmarks. A subset of 2,193 near-frontal images is provided with 194 control points. The database is available at: `https://biometrics.cie.put.poznan.pl/index.php?option=com_content&view=article&id=4&Itemid=2&lang=en`.

Databases in the wild:

BioID2001 - BioID database (Jesorsky et al, 2001) was recorded in an indoor lab environment, but "real world" conditions were used. This database contains 1,521 grey level face images of 23 subjects and each image is labeled with 20 landmarks. This database is available at: `http://www.bioid.com/index.php?q=downloads/software/bioid-face-database.html`.

LFW2007 - LFW database (Huang et al, 2007a) contains 13,233 face images of 5,749 subjects collected from the web. Each face in the database has been labeled with the name of the person pictured. 1,680 of the people pictured have two or more distinct photos in the data set. The constructors of this database did not provide manually labeled landmarks but there are other available sites: (Michal et al, 2012) `http://cmp.felk.cvut.cz/~uricamic/flandmark/`(7landmarks); (Dantone et al, 2012) `http://www.dantone.me/datasets/facial-features-lfw/`(10landmarks).

Annotated Facial Landmarks in the Wild 2011(AFLW) - AFLW database (Kostinger et al, 2011) is a large-scale, multi-view, real-world face database with annotated facial feature points. Images were collected from Flickr using a wide range of face relevant key words such as face, mugshot, and profile face. This database includes 25,993 images in total and each image is labeled with 21 landmarks. It is available at: `http://lrs.icg.tugraz.at/research/aflw/`.

Labeled Face Parts in the Wild 2011 (LFPW) - LFPW database (Belhumeur et al, 2011) is composed of 1,400 face images (1,100 as the training set and the other 300 images are taken as the testing set) downloaded from the web using simple text queries on websites such as Google.com, Flickr.com, and Yahoo.com. Due to copyright issues, the authors did not distribute image files but provided a list of image URLs. However, some image links are no longer available. 35 landmarks are labeled in total;29 of them are usually utilized in literatures. More information can be found at: `http://homes.cs.washington.edu/~neeraj/databases/lfpw/`.

Annotated Faces in the Wild 2012 (AFW) - AFW database (Zhu and Ramanan, 2012) contains 205 images with a highly cluttered background and large variations both in face scale and pose. Each image is labeled with 6 landmarks and the bounding box of the corresponding face. The dataset is available at: `http://www.ics.uci.edu/~xzhu/face/`.

Helen2012 - Helen database (Le et al, 2012) contains 2,300 high resolution face images collected from Flickr.com. Each face image is labeled with 194 landmarks. More information about this database can be found at: `http://www.ifp.illinois.edu/~vuongle2/helen/`.

300 Faces in-the-Wild Challenge (300-W) 2013 - 300-W database is a mixed database consisting of face images from several published databases (LFPW, Helen, AFW, and XM2VTS) and a new collected database IBUG. All these images are re-annotated with 68 landmarks. This database is published for the first Automatic Facial Landmark Detection in-the-Wild Challenge (300-W 2013) held in conjunction with the International Conference on Computer Vision 2013. This database is available at: `http://ibug.doc.ic.ac.uk/resources/300-W/`.

Caltech Occluded Faces in the Wild (COFW) 2013 - COFW database (Burgos-Artizzu et al, 2013) is composed of 1,007 face images showing large variations in shape and occlusions due to differences in pose, expression, use of accessories such as sunglasses and hats and interactions with objects (e.g. food, hands, microphones, *etc.*). 29 points are marked for each image. The major difference between this database and other ones is that each landmark is explicitly labeled whether it is occluded. This database presents a great challenging task for facial feature point detection due to the large amount and variety of occlusions and large shape variations. This database is available at: `http://www.vision.caltech.edu/xpburgos/ICCV13/`.

Table 4 Mean Normalized Error of Representative Methods on Various Databases

method	Database (#Training Images+#Testing Images)	MNE($\times 10^{(-2)}$)	#Landmarks
Belhumeur (Belhumeur et al, 2011)	LFPW (1100 + 300)	3.99	29
Cao (Cao et al, 2012)	LFPW(2000 + 500)	3.43	29
Xiong (Xiong and De la Torre, 2013)	LFPW (884 + 245)	3.47	29
Burgos-Artizzu (Burgos-Artizzu et al, 2013)	LFPW(845 + 194)	3.50	29
Xiong (Xiong and De la Torre, 2013)	LFW-A&C (Saragih, 2011) (604 + 512)	2.7	66
Xiong (Xiong and De la Torre, 2013)	Multi-PIE, LFW-A&C (training) + RU-FACS (Matthews and Baker, 2004) (test)	5.03	49
Sukno (Sukno et al, 2007)	XM2VTS (800 + 400)	2.03	64
Sukno (Sukno et al, 2007)	AR (180 + 90)	1.63	98
Le (Le et al, 2012)	MUCT+BioID 3755 + 1521	4.5	17
Le (Le et al, 2012)	Helen(2000 + 330)	9.1	194
Dantone (Dantone et al, 2012)	LFW (13000 + 1000)	6.985	10
Valstar (Valstar et al, 2010)	FERET(Phillips et al, 2000)+MMI(Valstar and Pantic, 2010) (360 + 40)	5.11	22
Martinez (Martinez et al, 2013)	MMI(Valstar and Pantic, 2010)+FERET(Phillips et al, 2000)+XM2VTS+BioID 946 + 1000	3.575	20
Ding (Ding and Martinez, 2008)	Americal Sign Lnguage Sentences (766 frames)	6.23	98
Ding (Ding and Martinez, 2010)	Collected training database+AR+XM2VTS(51664 + 1200)	8.4	98
Wu (Wu et al, 2013)	MMI(Valstar and Pantic, 2010)(196)	5.5275	26
Michal (Michal et al, 2012)	LFW (6919 + 2316)	5.4606	8

Table 5 Published Software Collection

Method	Website
Vukadinovic (Vukadinovic and Pantic, 2005)	http://ibug.doc.ic.ac.uk/resources/fiducial-facial-point-detector-20052007/
Milborrow (Miborrow and F., 2008)	http://www.milbo.users.sonic.net/stasm/
Inverse compositional AAM	http://sourceforge.net/projects/icaam/files/
Valstar (Valstar et al, 2010)	http://ibug.doc.ic.ac.uk/resources/facial-point-detector-2010/
Hansen (Hansen et al, 2011)	https://svn.imm.dtu.dk/AAMLab/svn/AAMLab/trunk/(username:guest,password:aamlab)
Saragih (Saragih et al, 2011)	https://github.com/kylemcdonald/FaceTracker
Tzimiropoulos (Tzimiropoulos et al, 2012)	http://ibug.doc.ic.ac.uk/resources/aoms-generic-face-alignment/
Rivera (Rivera and Martinez, 2012)	http://cbcsl.ece.ohio-state.edu/downloads.html
Zhu (Zhu and Ramanan, 2012)	http://www.ics.uci.edu/~xzhu/face/
Dantone (Dantone et al, 2012)	http://www.dantone.me/projects-2/facial-feature-detection/
Michal (Michal et al, 2012)	http://cmp.felk.cvut.cz/~uricamic/flandmark/
Sun (Sun et al, 2013)	http://mmlab.ie.cuhk.edu.hk/archive/CNN_FacePoint.htm
Asthana (Asthana et al, 2013)	https://sites.google.com/site/akshayasthana/clm-wild-code?
Xiong (Xiong and De la Torre, 2013)	www.humansensing.cs.cmu.edu/intraface
Martinez (Martinez et al, 2013)	http://ibug.doc.ic.ac.uk/resources/facial-point-detector-2010/
Yu (Yu et al, 2013)	http://www.research.rutgers.edu/?xiangyu/face_align.html
Tzimiropoulos (Tzimiropoulos and Pantic, 2013)	http://ibug.doc.ic.ac.uk/resources
Burgos-Artizzu (Burgos-Artizzu et al, 2013)	http://www.vision.caltech.edu/xpburgos/ICCV13/

Table 6 Mean Normalized Error of Eight Representative Methods on Three Databases: COFW, LFPW, and Helen

|		DRMF (Asthana et al, 2013)	OPM (Yu et al, 2013)	FF (Tzimiropoulos and Pantic, 2013)	CNN (Sun et al, 2013)	GM-99 (Zhu and Ramanan, 2012)	GM-1050 (Zhu and Ramanan, 2012)	Borman (Valstar et al, 2010)	SDM (Xiong and De la Torre, 2013)	RCPR (Burgos-Artizzu et al, 2013)
#landmarks		66	66	any	5	68	68	29	49	any
Face detection rate (%)	COFW	70.22	86.00	100	72.98	79.68	79.68	50.30	71.40	100
	LFPW	73.21	92.86	100	96.88	89.29	88.39	76.34	87.95	100
	Helen	63.03	89.39	100	95.76	92.73	92.42	65.15	93.64	100
Error (%)	COFW	9.3666	11.1453	12.2417	5.4457	12.3449	11.8249	12.8179	6.9927	8.7382
	LFPW	7.2202	10.3122	7.3907	5.7649	14.1085	14.5332	10.7461	5.3600	6.4350
	Helen	8.2878	11.5897	8.9364	3.9133	13.4897	13.4176	11.2004	5.8397	5.4654
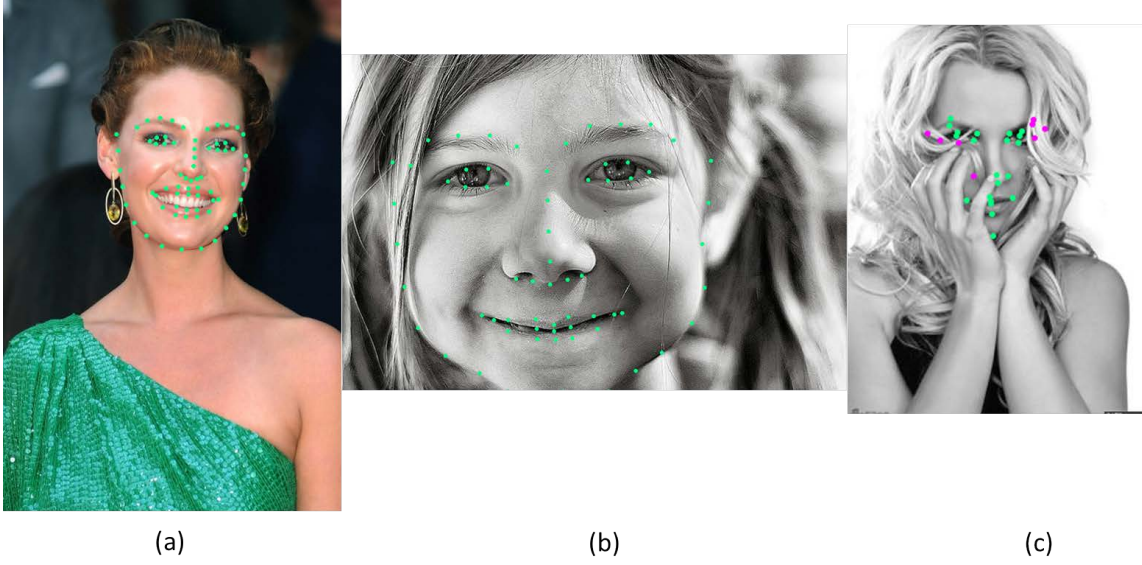
Fig. 7 Example faces from (a) LFPW, (b) Helen database and (c) COFW. The dots in deep red indicate corresponding points are occluded.
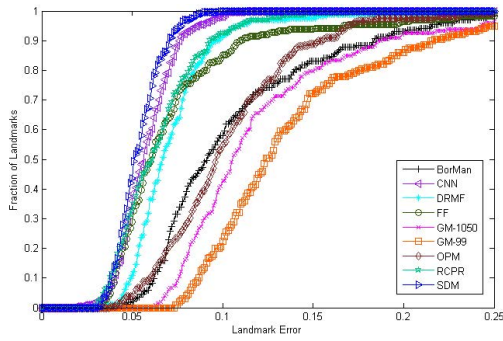

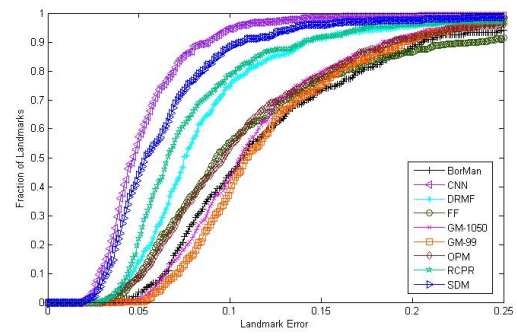**Fig. 8** Cumulative error curves on the LFPW database.
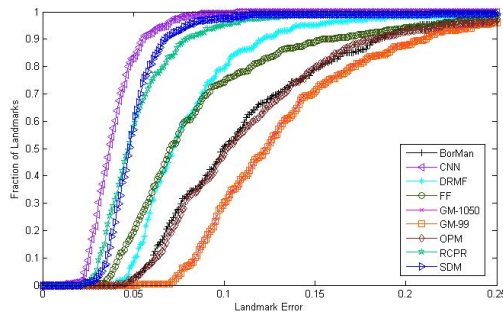

Fig. 9 Cumulative error curves on the Helen database.

Fig. 10 Cumulative error curves on the COFW database.

Table 7 Recalculation of mean normalized error RCPR method.

	COFW	LFPW	Helen
RCPR (Burgos-Artizzu et al, 2013) (%)	6.9557	5.0030	4.2730
SDM (Xiong and De la Torre, 2013) (%)	6.9927	5.3600	5.8397

faces in this database. The recomputed normalized error of RCPR on the COFW database was therefore calculated on the faces detected by SDM, and the details are shown in Table 7.

GM (Zhu and Ramanan, 2012) is trained on the Multi-PIE database (Gross et al, 2010), which is captured under laboratory conditions, but has inferior performance on real-world databases. Although the fast AAM fitting (FF) method (Tzimiropoulos and Pantic, 2013) achieves moderate performance, in our experience this method is very sensitive to the initialization. Of the four categories of methods, cascaded regression-based methods (e.g., (Cao et al, 2012; Xiong and De la Torre, 2013; Burgos-Artizzu et al, 2013)) and CNN (Sun et al, 2013) have the best performance.

Fig. 11, Fig. 12, and Fig. 13 show the detection errors for different landmarks or parts. Here, zero error means that the corresponding method does not detect

a point (part) or that the corresponding database does not label a landmark (part). From Fig. 11 and Fig. 12, it can be seen that landmarks around the outline of the face are the most difficult to accurately detect by all the tested methods. This is because the outline is easily affected by pose variation and occlusion. In contrast, the inner/outer corners of the eyes and the nose tips are relatively easy to localize, since these points are hardly affected by facial expressions, while the points around the mouth are heavily dependent on facial expressions.

Some methods are reported to have similar performance to human beings (Belhumeur et al, 2013; Burgos-Artizzu et al, 2013). However, occlusion and large shape variations in face images still provide significant challenges to successful and accurate detection, which is why in our experiments the above methods achieve better performance on the LFPW and Helen databases than on the COFW database.

It is also important to consider whether FFPD methods can detect facial landmarks in real-time. Model training is usually time-consuming in deep learning-based methods. The C++ implementation of CNN (Sun et al, 2013) took 0.12s to process a single image on a 3.30 GHz CPU, excluding face detection and image resizing. CLM-based methods generally take training time to learn local experts (e.g., learning weights using a linear SVM). State-of-the-art CLM methods (Saragih et al, 2011; Wang et al, 2008a; Gu and Kanade, 2008) are reported to take 0.120s, 0.098s, and 2.410s, respectively, on a 2.5 GHz Intel Core 2 Duo processor. Since publication of the seminal work in this area (Matthews and Baker, 2004), inverse composition fitting has significantly developed (Tzimiropoulos and Pantic, 2013), and this state-of-the-art fitting algorithm reaches near real-time performance on real-world databases. Recently, cascaded regression methods have attracted a lot of attention, not only due to their favorable performance, but also because of their training and detection speed. Cao et al. (Cao et al, 2012) reported that their method took only 20 minutes to train a model of 2000 images, with testing taking 0.015s with C++ implementation on an Intel Core i7 2.93 GHz CPU. RCPR (Burgos-Artizzu et al, 2013) has even better performance than Cao et al (2012), which is also a cascaded fern-based method.

7 Conclusion

Most existing methods improve the robustness dependent on carefully designed features such as pixel difference features (Cao et al, 2012; Burgos-Artizzu et al, 2013) and SIFT features (Xiong and De la Torre, 2013). Though these features achieve some success, they still
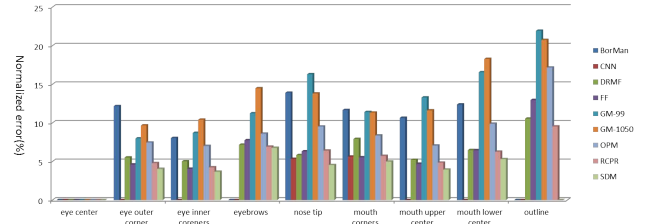

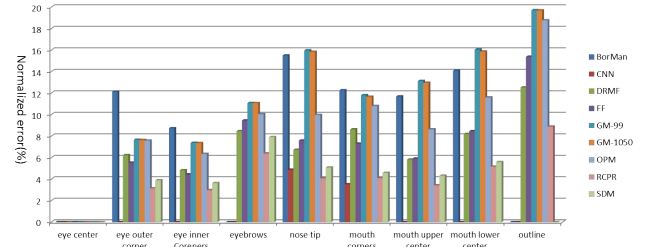**Fig. 11** Comparison of detection error of different landmarks (or parts): LFPW database.

Fig. 12 Comparison of detection error of different landmarks (or parts): Helen database.
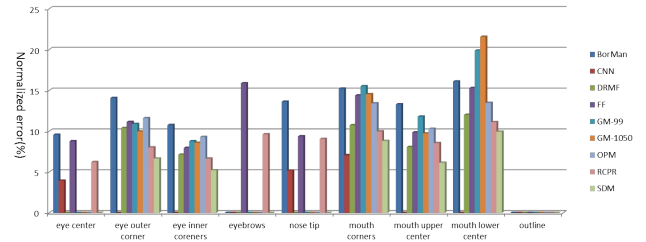

Fig. 13 Comparison of detection error of different landmarks (or parts): COFW database.

cannot adaptively deal with various shape variations and appearance variations. Recently, Ren et al (in press, 2014) presented an effective way to learn a set of local binary features to represent the facial image. Another promising way to adaptively learn features is by virtue of deep learning (Bengio et al, 2013) which achieves state-of-the-art performance on many computer vision tasks.

Besides feature learning, the model structure is another important issue related to the detection performance. Conventional ASM and AAM based methods assume that shape variations are statistically distribute as multivariate Gaussian, i.e. the linear PCA shape model. These explicit shape constraints actually have limited shape representation ability. Recent studies show that a cascaded set of simple linear regressors could achieve promising performance (Xiong and De la Torre, 2013; Ren et al, in press, 2014). Implicit shape constraint would be automatically hold if the initial shape is a legal face shape (Cao et al, 2012).

In this paper we reviewed FFPD methods, which can be grouped into four major categories: constrained local model-based, active appearance model-based, regression-based, and *other* methods. *Other* methods could be further divided into four minor categories: graphical model-based methods, joint face alignment methods, independent FFP detectors, and deep learning-based methods. By virtue of a comprehensive analysis and comparison of these methods, we found that cascaded regression-based methods achieved promising performance in the experimental setting. Although some state-of-the-art methods are ostensibly comparable to humans on some databases, there remain challenges in detecting occluded faces or those with large shape variation. Furthermore, most existing real-world databases are composed of frontal or near frontal images. Automatic FFP detection remains a distant promise.

References

Aizenberg I, Aizenberg N, Vandewalle J (2000) Multi-valued and universal binary neurons: theory, learning, applications. Kluwer Academic

Amberg B, Better T (2011) Optimal landmark detection using shape models and branch and bound. In: Proceedings of IEEE International Conference on Computer Vision, pp 455–462

Amberg B, Andrew B, Thomas V (2009) On compositional image alignment, with an application to active appearance models. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1714–1721

Anderson R, Stenger B, Cipolla R, Wan V (2013) Expressive visual text-to-speech using active appearance models. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 3382–3389

Arya S, Mount D, Silverman R, Wu A (1998) An optimal algorithm for approximate nearest neighbor searching. Journal of ACM 45(6):891–923

Ashraf A, Lucey S, Chen T (2010) Fast image alignment in the Fourier domain. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 2480–2487

Asteriadis S, Nikolaidis N, Pitas I (2009) Facial feature detection using distance vector fields. Pattern Recognition 42(7):1388–1398

Asthana A, Goecke R, Quadrianto N, Gedeon T (2009) Learning based automatic face annotation for arbitrary poses and expression from frontal images only. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1635–1642

Asthana A, Lucey S, Goecke R (2011) Regression based automatic face annotation for deformable model building. Pattern Recognition 44(10-11):2598–2613

Asthana A, Cheng S, Zafeiriou S, Pantic M (2013) Robust discriminative response map fitting with constrained local models. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 3444–3451

Baker S, Matthews I (2004) Lucas-kanade 20 years on: A unifying framework. International Journal of Computer Vision 56(1):221–255

Baker S, Gross R, Matthews I (2003) Lucas-kanade 20 years on: a unifying framework: part 3. Tech. rep., Carnegie Mellon University

Baltrusaitis T, Robinson P, Morency L (2012) 3D constrained local model for rigid and non-rigid facial tracking. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 2610–2617

Batur A, Hayes M (2003) A novel convergence scheme for active appearance models. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recogntion, pp 359–368

Batur A, Hayes M (2005) Adaptive active appearance models. IEEE Transactions on Image Processing 14(11):1707–1721

Belhumeur P, Jacobs D, Kriegman D, Kumar N (2011) Localizing parts of faces using a consensus of exemplars. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 545–552

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

Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(8):1798–1828

Bitouk D, Kumar N, Dhillon S, Belhumeur P, Nayar S (2008) Face swapping: automatically replacing faces in photographs. In: Proceedings of SIGGRAPH, pp 39.1–39.8

Blanz V, Vetter T (1999) A morphable model for the synthesis of 3d faces. In: Proceedings of SIGGRAPH, pp 187–194

Breiman L (1984) Classification and regression trees. Boca Raton: Chapman & Hall/CRC

Breiman L (2001) Random forests. Machine Learning 45(1):5–32

Burgos-Artizzu X, Perona P, Dollar P (2013) Robust face landmark estimation under occlusion. In: Proceedings of IEEE International Conference on Computer Vision, pp 1513–1520

Butakoff C, Frangi A (2006) A framework for weighted fusion of multiple statistical models of shape and appearance. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(11):1847–1857

Butakoff C, Frangi A (2010) Multi-view face segmentation using fusion of statistical shape and appearance models. Computer Vision and Image Understanding 114(3):311–321

Cao X, Wei Y, Wen F, Sun J (2012) Face alignment by explicit shape regression. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 2887–2894

Chen H, Xu Y, Shum H, Zhu S, Zheng N (2001) Example-based facial sketch generation with non-parametric sampling. In: Proceedings of IEEE International Conference on Computer Vision, pp 433–438

Chew S, Lucey P, Lucey S, Saragih J, Cohn J, Sridharan S (2011) Person-independent facial expression detection using constrained local models. In: Proceedings of Inteernational Conference on Automatic Face and Gesture Recognition, pp 915–920

Chow C, Liu C (1968) Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory 14(3):462–467

Cootes T, Taylor C (1992) Active shape models-'smart snakes'. In: Proceedings of British Machine Vision Conference, pp 266–275

Cootes T, Taylor C (1993) Active shape model search using local grey-level models: a quantitative evaluation. In: Proceedings of British Machine Vision Conference, pp 639–648

Cootes T, Taylor C (1999) A mixture model for representing shape variation. Image and Vision Computing 17(8):567–573

Cootes T, Taylor C (2001) Constrained active appearance models. In: Proceedings of IEEE International Conference on Computer Vision, pp 748–754

Cootes T, Taylor C (2006) An algorithm for tuning an active appearance model to new data. In: Proceedings of British Machine Vision Conference, pp 919–928

Cootes T, Taylor C, Lanitis A (1994) Active shape models: evaluation of a multi-resolution method for improving image search. In: Proceedings of British Machine Vision Conference, pp 327–336

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

Cootes T, Edwards G, Taylor C (1998a) Active appearance models. In: Proceedings of European Conference on Computer Vision, pp 484–498

Cootes T, Edwards G, Taylor C (1998b) A comparative evaluation of active appearance model algorithms. In:

Proceedings of British Machine Vision Conference, pp 680–689

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

Cootes T, Wheeler G, Walker K, Taylor C (2002) View-based active appearance models. Image and Vision Computing 20(9-10):657–664

Cootes T, Lonita M, Lindner C, Sauer P (2012) Robust and accurate shape model fitting using random forest regression voting. In: Proceedings of European Conference on Computer Vision, pp 278–291

Coughlan J, Ferreira S (2002) Finding deformable shapes using loopy belief propagation. In: Proceedings of European Conference on Computer Vision, pp 453–468

Criminisi A, Shotton J, Konukoglu E (2012) Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Foundations and Trends in Computer Graphics and Vision 7(2-3):81–227

Cristinacce D, Cootes T (2003) Facial feature detection using AdaBoost with shape constraints. In: Proceedings of British Machine Vision Conference, pp 24.1–24.10

Cristinacce D, Cootes T (2004) A comparison of shape constrained facial feature detectors. In: Proceedings of Inteernational Conference on Automatic Face and Gesture Recognition, pp 375–380

Cristinacce D, Cootes T (2006a) Facial feature detection and tracking with automatic template selection. In: Proceedings of Inteernational Conference on Automatic Face and Gesture Recognition, pp 429–434

Cristinacce D, Cootes T (2006b) Feature detection and tracking with constrained local models. In: Proceedings of British Machine Vision Conference, pp 929–938

Cristinacce D, Cootes T (2007) Boosted regression active shape models. In: Proceedings of British Machine Vision Conference, pp 1–10

Cristinacce D, Cootes T (2008) Automatic feature localization with constrained local models. Pattern Recognition 41(10):3054–3067

Cristinacce D, Cootes T, Scott I (2004) A multi-stage approach to facial feature detection. In: Proceedings of British Machine Vision Conference, pp 231–240

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 886–893

Dantone M, Gall J, Fanelli G, van Gool L (2012) Real-time facial feature detection using conditional regression forests. In: Proceedings of IEEE Conference on

Computer Vision and Pattern Recognition, pp 2578–2585

Dedeoglu G, Baker S, Kanade T (2006) Resolution-aware fitting of active appearance models to low resolution images. In: Proceedings of European Conference on Computer Vision, pp 83–97

Dedeoglu G, Kanade T, Baker S (2007) The asymmetry of image registration and its application to face tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(5):807–823

Ding L, Martinez A (2008) Precise detailed detection of faces and facial features. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1–7

Ding L, Martinez A (2010) Features versus context: an approach for precise and detailed detection and delineation of faces and facial features. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(11):2022–2038

Dollar P, Welinder P, Perona P (2010) Cascaded pose regression. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 1078–1085

Donner R, Reiter M, Georg L, 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

Duffy N (2002) Boosting methods for regression. Machine Learning 47(2-3):153–200

Everingham M, Sivic J, Zisserman A (2006) "hello! My name is ... Buffy"-automatic naming of characters in tv video. In: Proceedings of British Machine Vision Conference, pp 899–908

Fanelli G, Dantone M, Gool L (2013) Real time 3D face alignment with random forests-based active appearance models. In: Proceedings of Inteernational Conference on Automatic Face and Gesture Recognition, pp 1–8

Felzenszwalb P, Huttenlocher D (2005) Pictorial structures for object recognition. International Journal of Computer Vision 61(1):55–79

Fischler M, Bolles R (1981) Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. Communications of the ACM 24(6):381–395

Freund Y, Schapire R (1997) A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Science 55:119–139

Freund Y, Iyer R, Schapire R, Singer Y (2003) An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research 4(6):933–969

Friedman J (2001) Greedy function approximation: a gradient boosting machine. The Annals of Statistics 29(5):1189–1232

Friedman J, Hastie T, Tibshiani R (2000) Additive logistic regression: a statistical view of boosting. The Annals of Statistics 38(2):337–374

Gao H, Ekenel H, Fischer M, Stiefelhagen R (2011) Boosting pseudo census transform feature for face alignment. In: Proceedings of British Machine Vision Conference, pp 54.1–54.11

Gao H, Ekenel H, Stiefelhagen R (2012) Face alignment using a ranking model based on regression trees. In: Proceedings of British Machine Vision Conference, pp 118.1–118.11

Gao X, Su Y, Li X, Tao D (2010) A review of active appearance models. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews 40(2):145–158

van Ginneken B, Frangi A, Staal J, Romeny B, Viergever M (2002) Active shape model segmentation with optimal features. IEEE Transactions on Medical Imaging 21(8):924–933

Gonzalez-Mora J, De la Torre F, Murthi R, Guil N, Zapata E (2007) Bilinear acitve appearance models. In: Proceedings of IEEE International Conference on Computer Vision, pp 1–8

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

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

Grujic N, Ilic S, Lepetit V, Fua P (2008) 3D facial pose estimation by image retrieval. Tech. rep., Deutsche Telekom Laboratories

Gu L, Kanade T (2006) 3D alignment of face in a single image. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1305–1312

Gu L, Kanade T (2008) A generative shape regularization model for robust face alignment. In: Proceedings of European Conference on Computer Vision, pp 413–426

Gu L, Xing E, Kanade T (2007) Learning GMRF structures for spatial priors. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1–6

Hall P, D M, Martin R (2000) Merging and splitting eigenspace models. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(9):1042–1049

Hamsici O, Martinez A (2009) Active appearance models with rotation invariant kernels. In: Proceedings of IEEE International Conference on Computer Vision,

pp 1003–1009

Hansen M, Fagertun J, Larsen R (2011) Elastic appearance models. In: Proceedings of British Machine Vision Conference, pp 91.1–91.12

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

Hinton G, Osindero S, Teh Y (2006) A fast learning algorithm for deep belief nets. Neural Computation 18(7):1527–1554

Hou X, Li S, Zhang H, Cheng Q (2001) Direct appearance models. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 828–833

Huang C, Ding X, Fang C (2012) Pose robust face tracking by combining view-based AAMs and temporal filters. Computer Vision and Image Understanding 116(7):777–792

Huang G, Ramesh M, Berg T, Learned-Miller E (2007a) Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts

Huang Y, Liu Q, Metaxas D (2007b) A component based deformable model for generalized face alignment. In: Proceedings of IEEE International Conference on Computer Vision, pp 1–8

Jesorsky O, Kirchberg K, Frischholz R (2001) Robust face detection using the hausdoff distance. In: International Conference on Audio- and Video-based Biometric Person Authentication, pp 90–95

Kahraman F, Gokmen M, Darkner S, Larsen R (2007) An active illumination and appearance (AIA) model for face alignment. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1–7

Kasinski A, Florek A, Schmidt A (2008) The PUT face database. Image Processing and Communications 13(3):59–64

Kazemi V, Cullivan J (2011) Face alignment with part-based modeling. In: Proceedings of British Machine Vision Conference, pp 27.1–27.10

Kinoshita K, Konishi Y, Kawade M, Murase H (2012) Facial model fitting based on perturbation learning and it's evaluation on challenging real-world diversities images. In: Proceedings of European Conference on Computer Vision Workshop, pp 153–162

Kostinger M, Wohlhart P, Roth P, Bischof H (2011) Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: International Conference on Computer Vision WorkShops, pp 2144–2151

Kozakaya T, Shibata T, Takeguchi T, Nishiura M (2008a) Fully automatic feature localization for medical images using a global vector concentration approach. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1–6

Kozakaya T, Shibata T, Yuasa M, Yamaguchi O (2008b) Facial feature localization using weighted vector concentration approach. In: Proceedings of Interernational Conference on Automatic Face and Gesture Recognition, pp 1–6

Kozakaya T, Shibata T, Yuasa M, Yamaguchi O (2010) Facial feature localization using weighted vector concentration approach. Image and Vision Computing 28(5):772–780

La Cascia M, Sclaroff S, Athitsoss V (2000) Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3d models. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(4):322–336

Le V, Brandt J, Lin Z, Bourdev L, Huang T (2012) Interactive facial feature localization. In: Proceedings of European Conference on Computer Vision, pp 679–692

Learned-Miller E (2006) Data driven image models through continuous joint alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(2):236–250

Lee H, Kim D (2009) Tensor-based AAM with continuous variation estimation: application to variation-robust face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(6):1102–1116

Li Y, Gu L, Kanade T (2009) A robust shape model for multi-view car alignment. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 2466–2473

Li Y, Gu L, Kanade T (2011) Robustly aligning a shape model and its application to car alignment of unknown pose. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(9):1860–1876

Liang L, F W, Xu Y, Tang X, Shum H (2006a) Accurate face alignment using shape constrained Markov network. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1313–1319

Liang L, Wen F, Tang X, Xu Y (2006b) An integrated model for accurate shape alignment. In: Proceedings of European Conference on Computer Vision, pp 333–346

Liang L, Xiao R, Wen F, Sun J (2008) Face alignment via component-based discriminative search. In: Proceedings of European Conference on Computer Vision, pp 72–85

Liu X (2007) Generic face alignment using boosted appearance model. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1–8

Liu X (2009) Discriminative face alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(11):1941–1954

Liu X, Tu P, Wheeler F (2006) Face model fitting on low resolution images. In: Proceedings of British Machine Vision Conference, pp 1079–1088

Lowe D (2004) Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2):91–110

Lucey S, Wang Y, Cox M, Sridharan S, Cohn J (2009) Efficient constrained local model fitting for non-rigid face alignment. Image and Vision Computing 27(12):1804–1813

Lucey S, Navarathna R, Ashraf A, Sridharan S (2013) Fourier Lucas-Kanade algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(6):1383–1396

Luo P, Wang X, Tang X (2012) Hierarchical face parsing via deep learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 2480–2487

Martinez A, Benavente R (1998) The AR face database. Tech. rep., University of Barcelona

Martinez B, Valstar M, 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

Martins P, Caseiro R, Batista J (2010) Face alignment through 2.5D active appearance models. In: Proceedings of British Machine Vision Conference, pp 1–12

Martins P, Caseiro R, Henriques J, Batista J (2012a) Discriminative Bayesian active shape models. In: Proceedings of European Conference on Computer Vision, pp 57–70

Martins P, Caseiro R, Henriques J, Batista J (2012b) Let the shape speak-discriminative face alignment using conjugate priors. In: Proceedings of British Machine Vision Conference, pp 118.1–118.11

Martins P, Caseiro R, Batista J (2013) Generative face alignment through 2.5D active appearance models. Computer Vision and Image Understanding 117(3):250–268

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

Matthews I, Xiao J, Baker S (2007) 2D vs. 3D deformable face models: representational power, construction, and real-time fitting. International Journal of Computer Vision 75(1):93–113

Mei L, Figl M, Darzi A, Rueckert D, Edwards P (2008) Sample sufficiency and PCA dimension for statistical shape models. In: Proceedings of European Conference on Computer Vision, pp 492–503

Messer K, Matas J, Kittler J, Luettin J, Maitre G (1999) XM2VTSDB:the extended M2VTS database. In: International Conference on Audio- and Video-based Biometric Person Authentication, pp 72–77

Miborrow S, F N (2008) Locating facial features with an extended active shape model. In: Proceedings of European Conference on Computer Vision, pp 504–513

Miborrow S, Morkel J, Nicolls F (2010) The MUCT landmarked face database. In: Proceedings of Pattern Recognition Association of South Africa, pp 1–6

Michal U, Franc V, Hlaváč V (2012) Detector of facial landmarks learned by the structured output SVM. In: Proceedings of International Conference on Computer Vision Theory and Applications, pp 547–556

Navarathna R, Sridharan S, Lucey S (2011) Fourier active appearance models. In: Proceedings of IEEE International Conference on Computer Vision, pp 1919–1926

Nelder J, Mead R (1965) A simplex method for function minimization. Computer Journals 7(4):308–313

Nguyen M, Torre F (2008) Learning image alignment without local minima for face detection and tracking. In: Proceedings of Inteernational Conference on Automatic Face and Gesture Recognition, pp 1–7

Nguyen M, De la Torre F (2008) Local minima free parameterized appearance models. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1–8

Nguyen M, Torre F (2010) Metric learning for image alignment. International Journal of Computer Vision 88(1):69–84

Nordstrom M, Larsen M, Sierakowski J, Stegmann M (2004) The IMM face database-an annotated dataset of 240 face images. Tech. rep., Technical University of Denmark

Ojala T, Pietikainen M, Harwood D (1996) A comparative study of texute measures with classification based on featured distributions. Pattern Recognition 29(1):51–59

Ozuysal M, Calonder M, Lepetit V, Fua P (2010) Fast keypoint recognition using random ferns. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(3):448–461

Papandreou G, Maragos P (2008) Adaptive and constrained algorithms for inverse compositional active appearance model fitting. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1–8

Paquet U (2009) Convexity and Bayesian constrained local models. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1193–1199

Peng Y, Ganesh A, Wright J, Xu W, Ma Y (2012) RASL: robust alignment by sparse and low-rank decomposition for linearly correlated images. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(11):2233–2246

Peyras J, Bartoli A, Mercier H, Dalle P (2007) Segmented AAMs improve person-independent face fitting. In: Proceedings of British Machine Vision Conference, pp 1–10

Phillips P, Moon H, Rauss P, Rizvi S (2000) The FERET evaluation methodology for face recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(10):1090–1104

Ren S, Cao X, Wei Y, Sun J (in press, 2014) Face alignment at 3000 fps via regressing local binary features. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition

Rivera S, Martinez A (2012) Learning deformable shape manifolds. Pattern Recognition 45(4):1792–1801

Roberts M, Cootes T, Adams J (2007) Robust active appearance models with iteratively rescaled kernels. In: Proceedings of British Machine Vision Conference, pp 17.1–17.10

Roh M, Oguri T, Kanade T (2011) Face alignment robust to occlusion. In: Proceedings of Inteernational Conference on Automatic Face and Gesture Recognition, pp 239–244

Sanchez-Lozano E, De la Torre F, Gonzalez-Jimenez D (2012) Continuous regression for non-rigid image alignment. In: Proceedings of European Conference on Computer Vision, pp 250–263

Saragih J (2011) Principal regression analysis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 2881–2888

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: Proceedings of IEEE International Conference on Computer Vision, pp 1–8

Saragih J, Lucey S, Cohn J (2008) Deformable face fitting with soft correspondence constraints. In: Proceedings of Inteernational Conference on Automatic Face and Gesture Recognition, pp 1–8

Saragih J, Lucey S, Cohn J (2009a) Deformable model fitting with a mixture of local experts. In: Proceedings of IEEE International Conference on Computer Vision, pp 2248–2255

Saragih J, Lucey S, Cohn J (2009b) Face alignment through subspace constrained mean-shifts. In: Proceedings of IEEE International Conference on Computer Vision, pp 1034–1041

Saragih J, Lucey S, Cohn J (2009c) Probabilistic constrained adaptive local displacement experts. In: Proceedings of IEEE International Conference on Computer Vision Workshops, pp 288–295

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

Sauer P, Cootes T, Taylor C (2011) Accurate regression procedures for active appearance models. In: Proceedings of British Machine Vision Conference, pp 1–11

Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T (2007) Robust object recognition with cortex-like mechanisms. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(3):411–426

Shen X, Lin Z, Brandt J, Wu Y (2013) Detecting and aligning faces by image retrieval. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 3460–3467

Sivic J, Everingham M, Zisserman A (2009) "who are you?"-learning person specific classifiers from video. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1145–1152

Smith B, Zhang L (2012) Joint face alignment with non-parametric shape models. In: Proceedings of European Conference on Computer Vision, pp 43–56

Smith B, Zhang L, Brandt J, Lin Z, Yang J (2013) Exemplar-based face parsing. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 3484–3491

Sozou P, Cootes T, Taylor C, Mauro E (1995) A nonlinear generalization of point distribution models using polynomial regression. Image and Vision Computing 13(5):451–457

Sozou P, Cootes T, Taylor C, Mauro E (1997) Nonlinear point distribution modeling using a multi-layer perceptron. Image and Vision Computing 15(6):457–463

Stegmann M, Ersboll B, Larsen R (2003) FAME-a flexible appearance modeling environment. IEEE Transactions on Medical Imaging 22(10):1319–1331

Sukno F, Ordas S, Butakoff C, Cruz S, Frangi A (2007) Active shape models with invariant optimal features: application to facial analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(7):1105–1117

Sun Y, Wang X, Tang X (2013) Deep convolutional network cascade for facial point detection. In: Proceed-

ings of IEEE Conference on Computer Vision and Pattern Recognition, pp 3476–3483

Sung J, Kim D (2008) Pose-robust facial expression recognition using view-based 2d+3d AAM. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 38(4):852–866

Sung J, Kanade T, Kim D (2007) A unified gradient-based approach for combining ASM into AAM. International Journal of Computer Vision 75(2):297–309

Sung J, Kanade T, Kim D (2008) Pose robust face tracking by combining active appearance models and cylinder head models. International Journal of Computer Vision 80(2):260–274

Tong Y, Liu X, Wheeler F, Tu P (2009) Automatic facial landmark labeling with minimal supervision. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 2097–2104

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

De la Torre F, Nguyen M (2008) Parameterized kernel principal component analysis: theory and applications to supervised and unsupervised image alignment. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1–8

Tresadern P, Bhaskar H, Adeshina S, Taylor C, Cootes T (2009) Combining local and global shape models for deformable object matching. In: Proceedings of British Machine Vision Conference, pp 1–12

Tresadern P, Sauer P, Cootes T (2010) Additive update predictors in active appearance models. In: Proceedings of British Machine Vision Conference, pp 1–12

Tresadern P, Ionita M, Cootes T (2012) Real-time facial feature tracking on a mobile device. International Journal of Computer Vision 96(3):280–289

Tzimiropoulos G, Pantic M (2013) Optimization problems for fast AAM fitting in-the-wild. In: Proceedings of IEEE International Conference on Computer Vision, pp 593–600

Tzimiropoulos G, Zafeiriou S, Pantic M (2011) Robust and efficient parametric face alignment. In: Proceedings of IEEE International Conference on Computer Vision, pp 1847–1854

Tzimiropoulos G, Alabort-i Medina J, Zafeiriou S, Pantic M (2012) Generic active appearance models revisited. In: Proceedings of Asian Conference on Computer Vision, pp 650–663

Valstar M, Pantic M (2010) Induced disgust, happiness and surprise: an addition to the MMI facial expression database. In: Proceedings of International Conference on Language Resources and Evaluation, Workshop EMOTION, pp 65–70

Valstar M, Martinez B, Binefa X, Pantic M (2010) Facial point detection using boosted regression and graph models. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 2729–2736

Viola P, Jones M (2004) Robust real-time face detection. International Journal of Computer Vision 57(2):137–154

Vogler C, Li Z, Kanaujia A (2007) The best of both worlds: combining 3d deformable models with active shape models. In: Proceedings of IEEE International Conference on Computer Vision, pp 1–7

Vukadinovic D, Pantic M (2005) Fully automatic facial feature point detection using Gabor feature based boosted classifiers. In: Proceedings of International Conference on Systems, Man, and Cybernetics, pp 1692–1698

Wang N, Tao D, Gao X, Li X, Li J (2014) A comprehensive survey to face hallucination. International Journal of Computer Vision 106(1):9–30

Wang Y, Lucey S, Cohn J (2008a) Enforcing convexity for improved alignment with constrained local models. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1–8

Wang Y, Lucey S, Cohn J, Saragih J (2008b) Non-rigid face tracking with local appearance consistency constraint. In: Proceedings of Inteernational Conference on Automatic Face and Gesture Recognition, pp 1–8

Weise T, Bouaziz S, Li H, Pauly M (2011) Realtime performance-based facial animation. In: Proceedings of SIGGRAPH, pp 77.1–77.9

Wimmer M, Stulp F, Pietzsch S, Radig B (2008) Learning local objective functions for robust face model fitting. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(8):1357–1370

Wu H, Liu X, Doretto G (2008) Face alignment via boosted ranking model. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp 1–8

Wu Y, Wang Z, Ji Q (2013) Facial feature tracking under varying facial expressions and face poses based on restricted Boltzmann machine. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 3452–3459

Xiao J, Baker S, Matthews I, Kanade T (2004) Real-time combined 2D+3D active appearance models. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recogntion, pp 535–542

Xiong X, De la Torre F (2013) Supervised descent method and its application to face alignment. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 532–539

Yang H, Patras I (2012) Face parts localization using structured output regression forests. In: Proceedings of Asian Conference on Computer Vision, pp 667–679

Yang H, Patras I (2013) Sieving regression forest votes for facial feature detection in the wild. In: Proceedings of IEEE International Conference on Computer Vision, pp 1936–1943

Yang M, Kriegman D, Ahuja N (2002) Detecting faces in images: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(1):34–58

Yu X, Huang J, Zhuang S, Yan W, Metaxas D (2013) Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In: Proceedings of IEEE International Conference on Computer Vision, pp 1944–1951

Zhang H, Liu D, Poel M, Nijholt A (2009) Face alignment using boosting and evolutionary search. In: Proceedings of Asian Conference on Computer Vision, pp 110–119

Zhao C, Cham W, Wang X (2011) Joint face alignment with a generic deformable face model. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 561–568

Zhao X, Chai X, Shan S (2012) Joint face alignment: rescue bad alignments with good ones by regularized re-fitting. In: Proceedings of European Conference on Computer Vision, pp 616–630

Zhao X, Shan S, Chai X, Chen X (2013) Cascaded shape space pruning for robust facial landmark detection. In: Proceedings of IEEE International Conference on Computer Vision, pp 1033–1040

Zheng Y, Zhou X, Georgescu B, Zhou S, Comaniciu D (2006) Example based non-rigid shape detection. In: Proceedings of European Conference on Computer Vision, pp 423–436

Zhou S, Comaniciu D (2007) Shape regression machine. In: Information Proceeding in Medical Imaging, pp 13–25

Zhou X, Comaniciu D, Gupta A (2005) An information fusion framework for robust shape tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(1):115–129

Zhou Y, Gu L, Zhang H (2003) Bayesian tangent shape model: estimating shape and pose parameters via Bayesian inference. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 109–116

Zhu M, Martinez A (2006) Subclass discriminant analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(8):1274–1286

Zhu X, Ramanan D (2012) Face detection, pose estimation, and landmark localization in the wild. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 2879–2886