Computer Vision and Image Understanding 162 (2017) 1–22

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Face alignment in-the-wild: A Survey

Xin Jin[a,b], Xiaoyang Tan[a,b,*]

[a] Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, #29 Yudao Street, Nanjing 210016, PR China
[b] Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

ARTICLE INFO

Article history:
Received 14 August 2016
Revised 20 July 2017
Accepted 16 August 2017
Available online 24 August 2017

Keywords:
Face alignment
Active appearance model
Constrained local model
Cascaded regression
Deep convolutional neural networks

ABSTRACT

Over the last two decades, face alignment or localizing fiducial facial points on 2D images has received increasing attention owing to its comprehensive applications in automatic face analysis. However, such a task has proven extremely challenging in unconstrained environments due to many confounding factors, such as pose, occlusions, expression and illumination. While numerous techniques have been developed to address these challenges, this problem is still far away from being solved. In this survey, we present an up-to-date critical review of the existing literatures on face alignment, focusing on those methods addressing overall difficulties and challenges of this topic under uncontrolled conditions. Specifically, we categorize existing face alignment techniques, present detailed descriptions of the prominent algorithms within each category, and discuss their advantages and disadvantages. Furthermore, we organize special discussions on the practical aspects of face alignment *in-the-wild*, towards the development of a robust face alignment system. In addition, we show performance statistics of the state of the art, and conclude this paper with several promising directions for future research.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

Fiducial facial points refer to the predefined landmarks on a face graph, which are mainly located around or centered at the facial components such as eyes, mouth, nose and chin (see Fig. 1). Localizing these facial points, which is also known as face alignment, has recently received significant attention in computer vision, especially during the last decade. At least two reasons account for this. Firstly, many important tasks, such as face recognition, face tracking, facial expression recognition, head pose estimation, can benefit from precise facial point localization. Secondly, although some level of success has been achieved in recent years, face alignment in unconstrained environments is so challenging that it remains an open problem in computer vision, and continues to attract researchers to attack it.

While face detection is generally regarded as the starting point for all face analysis tasks (Ding and Martinez, 2010; Zafeiriou et al., 2015), face alignment can be regarded as *an important and essential intermediary step* for many subsequent face analyses that range from biometric recognition to mental state understanding. Concrete tasks may differ in the number and type of the needed facial points, as well as the way these points are used. Below we give some details on three typical tasks where face alignment plays a prominent role:

- *Face recognition:* Face alignment is widely used by face recognition algorithms to improve their robustness against pose variations. For example, in the stage of face registration, the first step is usually to locate some major facial points and use them as anchor points for affine warping, while other face recognition algorithms, such as feature-based (structural) matching (Campadelli et al., 2003; Zhao et al., 2003), rely on accurate face alignment to build the correspondence among local features (e.g, eyes, nose, mouth, etc.) to be matched.
- *Attribute computing:* Face alignment is also beneficial to facial attribute computing, since many facial attributes such as eyeglasses and nose shape are closely related to specific spatial positions of a face. In Kumar et al. (2009), six facial points are localized to compute qualitative attributes and similes that are then used for robust face verification in unconstrained conditions.
- *Expression recognition:* The configurations of facial points (typically between 20–60) are reliable indicative of the deformations caused by expressions, and the subsequent analysis will reveal the particular type of expression that may lead to such deformation. Many works (Bailenson et al., 2008; Li et al., 2015; Rudovic et al., 2010; Senechal et al., 2011; Valstar and Pantic,

* Corresponding author at: Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, #29 Yudao Street, Nanjing 210016, PR China.
E-mail addresses: x.jin@nuaa.edu.cn (X. Jin), x.tan@nuaa.edu.cn (X. Tan).

http://dx.doi.org/10.1016/j.cviu.2017.08.008
1077-3142/© 2017 Elsevier Inc. All rights reserved.

Fig. 1. Illustration of some example face images with 68 manually annotated points from the IBUG database (Sagonas et al., 2013a).

2012) follow this idea and use various features extracted from these points for expression recognition.

The above-mentioned applications, as well as numerous ones yet to be conceived, urge the need for developing robust and accurate face alignment techniques in real-life scenarios.

Under constrained environments or on less challenging databases, the problem of face alignment has been well addressed, and some algorithms even achieve performance that is close to that of human beings (Belhumeur et al., 2011; Dantone et al., 2012). Under unconstrained conditions, however, this task is extremely challenging and far from being solved, due to the high degree of facial appearance variability caused either by intrinsic dynamic features of the facial components such as eyes and mouth, or by ambient environment changes. In particular, the following factors have significant influence on facial appearance and the states of local facial features:

- *Pose:* The appearance of local facial features differ greatly between different camera-object poses (e.g., frontal, profile, upside down), and some facial components such as the one side of the face contour, can even be completely occluded in a profile face.
- *Occlusion:* For face images captured in unconstrained conditions, occlusion frequently happens and brings great challenges to face alignment. For example, the eyes may be occluded by hair, sunglasses, or myopia glasses with black frames.
- *Expression:* Some local facial features such as eyes and mouth are sensitive to the change of various expressions. For example, laughing may cause the eyes to close completely, and largely deform the shape of the mouth.
- *Illumination:* Lighting (varying in spectra, source distribution, and intensity), may significantly change the appearance of the whole face, and make the detailed textures of some facial components missing.

These challenges are illustrated in Fig. 2 by the IBUG database (Sagonas et al., 2013a). An ideal face alignment system should be robust to these facial variations on one hand; while on the other hand, as efficient as possible to satisfy the need of practical applications (e.g., real-time face tracking).

Over the last two decades, numerous techniques have been developed for face alignment with varying degrees of success. Çeliktutan et al. (2013) surveyed many traditional methods, but some recent state-of-the-art methods are not covered. Wang et al. (2014) gave a more comprehensive survey of face alignment methods over the last two decades, but the overall difficulties and challenges in unconstrained environments have not been highlighted. More recently, Yang et al. (2015) provided an empirical study of recent face alignment methods, aiming to draw some empirical yet useful conclusions and make insightful suggestions for practical applications.

The significant contribution of this paper is to give a comprehensive and critical survey of the ad hoc face alignment methods on 2D images, addressing the difficulties and challenges in unconstrained environments. We believe that it would be a useful complement to Çeliktutan et al. (2013), Wang et al. (2014) and Yang et al. (2015). But to be self-contained, some traditional methods covered in Çeliktutan et al. (2013) and Wang et al. (2014) are also included. However, contrary to the previous works, we add some state-of-the-art algorithms emerged recently (e.g., 3D face alignment methods), and pay special attention to study and summarize the motivation and successful experiences behind the state-of-the-art. Furthermore, we organize special discussions on the practical aspects of constructing a face alignment system, which in our opinion is a very important topic in practice, but is mostly ignored in previous studies. In addition, we show comparative performance statistics of the state of the art, and propose several promising directions for future research.

In Section 2, we briefly describe the main idea of face alignment and categorize existing methods into two main categories. Then, the prominent methods within each category are reviewed and analyzed in Sections 3 and 4. In Section 5, we investigate some practical aspects of developing of a robust face alignment system. In Section 6, we discuss a few issues concerning performance evaluation. Finally, we conclude this paper with a discussion of several promising directions for further research in Section 7.

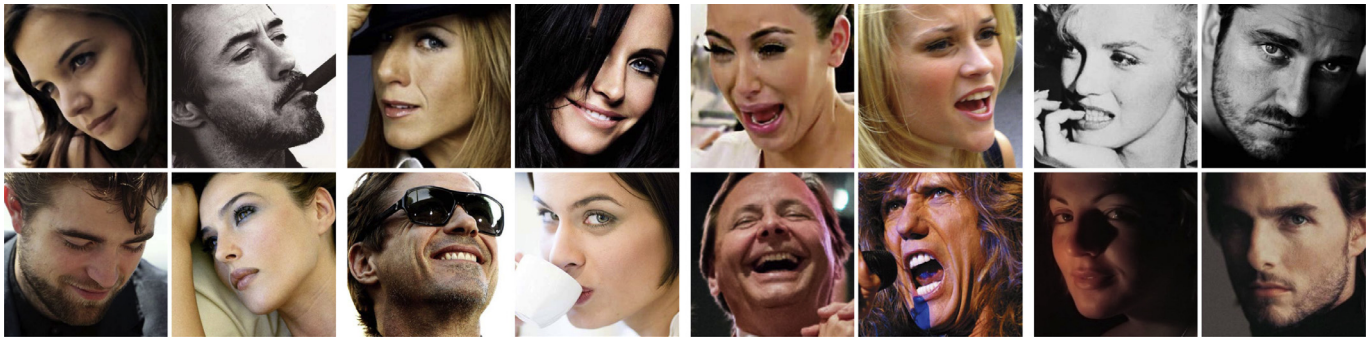
2. Overview

The problem of face alignment on 2D images has a long history in computer vision. A large number of approaches have been proposed to tackle it with varying degrees of success. From an overall perspective, face alignment can be formulated as a problem of searching over a face image for the pre-defined facial points (also called facial landmarks, or face shape). It typically starts from a coarse initial shape, and proceeds by refining the shape estimate step by step until convergence. During the search process, two different sources of information are typically used: facial appearance and shape information. The latter aims to explicitly model the spatial relations between the locations of facial points to ensure that the estimated facial points can form a valid face shape. Although some methods make no explicit use of the shape information, it is common to combine these two sources of information.

Before describing specific and prominent algorithms, a clear and high-level categorization will help to provide a holistic understanding of the commonality and differences of existing methods in using the appearance and shape information. For this, we follow the basic modeling principles in pattern recognition, and roughly divide existing methods into two categories: *generative* and *discriminative*.

- *Generative methods:* These methods build generative models for both the face shape and appearance. They typically formulate face alignment as an optimization problem to find the shape and appearance parameters that generate an appearance model instance giving best fit to the test face. Note that the facial appearance can be represented either by the whole (warped) face, or by the local image patches centered at the facial points.
- *Discriminative methods:* These methods directly infer the target location from the facial appearance. This is typically done by learning independent local detector or regressor for each facial point and employing a global shape model to regularize their predictions, or by directly learning a vectorial regression function to infer the whole face shape, during which the shape constraint is implicitly encoded.

Table 1 summarizes algorithms and representative works for face alignment, where we further divide the generative methods and discriminative methods into several subcategories. A few methods overlap category boundaries, and are discussed at the end of the section where they are introduced. Below, we discuss the



(a) Pose (b) Occlusion (c) Expression (d) Illumination

Fig. 2. An illustration of the great challenges of face alignment in the wild (IBUG (Sagonas et al., 2013a)), from left to right (every two columns): variations in pose, occlusion, expression and illumination.

Table 1
Categorization of the popular approaches for face alignment.

Approach	Representative works
Generative methods	
Active appearance models (AAMs)	
Regression-based fitting	Original AAM (Cootes et al., 2001); Boosted Appearance Model (Liu, 2007); Nonlinear discriminative approach (Saragih and Goecke, 2007); Accurate regression procedures for AMMs (Sauer et al., 2011)
Gradient descent-based fitting	Project-out inverse compositional (POIC) algorithm (Matthews and Baker, 2004); Simultaneous inverse compositional (SIC) algorithm (Gross et al., 2005); Fast AAM (Tzimiropoulos and Pantic, 2013); 2.5D AAM (Martins et al., 2013); Active Orientation Models (Tzimiropoulos et al., 2014)
Part-based generative models	Original Active Shape Model (ASM) (Cootes et al., 1995); Gauss-Newton deformable part model (Tzimiropoulos and Pantic, 2014); Project-out cascaded regression (Tzimiropoulos, 2015); Active pictorial structures (Antonakos et al., 2015b)
Discriminative methods	
Constrained local models (CLMs)[a]	
PCA shape model	Regularized landmark mean-shift (Saragih et al., 2011); Regression voting-based shape model matching (Cootes et al., 2012); Robust response map fitting (Asthana et al., 2013); Constrained local neural field (Baltrusaitis et al., 2013)
Exemplar shape model	Consensus of exemplar (Belhumeur et al., 2011); Exemplar-based graph matching (Zhou et al., 2013b); Robust Discriminative Hough Voting (Jin and Tan, 2016)
Other shape models	Gaussian Process Latent Variable Model (Huang et al., 2007b); Component-based discriminative search (Liang et al., 2008); Deep face shape model (Wu and Ji, 2015)
Constrained local regression	Boosted regression and graph model (Valstar et al., 2010); Local evidence aggregation for regression (Martinez et al., 2013); Guided unsupervised learning for model specific models (Jaiswal et al., 2013)
Deformable part models (DPMs)	Tree structured part model (Zhu and Ramanan, 2012); Structured output SVM (Uřičář et al., 2012); Optimized part model (Yu et al., 2013); Regressive Tree Structured Model (Hsu et al., 2015)
Ensemble regression-voting	Conditional regression forests (Dantone et al., 2012); Privileged information-based conditional regression forest (Yang and Patras, 2013a); Sieving regression forest votes(Yang and Patras, 2013b); Nonparametric context modeling (Smith et al., 2014)
Cascaded regression	
Two-level boosted regression	Explicit shape regression (Cao et al., 2012); Robust cascaded pose regression (Burgos-Artizzu et al., 2013); Ensemble of regression trees (Kazemi and Josephine, 2014); Gaussian process regression trees (Lee et al., 2015);
Cascaded linear regression	Supervised descent method (Xiong and De la Torre, 2013); Multiple hypotheses-based regression (Yan et al., 2013); Local binary feature (Ren et al., 2014); Incremental face alignment (Asthana et al., 2014); Coarse-to-fine shape search (Zhu et al., 2015)
Deep neural networks[b]	
Deep CNNs	Deep convolutional network cascade (Sun et al., 2013); Tasks-constrained deep convolutional network (Zhang et al., 2014c); Deep Cascaded Regression(Lai et al., 2015)
Other deep networks	Coarse-to-fine Auto-encoder Networks (CFAN) (Zhang et al., 2014a); Deep face shape model (Wu and Ji, 2015); RMnemonic Descent Method (Trigeorgis et al., 2016)
3D alignment methods[c]	
3D shape regression	3D face shape regression (Tulyakov and Sebe, 2015); Pose-invariant 3D face alignment (Jourabloo and Liu, 2015); Two-Stage convolutional part heatmap regression (Bulat and Tzimiropoulos, 2016)
Dense 3D model fitting	Displaced dynamic expression regression (Cao et al., 2014a); Dense 3D face alignment from 2D videos (Jeni et al., 2015); CNN-based dense 3D model fitting (Jourabloo and Liu, 2016); 3D dense face alignment (Zhu et al., 2016)

[a] Classic Constrained Local Models (CLMs) typically refer to the combination of local detector for each facial point and the parametric Point Distribution Model (Cristinacce and Cootes, 2006; Saragih et al., 2011; Wang et al., 2008). Here we extend the range of CLMs by including some methods based on other shape models (i.e., exemplar-based model (Belhumeur et al., 2011)). In particular, we will show that the exemplar-based method (Belhumeur et al., 2011) can also be interpreted under the conventional CLM framework.

[b] We note that some deep learning-based systems can also be placed in other categories. For instance, some systems are constructed in a cascade manner (Lai et al., 2015; Trigeorgis et al., 2016; Zhang et al., 2014a), and hence can be naturally categorized as cascaded regression. However, to highlight the increasing important role of deep learning techniques for face alignment, we organize them together for more systematic introduction and summarization.

[c] 3D face alignment refer to 3D alignment from 2D images in this paper, rather than the alignment of 3D faces. Since current 3D alignment methods basically employ discriminative regression techniques (e.g., cascaded regression), we categorized them as discriminative methods.

motivation and general approach of each category first, and then, give the review of prominent algorithms within each category, discussing their advantages and disadvantages.

3. Generative methods

Typically, faces are modelled as deformable objects which can vary in terms of shape and appearance. Generative face alignment methods construct parametric models for both face shape and appearance, and seek to find the best model parameters that can reconstruct the test face well during testing. This is similar to the EigenFace algorithm that employs Principle Component Analysis (PCA) to learn a set of liner appearance bases from training images, and uses these learned bases to reconstruct the new images during testing (Turk and Pentland, 1991). However, generative face alignment methods take into account the deformation of face shape, and build appearance model in a canonical reference frame where the shape variations have been removed.

According to the type of facial representation, generative methods can be further divided into two categories: Active Appearance Models that use holistic representation, and part-based generative models that use part-based representation.

3.1. Active appearance models

Active appearance models (AAMs), proposed by Cootes et al. (2001), are *linear* statistical models of both the shape and the appearance of the deformable object. AAMs been widely used in many computer vision tasks, such as face recognition (Lanitis et al., 1997), object tracking (Stegmann and Olsen, 2001), 3D modeling (Hamsici and Martinez, 2009; Xiao et al., 2004) and medical image analysis (Stegmann et al., 2003). In the field of face alignment, AAMs are arguably the most well-known family of generative methods that have been extensively studied during the last 20 years (Cootes et al., 2001; Gross et al., 2005; Matthews and Baker, 2004; Tzimiropoulos and Pantic, 2013).

In the following, we briefly introduce the basic AAM algorithm first, and then describe some recent advances on AAM research, and present some discussions about the advantages and disadvantages of AAMs.

3.1.1. Basic AAM algorithm: modeling and fitting

AAM modeling. An AAM is defined by three components, i.e., *shape model, appearance model*, and *motion model*. The shape model, which is coined Point Distribution Model (PDM) (Cootes and Taylor, 1992), is built from a collection of manually annotated facial points $\mathbf{s} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T)^T$ describing the face shape, where $\mathbf{x}_i = (x_i, y_i)$ is the 2-D location of the ith point. To learn the shape model, the training face shapes are normalized with respect to a global similarity transform (typically using Procrustes Analysis (Gower, 1975)) and PCA is applied to obtained a set of linear shape bases. The shape model can be mathematically expressed as:

$$\mathbf{s}(\mathbf{p}) = \mathbf{s_0} + \mathbf{Sp}, \tag{1}$$

where $\mathbf{s}_0 \in \mathbb{R}^{\{2N,1\}}$ is the mean shape, $\mathbf{S} \in \mathbb{R}^{\{2N,n\}}$ and $\mathbf{p} \in \mathbb{R}^n$ is the shape eigenvectors and parameters. Furthermore, this shape model need to be composed with a 2D global similarity transform, in order to position a particular shape model instance arbitrarily on the image frame. For this, using the re-orthonormalization procedure described in Matthews and Baker (2004), the final expression for the shape model can be compactly written using (1) by appending $\mathbf{S}$ with 4 similarity eigenvectors. The first row in Fig. 3 illustrates the mean shape and first two shape eigenvectors.

The appearance model is obtained by warping the training faces onto a common reference frame (typically defined by the mean shape), and applying PCA onto the warped appearances. Mathematically, the texture model is defined as follows:

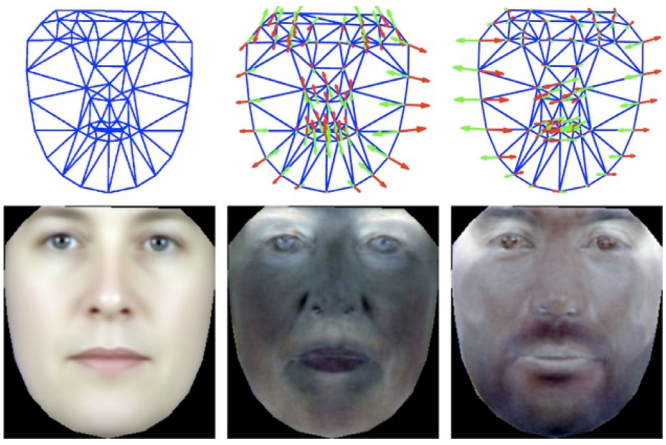


Fig. 3. First row: mean shape and first two shape eigenvectors. Second row: a face warped onto the canonical frame, and first two appearance eigenvectors.

$$\mathbf{A}(\mathbf{c}) = \mathbf{a_0} + \mathbf{Ac}, \tag{2}$$

where $\mathbf{a}_0 \in \mathbb{R}^{\{F,1\}}$ is the mean appearance, $\mathbf{A} \in \mathbb{R}^{\{F,m\}}$ and $\mathbf{c} \in \mathbb{R}^{\{m,1\}}$ is the appearance eigenvectors and parameters respectively. The second row in Fig. 3 illustrates the mean warped appearance and first two appearance eigenvectors.

To produce the shape-free textures, the motion model plays a role as a bridge between the image frame and the canonical reference frame. Typically, it is a warp function $\mathcal{W}$ that defines how, given a shape, the image should be warped into a canonical reference frame. Popular motion models include piece-wise affine warp (Matthews and Baker, 2004; Tzimiropoulos and Pantic, 2013) and Thin-Plate Splines warp (Baker and Matthews, 2001).

AAM fitting. Given an test image $\mathbf{I}$, AAM fitting aims to find the optimal parameters $\mathbf{p}$ and $\mathbf{c}$ so that the synthesized appearance model instance gives best fit to the test image in the reference frame. Formally, let $\mathbf{I}[\mathbf{p}] = \mathbf{I}(\mathcal{W}(\mathbf{p}))$ denote the vectorized version of the warped test image, then AAM fitting can be formulated as the following optimization problem,

$$\arg\min_{\mathbf{p},\mathbf{c}} ||\mathbf{I}[\mathbf{p}] - \mathbf{a}_0 - \mathbf{Ac}||^2. \tag{3}$$

Solving (3) is an iterative process that at each iteration an update of the current model parameters is estimated. In general, there are two main approaches for AAM fitting.

The first approach is to assume a *fixed* relationship between the residual image and parameter increments, and learn this relationship via *regression*. For example, in the original AAM paper (Cootes et al., 1998), this relationship is assumed linear and learned by linear regression, while in Saragih and Goecke (2007) a nonlinear repressor is learned via boosting. However, because of the incorrect assumption of *fixed* relationship (Matthews and Baker, 2004), the regression-based fitting strategies are efficient but approximate.

The second approach is to linearize with respect to $\mathbf{p}$ and then solve (3) iteratively in a Gauss–Newton fashion, as done by the Lukas–Kanade (LK) image alignment algorithm (Matthews and Baker, 2004)[1]. However, standard LK algorithm is inefficient when applied to AAMs, because the partial derivatives, Hessian and

[1] The standard Lucas–Kanade image alignment algorithm (Lucas et al., 1981) aims to find the locally best alignment between a constant template image and the input image with respect to the warp (shape) parameters, while Matthews and Baker (2004) replace the constant template image with a parameterized appearance model for AAM fitting.

gradient direction all need to be recomputed at each iteration. This problem is addressed by Matthews and Baker (2004) with an efficient inverse compositional image alignment algorithm (Gross et al., 2003), where the Jacobian and Hessian matrix can be pre-computed during fitting[2]. However, although extremely fast, the inverse compositional algorithm is also known to generalize poorly to unseen images (Gross et al., 2005). Several attempts to improve the generalization ability under the inverse compositional framework have been proposed in literature (Gross et al., 2003; Papandreou and Maragos, 2008), but are at a cost of increasing computational burden.

3.1.2. Recent advances on AAMs

Recently, some extensions and improvements of AAMs have been proposed to make this classic algorithm better adapted to the task of face alignment *in-the-wild*. In general, recent advances on AAMs mainly focus on three aspects: (1) unconstrained training data (Tzimiropoulos and Pantic, 2013), (2) feature-based representations (Antonakos et al., 2014; Tzimiropoulos et al., 2012) and (3) advanced fitting strategies (Tzimiropoulos et al., 2012; Tzimiropoulos and Pantic, 2013).

Unconstrained training data. Although some AAM fitting algorithms are known to perform well on constrained face databases (Gross et al., 2003), their performance has not been assessed on in-the-wild databases until recently. Tzimiropoulos and Pantic (2013) showed that, when trained in-the-wild, AAMs can generalize well to unseen images only using raw un-normalized pixel intensities as features.

Feature-based representations. Pixel-based image representation is typically considered to be sensitive to global lighting (Cootes et al., 1998; Matthews and Baker, 2004; Tzimiropoulos and Pantic, 2013). Therefore, a natural way to improve the robustness of AAMs is to use the feature-based representation such as HOG (Dalal and Triggs, 2005), SIFT (Lowe, 2004) and SURF (Bay et al., 2008), and this has been confirmed by some recent works on AAMs (Antonakos et al., 2015a; Tzimiropoulos et al., 2012; 2014).

Advanced fitting strategies. Finding a good trade-off between efficiency and accuracy is important for AAM fitting. However, most of traditional algorithms only pursue either efficiency (Matthews and Baker, 2004) or accuracy (Gross et al., 2003), but not both of them. Recently, some advanced fitting algorithms have been proposed to fill this gap (Papandreou and Maragos, 2008; Tzimiropoulos and Pantic, 2013). For example, by using a standard result from optimization theory, Tzimiropoulos and Pantic (2013) dramatically reduced the dominant cost in Gross et al. (2003) and the standard Lukas–Kanade algorithm in Matthews and Baker (2004), while achieving promising fitting accuracy.

3.1.3. Discussion

We have described the basic AAM algorithm and recent advances on AAMs. Despite the popularity, AMMs have been traditionally criticized for the limited representational power of their holistic representation, especially when used in wild conditions. However, recent works on AAMs (Antonakos et al., 2014; Lucey et al., 2013; Tzimiropoulos et al., 2012) suggest that this limitation might have been over-stressed in the literature and that AAMs can produce highly accurate results if appropriate training data (Tzimiropoulos and Pantic, 2013), image representations (Antonakos et al., 2014; Tzimiropoulos et al., 2012) and fitting strategies (Tzimiropoulos et al., 2012; Tzimiropoulos and Pantic, 2013) are employed.

Despite this, the partial occlusion cannot be easily handled by the holistic appearance model. One possible way to overcome this is to use part-based representations, based on the observation that local features are generally not as sensitive as global features to lighting and occlusion.

3.2. Part-based generative models

Part-based generative methods build generative appearance models for facial parts, typically with a shape model to govern the deformations of the face shapes. In general, there are two approaches to construct generative part models.

The first is to construct individual appearance model for each facial part, and a notable example is the well-known active shape models (Cootes and Taylor, 1992; Cootes et al., 1995) that combine the generative appearance model for each facial part and the Point Distribution Model for global shapes. However, a more natural and popular way is to model individual facial part is the *discriminatively* trained local detector (Asthana et al., 2013; Cootes et al., 2012; Cristinacce and Cootes, 2007; Saragih et al., 2011), as adopted by a very successful family of methods coined constrained local models (CLMs) (Asthana et al., 2013; Saragih et al., 2011). Actually, ASMs can be regarded as the predecessors of CLMs, and we refer the reader to Section 4.1 for more details about ASMs under the CLM framework.

The second approach is to construct generative models for all facial parts simultaneously. For example, the Gauss–Newton Deformable Part Model (GN-DPM) (Tzimiropoulos and Pantic, 2014) build linear statistical model for both the concatenated facial parts and the shape using PCA. With the part-based representation, the motion model of GN-DPM degenerates to similarity transformation, rather than the affine warp of AAMs. In the fitting phase, GN-DPM formulates and solves the non-linear least squares optimization problem similar to AAMs (Matthews and Baker, 2004; Tzimiropoulos and Pantic, 2013), jointly optimizing the appearance model and shape model in a Gauss-Newton fashion (Tzimiropoulos and Pantic, 2013). Apart from the PCA-based appearance model, Antonakos et al. (2015b) propose to model the appearance of facial parts using multiple pairwise distributions based on the edges of a graph (GMRF), and show that this outperforms the commonly used PCA model under an inverse Gauss–Newton optimization framework.

Compared to AAMs, the part-based generative models mainly have the advantages from part-based representation, i.e., more robust to global lighting and occlusion in wild conditions. Extensive experiments on wild face databases (Belhumeur et al., 2013; Le et al., 2012; Zhu and Ramanan, 2012) demonstrate that the part-based GN-DPM (Tzimiropoulos and Pantic, 2014) can outperform AAMs by a large margin.

3.3. Summary and discussion

We have reviewed generative methods for face alignment in two categories, i.e., Active Appearance Models that use the holistic representation and the part-based generative models that use the part-based representation. Recent results show that generative methods can produce high fitting accuracy for face alignment *in-the-wild*, if unconstrained training data (Tzimiropoulos and Pantic, 2013), robust image representations (Antonakos et al., 2014; Tzimiropoulos et al., 2012) and appropriate fitting strategies (Tzimiropoulos, 2015; Tzimiropoulos et al., 2012; Tzimiropoulos and Pantic, 2013; 2014) are employed. These results suggest that the limitations of generative methods might have been overstressed in the literature. In addition, generative methods typically

[2] The inverse compositional algorithm that "projects out" the appearance variation during fitting is considered as a seminal work in AAMs fitting (Matthews and Baker, 2004). However, because AAM fitting is not the key concern of this survey, we do not give detailed description about this algorithm and refer the readers to Matthews and Baker (2004) for more details.

have the advantage of requiring fewer training examples than the discriminative methods to perform well (Antonakos et al., 2015b).

However, with recent development of unconstrained facial databases with an abundance of annotated facial data captured, the discriminative methods, which are capable of effectively leveraging large bodies of training data, are now playing a more and more prominent role in face alignment.

4. Discriminative methods

Discriminative face alignment methods seek to learn a (or a set of) discriminative function that directly maps the facial appearance to the target facial points. In general, there are two main lines of research for discriminative methods. The first line is to follow the "divide and conquer" strategy by learning discriminative local appearance model for each facial point, and a shape model to impose global constraints on these local models. This line can be further subdivided into three classes: (1) *Constrained Local Models* that learn independent local detector for each facial point, with a shape model to regularize the detection responses of these local detectors. (2) *constrained local regression* methods that learn independent local regressor for each point and use a graph model to guide the search of these local regressors, and (3) *deformable part models* that learn the local appearance model and the tree structured shape model jointly in a discriminative framework.

The second line is to directly learn a vectorial regression function to infer the *whole* face shape, during which the shape constraint is implicitly encoded. This line can also be further subdivided into four classes: (1) *ensemble regression-voting* methods that cast votes for all facial points from local regions via regression, and ensemble the votes from different regions to form a robust prediction, (2) *cascaded regression* methods that learn a vectorial regression function in a cascade manner to estimate the face shape stage-by-stage, (3) *deep neural networks* that employ deep convolutional networks (Sun et al., 2013; Zhang et al., 2014c) or autoencoder networks (Zhang et al., 2014a) to model the nonlinear relationship between the facial appearance and the shape update, and (4)*3D alignment methods* that treat the face as a 3D object, and aim to recover the 3D locations of facial points from 2D images, typically through discriminative regression techniques (e.g., cascaded regression).

Table 2 gives a overview of the seven classes of discriminative methods in our taxonomy, where the appearance model, shape model and highlights of them are listed respectively to show the differences and relations between them.

4.1. Constrained local models

Constrained local models (CLMs), which can date back to the seminal work of Active Shape Model (ASM) (Cootes et al., 1995), are a relatively mature approach for face alignment (Asthana et al., 2013; Baltrusaitis et al., 2013; Cootes et al., 2012; Cristinacce and Cootes, 2006; Saragih et al., 2011). In the training phase, CLMs learn independent local detector for each facial point, and a prior shape model to characterize the deformation of face shapes. In testing, face alignment is typically formulated as an optimization problem to find the best fit of the shape model to the test image. We classify CLMs as the discriminative methods because of the discriminative nature of usual local detectors.

While the seminal work of Saragih et al. (2011) unifies various CLM approaches in a probabilistic framework, it only focuses on the CLMs using the PDM-based shape model. However, we note that some methods using other shape model (i.e., the exemplar shape model (Belhumeur et al., 2011)) are also close to Saragih et al. (2011) in methodology. Hence, in this paper we refer to those methods combining independent local detector and any kind of shape model collectively as *constrained local models*.

In the following, we will first briefly introduce the basic PDM-based CLM algorithm, then describe recent advances on CLMs in handling unconstrained challenges. In particular, we will show that exemplar-based method (Belhumeur et al., 2011) can also be interpreted under the conventional CLM framework (Saragih et al., 2011). Finally, we discuss the advantages and disadvantages of CLMs.

4.1.1. Basic CLM algorithm: modeling and fitting

CLM modeling. A CLM consists of two important components: *local detector* for each facial point, and the *shape model* that captures the deformations of valid face shapes. The task of local detector is to compute a pseudo probability (likelihood) that the target point occurs at a particular position. Existing local detectors can be broadly categorized into three groups.

- *Generative approach:* Generative approaches can be use to model local image patches centered at the annotated facial points. For example, Cootes and Taylor (1992); 1993) assume that the local appearance is multivariate Gaussian distributed, and use the Mahalanobis distance as the fitting response for a new image patch.
- *Discriminative classifier:* Discriminative classifier-based approach learns a binary classifier for each point with annotated image patches to discriminate whether the target point is aligned or not when testing. To cast various CLM fitting strategies in a unified probabilistic framework, the output of these classifiers are typically transformed into pseudo probabilities. Different types of classifiers have been exploited in literature, e.g., logistic regression (Saragih et al., 2011), SVM (Asthana et al., 2013; Belhumeur et al., 2011), and local neural field (LNF) (Baltrusaitis et al., 2013).
- *Regression-voting approach:* The regression-voting approach casts votes for the target point from a nearby region, then compute the pseudo probabilities by accumulating votes from different regions (Cootes et al., 2012; Cristinacce and Cootes, 2007). The regression-voting approach has the potential to be more efficient since a locally exhaustive search is avoided.

Due to the local patch support and large variations in training, the local detectors are typically imperfect, and the correct location will not always be at the location with the highest detection response. Therefore, a global shape model is typically employed to regularize the detection of these local detectors. For this, conventional CLMs use the PDM that simply models the normalized face shapes as multivariate Gaussian and approximates them using PCA (see Eq. (1)).

CLM fitting. Overall, give an image $\mathbf{I}$, the goal of PDM-based CLMs is to find the optimal shape parameter $\mathbf{p}$ that maximizes the probability of its points corresponding to consistent locations of the facial features. By assuming that the local search of each facial point is conditionally independent, the fitting objective of PDM-based CLMs can be written as:

$$\begin{aligned}\mathbf{p}^* &= \arg\max_{\mathbf{p}} p(\mathbf{p}|\{l_i=1\}_{i=1}^N, \mathbf{I}) \\ &= \arg\max_{\mathbf{p}} p(\mathbf{p}) \prod_{i=1}^{N} p(l_i=1|\mathbf{x}_i(\mathbf{p}), \mathbf{I}),\end{aligned} \tag{4}$$

where $\mathbf{x}_i(\mathbf{p})$ is the location of the ith point generated by the shape model, $l_i \in \{1, -1\}$ is a discrete random variable denoting whether the ith facial point is aligned or not, and $p(\mathbf{p})$ is the prior distribution of $\mathbf{p}$ that can be estimated from the training data.

CLM fitting based on (4) is an iterative process that entails (1) convolving the local detectors with the image to generate response

Table 2
Overview of the six classes of discriminative methods in our taxonomy.

	Appearance model	Shape model	Highlights of the method
Constrained local models	Independently trained local detector that computes a pseudo probability of the target point occurring at a particular position.	Point distribution model; Exemplar model, etc[a].	The local detectors are first correlated with the image to yield a filter response for each facial point, and then shape optimization is performed over these filter responses.
Constrained local regression	Independently trained local regressor that predicts a distance vector relating to a patch location.	Markov random fields to model the relations between relative positions of pairs of points.	Graph model is used to constrain the search space of local regressors by exploiting the constellations that facial points can form.
Deformable part models	Part-based appearance model that computes the appearance evidence for placing a template for a facial part.	Tree-structured models that are easier to optimize than dense graph structures.	All parameters of the appearance model and shape model are discriminatively learned in a max-margin structured prediction framework; efficient dynamic programming algorithms can be used to find globally optimal solutions.
Ensemble regression-voting	Image patches to cast votes for all facial points relating to the patch centers; Local appearance features centered at facial points.	Implicit shape constraint that is naturally encoded into the multi-output function (e.g., regression tree).	Votes from different regions are ensembled to form a robust prediction for the face shape.
Cascaded regression	Shape-indexed feature that is related to current shape estimate (e.g., concatenated image patches centered at the facial points).	Implicit shape constraint that is naturally encoded into the regressor in a cascaded learning framework.	Cascaded regression typically starts from an initial shape (e.g., mean shape), and refines the holistic shape through sequentially trained regressors.
Deep neural networks	Whole face region that is typically used to estimate the whole face shape jointly; Shape-indexed feature[b].	Implicit shape constraint that is encoded into the networks since all facial points are predicted simultaneously.	Deep network is a good choice to model the nonlinear relationship between the facial appearance and the shape update. Among others, deep CNNs have the capacity to learn highly discriminative features for face alignment.
3D alignment methods	Shape-indexed feature[c]; Specially designed features that are more appropriate for 3D regression.	3D point distribution model (PDM) that extends the traditional 2D PDM to the 3D space; Implicit shape constraint that is encoded into the multi-ouput regressors.	3D methods have strong advantages over 2D with respect to the robustness to pose, as they can accommodate a widely range of views. Currently, most of the 3D face alignment methods employ regression techniques, e.g., cascaded regression, CNN-based regression.

[a] Constrained local models (CLMs) typically employ a parametric (PCA-based) shape model (Saragih et al., 2011), but we will show that the exemplar-based method (Belhumeur et al., 2011) can also be derived from the CLM framework. Furthermore, we extend the range of CLMs by including some methods that combine independently local detector and other face shape model (Huang et al., 2007b; Liang et al., 2008; Wu and Ji, 2015).
[b] Some deep network-based systems follow the cascaded regression framework, and use the shape-indexed feature (Zhang et al., 2014a).
[c] Some 3D alignment methods (Jourabloo and Liu, 2015; Tulyakov and Sebe, 2015) also follow the cascaded regression framework using shaped indexed features.

Table 3
Different approximation strategies of response map.

	Approximation of response map
Isotropic Gaussian estimator (Cootes et al., 1995)	$\mathcal{N}(\mathbf{x}_i(\mathbf{p}); \boldsymbol{\mu}_i, \sigma_i^2\mathbf{E})$
Anisotropic Gaussian estimator (Wang et al., 2008)	$\mathcal{N}(\mathbf{x}_i(\mathbf{p}); \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$
Gaussian mixture model (Gu and Kanade, 2008)	$\sum_{k=1}^{K_i} \pi_{ik}\mathcal{N}(\mathbf{x}_i(\mathbf{p}); \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik})$
Gaussian kernel estimation (Saragih et al., 2011)	$\sum_{\mathbf{y}_j \in \Psi_{\mathbf{x}_i}} \pi_{\mathbf{y}_j}\mathcal{N}(\mathbf{x}_i(\mathbf{p}); \mathbf{y}_j, \rho^2\mathbf{E})$

maps, and (2) performing a global shape optimization procedure over these response maps. To make optimization efficient and numerically stable, a common choice of existing optimization strategies is to replace the true response maps with some approximate forms and then perform Gauss–Newton optimization over them instead of the original response maps.

The seminal framework of Saragih et al. (2011) unifies various approximation strategies for the true response maps. As listed in Table 3, they are (1) the isotropic Gaussian estimators used by original ASMs (Cootes and Taylor, 1992; Cootes et al., 1995), where $\boldsymbol{\mu}_i$ is the location of the maximum filter response within the ith response map, and σ_i^{-2} is the detection confidence over peak response coordinate, (2) a full covariance anisotropic Gaussian estimators used in Wang et al. (2008), where $\boldsymbol{\Sigma}_i$ is the anisotropic covariance matrix of Gaussian distribution, (3) Gaussian mixture model (GMM) used in Gu and Kanade (2008), where K_i denotes the number of modes and $\{\pi_{ik}\}_{k=1}^{K_i}$ are the mixing coefficients for the GMM of the ith point, and (4) a homoscedastic isotropic Gaussian kernel estimation (KDE) used by Saragih et al. (2011), where $\pi_{\mathbf{y}_j} = p(l_i = 1|\mathbf{y}_j, \mathbf{I})$ denotes the likelihood that the ith point is aligned at location $\mathbf{y}_j$, and ρ^2 denotes the variance of the noise on facial point locations, $\mathbf{E}$ is the identity matrix. Among them, the nonparametric Gaussian kernel estimation (KDE) method (Saragih et al., 2011) is considered to achieve a good tradeoff between representation power and the computational complexity. This method is known as Regularized Landmark Mean-Shift (RLMS) fitting, as the resulting update equations based on this nonparametric approximation are reminiscent of the well known mean-shift (Fukunaga and Hostetler, 1975) over the facial point but with regularization imposed by the PDM. Baltrušaitis et al. (2012) explored the information of depth images, and extend the RLMS (Saragih et al., 2011) algorithm to a 3D vision. Unlike aforementioned approximations to response maps, Asthana et al. (2013) proposes a novel discriminative regression based approach to directly estimate the parameter update, and results in significant performance improvement.

4.1.2. Recent advances on CLMs

Recently, some improvements of the conventional CLMs have been proposed to better handle various challenges in-the-wild. In general, recent advances on CLMs mainly focus on three aspects: (1) better local detectors, (2) discriminative fitting, and (3) other shape models.

Better local detectors. Conventional CLMs typically use logistic regression (Saragih et al., 2011) or SVM (Asthana et al., 2013; Bel-

humeur et al., 2011) to train local detector, while recently some advanced local detectors have been proposed, such as the Minimum Output Sum of Squared Errors (MOSSE) filters (Martins et al., 2014) and the local neural field (LNF) patch expert. These detectors are able to capture more complex information and exploit spatial relationships between pixels, and hence can achieve better detection results.

Discriminative fitting. It is widely acknowledged that the formulation based on CLMs is non-convex, and in general prone to local minima. As an alternative, Asthana et al. (2013) proposed a novel Discriminative Response Map Fitting (DRMF) method for the CLM fitting that outperforms the RLMS fitting method (Saragih et al., 2011) in wild databases. Unlike the RLMS method that performs Gaussian–Newton optimization, the DRMF method directly estimates the parameters of the DPM via discriminative regression using a response map based texture model.

Other shape models. One problem with the PDM is that its the model flexibility is heuristically determined by PCA dimension. To overcome this, some other shape models are investigated under the CLM framework (Belhumeur et al., 2011; Huang et al., 2007b; Wu and Ji, 2015). In particular, we will show that the exemplar-based method (Belhumeur et al., 2011) can be derived and well interpreted under the conventional CLM framework (Saragih et al., 2011).

The exemplar-based method (Belhumeur et al., 2011) assumes that the face shape $\mathbf{s} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$ in the test image is generated by one of the transformed exemplar shapes (global models). Let $\mathbf{s}_{k,\,t}$ $(k = 1, \ldots, D)$ denote locations of all facial points in the kth of the D exemplars that transformed by some similarity transformation t, and let $\mathbf{x}_{i,\,k,\,t}$ denote location of the ith facial point of the transformed exemplar $\mathbf{s}_{k,\,t}$. By assuming that conditioned on the global model $\mathbf{s}_{k,\,t}$, the location of each facial point $\mathbf{x}_i$ is conditionally independent of one another, the exemplar-based shape model $p(\mathbf{s})$ can be written as follows:

$$\begin{aligned} p(\mathbf{s}) &= \sum_{k=1}^{D} \int_{t\in T} p(\mathbf{s}, \mathbf{s}_{k,t}) dt \\ &= \sum_{k=1}^{D} \int_{t\in T} \prod_{i=1}^{N} p(\mathbf{x}_i|\mathbf{x}_{i,k,t}) p(\mathbf{s}_{k,t}) dt, \end{aligned} \tag{5}$$

where $p(\mathbf{x}_i|\mathbf{x}_{i,\,k,\,t})$ is modeled as a Gaussian distribution centered at $\mathbf{x}_{i,\,k,\,t}$, and the prior of the global model $p(\mathbf{s}_{k,\,t})$ is assumed as an uniform distribution. Then, by replacing the shape model $p(\mathbf{p})$ in conventional CLM framework (4) with above exemplar-based model $p(\mathbf{s})$, we derive the objective function of Belhumeur et al. (2011) (difference in notations) as follows:

$$\mathbf{s}^* = \arg\max_{\mathbf{s}} \sum_{k=1}^{D} \int_{t\in T} \prod_{i=1}^{N} p(\mathbf{x}_i|\mathbf{x}_{i,k,t}) p(l_i = 1|\mathbf{x}_i, \mathbf{I}) dt. \tag{6}$$

This function is optimized by employing RANSAC to sample global models. Due to the use of RANSAC, the exemplar-based method (Belhumeur et al., 2011) has two advantages over conventional CLMs: (1) independent of shape initialization, and (2) robust to partial occlusion, and achieves state-of-the-art performance on the wild LFPW database (Belhumeur et al., 2011) at that time.

The global models in Belhumeur et al. (2011) are scored and selected by the global likelihood, i.e., multiplying the detection response of each local detector. However, as pointed by Jin and Tan (2016), this global likelihood score function ignores the difference between local detectors, while in fact, an eye detector is typically more reliable than a chin detector. In Jin and Tan (2016), a discriminatively trained score function is proposed to evaluate the goodness of a global model, which weighs the importance of different local detectors. Furthermore, an efficient pipeline was proposed in Jin and Tan (2016) to alleviate the effect of inaccurate anchor points for generating global models.

4.1.3. Discussion

We have reviewed the basic CLM algorithm and recent advances. In general, CLMs are considered to be more robust to partial occlusion and global lighting than the holistic approaches (e.g., AAMs) (Saragih et al., 2011), due to their part-based modeling. However, the local detectors of CLMs are imperfect and have been shown to result in detection ambiguities in testing. Since the global shape optimization is performed on the response maps, the detection ambiguities may lead to performance bottleneck, when facing various challenges in unconstrained conditions.

CLMs perform expensive locally exhaustive search for each facial point. To reduce the computational cost, one way is to use a displacement expert (local regressor, i.e., estimate the relative position of the target point with respect to the given patch.

4.2. Constrained local regression

Besides CLMs, another discriminative local approach is to train independent local *regressor* for each point, and employ a global shape model to restrict the search of these local regressors to anthropomorphically consistent regions (Martinez et al., 2013; Valstar et al., 2010). Since this idea is similar to CLMs, we refer to this approach as *constrained local regression*.

A representative work of this group is the Boosted Regression coupled with Markov Networks (Valstar et al., 2010) (BoRMaN) method, which iteratively uses support vector regression (SVR) to provide an initial prediction for all points, and then applies the Markov Network to ensure that the new locations sampled to apply the local regressors are from correct point constellations. BoRMaN let each node in the graph associated to a spatial relation between two points and define pairwise relations between nodes, which allows a representation that is invariant to in-plane rotations, scale changes and translations. Essentially, BoRMaN performs an iterative sequential refinement of the estimate, where the previous target estimate becomes the test location at the next iteration. Martinez et al. (2013) argue that this sequential estimation approach has a series of drawbacks, for example, sensitive to the starting point and any errors in the estimation process. To improve the robustness of BoRMaN, Martinez et al. (2013) proposes to detect the target location by aggregating the estimates obtained from stochastically selected local appearance information into a single robust prediction.

The main advantage of constrained local regression approach is that combining local regressors with MRF may drastically reduce the time needed to search for point location, while its disadvantages are: (1) similar to CLMs, its performance is limited by the detection ambiguities of the independently trained local regressors, and (2) globally optimizing MRF is intractable. An alternative choice to the graph-based MRF are the tree-structured models, which are also effective to capture global elastic deformation, but easier to optimize than MRF.

4.3. Deformable part models

The tree-structured models are a natural and effective choice to model deformable objects (Yang and Ramanan, 2013; Zhu and Ramanan, 2012), and can find globally optimal solutions using an efficient dynamic programming algorithms (Felzenszwalb and Huttenlocher, 2005). Discriminatively trained tree-structured models have been successfully explored in many computer vision tasks, such as object detection (Felzenszwalb et al., 2010), human pose estimation (Yang and Ramanan, 2013), and recently in face alignment (Hsu et al., 2015; Uřičář et al., 2012; Zhu and Ramanan, 2012). We

follow the nomenclature of Felzenszwalb et al. (2010) and refer to them collectively as *deformable part models* (DPMs).

The main challenges of applying tree-structured model for face alignment may lie in the fact that a single tree-structured pictorial structure, perhaps, is insufficient to capture various shape deformations due to viewpoint. This problem is addressed by the seminal work of Zhu and Ramanan (2012), with a unified framework for face detection, pose estimation and face alignment. They modeled every facial point as a part and used mixtures of trees to capture the global topological changes due to viewpoint; a part will only be visible in certain mixtures/views. Formally, let $T_m = (\mathcal{V}_m, \mathcal{E}_m)$ be a linearly-parameterized, tree-structured pictorial structure for the mth mixture. Then, given image $\mathbf{I}$ and a face shape $\mathbf{s} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$, the tree structured part model of view m scores $\mathbf{s}$ as:

$$\begin{aligned} \mathcal{S}(\mathbf{I}, \mathbf{s}, m) &= \mathrm{App}_m(\mathbf{I}, \mathbf{s}) + \mathrm{Shape}_m(\mathbf{s}) + \alpha^m \\ \mathrm{App}_m(\mathbf{I}, \mathbf{s}) &= \sum_{i \in \mathcal{V}_m} \mathbf{w}_i^m \cdot \phi(\mathbf{I}, \mathbf{x}_i) \\ \mathrm{Shape}_m(\mathbf{s}) &= \sum_{ij \in \mathcal{E}_m} a_{ij}^m dx^2 + b_{ij}^m dx + c_{ij}^m dy^2 + d_{ij}^m dy, \end{aligned} \tag{7}$$

where $\mathrm{App}_m(\mathbf{I}, \mathbf{s})$ sums the appearance evidence at each part in $\mathbf{s}$, $\mathrm{Shape}_m(\mathbf{s})$ scores the mixture-specific spatial arrangement of $\mathbf{s}$, and α^m is a scalar bias associated with view point mixture m. Since parts may look consistent across some changes in viewpoint, (Zhu and Ramanan, 2012) allows different mixtures to share part templates to reduce the computational complexity.

To learn above mixtures of tree structured part models, the Chow–Liu algorithm (Chow and Liu, 1968) is first used to find the maximum likelihood tree structure that best explains the face shape for a given mixture. Then, for each view, all the model parameters in Eq. (7) are discriminatively learned in a max-margin structured prediction framework. In the testing phase, the input image is scored by all tree structures $T_m = (\mathcal{V}_m, \mathcal{E}_m)$ respectively, and the globally optimal shape $\mathbf{s}$ is efficiently solved with dynamic programming algorithm (Felzenszwalb and Huttenlocher, 2005).

Due to its simplicity and effectiveness, the tree structured part model (Zhu and Ramanan, 2012) has been extensively investigated and improved for face alignment. Uřičář et al. (2016) argue that the learning algorithm of Zhu and Ramanan (2012) is a variant of a two-class Support Vector Machines, which optimizes the detection rate of resulting face detector while the facial point locations serve only as latent variables not appearing in the loss function. In contrast, Uřičář et al. (2016) directly optimizes the average face alignment error with a novel objective function using the Structured Output SVMs algorithm, which leads to a significant improvement in alignment accuracy. Yu et al. (2013) presented a two-stage cascaded deformable shape model for face alignment, where a group sparse learning method is proposed to automatically select the optimized anchor points to achieve robust initialization based on the part mixture model of Zhu and Ramanan (2012). Hsu et al. (2015) proposed to improve the run-time speed and localization accuracy of Zhu and Ramanan (2012) with the Regressive Tree Structure Model (RTSM), where the tree structured model is applied on images with increasing resolution.

In general, the tree structured part model is effective at capturing global elastic deformation, while being easy to optimize unlike dense graph structure. Furthermore, it provide an unified framework to solve three tasks, namely face detection, face alignment and pose estimation, which is very appealing in automatic face analysis. However, its sluggish runtime impedes the potential for real-time facial point tracking; and perhaps due to the fact that the tree-based shape models allow for the non-face like structures to occur frequently, the performance of the tree structured part model (Zhu and Ramanan, 2012) is reported to be slightly inferior to that of the CLMs (Asthana et al., 2013; Saragih et al., 2011).

A common limitation of above part-based discriminative methods (i.e., CLMs, constrained local regression, and DPMs), however, is that their performance is greatly constrained by the ambiguity of the local appearance models. To break this bottleneck, many researchers have proposed to jointly estimate the whole face shape from the image, as described in the following sections.

4.4. Ensemble regression-voting

Apart from above discriminative local methods, another main stream of discriminative methods is to jointly estimate the whole face shape from the image. A simple way for this is to employ a vectorial function to cast votes for the face shape from image patches, during which the shape constraint is implicitly encoded. Because voting from a single region is rather weak, a robust prediction is typically obtained by ensembling votes from different regions. We refer to these methods as *ensemble regression-voting*.

Regression forests (Breiman, 2001) are a natural choice to perform regression-voting due to their simplicity and low computational complexity. Cootes et al. (2012) use random forest regression-voting to produce accurate response map for each facial point, which is then combined with the CLM fitting for robust prediction. Dantone et al. (2012) pointed out that conventional regression forests may lead to a bias to the mean face, because a regression forest is trained with image patches on the entire training set and averages the spatial distributions over all trees in the forest. Therefore, they extended the concept of regression forests to conditional regression forests. A conditional regression forest consists of multiple forests that are trained on a subset of the training data specified by global face properties (e.g., head pose used in Dantone et al. (2012)). During testing, the head pose is first estimated by a specialized regression forest, then trees of the various conditional forests are selected to estimate the facial points. Due to the high efficiency of random forests, Dantone et al. (2012) achieves close-to-human accuracy while processing images in real-time on the labeled facesin the wild (LFW) database (Huang et al., 2007a). After that, Yang and Patras (2013a) extended Dantone et al. (2012) by exploiting the information provided by global properties to improve the quality of decision trees, and later deployed a cascade of sieves to refine the voting map obtained from random regression forests (Yang and Patras, 2013b). Apart from the regression forests (Dantone et al., 2012; Yang and Patras, 2012; 2013a; 2013b), Smith et al. (2014) used each local feature surrounding the facial point to cast a weighted vote to predict facial point locations in a nonparametric manner, where the weight is pre-computed to take into account the feature's discriminative power.

In general, the ensemble regression-voting approach is more robust than previous local detector-based methods, and we conjecture that this robustness mainly stems from the combination of votes from different regions. However, current ensemble regression-voting approach, arguably, have not achieved a good balance between accuracy and efficiency for face alignment *in-the-wild*. The random forests approach (Dantone et al., 2012; Yang and Patras, 2012; 2013a; 2013b) is very efficient but can hardly cast precise votes for those unstable facial points (e.g., face contour), while on the other hand, the nonparametric feature voting approach based on facial part features (Smith et al., 2014) is more accurate but suffers from very high computational burden.

4.5. Cascaded regression

Recently, cascaded regression has established itself as one of the most popular and state-of-the-art methods for face alignment, due

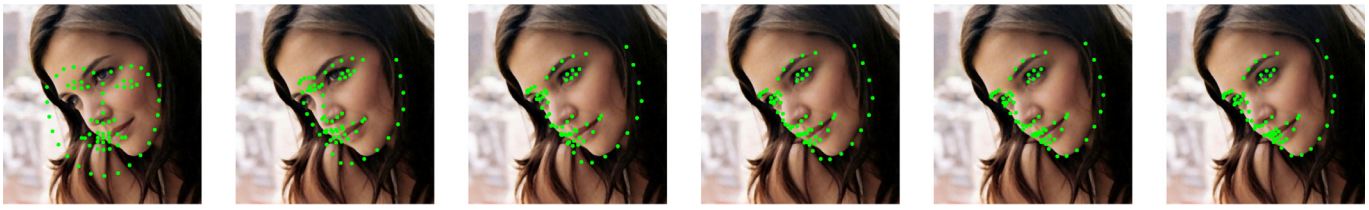


Fig. 4. Illustration of face alignment results in different stages of cascaded regression by Kazemi and Josephine (2014). The shape estimate is initialized and iteratively updated through a cascade of regression trees: (a) initial shape estimate, (b)-(f) shape estimates at different stages.

to its high accuracy and speed (Cao et al., 2012; Ren et al., 2014; Sun et al., 2013; Xiong and De la Torre, 2013; Zhu et al., 2015). The motivation behind this approach is that, since performing regression from image features to face shape in one step is extremely challenging, we can divide the regression process into stages, and learn a cascade of vectorial regressors.

Formally, given an image $\mathbf{I}$ and an initial shape $\mathbf{s}^0$, the face shape $\mathbf{s}$ is progressively refined by estimating a shape increment $\Delta\mathbf{s}$ stage-by-stage. In a generic form, a shape increment $\Delta\mathbf{s}$ at stage t is regressed as:

$$\Delta\mathbf{s}^t = \mathcal{R}^t\left(\Phi^t(\mathbf{I}, \mathbf{s}^{t-1})\right), \tag{8}$$

where $\mathbf{s}^{t-1}$ is the shape estimated in the previous stage, Φ^t is the feature mapping function, and $\mathcal{R}^t$ is the stage regressor. Note that $\Phi^t(\mathbf{I}, \mathbf{s}^{t-1})$ is referred to as *shape-indexed* feature (Burgos-Artizzu et al., 2013; Cao et al., 2012) that depends on the current shape estimate, and can be either designed by hand (Xiong and De la Torre, 2013; Zhu et al., 2015) or by learning (Cao et al., 2012; Kazemi and Josephine, 2014; Ren et al., 2014). Fig. 4 illustrates the alignment results in different stages. In the training phase, these stage regressors $(\mathcal{R}^1, \ldots, \mathcal{R}^T)$ are sequentially learnt to reduce the alignment errors on training set, during which geometric constraints among points are *implicitly* encoded. .

Existing cascaded regression methods mainly differ in the specific form of the stage regressor $\mathcal{R}^t$ and the feature mapping function Φ^t. Here, according to the type of the stage regressor $\mathcal{R}^t$, we roughly divide existing cascaded regression methods into two categories, i.e., *two-level boosted regression*, and *cascaded linear regression*.

4.5.1. Two-level boosted regression

Cascaded regression is first introduced into face alignment by Cao et al. (2012) in their seminal work called explicit shape regression (ESR). They design a two-level boosted regression framework by again investigating boosted regression as the stage regressor $\mathcal{R}^t$. More specifically, they use a cascade of random ferns as $\mathcal{R}^t$ to regress the *fixed* shape-indexed pixel difference feature at each stage, and adopt a correlation-based feature selection strategy to learn task-specific features. This combination makes ESR a break-through face alignment method in both accuracy and efficiency, and is widely adapted ever since.

Burgos-Artizzu et al. (2013) also use the fern primitive regressor under the two-level boosted regression framework, but improve (Cao et al., 2012) by explicitly incorporating the occlusion information into the regression target to better handle occlusions. Instead of random ferns used by Cao et al. (2012) and Burgos-Artizzu et al. (2013), Kazemi and Josephine (2014) present a general framework based on gradient boosting for learning an ensemble of regression trees, achieving super-realtime performance with high quality predictions and naturally handling missing or partially labelled data. Lee et al. (2015) propose to use the Gaussian process regression tree (GPRT) to fit the primitive regressor under the two-level boosted regression framework, where GPRT is a Gaussian process with a kernel defined by a set of trees.

4.5.2. Cascaded linear regression

Besides the two-level boosted regression framework (Burgos-Artizzu et al., 2013; Cao et al., 2012; Kazemi and Josephine, 2014; Lee et al., 2015), any kind of stage regressor $\mathcal{R}^t$ with strong fitting capacity will be desirable. A notable example is the cascaded linear regression proposed by Xiong and De la Torre (2013) using strong hand-craft SIFT (Lowe, 2004) feature.

The primary innovation of the cascaded linear regression method (Xiong and De la Torre, 2013) is a supervised gradient descent method (SDM) that gives a mathematically sound explanation of the cascaded linear regression by placing it in the context of Newton optimization for non-linear least squares problem. SDM shows that a Newton update for the non-linear least squares alignment error function can be expressed as a linear combination of the facial feature differences between the one extracted at current shape and the ground truth template, resulting in a linear update function $\mathcal{R}^t$ at each stage, i.e.,

$$\mathcal{R}^t : \Delta\mathbf{s}^t \leftarrow \mathbf{W}^t\left(\Phi^t(\mathbf{I}, \mathbf{s}^{t-1})\right) + \mathbf{b}^t, \tag{9}$$

where Φ^t is the SIFT operator that extract SIFT feature at each facial point, and $\mathbf{W}^t$ is the *averaged* descent direction on the training set.

Actually, SDM bears some similarities to AAMs discriminatively trained with linear regression (Cootes et al., 2001), but differs from them in three aspects: (1) SDM is non-parametric in both shape and appearance; (2) SDM uses the part-based representation; (3) SDM learns different regressors $\mathcal{R}^t$ at different stages, while the original AAM (Cootes et al., 2001) learns a constant regressor $\mathcal{R}$ for all stages.

Due to its concise formulation and state-of-the-art performance, SDM has been extensively investigated and extended. Xiong and De la Torre (2015) pointed out that SDM is a local algorithm that is likely to average conflicting gradient directions, and proposed an extension of SDM called Global SDM (GSDM) that divides the search space into regions of similar gradient directions. Yan et al. (2013) proposed to generate multiple hypotheses, and then learn to rank or combine these hypotheses to get the final result. Asthana et al. (2014) proposed an incremental formulation for the cascaded linear regression framework, and presented multiple ways for incrementally updating a cascade of regression functions in an efficient manner. Ren et al. (2014) propose to learn a set of highly discriminative local binary features for each facial point independently, and then uses the learned features jointly to learn a linear regression for the final prediction. Since regressing the local binary feature is very cheap, this approach is highly efficient and achieves very accurate performance. Zhu et al. (2015) designed a cascaded regression framework that begins with a coarse search over a shape space that contains diverse shapes, and employs the coarse solution to constrain subsequent finer search of

Table 4
A list of sources of wild databases for face alignment.

Databases	Year	#Images	#Training	#Test	#Point	Links
LFW (Huang et al., 2007a)	2007	13,233	1100	300	10	http://www.dantone.me/datasets/facial-features-lfw/
LFPW (Belhumeur et al., 2011)	2011	1,432[a]	–	–	35[b]	http://homes.cs.washington.edu/~neeraj/databases/lfpw/
AFLW (Köstinger et al., 2011)	2011	25,993	–	–	21	http://lrs.icg.tugraz.at/research/aflw
AFW (Zhu and Ramanan, 2012)	2012	205	–	–	6	http://www.ics.uci.edu/~xzhu/face/
HELEN (Le et al., 2012)	2012	2330	2000	300	194	http://www.ifp.illinois.edu/~vuongle2/helen/
300-W (Sagonas et al., 2013)	2013	3,837	3148	689	68	http://ibug.doc.ic.ac.uk/resources/300-W/
COFW (Burgos-Artizzu et al., 2013)	2013	1,007	–	–	29	http://www.vision.caltech.edu/xpburgos/ICCV13/
MTFL (Zhang et al., 2014c)	2014	12,995	–	–	5	http://mmlab.ie.cuhk.edu.hk/projects/TCDCN.html
MAFL (Zhang et al., 2016)	2016	20,000	–	–	5	http://mmlab.ie.cuhk.edu.hk/projects/TCDCN.html

[a] LFPW is shared by web URLs, but some URLs are no longer valid.
[b] Each face image in LFPW is annotated with 35 points, but only 29 points defined in Belhumeur et al. (2011) are used for the face alignment.

Table 5
A list of published software of face alignment.

Methods	Year	#Points	Links
Boosted regression with Markov networks (BoRMaN) (Valstar et al., 2010)	2010	22	http://ibug.doc.ic.ac.uk/resources/facial-point-detector-2010/
Constrained local model (CLM) (Saragih et al., 2011)	2011	66	https://github.com/kylemcdonald/FaceTracker
Tree structured part model (TSPM) (Zhu and Ramanan, 2012)	2012	68	http://www.ics.uci.edu/~xzhu/face/
Conditional random forests (CRF) (Dantone et al., 2012)	2012	10	http://www.dantone.me/projects-2/facial-feature-detection/
Structured output SVM (Uřičář et al., 2012)	2012	7	http://cmp.felk.cvut.cz/~uricamic/flandmark/
Cascaded CNN (Sun et al., 2013)	2013	5	http://mmlab.ie.cuhk.edu.hk/archive/CNN_FacePoint.htm
Discriminative response Map Fitting (DRMF) (Asthana et al., 2013)	2013	66	https://sites.google.com/site/akshayasthana/clm-wild-code?
Supervised descent method (SDM) (Xiong and De la Torre, 2013)	2013	49	www.humansensing.cs.cmu.edu/intraface
Robust cascaded pose regression (RCPR) (Burgos-Artizzu et al., 2013)	2013	29	http://www.vision.caltech.edu/xpburgos/ICCV13/
Optimized part mixtures (OPM) (Yu et al., 2013)	2013	68	http://www.research.rutgers.edu/~xiangyu/face_align/face_align_iccv_1.1.zip
Continuous Conditional Neural Fields (CCNF) (Baltrušaitis et al., 2014)	2014	68	https://github.com/TadasBaltrusaitis/CCNF
Coarse-to-fine Shape Searching (CFSS) (Zhu et al., 2015)	2015	68	http://mmlab.ie.cuhk.edu.hk/projects/CFSS.html
Project-Out Cascaded Regression (PO-CR) (Tzimiropoulos, 2015)	2015	68	http://www.cs.nott.ac.uk/~yzt/
Active pictorial structures (APS) (Antonakos et al., 2015b)	2015	68	https://github.com/menpo/menpo
Tasks-Constrained Deep Convolutional Network (TCDCN) (Zhang et al., 2016)	2016	68	http://mmlab.ie.cuhk.edu.hk/projects/TCDCN.html

shape, which improves the robustness of cascaded linear regression in coping with large pose variations.

4.5.3. Discussion

Arguably, cascaded regression is playing a prominent role among the state-of-the-art methods for face alignment *in-the-wild*. This is primarily because it has some distinct characteristics. (1) It is capable of effectively leveraging large bodies of training data, which are typically generated by using multiple initial shapes for one image. (2) The shape constraints are encoded into regressors adaptively, which is more flexible than the parametric shape model that heuristically determines the model flexibility (e.g.,PCA dimension). (3) The cascaded regression framework is simple and generalizable, which allows different choices for the stage regressor $\mathcal{R}^t$ and incorporation of feature learning techniques.

Although cascaded regression has achieved great success in face alignment, it is still not easy to perform regression from texture features to the whole shape update for some challenging faces with extreme expression or pose variation. This limitation can be partially confirmed by the fact that for some more flexible part localization task such as human pose estimation, the part detector-based methods still play a dominant role at present (Liu et al., 2015; Yang and Ramanan, 2013), rather than cascaded regression.

4.6. Deep neural networks

Deep neural networks, especially the deep convolutional network that can extract high-level image features, have been successfully utilized in many computer vision tasks, such as face verification (Sun et al., 2014; Taigman et al., 2014), image classification (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Szegedy et al., 2015), and object detection (Girshick et al., 2014). Naturally, they are also an effective choice to model the nonlinear relationship between the facial appearance and the face shape (or shape update).

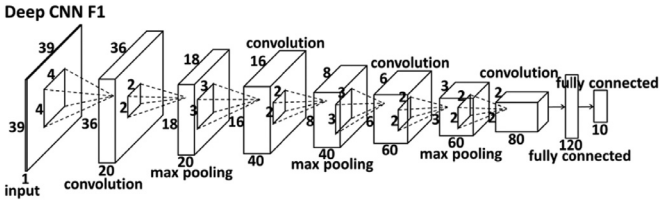


Fig. 5. One of the first-level convolutional neural network structures used in Sun et al. (2013) to predict five major facial points. Sizes of input, convolution, and max pooling layers are illustrated by cuboids whose length, width, and height denote the number of maps, and the size of each map. Local receptive fields of neurons in different layers are illustrated by small squares in the cuboids.

However, applying deep network directly to face alignment is nontrivial due to the follwoing reasons: (1) While fine-tuning an existing CNN architecture (e.g., AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015)) to adapt it to the task at hand is very popular in computer vision (Girshick et al., 2014; Zhang et al., 2014b), such a strategy can hardly be applied for face alignment because the off-the-shelf large networks are typically trained for image classification while face alignment is a structural prediction problem. (2) Constructing a deep network-based system from scratch for face alignment should take into account the issue of over-fitting, and hence the network structures at each stage need to be carefully designed according to the task of this stage and the complexity involved.

Focusing on above issues, Sun et al. (2013) were pioneers in this area with their work called Deep Convolutional Network Cascade. They handled the face alignment task with three-level carefully designed convolutional networks, and fuse the outputs of multiple networks at each level for robust prediction (Fig. 5

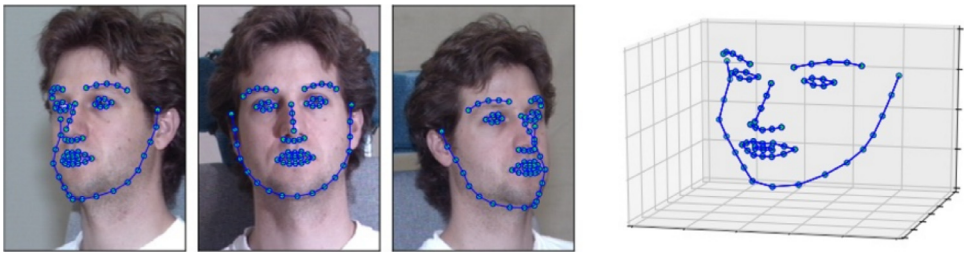


Fig. 6. An example of MultiPIE (Gross et al., 2010) recording annotated with 3D ground truth. 3D annotation can accommodate a full range of head rotation while still maintaining the correspondences.

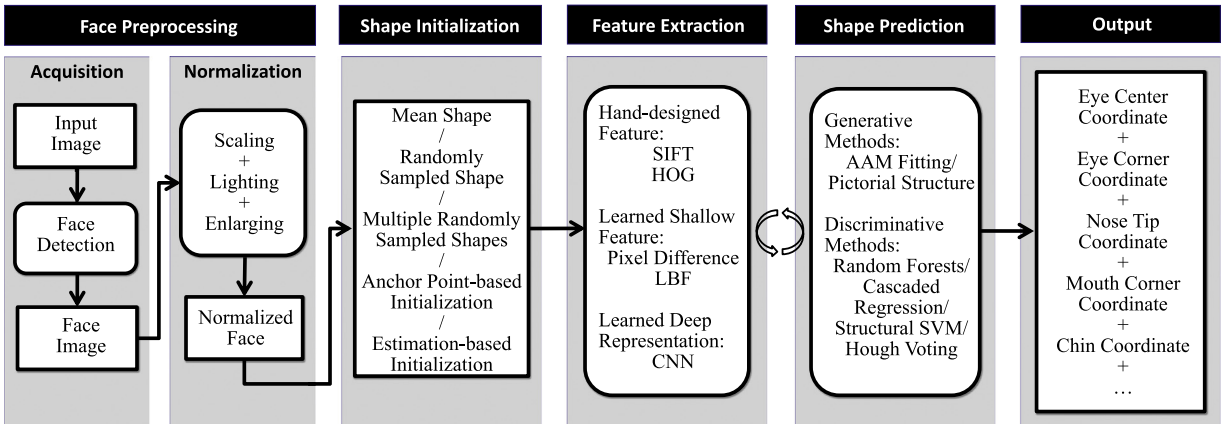


Fig. 7. A global system architecture for face alignment.

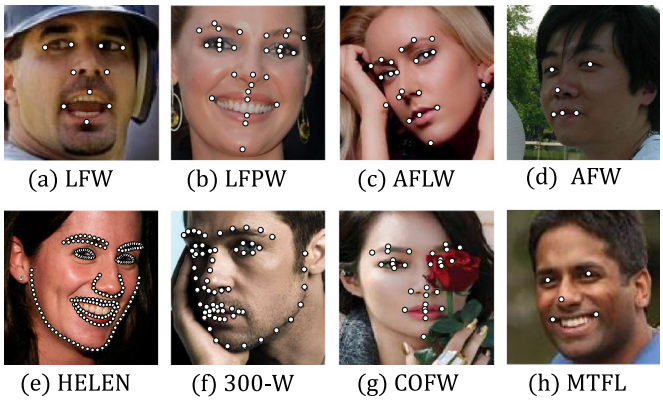


Fig. 8. Illustration of the example face images from eight wide face databases with original annotation.

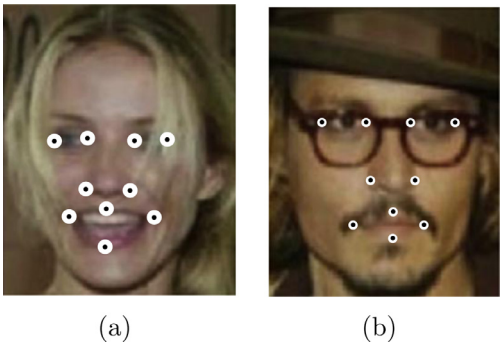


Fig. 9. Two images of the LFW database annotated with 10 facial feature points. The white circles show the disturbance range from the ground truth (black points), 10% of the inter-ocular distance in (a) while 5% in (b), which aims to give a intuitive feeling of the localization error listed in Table 6.

illustrates one of the first-level CNN structures). The first level network takes the whole face image as input to predict the initial estimates of the holistic face shape, during which the shape constraints are implicitly encoded. Then, the following two level networks refine the position of each point to achieve higher accuracy. Several network structures critical for face alignment are investigated in Sun et al. (2013), providing some important principles on the choice of convolutional network structures. For example, convolutional networks at the first level should be deeper than the following networks, since predicting facial points from large input regions is a high-level task.

Ever since the work of Zhang et al. (2014c), deep CNNs have been widely exploited for face alignment. Similar to Zhang et al. (2014c), Zhou et al. (2013a) designed a four-level convolutional network cascade to tackle the face alignment problem in a coarse-to-fine manner, where each network level is trained to locally refine a subset of facial points generated by previous network levels. Zhang et al. (2014c) extended the work of Sun et al. (2013) by jointly learning auxiliary attributes along with face alignment. Their work confirms that some heterogeneous but subtly correlated tasks, e.g., head pose estimation and facial attribute inference can aid the face alignment task through multi-task learning. Lai et al. (2015) proposed an end-to-end CNN architecture to learn highly discriminative shape-indexed features, by encoding the image into high-level feature maps in the same size of the image, and then extracting deep features from these high level descriptors through a novel “Shape-Indexed Pooling” method. Despite of the great popularity and success, as mentioned before, we should take into account the tradeoff between the model complexity and training data size, since some deep models have been reported to be pre-trained with enormous quantity of external data sources (Sun et al., 2013; Zhang et al., 2014c).

4.7. 3D alignment methods

Most of existing face alignment methods focus on 2D alignment that treats the face as a 2D object. However, 2D methods are mainly designed for faces in small to medium poses (below 45°) - when face orientation varies from frontal to profile, labeling 2D facial points becomes extremely challenging since the invisible points have to be guessed, and some annotated points (e.g., cheek landmarks) may lose correspondence. In contrast, 3D annotation can accommodate a full range of head rotation while still maintaining the correspondences (see Fig. 6). Besides, it also provides more information for estimating the head pose (Tulyakov and Sebe, 2015) and facial point visibility (Jourabloo and Liu, 2015). Due to these reasons, 3D face alignment from 2D images has recently emerged as a promising direction to address the large pose face alignment problem (Bulat and Tzimiropoulos, 2016; Jourabloo and Liu, 2015; 2016; Tulyakov and Sebe, 2015). In particular, the first 3D Face Alignment in the Wild (3DFAW) Challenge was held in conjunction with the 14th European Conference on Computer Vision to encourage the study of 3D methods (Jeni et al., 2016). Fig. 6 shows the 3D annotation used by 3DFAW.

3D alignment from 2D images requires 3D annotation for facial points. However, most existing face alignment databases only contain 2D annotations, with no associated 3D information. Fortunately, because the 2D face can be regarded as a projection of the 3D face on the image plane (Xiao et al., 2004), the 3D face shape can be reconstructed from the annotated 2D shape using a model-based structure-from-motion technique (Jeni et al., 2016; Jourabloo and Liu, 2015). Through this, the vast amount of existing 2D face datasets can be leveraged for 3D face alignment.

Currently, 3D alignment methods are mainly based on discriminative regression techniques (e.g., cascaded regression), and thus we categorized them as discriminative methods. Furthermore, according to whether the method estimates the dense 3D face surface, we categorize existing methods into two groups: *3D shape regression* and *dense 3D model fitting*.

4.7.1. 3D shape regression

Motivated by the success of 2D shape regression (Cao et al., 2012; Ren et al., 2014; Sun et al., 2013; Xiong and De la Torre, 2013; Zhu et al., 2015), researchers further exploit 3D shape regression[3] for 3D face alignment, building upon the state-of-the-art 2D regression techniques, such as cascaded regression and CNN-based regression. 3D shape regression inherits the merits of 2D shape regression (e.g., high accuracy and speed), and is currently the mainstream approach to 3D face alignment. In general, there are three ways to regress a 3D face shape from a single image.

The first way is to extend the 2D regression methods directly by augmenting the output vector with a depth dimension. For example, Tulyakov and Sebe (2015) follow the cascaded regression approach, and build tree-based regressors to produce a 3D shape increment using the 3D-invariant features. They also show that regressing a 3D shape can improve the accuracy even if we are only interested in 2D facial points in the image plane.

The second way is to decompose the 3D face alignment problem into two steps: X, Y (2D) regression and Z (depth) regression (Bulat and Tzimiropoulos, 2016; Gou et al., 2016; Zhao et al., 2016). In general, any state-of-the-art 2D face alignment method, such as cascaded regression (Gou et al., 2016) and deep CNN regression (Zhao et al., 2016), can be employed as the first step. The resulting 2D face shape is then used to guide the estimation of the depth information. For example, Gou et al. (2016) propose to recover 3D face shape by fitting a 3D PDM to the image, with the estimated 2D landmarks as a solid constraint. Zhao et al. (2016) use a multi-layer neural network to model the mapping from X, Y locations to the depth information. In Bulat and Tzimiropoulos (2016), a deep residual network (He et al., 2016) guided by the heatmaps produced by the 2D regression subnetwork is introduced to estimate the depth information.

The third way is to estimate the 2D and 3D facial landmarks jointly, by regarding the 2D face as a projection of the 3D face. For this purpose, Jourabloo and Liu (2015) develop a coupled-regressor approach to estimate both the camera projection matrix and 3D facial landmarks under the cascaded regression framework. Besides, they also propose to estimate the visibility of facial landmarks via 3D surface normal.

4.7.2. Dense 3D model fitting

Another line of 3D alignment research is to consider it as a part of 3D face surface reconstruction, and tackle it by fitting a dense 3D morphable model (3DMM) to the image (Jourabloo and Liu, 2016; Zhu et al., 2016). The 3DMM represents the dense 3D shape of a face with PCA:

$$\mathbf{A} = \overline{\mathbf{A}} + \mathbf{A}_{id}\alpha_{id} + \mathbf{A}_{exp}\alpha_{exp}, \tag{10}$$

where $\mathbf{A}$ is a dense 3D face, $\overline{\mathbf{A}}$ is the mean shape, $\mathbf{A}_{id}$ is the identity basis, and $\mathbf{A}_{\text{exp}}$ is the expression basis. In both Jourabloo and Liu (2016) and Zhu et al. (2016), the Basel 3D face model (Paysan et al., 2009) is used as the identity bases, and the face wearhouse (Cao et al., 2014b) is used as expression bases.

Traditional 3DMM fitting follows the analysis-by-synthesis principle (Blanz and Vetter, 1999; 2003), which however is inefficient and requires pre-located 2D facial landmarks. This motivates researchers to investigate discriminative fitting methods that directly learn the mapping from a 2D face image to the 3DMM and projection matrix (Jourabloo and Liu, 2016; Zhu et al., 2016). In particular, since this mapping is intrinsically non-linear, cascaded convolutional neural networks (CNN) are commonly used for 3DMM fitting. To facilitate the CNN regressor learning, both the global (Zhu et al., 2016) and local (Jourabloo and Liu, 2016) pose-invariant features have been exploited in literature.

In contrast to 3D shape regression approach, dense 3D model fitting can uncover the complete 3D shape of a face, rather than only localize a sparse set of points such as facial landmarks. The number of estimated landmarks is bounded by the number of 3D vertexes of the 3DMM, allowing us to estimate much more facial landmarks than conventional methods.

4.7.3. 3D face alignment in the wild challenge

To enable the comparison among different 3D methods, the first 3D Face Alignment in the Wild (3DFAW) Challenge (Jeni et al., 2016) is held in conjunction with the 14th European Conference on Computer Vision. The 3DFAW creates a dataset consisting of over 23,000 multi-view images with 3D annotation, makes the training and validation set available to participants, and tests their algorithms on an independent test-set. Eight teams participate in this challenge, but only four of them provide necessary technique descriptions (de Bittencourt Zavan et al., 2016; Bulat and Tzimiropoulos, 2016; Gou et al., 2016; Zhao et al., 2016). Next, we will briefly describe these four methods, and discuss what their results mean for current and future research.

Gou et al. (2016) propose to first estimate 2D facial landmarks via cascaded shape regression, and then recover 3D face shape by fitting a 3D Point Distribution Model. de Bittencourt Zavan et al. (2016) first detect the nose of the face with a Faster R-CNN (Ren et al., 2015), and then estimate the orientation of the face and place the average face landmark onto the face according to the face orientation. Zhao et al. (2016) propose a two-stage deep

[3] While 2D shape regression mainly refers to cascaded regression, we define 3D shape regression as the general regression techniques that can regress the 3D facial landmarks from a single image.

Table 6
Lists of face alignment performance evaluated on various wild face databases.

Databases	Challenges	#Test	#Points	Methods	Error (%)	FPS
LFW (Huang et al., 2007a)	Low resolution, large variations in illuminations, expressions and poses	13,233[a]	10	Conditional random forests (CRF) (Dantone et al., 2012)	7.00	10 (c++)
				Explicit shape regression (ESR) (Cao et al., 2012)	5.90	11 (Matlab)
				Robust Cascaded Pose Regression (RCPR) (Burgos-Artizzu et al., 2013)	5.30	15 (Matlab)
			55[b]	Consensus of Exemplar (CoE) (Belhumeur et al., 2013)	5.18	-
LFPW (Belhumeur et al., 2011)	Large variations in illuminations, expressions, poses and occlusion	224 ~ 300[c]	21	Consensus of Exemplar (CoE) (Belhumeur et al., 2011)	3.99	≈ 1 (C++)
				Explicit shape regression (ESR) (Cao et al., 2012)	3.47	220 (C++)
				Robust cascaded pose regression (RCPR) (Burgos-Artizzu et al., 2013)	3.50	12 (Matlab)
				Supervised descent method (SDM) (Xiong and De la Torre, 2013)	3.49	160 (C++)
				Exemplar-based graph matching (EGM) (Zhou et al., 2013b)	3.98	< 1
				Local binary feature (LBF) (Ren et al., 2014)	3.35	460 (C++)
				Fast local binary feature (LBF fast) (Ren et al., 2014)	3.35	4600 (C++)
			68[d]	Tree Structured Part Model (TSPM) (Zhu and Ramanan, 2012)	8.29	0.04 (Matlab)
				Discriminative Response Map Fitting (DRMF) (Asthana et al., 2013)	6.57	1 (Matlab)
				Robust Cascaded Pose Regression (RCPR) (Burgos-Artizzu et al., 2013)	6.56	12 (Matlab)
				Supervised descent method (SDM) (Xiong and De la Torre, 2013)	5.67	70 (C++)
				Gauss-Newton Deformable Part Model (GN-DPM) (Tzimiropoulos and Pantic, 2014)	5.92	70
				Coarse-to-fine Auto-encoder Networks (CFAN) (Zhang et al., 2014a)	5.44	20
				Coarse-to-fine Shape Searching (CFSS) (Zhu et al., 2015)	4.87	–
				CFSS Practical (Zhu et al., 2015)	4.90	–
				Deep Cascaded Regression (DCR) (Lai et al., 2015)	4.57	–
HELEN (Le et al., 2012)	Computation burden due to the dense annotation, large variations in expressions, poses and occlusion	330	194	Stacked Active Shape Model (STASM) (Milborrow and Nicolls, 2008)	11.10	–
				Component-based ASM (ComASM) (Le et al., 2012)	9.10	-
				Explicit Shape Regression (ESR) (Cao et al., 2012)	5.70	70 (C++)
				Robust Cascaded Pose Regression (RCPR) (Burgos-Artizzu et al., 2013)	6.50	6 (Matlab)
				Supervised Descent Method (SDM) (Xiong and De la Torre, 2013)	5.85	21 (C++)
				Ensemble of Regression Trees (ERT) (Kazemi and Josephine, 2014)	4.9	1000
				Local Binary Feature (LBF) (Ren et al., 2014)	5.41	200 (C++)
				Fast Local Binary Feature (LBF fast) (Ren et al., 2014)	5.80	1500 (C++)
				Coarse-to-Fine Shape Searching (CFSS) (Zhu et al., 2015)	4.74	-
				CFSS Practical (Zhu et al., 2015)	4.84	-
				cascade Gaussian Process Regression Trees (cGPRT) (Lee et al., 2015)	4.63	-
			68[d]	Tree Structured Part Model (TSPM) (Zhu and Ramanan, 2012)	8.16	0.04 (Matlab)
				Discriminative Response Map Fitting (DRMF) (Asthana et al., 2013)	6.70	1 (Matlab)
				Robust Cascaded Pose Regression (RCPR) (Burgos-Artizzu et al., 2013)	5.93	12 (Matlab)
				Supervised Descent Method (SDM) (Xiong and De la Torre, 2013)	5.67	70 (C++)
				Gauss-Newton Deformable Part Model (GN-DPM) (Tzimiropoulos and Pantic, 2014)	5.69	70
				Coarse-to-fine Auto-encoder Networks (CFAN) (Zhang et al., 2014a)	5.53	20
				Coarse-to-Fine Shape Searching (CFSS) (Zhu et al., 2015)	4.63	-
				CFSS Practical (Zhu et al., 2015)	4.72	-
				Deep Cascaded Regression (Lai et al., 2015)	4.25	-
300-W (Sagonas et al., 2013)	Large variations in illuminations, expressions, poses and occlusion	689	68	Tree Structured Part Model (TSPM) (Zhu and Ramanan, 2012)	12.20	0.04 (Matlab)
				Discriminative Response Map Fitting (DRMF) (Asthana et al., 2013)	9.10	1 (Matlab)
				Explicit Shape Regression (ESR) (Cao et al., 2012)	5.28	120 (C++)
				Robust Cascaded Pose Regression (RCPR) (Burgos-Artizzu et al., 2013)	8.35	-
				Supervised Descent Method (SDM) (Xiong and De la Torre, 2013)	7.50	70 (C++)
				Ensemble of Regression Trees (ERT) (Kazemi and Josephine, 2014)	6.4	1000
				Local Binary Feature (LBF) (Ren et al., 2014)	6.32	320 (C++)
				Fast Local Binary Feature (LBF fast) (Ren et al., 2014)	7.37	3100 (C++)
				Coarse-to-Fine Shape Searching (CFSS) (Zhu et al., 2015)	5.76	25
				CFSS Practical (Zhu et al., 2015)	5.99	25
				cascade Gaussian Process Regression Trees (cGPRT) (Lee et al., 2015)	5.71	93
				fast cGPRT (Lee et al., 2015)	6.32	871
				Tasks-Constrained Deep Convolutional Network (TCDCN) (Zhang et al., 2016)	5.54	59
				Deep Cascaded Regression (DCR) (Lai et al., 2015)	5.02	-
				Megvii-Face++ (Huang et al., 2015)	4.54	-

(*continued on next page*)

Table 6 (*continued*)

Databases	Challenges	#Test	#Points	Methods	Error (%)	FPS
IBUG (Sagonas et al., 2013)	Extremely large variations in illuminations, expressions, poses and occlusion	135	68	Tree Structured Part Model (TSPM) (Zhu and Ramanan, 2012)	18.33	0.04 (Matlab)
				Discriminative Response Map Fitting (DRMF) (Asthana et al., 2013)	19.79	1 (Matlab)
				Explicit Shape Regression (ESR) (Cao et al., 2012)	17.00	120 (C++)
				Robust Cascaded Pose Regression (RCPR) (Burgos-Artizzu et al., 2013)	17.26	-
				Supervised Descent Method (SDM) (Xiong and De la Torre, 2013)	15.40	70 (C++)
				Local Binary Feature (LBF) (Ren et al., 2014)	11.98	320 (C++)
				Fast Local Binary Feature (LBF fast) (Ren et al., 2014)	15.50	3100 (C++)
				Robust Discriminative Hough Voting (RDHV) (Jin and Tan, 2016)	11.32	< 1 (Matlab)
				Coarse-to-Fine Shape Searching (CFSS) (Zhu et al., 2015)	9.98	25
				CFSS Practical (Zhu et al., 2015)	10.92	25
				Tasks-Constrained Deep Convolutional Network (TCDCN) (Zhang et al., 2016)	8.60	59
				Deep Cascaded Regression (DCR) (Lai et al., 2015)	8.42	-
				Megvii-Face++ (Huang et al., 2015)	7.46	-

[a] For LFW, the reported performance of Burgos-Artizzu et al. (2013); Cao et al. (2012); Dantone et al. (2012) follows the evaluation procedure proposed in Dantone et al. (2012), consisting of a ten-fold cross validation using each time 1500 training images and the rest for testing. In Belhumeur et al. (2013), the model is trained on Columbia's PubFig (Kumar et al., 2009), and tested on all 13,233 images of LFW.

[b] Although used by Belhumeur et al. (2013), the 55 point annotation of LFW is not shared.

[c] LFPW is shared by web URLs, but some URLs are no longer valid. So both the training and test images downloaded by other authors are less than the original version (1,100 training images and 300 test images).

[d] LFPW and HELEN are originally annotated with 29 and 194 points respectively, while later Sagonas et al. (2013)) re-annotate them with 68 points. Some authors reported their performance on the 68 points version of these databases.

learning approach that first estimates 2D facial landmarks with a deep CNN, and then estimates the depth of the landmarks with another deep neural network. Bulat and Tzimiropoulos (2016) proposed a two-stage alignment method using deep residual network (He et al., 2016). It first calculates heat-maps of 2D landmarks using convolutional part heat-map regression, then uses these heat-maps along with the original RGB image as an input to a very deep residual network to regress the depth information.

From above four reported methods, we can observe that cascaded regression and CNN-based regression, which are popular in 2D alignment, are also the mainstream technologies for 3D face alignment. In particular, most researchers decompose the 3D face alignment problem into two steps: 2D estimation and depth estimation, allowing us to efficiently leverage existing state-of-the-art 2D face alignment techniques. We note that the two deep learning-based approaches (Bulat and Tzimiropoulos, 2016; Zhao et al., 2016) have achieved the top 2 performance in 3DFW challenge, which confirms the power of feature learning in 3D face alignment. Bulat and Tzimiropoulos (2016) is the top 1 performer in the 3DFAW Challenge, surpassing the second best result by more than 22%. This may be attributed to two factors: the powerful residual learning, and the part heatmap regression that learns where to "look" during depth estimation by explicitly exploiting information about the 2D location of the landmarks.

4.8. Summary and discussion

We have reviewed discriminative methods for face alignment in seven groups, i.e., *CLMs, constrained local regression, DPMs, ensemble regression-voting, cascaded regression, deep neural networks* and *3D alignment methods*. Among them, CLMs, constrained local regression and DPMs follow the "divide and conquer" principle to simplify the face alignment task by constructing individual local appearance model for each facial point. However, due to their small patch support and large appearance variation in training, these local appearance models are typically plagued by the problem of ambiguity. Since the further inference (global shape optimization) is based on the detection responses of these local appearance models, the problem of ambiguity may create the most serious performance bottleneck.

To break this bottleneck, another main stream in face alignment is to jointly estimate the whole face shape from image, implicitly exploiting the spatial constraints among facial points. In this line, we have first reviewed the *ensemble regression-voting* and *cascaded regression* methods, which learn a vectorial regression function to infer the whole face shape in an ensemble or cascaded manner. In particular, cascaded regression has emerged as one of the most popular and state-of-the-art methods, due to its speed, accuracy and robustness. Then, we briefly reviewed the deep learning-based approach for face alignment, which have the advantage of learning highly discriminative task-specific features, but should take into account the issue of over-fitting. Finally, we reviewed the 3D alignment methods that treat the face as a 3D object, which has the advantage in solving the large pose face alignment problem.

It is worth noting that some methods involve techniques motivated by different principles, which clearly overlap our category boundaries. For example, we classify the regression voting-based shape model matching method (Cootes et al., 2012) as CLM, since they fit a parametric shape model to a new image based on the response map for each facial point. However, since the response maps in Cootes et al. (2012) are generated by random forest regression-voting, it can also be considered as an ensemble regression-voting method. Furthermore, some deep learning-based methods can also be classified as cascaded regression due to their cascaded structure (Lai et al., 2015; Zhang et al., 2014a).

5. Towards the development of a robust face alignment system

Face alignment *in-the-wild* is very challenging due to many kinds of undesirable appearance variations, and hence it is often the case that no single modality is enough. In this section, we will focus on the practical aspects of constructing a robust face alignment system, which is mostly ignored in previous studies. Specifically, we first present a global system architecture for face alignment, and then have a close look at possible strategies to improve the robustness of face alignment under this architecture.

5.1. The global system architecture for face alignment

Inspired by Song et al. (2013) and Fasel and Luettin (2003), we give a global system architecture for face alignment, where a complicated system is divided into several substages. As shown in Fig. 7, the architecture can be roughly divided into three parts: face preprocessing, shape initialization, and the iterative process of feature extraction and shape prediction. We note that this architecture is only to illustrate a general pipeline for face alignment, while in practical not all components are mandatory. For example, the consensus of exemplar method (Belhumeur et al., 2011) do not involve the shape initialization step.

While the feature extraction and shape prediction process have drawn a great deal of attention in literature, the face preprocessing and shape initialization steps are often ignored. Meanwhile, problems such as training data augmentation, and the accuracy and efficiency tradeoff are also essential for any practical face alignment system. In the following, we will have a closer look at these issues.

5.2. Training data augmentation

Due to the difficulty and cost of manual annotation, the number of training samples we *actually* have is often much smaller than that we *supposedly* have. In such a case, artificial data augmentation, which is usually done by label-preserving transforms, is the easiest and most common method to reduce over-fitting.

In general, there are four distinct forms of data augmentation to enlarge the training set: (1) generating image rotations from a small interval (e.g., $[-15^\circ, +15^\circ]$ used in Belhumeur et al. (2013)); (2) synthesizing images by left-right flip to double the training set; (3) disturbing the bounding boxes by randomly scaling and translating the bounding box for each image, which also increases the robustness of face alignment algorithms to the bounding boxes; (4) sampling multiple initialization for each training image, which is typically used by cascaded regression methods.

5.3. Face preprocessing

For the task of face alignment, it is useful to remove the scaling variations of the detected faces, and enlarge the face region to ensure that all predefined facial points are enclosed.

5.3.1. Handling scaling variations

Typically, for a face analysis system, the training and test faces are required to be roughly the same scale, by rescaling the bounding box produced by the face detector. We note that to help preserve more detailed texture information, the size of the normalized bounding box for high-resolution face databases is typically chosen to be larger than that for low-resolution face databases. For example, Belhumeur et al. (2013) rescale the high-resolution images from the LFPW database so that the faces have an inter-ocular distance of roughly 55 pixels, while Dantone et al. (2012) choose to rescale the bounding box of the low-resolution faces from the LFW database (Huang et al., 2007a) to {100,100}, which is slightly smaller than the size chosen by Belhumeur et al. (2013).

5.3.2. Enlarging face areas

The output of a face detector is a rough face region that might miss some facial points (e.g., the chin). This has little impact on cascaded regression, for which the bounding box only serves to rescale the face and compute the initial shape. However, for those methods based on exhaustive search or feature voting, it is necessary to enlarge the face bounding box to enclose all the facial points, or define the sampling region of image patches to cast votes. For this, Dantone et al. (2012) suggest to enlarge the face bounding box by 30%, and we believe that this strategy may satisfy the requirements of all face alignment algorithms.

5.4. Shape initialization

Most face alignment methods start from a rough initialization, and then refine the shape iteratively until convergence. The initialization step typically has great influence on the final result, and an initial shape far from the ground truth might lead to very bad alignment results.

The most common choice is to use the *mean shape* for initialization (Kazemi and Josephine, 2014; Ren et al., 2014; Xiong and De la Torre, 2013). However, sometimes, the mean shape is likely to be far from the target shape, and leads to bad result. As an alternative, Cao et al. (2012) propose to run the algorithm several times using different initialisations *randomly* sampled from the training shapes, and take the median result as the final estimation to improve robustness. Burgos-Artizzu et al. (2013) proposed a smart restart method to further improve the multiple initialization strategy in Cao et al. (2012) by checking the variance between the predictions using different initializations.

Recently, some authors proposed to estimate an initial shape that is tailored to the input face. Zhang et al. (2014c) showed that the five major facial points localized by their deep model can serve as anchor points to apply similarity transform to randomly sampled training shapes. Through this, very accurate initial shapes can be generated for other algorithms (e.g., Burgos-Artizzu et al. (2013)) and lead to promising performance improvement. Zhang et al. (2014a) and Sun et al. (2013) proposed to directly estimate a rough initial shape from the global image, which in general produces good initial shape that aids following alignment.

5.5. Accuracy and efficiency tradeoffs

Face alignment in real time is crucial to many practical applications. The efficiency mainly depends on the feature extraction and shape prediction steps. In general, strong hand-designed feature (e.g., SIFT Lowe (2004)) captures detailed texture information that may aid detection, but have higher computational cost compared to simpler features (e.g., BRIEF Calonder et al. (2010)). Zhu et al. (2015) identified this phenomenon under the cascaded regression framework, and proposed to exploit different types of features at different stages to achieve a good trade-off between accuracy and efficiency, i.e., employ less accurate but computationally efficient BRIEF feature at the early stages, and use more accurate but relatively slow SIFT feature at later stages. Besides this hybrid strategy, a better choice is to learn highly efficient and discriminative features (Cao et al., 2012; Kazemi and Josephine, 2014; Ren et al., 2014). In particular, Ren et al. (2014) propose to learn a set of highly discriminative local binary features for each facial point independently. Because extracting and regressing local binary features is computationally very cheap, Ren et al. (2014) achieves over 3000 FPS while obtaining accurate alignment result.

In term of shape prediction, the regression-based methods in general are very efficient, while the exhaustive search based methods typically suffer from high computational cost (Belhumeur et al., 2011; Zhou et al., 2013b). Dibeklioğlu et al. (2012) propose to mitigate this issue through a coarse-to-fine search strategy. In Dibeklioğlu et al. (2012), a three-level image pyramid from the cropped high-resolution face images is designed to reduce the search region, where the coarse-level images have lower resolution but much smaller size.

6. System evaluation

In this section, we first review the major wild face databases and evaluation metric in the literature, then summarize and discuss some of reported performance of current state-of-the-art, on

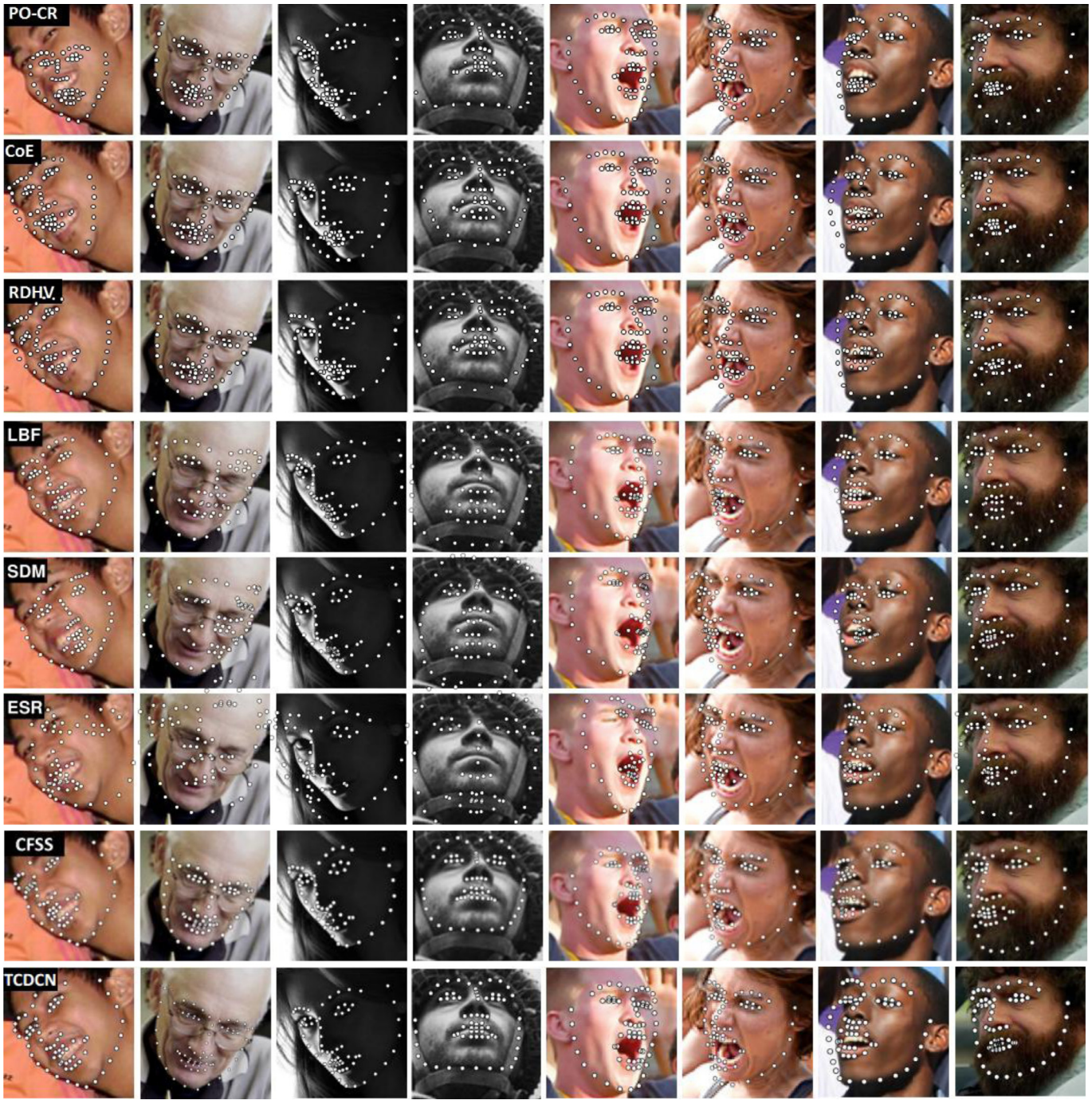


Fig. 10. Example results on IBUG database (Sagonas et al., 2013a) by eight state-of-the-art methods. These images are extremely difficult due to the mixing of large head poses, extreme lighting, and partial occlusions. From top to bottom, results are produced by the Project-Out Cascaded Regression (PO-CR) method (Tzimiropoulos, 2015), Consensus of Exemplar (CoE) method (Belhumeur et al., 2013), Robust Discriminative Hough Voting (RDHV) method (Jin and Tan, 2016), Local Binary Feature (LBF) method (Ren et al., 2014), Supervised Descent Method (SDM) (Xiong and De la Torre, 2013), Explicit Shape Regression (ESR) method (Cao et al., 2012), Coarse-to-Fine Shape Searching (CFSS) method (Zhu et al., 2015), Tasks-Constrained Deep Convolutional Network (TCDCN) method (Zhang et al., 2016). Among these methods, we implement the Consensus of Exemplar (CoE) (Belhumeur et al., 2013) and Robust Discriminative Hough Voting (RDHV) (Jin and Tan, 2016) methods and test them on these images, while other results are obtained from the published papers.

the several popular wild face databases using the same evaluation metric for reference. Note that we do not include the evaluation of 3D face alignment methods emerged recently, as the common dataset 3DFAW (Jeni et al., 2016) has not been available to the public. For this, we refer the readers to the paper of the first 3D Face Alignment in the Wild (3DFAW) Challenge (Jeni et al., 2016).

6.1. Databases and metric

6.1.1. Databases

There have been many face databases developed for face alignment, with the ground truth facial points labelled manually by employing workers or through the tools such as Amazon mechanical turk (MTurk). Among them, some databases are col-

lected under controlled laboratory conditions with normal lighting, neutral expression and high image quality, including the Extended M2VTS database (XM2VTS) (Messer et al., 1999), BioID face database (Jesorsky et al., 2001), PUT (Kasinski et al., 2008), Multi-Pie (Gross et al., 2010), etc.

However, the goal of this paper is to investigate the problem of face alignment *in-the-wild*, so we are more concerned with the *uncontrolled* databases that exhibit large facial variations due to pose, expressions, lighting, occlusion and image quality. These uncontrolled databases are typically collected from social network such as google.com, flickr.com, facebook.com, which are more realistic and challenging for face alignment. In Table 4, we list the basic information of 9 wild face databases, including LFW (Huang et al., 2007a), LFPW (Belhumeur et al., 2011), AFLW (Köstinger et al., 2011), AFW (Zhu and Ramanan, 2012), HELEN (Le et al., 2012), 300-W (Sagonas et al., 2013), COFW (Burgos-Artizzu et al., 2013), MTFL (Zhang et al., 2014c), and MAFL (Zhang et al., 2016), and also provide links to download them. The example face images from these databases with original annotation are illustrated in Fig. 8. It is worth noting that the LFPW, AFW and HELEN databases are re-annotated by Sagonas et al. (2013)) with 68 points.

6.1.2. Evaluation metric

There have been several evaluation metrics for the alignment accuracy in the literature. For example, many authors reported the inter-pupil distance normalized facial point error averaged over all facial points and images for each database (Burgos-Artizzu et al., 2013; Kazemi and Josephine, 2014; Lee et al., 2015; Ren et al., 2014; Zhu et al., 2015). Specifically, the inter-ocular distance normalized error for facial point i is defined as:

$$e_i = \frac{||\mathbf{x}_i - \mathbf{x}_i^*||_2}{d_{IOD}}, \tag{11}$$

where $\mathbf{x}_i$ is the automatically localized facial point location, $\mathbf{x}_i^*$ is the manually annotated location, and d_{IOD} is the inter-ocular distance. The normalization term d_{IOD} in this formulation can eliminate unreasonable measurement variations caused by variations of face scales.

The cumulative errors distribution (CED) curve is also often chosen to illustrate the comparative performance, showing the proportion of the test images or facial points with the increase of the normalized error (Belhumeur et al., 2011; Saragih et al., 2011; Tzimiropoulos, 2015; Tzimiropoulos and Pantic, 2014; Zhu et al., 2015). Some other evaluation metric can also been found in literature, such as the facial point error normalized by face size (Yu et al., 2013), the percentage of the test images or facial points less than given relative error level (Dibeklioğlu et al., 2012; Yu et al., 2013), and the percentage of accuracy improvement over other algorithm (Cao et al., 2012).

Besides the accuracy, the efficiency is another important performance indicator of face alignment algorithms, which is typically measured by frames per second (FPS).

6.2. Evaluation and discussion

We choose four common wild databases, i.e., LFW, LFPW, HELEN, 300 W and IBUG (challenging subset of 300 W) databases, to show comparative performance statistics of the state of the art. Table 5 lists some softwares published online, and Table 6 summarizes the reported performance on above databases. Fig. 10 shows some challenging images from IBUG aligned by eight state-of-the-art methods respectively.

For performance evaluation, we are mainly concerned with two key performance indicators, i.e., accuracy and efficiency. The former is measured by the normalized facial point error (cf. Eq. (11)) averaged over all facial points and images for each database, while the later is measured by frames per second (FPS).

6.2.1. Accuracy

As shown in Table 6, the localization error on all these databases has been reduced to less than 10% of the inter-ocular distance by current state-of-the-art. Except for the extremely challenging IBUG database, the best performance on other databases is about 5% of the inter-ocular distance. To have an intuitive feeling of the extent of localization error, we exemplify the error range of 10% and 5% of the inter-ocular distance respectively in Fig. 9 (a) and (b). This implies that most of the localized facial points by the state-of-the-art may lie in the error range depicted by the white circles in Fig. 9 (a), while on LFPW annotated with 29 points, the mean error range goes to the white circles in Fig. 9 (b). Besides the statistics listed in Table 6, some authors also compared their methods with human beings and reported close to human performance on LFPW (Belhumeur et al., 2011; Burgos-Artizzu et al., 2013) and LFW (Dantone et al., 2012).

From Table 6, we can observe that although generative methods (e.g., the GN-DPM (Tzimiropoulos and Pantic, 2014)) can produce good performance for face alignment *in-the-wild*, discriminative methods, especially those based on cascaded regression (Burgos-Artizzu et al., 2013; Cao et al., 2012; Huang et al., 2015; Kazemi and Josephine, 2014; Lai et al., 2015; Ren et al., 2014; Xiong and De la Torre, 2013; Zhu et al., 2015), have been playing a dominate role for this task, partially due to recent development of large unconstrained databases. Furthermore, the deep learning-based approach (Huang et al., 2015; Sun et al., 2013; Zhang et al., 2014c; 2016) have recently emerged as a popular and state-of-the-art method due to their strong feature learning capability, achieving very accurate (even the best) performance on the challenging 300-W and IBUG databases (Sagonas et al., 2013).

Fig. 10shows some extremely challenging cases on IBUG aligned by eight state-of-the-art methods, from which we can observe that large head poses, extreme lighting, and partial occlusions may pose major challenges for many advanced face alignment algorithms, but good results can still be achieved by some state-of-the-art, for example, by the Tasks-Constrained Deep Convolutional Network (TCDCN) method (Zhang et al., 2016). Furthermore, we find the Fig. 10 that: (1) Compared to other facial points, the points around the outline of the face are much more difficult to localize, due to the lack of distinctive local texture. (2) As the points around the mouth are heavily dependent on facial expressions, they are more difficult to localize than those points insensitive to facial expressions, such as the points along the eyebrows, outer corners of the eyes, and the nose tips.

Finally, we have to highlight that the accuracy statistics listed in Table 6 may not fully characterize the behavior of these algorithms, since several factors can complicate the assessment. First, even for the same algorithm, different experimental details and programming skills may results in different performance. Secondly, while the number and variety of training examples have a direct effect on the final performance, the training data of some released software is not clear. Thirdly, as pointed by Yang et al. (2015), the performance of many algorithms is sensitive to the face detection variation, but different systems may employ different face detectors. For example, SDM (Xiong and De la Torre, 2013) employs the Viola Jones detector (Viola and Jones, 2004), while GN-DPM (Tzimiropoulos and Pantic, 2014) uses the in-house face detector of the IBUG group.

6.2.2. Efficiency

Besides accuracy, efficiency is another key performance indicator of face alignment algorithms. In the last column of Table 6, we report the efficiency of some algorithms, and highlight the implementation types of them (Matlab or C++). In general, the running time listed here is consistent with the algorithm's complexity. For example, algorithms that involves an exhaustive search of

local detectors typically have a high time cost (Belhumeur et al., 2011; Zhou et al., 2013b; Zhu and Ramanan, 2012), while the cascaded regression methods are extremely fast since both the shape-index feature and the stage regression are very efficient to compute (Burgos-Artizzu et al., 2013; Cao et al., 2012; Xiong and De la Torre, 2013). It is worth noting that impressive speed (more than 1000 FPS for 194 points on HELEN) has been achieved by the local binary feature (LBF) (Ren et al., 2014) and ensemble of regression trees (ERT) (Kazemi and Josephine, 2014), using learning-based features.

7. Conclusion and prospect

Face alignment is an important and essential intermediary step for many face analysis applications. Such a task is extremely challenging in unconstrained environments due to the complexity of facial appearance variations. However, extensive studies on this problem have resulted in a great amount of achievements, especially during the last few years.

In this paper, we have focused on the overall difficulties and challenges in unconstrained environments, and provide a comprehensive and critical survey of the current state of the art in dealing with these challenges. Furthermore, we hope that the practical aspects of face alignment we organized can provide further impetus for high-performance, real-time, real-life face alignment systems. Finally, it is worth mentioning that some closely related problems are deliberately ignored in this paper, such as facial feature tracking in videos (Ahlberg, 2001; Kapoor and Picard, 2002), which are also very important in practice.

Despite of many efforts devoted to face alignment during the last two decades, we have to admit that this problem is far from being solved, and several general promising research directions could be suggested.

- *Challenging databases collection:* Besides new methodologies, another notable development in the field of face alignment has been the collection and annotation of large facial datasets captured *in-the-wild* (cf., Table 4). But even so, we argue that the collection of challenging databases is still important and has the potential to boost the performance of existing methods. This argument can be partially supported by the fact that: the performance of most algorithms on IBUG is inferior to that on other databases such as LFPW and HELEN, as the training set of these algorithms is typically less challenging compared to IBUG.
- *Feature learning:* One of the holy grails of machine learning is to automate more and more of the feature engineering process (Domingos, 2012), i.e., to learn task-specific features in a data-driven manner. In the field of face alignment, many approaches that employ feature learning techniques, including both shallow feature learning (Burgos-Artizzu et al., 2013; Cao et al., 2012; Ren et al., 2014) and deep learning (Huang et al., 2015; Sun et al., 2013) methods, have achieved state-of-the-art performances. We believe that, with the assistance of abundant manually labeled images, automatic feature learning techniques can be a powerful weapon for triumphing over various challenges of face alignment in the wild, and deserve the efforts and smarts of researchers.
- *Multi-task learning:* Multi-task learning aims to improve the generalization performance of multiple related tasks by learning them jointly, which has proven effective in many computer vision problems (Yuan et al., 2012; Zhang et al., 2013). For face alignment *in-the-wild*, on the one hand, many factors such as pose, expression and occlusion may pose great challenges; while on the other hand, these factors can be considered jointly with face alignment to expect an improvement of robustness. This has been confirmed by the work of Zhang et al. (2014c), which proposes to exploit the power of multi-task learning under the deep convolutional network architecture, leading to a better performance compared to single task-based deep model. Although some attempts have been proposed, we believe that multi-task learning remains a meaningful and promising direction for face alignment in future.
- *3D face alignment:* 2D face alignment has be extensively studied in literature. But as mentioned in Section 4.7, 2D methods are mainly designed for faces in small to medium poses (below 45°). As face orientation varies from frontal to profile, 2D annotations (e.g., cheek landmarks) may lose correspondence. In this setting, 3D face alignment from 2D images has been proposed as a potential solution. Although a number of promising 3D methods have be proposed recently (Bulat and Tzimiropoulos, 2016; Jourabloo and Liu, 2015; 2016; Tulyakov and Sebe, 2015), we believe that 3D face alignment is a novel and important topic that deserves ongoing effort.

We believe that face alignment *in-the-wild* is a very exciting line of research due to its inherent complexity and wide practical applications, and will draw increasing attention from computer vision, pattern recognition and machine learning.

Acknowledgments

This work is partially supported by National Science Foundation of China (61373060,61672280), Qing Lan Project, and the Funding of Jiangsu Innovation Program for Graduate Education (KYLX_0289).

References

Ahlberg, J., 2001. Using the active appearance algorithm for face and facial feature tracking. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. IEEE, pp. 68–72.
Antonakos, E., Alabort-i Medina, J., Tzimiropoulos, G., Zafeiriou, S., 2014. Hog active appearance models. In: Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, pp. 224–228.
Antonakos, E., Alabort-i Medina, J., Tzimiropoulos, G., Zafeiriou, S.P., 2015. Feature-based lucas–kanade and active appearance models. Image Process. IEEE Trans. 24 (9), 2617–2632.
Antonakos, E., Alabort-i Medina, J., Zafeiriou, S., 2015. Active pictorial structures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5435–5444.
Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M., 2013. Robust discriminative response map fitting with constrained local models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3444–3451.
Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M., 2014. Incremental face alignment in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1859–1866.
Bailenson, J.N., Pontikakis, E.D., Mauss, I.B., Gross, J.J., Jabon, M.E., Hutcherson, C.A., Nass, C., John, O., 2008. Real-time classification of evoked emotions using facial feature tracking and physiological responses. Int. J. Hum. Comput. Stud. 66 (5), 303–317.
Baker, S., Matthews, I., 2001. Equivalence and efficiency of image alignment algorithms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1. IEEE, pp. I–1090.
Baltrušaitis, T., Robinson, P., Morency, L.-P., 2012. 3d constrained local model for rigid and non-rigid facial tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2610–2617.
Baltrusaitis, T., Robinson, P., Morency, L.-P., 2013. Constrained local neural fields for robust facial landmark detection in the wild. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 354–361.
Baltrušaitis, T., Robinson, P., Morency, L.-P., 2014. Continuous conditional neural fields for structured regression. In: European Conference on Computer Vision. Springer, pp. 593–608.
Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. Speeded-up robust features (surf). Comput. Vision Image Understand. 110 (3), 346–359.
Belhumeur, P., Jacobs, D., Kriegman, D., Kumar, N., 2013. Localizing parts of faces using a consensus of exemplars.. IEEE Trans. Pattern Anal. Mach. Intell. 35 (12), 2930–2940.
Belhumeur, P.N., Jacobs, D.W., Kriegman, D., Kumar, N., 2011. Localizing parts of faces using a consensus of exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 545–552.
de Bittencourt Zavan, F.H., Nascimento, A.C., e Silva, L.P., Bellon, O.R., Silva, L., 2016. 3d face alignment in the wild: a landmark-free, nose-based approach. In: European Conference on Computer Vision. Springer, pp. 581–589.

Blanz, V., Vetter, T., 1999. A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co., pp. 187–194.

Blanz, V., Vetter, T., 2003. Face recognition based on fitting a 3d morphable model. IEEE Trans. Pattern Anal. Mach. Intell. 25 (9), 1063–1074.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Bulat, A., Tzimiropoulos, G., 2016. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In: European Conference on Computer Vision. Springer, pp. 616–624.

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

Calonder, M., Lepetit, V., Strecha, C., Fua, P., 2010. Brief: binary robust independent elementary features. In: European Conference on Computer Vision, pp. 778–792.

Campadelli, P., Lanzarotti, R., Savazzi, C., 2003. A feature-based face recognition system. In: Image Analysis and Processing, 2003. Proceedings. 12th International Conference on. IEEE, pp. 68–73.

Cao, C., Hou, Q., Zhou, K., 2014. Displaced dynamic expression regression for real–time facial tracking and animation. ACM Trans. Graph. 33 (4), 43.

Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K., 2014. Facewarehouse: a 3d facial expression database for visual computing. IEEE Trans. Vis. Comput. Graph. 20 (3), 413–425.

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

Çeliktutan, O., Ulukaya, S., Sankur, B., 2013. A comparative study of face landmarking techniques. EURASIP J. Image Video Process. 2013 (1), 13.

Chow, C., Liu, C., 1968. Approximating discrete probability distributions with dependence trees. Inf. Theory IEEE Trans. 14 (3), 462–467.

Cootes, T.F., Edwards, G.J., Taylor, C.J., 1998. Active appearance models. In: European Conference on Computer Vision. Springer, pp. 484–498.

Cootes, T.F., Edwards, G.J., Taylor, C.J., 2001. Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. 23 (6), 681–685.

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

Cootes, T.F., Taylor, C.J., 1992. Active shape modelssmart snakes. In: BMVC92. Springer, pp. 266–275.

Cootes, T.F., Taylor, C.J., 1993. Active shape model search using local grey-level models: a quantitative evaluation.. In: BMVC, 93. Citeseer, pp. 639–648.

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models-their training and application. Comput. Vis. Image Understand. 61 (1), 38–59.

Cristinacce, D., Cootes, T.F., 2006. Feature detection and tracking with constrained local models.. In: BMVC, 2, p. 6.

Cristinacce, D., Cootes, T.F., 2007. Boosted regression active shape models.. In: BMVC, pp. 1–10.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1. IEEE, 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 the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2578–2585.

Dibeklioğlu, H., Salah, A.A., Gevers, T., 2012. A statistical method for 2-d facial landmarking. Image Process. IEEE Trans. 21 (2), 844–858.

Ding, L., Martinez, A.M., 2010. Features versus context: an approach for precise and detailed detection and delineation of faces and facial features. IEEE Trans. Pattern Anal. Mach. Intell. 32 (11), 2022–2038.

Domingos, P., 2012. A few useful things to know about machine learning. Commun. ACM 55 (10), 78–87.

Fasel, B., Luettin, J., 2003. Automatic facial expression analysis: a survey. Pattern Recognit. 36 (1), 259–275.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. Pattern Anal. Mach. Intell. IEEE Trans. 32 (9), 1627–1645.

Felzenszwalb, P.F., Huttenlocher, D.P., 2005. Pictorial structures for object recognition. Int. J. Comput. Vis. 61 (1), 55–79.

Fukunaga, K., Hostetler, L.D., 1975. The estimation of the gradient of a density function, with applications in pattern recognition. Inf. Theory IEEE Trans. 21 (1), 32–40.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587.

Gou, C., Wu, Y., Wang, F.-Y., Ji, Q., 2016. Shape augmented regression for 3d face alignment. In: European Conference on Computer Vision. Springer, pp. 604–615.

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

Gross, R., Matthews, I., Baker, S., 2003. Lucas-kanade 20 years on: a unifying framework: part 3. Cmu-ri-tr-03-05, CMU.

Gross, R., Matthews, I., Baker, S., 2005. Generic vs. person specific active appearance models. Image Vis. Comput. 23 (12), 1080–1093.

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

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

Hamsici, O.C., Martinez, A.M., 2009. Active appearance models with rotation invariant kernels. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE, pp. 1003–1009.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hsu, G.-S., Chang, K.-H., Huang, S.-C., 2015. Regressive tree structured model for facial landmark localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3855–3861.

Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E., 2007. E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments.

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

Huang, Z., Zhou, E., Cao, Z., 2015. Coarse-to-fine face alignment with multi-scale local patch regression. arXiv preprint arXiv:1511.04901.

Jaiswal, S., Almaev, T., Valstar, M., 2013. Guided unsupervised learning of mode specific models for facial point detection in the wild. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 370–377.

Jeni, L.A., Cohn, J.F., Kanade, T., 2015. Dense 3d face alignment from 2d videos in real-time. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, 1. IEEE, pp. 1–8.

Jeni, L.A., Tulyakov, S., Yin, L., Sebe, N., Cohn, J.F., 2016. The first 3d face alignment in the wild (3dfaw) challenge. In: European Conference on Computer Vision. Springer, pp. 511–520.

Jesorsky, O., Kirchberg, K.J., Frischholz, R.W., 2001. Robust face detection using the hausdorff distance. In: Audio-and video-based biometric person authentication. Springer, pp. 90–95.

Jin, X., Tan, X., 2016. Face alignment by robust discriminative hough voting. Pattern Recognit. 60, 318–333.

Jourabloo, A., Liu, X., 2015. Pose-invariant 3d face alignment. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3694–3702.

Jourabloo, A., Liu, X., 2016. Large-pose face alignment via cnn-based dense 3d model fitting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4188–4196.

Kapoor, A., Picard, R.W., 2002. Real-time, fully automatic upper facial feature tracking. In: Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on. IEEE, pp. 8–13.

Kasinski, A., Florek, A., Schmidt, A., 2008. The put face database. Image Process. Commun. 13 (3–4), 59–64.

Kazemi, V., Josephine, S., 2014. One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE.

Köstinger, M., Wohlhart, P., Roth, P.M., Bischof, H., 2011. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. IEEE, pp. 2144–2151.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105.

Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K., 2009. Attribute and simile classifiers for face verification. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE, pp. 365–372.

Lai, H., Xiao, S., Cui, Z., Pan, Y., Xu, C., Yan, S., 2015. Deep cascaded regression for face alignment. arXiv preprint arXiv:1510.09083.

Lanitis, A., Taylor, C.J., Cootes, T.F., 1997. Automatic interpretation and coding of face images using flexible models. Pattern Anal. Mach. Intell. IEEE Trans. 19 (7), 743–756.

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

Lee, D., Park, H., Yoo, C.D., 2015. Face alignment using cascade gaussian process regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4204–4212.

Li, H., Ding, H., Huang, D., Wang, Y., Zhao, X., Morvan, J.-M., Chen, L., 2015. An efficient multimodal 2d+ 3d feature-based approach to automatic facial expression recognition. Comput. Vis. Image Understand. 140, 83–92.

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

Liu, J., Li, Y., Allen, P., Belhumeur, P., 2015. Articulated pose estimation using hierarchical exemplar-based models. AAAI Conference on Artificial Intelligence. arXiv preprint arXiv:1512.04118.

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

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60 (2), 91–110.

Lucas, B.D., Kanade, T., et al., 1981. An iterative image registration technique with an application to stereo vision. IJCAI.

Lucey, S., Navarathna, R., Ashraf, A.B., Sridharan, S., 2013. Fourier lucas-kanade algorithm. Pattern Anal. Mach. Intell. IEEE Trans. 35 (6), 1383–1396.

Martinez, B., Valstar, M.F., Binefa, X., Pantic, M., 2013. Local evidence aggregation for regression-based facial point detection. Pattern Anal. Mach. Intell. IEEE Trans. 35 (5), 1149–1163.

Martins, P., Caseiro, R., Batista, J., 2013. Generative face alignment through 2.5 d active appearance models. Comput. Vis. Image Understand. 117 (3), 250–268.

Martins, P., Caseiro, R., Batista, J., 2014. Non-parametric bayesian constrained local models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1797–1804.

Matthews, I., Baker, S., 2004. Active appearance models revisited. Int. J. Comput. Vis. 60 (2), 135–164.
Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G., 1999. Xm2vtsdb: The extended m2vts database. In: Second international conference on audio and video-based biometric person authentication, 964. Citeseer, pp. 965–966.
Milborrow, S., Nicolls, F., 2008. Locating facial features with an extended active shape model. In: European Conference on Computer Vision. Springer, pp. 504–513.
Papandreou, G., Maragos, P., 2008. Adaptive and constrained algorithms for inverse compositional active appearance model fitting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8.
Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T., 2009. A 3d face model for pose and illumination invariant face recognition. In: Advanced Video and Signal based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on. IEEE, pp. 296–301.
Ren, S., Cao, X., Wei, Y., Sun, J., 2014. Face alignment at 3000 fps via regressing local binary features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE.
Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99.
Rudovic, O., Patras, I., Pantic, M., 2010. Coupled gaussian process regression for pose-invariant facial expression recognition. In: European Conference on Computer Vision. Springer, pp. 350–363.
Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M., 2013. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. IEEE, pp. 397–403.
Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M., 2013. A semi-automatic methodology for facial landmark annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE, pp. 896–903.
Saragih, J., Goecke, R., 2007. A nonlinear discriminative approach to aam fitting. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE, pp. 1–8.
Saragih, J.M., Lucey, S., Cohn, J.F., 2011. Deformable model fitting by regularized landmark mean-shift. Int. J. Comput. Vis. 91 (2), 200–215.
Sauer, P., Cootes, T.F., Taylor, C.J., 2011. Accurate regression procedures for active appearance models.. In: BMVC, pp. 1–11.
Senechal, T., Rapp, V., Salam, H., Seguier, R., Bailly, K., Prevost, L., 2011. Combining aam coefficients with lgbp histograms in the multi-kernel svm framework to detect facial action units. In: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on. IEEE, pp. 860–865.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
Smith, B.M., Brandt, J., Lin, Z., Zhang, L., 2014. Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1741–1748.
Song, F., Tan, X., Chen, S., Zhou, Z.-H., 2013. A literature survey on robust and efficient eye localization in real-life scenarios. Pattern Recognit. 46 (12), 3157–3173.
Stegmann, M.B., Ersbøll, B.K., Larsen, R., 2003. Fame-a flexible appearance modeling environment. Med. Imag. IEEE Trans. 22 (10), 1319–1331.
Stegmann, M. B., Olsen, S. I., 2001. Object tracking using active appearance models.
Sun, Y., Wang, X., Tang, X., 2013. Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3476–3483.
Sun, Y., Wang, X., Tang, X., 2014. Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1891–1898.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.
Taigman, Y., Yang, M., Ranzato, M., Wolf, L., 2014. Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708.
Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S., 2016. Mnemonic descent method: a recurrent process applied for end-to-end face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4177–4187.
Tulyakov, S., Sebe, N., 2015. Regressing a 3d face shape from a single image. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3748–3755.
Turk, M.A., Pentland, A.P., 1991. Face recognition using eigenfaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 586–591.
Tzimiropoulos, G., 2015. Project-out cascaded regression with an application to face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3659–3667.
Tzimiropoulos, G., Alabort-i Medina, J., Zafeiriou, S., Pantic, M., 2012. Generic active appearance models revisited. In: Computer Vision–ACCV 2012. Springer, pp. 650–663.
Tzimiropoulos, G., Alabort-i Medina, J., Zafeiriou, S., Pantic, M., 2014. Active orientation models for face alignment in-the-wild. Inf. Forensics Security IEEE Trans. 9 (12), 2024–2034.
Tzimiropoulos, G., Pantic, M., 2013. Optimization problems for fast aam fitting in-the-wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 593–600.
Tzimiropoulos, G., Pantic, M., 2014. Gauss-newton deformable part models for face alignment in-the-wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1851–1858.
Uřičář, M., Franc, V., Hlaváč, V., 2012. Detector of facial landmarks learned by the structured output svm. VISAPP 12, 547–556.
Uřičář, M., Franc, V., Thomas, D., Sugimoto, A., Hlaváč, V., 2016. Multi-view facial landmark detector learned by the structured output svm. Image Vis. Comput. 47, 45–59.
Valstar, M., Martinez, B., Binefa, X., Pantic, M., 2010. Facial point detection using boosted regression and graph models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2729–2736.
Valstar, M.F., Pantic, M., 2012. Fully automatic recognition of the temporal phases of facial actions. Syst. Man Cybern. Part B 42 (1), 28–43.
Viola, P., Jones, M.J., 2004. Robust real-time face detection. Int. J. Comput. Vis. 57 (2), 137–154.
Wang, N., Gao, X., Tao, D., Li, X., 2017. Facial feature point detection: a comprehensive survey. Neurocomputing. arXiv preprint arXiv:1410.1037.
Wang, Y., Lucey, S., Cohn, J.F., 2008. Enforcing convexity for improved alignment with constrained local models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8.
Wu, Y., Ji, Q., 2015. Discriminative deep face shape model for facial point detection. Int. J. Comput. Vis. 113 (1), 37–53.
Xiao, J., Baker, S., Matthews, I., Kanade, T., 2004. Real-time combined 2d+ 3d active appearance models. In: CVPR (2), pp. 535–542.
Xiong, X., De la Torre, F., 2013. Supervised descent method and its applications to face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 532–539.
Xiong, X., De la Torre, F., 2015. Global supervised descent method. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2664–2673.
Yan, J., Lei, Z., Yi, D., Li, S., 2013. Learn to combine multiple hypotheses for accurate face alignment. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 392–396.
Yang, H., Jia, X., Loy, C. C., Robinson, P., 2015. An empirical study of recent face alignment methods. arXiv preprint arXiv:1511.05049.
Yang, H., Patras, I., 2012. Face parts localization using structured-output regression forests.. In: ACCV (2), pp. 667–679.
Yang, H., Patras, I., 2013. Privileged information-based conditional regression forest for facial feature detection. In: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. IEEE, pp. 1–6.
Yang, H., Patras, I., 2013. Sieving regression forest votes for facial feature detection in the wild. In: Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, pp. 1936–1943.
Yang, Y., Ramanan, D., 2013. Articulated human detection with flexible mixtures of parts. Pattern Anal. Mach. Intell. IEEE Trans. 35 (12), 2878–2890.
Yu, X., Huang, J., Zhang, S., Yan, W., Metaxas, D.N., 2013. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In: Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, pp. 1944–1951.
Yuan, X.-T., Liu, X., Yan, S., 2012. Visual classification with multitask joint sparse representation. Image Process. IEEE Trans. 21 (10), 4349–4360.
Zafeiriou, S., Zhang, C., Zhang, Z., 2015. A survey on face detection in the wild: past, present and future. Comput. Vis. Image Understand. 138, 1–24.
Zhang, J., Shan, S., Kan, M., Chen, X., 2014. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In: European Conference on Computer Vision. Springer, pp. 1–16.
Zhang, N., Donahue, J., Girshick, R., Darrell, T., 2014. Part-based r-cnns for fine–grained category detection. In: European Conference on Computer Vision. Springer, pp. 834–849.
Zhang, T., Ghanem, B., Liu, S., Ahuja, N., 2013. Robust visual tracking via structured multi-task sparse learning. Int. J. Comput. Vis. 101 (2), 367–383.
Zhang, Z., Luo, P., Loy, C.C., Tang, X., 2014. Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision. Springer, pp. 94–108.
Zhang, Z., Luo, P., Loy, C.C., Tang, X., 2016. Learning deep representation for face alignment with auxiliary attributes. IEEE Trans. Pattern Anal. Mach. Intell. 38 (5), 918–930.
Zhao, R., Wang, Y., Benitez-Quiroz, C.F., Liu, Y., Martinez, A.M., 2016. Fast and precise face alignment and 3d shape reconstruction from a single 2d image. In: European Conference on Computer Vision. Springer, pp. 590–603.
Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A., 2003. Face recognition: a literature survey. Acm Comput. Surv. 35 (4), 399–458.
Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q., 2013. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 386–391.
Zhou, F., Brandt, J., Lin, Z., 2013. Exemplar-based graph matching for robust facial landmark localization. In: Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, pp. 1025–1032.
Zhu, S., Li, C., Loy, C.C., Tang, X., 2015. Face alignment by coarse-to-fine shape searching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE.

Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z., 2016. Face alignment across large poses: a 3d solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 146–155.

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