# Adaptive Cascade Regression Model for Robust Face Alignment

Qingshan Liu, *Senior Member, IEEE*, Jiankang Deng, Jing Yang, Guangcan Liu, *Member, IEEE*, and Dacheng Tao, *Fellow, IEEE*

***Abstract*— Cascade regression is a popular face alignment approach, and it has achieved good performances on the wild databases. However, it depends heavily on local features in estimating reliable landmark locations and therefore suffers from corrupted images, such as images with occlusion, which often exists in real-world face images. In this paper, we present a new adaptive cascade regression model for robust face alignment. In each iteration, the shape-indexed appearance is introduced to estimate the occlusion level of each landmark, and each landmark is then weighted according to its estimated occlusion level. Also, the occlusion levels of the landmarks act as adaptive weights on the shape-indexed features to decrease the noise on the shape-indexed features. At the same time, an exemplar-based shape prior is designed to suppress the influence of local image corruption. Extensive experiments are conducted on the challenging benchmarks, and the experimental results demonstrate that the proposed method achieves better results than the state-of-the-art methods for facial landmark localization and occlusion detection.**

***Index Terms*— Robust face alignment, cascade regression model, shape-indexed appearance, adaptive shape prior.**

## I. Introduction

FACE alignment has been an active research topic over the last two decades [1], because it potentially has significance for many face-oriented applications, such as face recognition [2]–[5], expression analysis [6], [7], face animation [8], face synthesis [9], and 3D face modeling [10], [11]. A large number of facial landmark localization methods have been proposed in the past two decades [12], and the most popular solution is to take the ensemble of facial landmarks as a whole shape and learn a general face shape model from labeled training images [13]. In respect of this shape model, the previous works can be categorized as explicit shape model-based methods and implicit shape model-based methods.

Manuscript received May 6, 2016; revised October 3, 2016 and November 14, 2016; accepted November 26, 2016. Date of publication November 30, 2016; date of current version December 12, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61532009 and Grant 61272223, in part by the Natural Science Foundation of Jiangsu province under Grant BK2012045, and in part by the Australian Research Council under Project DP-140102164 and Project FT-130101457. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yonggang Shi.

Q. Liu, J. Deng, J. Yang, and G. Liu are with the B-DAT Laboratory, Department of Information and Control, Nanjing University of Information and Technology, Nanjing 210014, China.

D. Tao is with the Center for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology, Sydney, NSW 2007, Australia.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2633939

Most early works on this topic address the face alignment problem by employing explicit shape constraints, and they learn a parametric shape model from the labeled training data. Representative works are Active Shape Model (ASM) [14] and Active Appearance Model (AAM) [15]–[17], in which the variation in face shape is modeled by Principal Component Analysis (PCA) [14], [15]. Other methods include Markov Random Field (MRF)-based modeling [18], [19], Graph-based model [20], [21], and exemplar-based modeling [21], [22]. In the context of medical image analysis, Zhang *et al.* [23] developed an Adaptive Shape Composition method (ASC) to model shapes and implicitly incorporate the shape prior constraint effectively by utilizing sparse representation on the shape dictionary. ASC is able to handle non-Gaussian errors, model multi-modal distribution of shapes and recover local details. The problem is efficiently solved by an EM type of framework and an efficient convex optimization algorithm. Inspired by ASC, Liu *et al.* [24] proposed a dual sparse constrained cascade regression model (DSC-CR) for robust facial landmark localization. During the regressor training, a sparse constraint is incorporated by Lasso [25], which can select the robust features and compress the size of the model. Another sparse shape constraint is incorporated between the regressors to suppress the ambiguity in the local features. Due to the limited capacity of explicit shape models, they tend to under-perform on faces that have extreme variations in pose and expression [26].

In recent years, implicit shape constraints have attracted much attention. Their main objective is to learn shape regression functions that directly map the face image to the landmark coordinates without a parametric shape model, and good performances have been achieved on some standard benchmarks [27], [28], as a result of their ability to integrate contextual information and their flexibility in building the relationship between landmark points. There are two popular ways to learn such a regression function. One is based on deep network learning [28], [29], and the cascade regression model is another popular implicit shape model. Our work focuses on the cascade regression model, which aims to learn a series of face shape regressors and combine them in an additive manner to approximate the complex nonlinear mapping between the initial shape and the true shape [27]. However, the cascade regression model is sensitive to large occlusion, because occlusion not only affects the location updates around occluded regions but also has an effect on the location updates in non-occluded regions during shape regressor iterations [30].

In this paper, we present a new adaptive cascade regression model for robust face alignment. In contrast to previous works, we first use shape-indexed appearance, the normalised face appearance [15] at the current face shape, to estimate the occlusion level of landmarks in each iteration. Based on the estimated occlusion levels, an adaptive weighting scheme is applied to the shape-indexed features [31] to decrease the influence of corrupted landmarks. An exemplar-based shape prior model is also incorporated to smooth the updated shape adaptively. In contrast to the sparse shape constraint method proposed in [24], the proposed adaptive exemplar-based shape prior utilizes the occlusion level information, and the sparse reconstruction is only performed on the visible landmarks. The proposed method is evaluated on four challenging benchmarks, and the experimental results demonstrate its advantages over state-of-the-art methods for facial landmark localization and occlusion detection.

Our contributions are as follows: 1) Shape-indexed appearance is innovatively utilized to estimate the occlusion level for each landmark. 2) Based on the estimated occlusion levels, an adaptive weighting scheme is designed to suppress the influence of noise corruption efficiently. 3) We propose an efficient exemplar-based shape constraint to suppress the influence of local image corruption. 4) We conduct the experiments from face detection to face alignment, and the experimental results on challenging benchmarks show the power of our model for facial landmark localization and occlusion detection.

## II. Related Work

Cascade regression is a popular implicit shape model, which relies on shape-indexed local features and stacked regressors. The idea of regression was first proposed to estimate pose in [31]. In [26], an explicit shape regression (ESR) method was designed for facial landmark localization. In [27], a supervised Descent Method (SDM) was proposed to learn cascade regressors with fast SIFT features, and the cascade regression procedure was interpreted from the perspective of gradient descent. To reduce the influence of inaccurate initializations, Yan *et al.* [32] utilized the strategies of learn to rank and learn to combine from multiple hypotheses with a structural SVM framework. The Incremental Parallel Cascade Linear Regression (iPar-CLR) method was proposed in [33], which incrementally updates all the linear regressors in a parallel way instead of the traditional sequential manner. Each level is trained independently, using only the statistics of the previous level, and the generative model is gradually turned into a person-specific model by a recursive linear least-squares method. An $\ell_1$-induced Stagewise Relational Dictionary (SRD) model was proposed in [34] to learn consistent and coherent relationships between face appearance and shape for face images with large variations in view. In recent years, cascade regression has attracted much attention because its effectiveness has been demonstrated by extensive comparison [1], [27], [35].

In cascade regression, shape-indexed features have an important role, and many local feature descriptors have been successfully applied including Haar wavelets [36], random ferns [37], Local Binary features (LBF) [35], [38], SIFT [27], and HoG [32]. Since almost all the local feature descriptors are sensitive to occlusion, cascade regression cannot handle face images with large occlusion well. To overcome this shortcoming, Roh *et al.* [39] utilized over-sufficient facial feature detectors and the RANSAC-based method to infer occlusion. In [40], occlusion was modeled as a sparse outlier; however, the sparse error could occur from either the influence of the occluded landmarks or the perturbation of the visible landmarks. Robust Cascade Pose Regression (RCPR) [41] explicitly predicts the likelihood of landmark occlusion using a fixed occlusion prior on the divided blocks. The occlusion dictionary was deployed in [34] to deal with different kinds of partial face occlusion. In [42], a hierarchical deformable part model was proposed to model the occlusion of facial parts explicitly. Yu *et al.* [30] proposed an occlusion-robust regression method by forming a consensus from a set of occlusion-specific regressors. In this paper, we propose a new method to estimate the occlusion levels around the landmarks by shape-indexed appearances, which in turn act as adaptive weights on both shape-indexed features and nearest exemplar-based shape constraint.

## III. Adaptive Cascade Regression Model

Cascade regression combines a sequence of regressors in an additive manner, and each regression function $f$ is learnt by minimizing the mean square error [27]:

$$f = \arg\min_{f} \sum_{i=1}^{M} \left\| \left(S_i^* - S_i^0\right) - f(I_i, S_i^0) \right\|_2^2, \tag{1}$$

where $M$ is the number of training samples, $I_i$ is the face image, $S_i^*$ is the ground truth shape, $S_i^0$ is the initialization of face shape, and $f$ is a single step regression function. Due to the complex variation of the human face, one step of regression $f$ is insufficient. Cascade regression combines a series of simple regression function $f_t, t = 1, \ldots, T$ to approximate a complex nonlinear mapping between the initial shape $S_i^0$ and the ground truth shape $S_i^*$.

$$\arg\min_{R^t} \sum_{i=1}^{M} \left\| \left(S_i^* - S_i^{t-1}\right) - R^t \Phi(I_i, S_i^{t-1}) \right\|_2^2, \tag{2}$$

where $\Phi$ is the nonlinear feature descriptor and $R^t, t = 1, \ldots, T$ is the linear transform matrix which iteratively maps the current feature vectors $\Phi(I_i, S_i^{t-1})$ to the updated landmark location $(S_i^* - S_i^{t-1})$. The equation indicates that the displacement of each facial landmark is related to all other fiducial points, so in this way the shape constraint is incorporated implicitly. Since this is a linear least squares problem, $R_t$ has a close-form solution

$$\left(S_i^* - S_i^{t-1}\right) \left(\Phi(I_i, S_i^{t-1})\right)^{\mathrm{T}} \left(\Phi(I_i, S_i^{t-1}) \left(\Phi(I_i, S_i^{t-1})\right)^{\mathrm{T}}\right)^{-1}.$$

Although cascade regression achieves good performance in face alignment, it is sensitive to occlusion because it depends heavily on the local features around each landmark. To successfully overcome this issue, we propose an adaptive cascade regression model which utilizes the occlusion levels $w_j, j = 1, \cdots, N$ of each landmark to adjust the

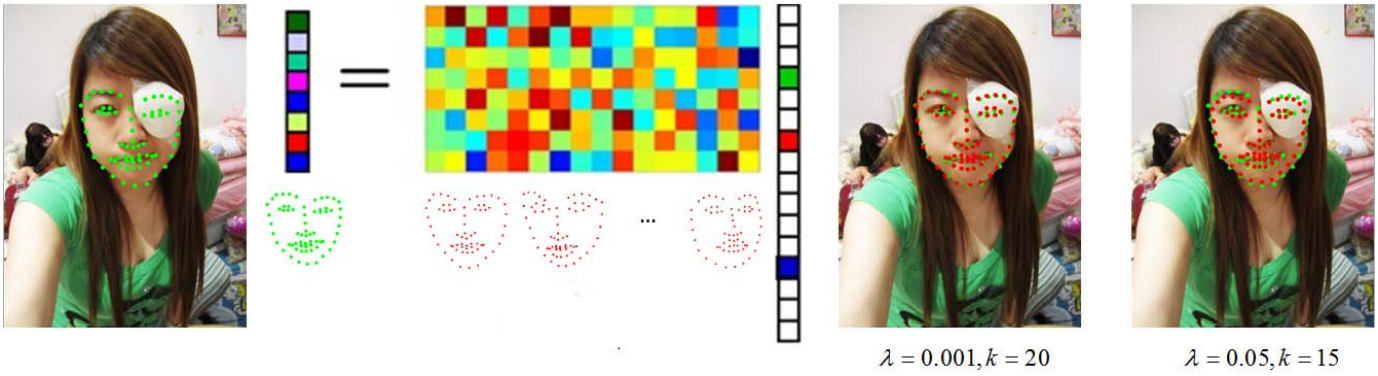



Fig. 1. Sparse shape constraint under difference regularization parameters.

shape-indexed features $\Phi(I_i, S_i^{t-1})$ and adaptively constrain the intermediate shapes by exemplar shapes during cascade regression.

$$\arg\min_{R^t} \sum_{i=1}^{M} \left\| \left( S_i^* - \Psi\left( D_S \alpha_i^{t-1}, \gamma_i^{t-1} \right) \right) - R^t \left( W_i^F \Phi\left( I_i, \Psi\left( D_S \alpha_i^{t-1}, \gamma_i^{t-1} \right) \right) \right) \right\|_2^2,$$

where $W_i^F = diag(w_{i1}^F, w_{i2}^F, \ldots, w_{iN}^F, 1)$, $w_{ij}^F = w_{ij} I_{P\times P}$, $w_{ij}$ indicates the occlusion level for the landmark $j$, and $P$ is the dimension of the landmark feature vector. The goal of $W_i^F$ is to allocate high weights to the uncorrupted landmarks and small or even zero weights to the occluded landmarks, which can efficiently restrain the influence of the occluded regions. $\Psi\left( D_S \alpha_i^{t-1}, \gamma_i^{t-1} \right)$ is the output of the exemplar-based shape constraint on $S_i^{t-1}$. $D_S \in \mathbb{R}^{2N\times M}$ is the exemplar shape dictionary, $\alpha_i^{t-1} \in \mathbb{R}^{M\times 1}$ is the reconstruction coefficient, $\Psi$ is the similarity transformation, and $\gamma_i^{t-1}$ is the similarity transformation coefficient.

### *A. Adaptive Exemplar-Based Shape Model*

The sparse shape model is widely used as a shape constraint because it is able to correct gross errors of the input shape and preserve shape details, even if they are not statistically significant in the training set [23], [24].

$$\min_{\alpha} \|S - D_S\alpha\|_2^2 + \lambda\|\alpha\|_1, \tag{3}$$

where $S = (x_1, \ldots x_N, y_1, \ldots y_N)^T$ is the observed normalized shape, $D_S \in \mathbb{R}^{2N\times M}$ is the normalized shape dictionary, $\alpha \in \mathbb{R}^{M\times 1}$ is the sparse reconstruction coefficient, and $\lambda$ is the regularization parameter. The sparse shape model has two drawbacks, however: 1) It is very time-consuming due to the high dimension $l_1$ optimization problem, i.e., so-called Lasso optimization [25]; 2) All the landmark points in the shape are treated equally, including the corrupted landmarks [23], [24]. If some landmarks are occluded, the errors from the misaligned landmarks will spread to all the other landmarks to some extend, due to the residual minimization used in the procrustes analysis [43] and shape reconstruction [44].

As can be seen from Fig. 1, there is an occlusion above the left eye which causes the alignment result on the corresponding area to be inaccurate. We take the normalized face shapes from the HELEN [45] training set as the shape dictionary and give two sparse reconstruction results under difference regularization parameters. When we set $\lambda$ as 0.001, the number of non-zero coefficients is 20, and the gross error on the left eye is removed. However, the gross error on the left eyebrow is still there. As we increase $\lambda$ to 0.05, the constrained shape is smoother, but the accuracy of the non-occluded area is sacrificed. If we decrease $\lambda$, the constrained shape will be almost the same as the input face shape without the ability to correct the gross error, due to the minimization of the reconstruction error. Thus, sometimes the sparse shape constraint has limited ability in fixing the gross error.

To overcome the above issue, we propose a new exemplar-based shape model as,

$$\min_{\alpha} \left\| W^S S - \left( W^S S \odot W^S D_S \right) \alpha \right\|_2^2, \tag{4}$$

where $W^S = diag(w_1, \ldots, w_N, w_1, \ldots, w_N)$ is the weighting matrix, and $w_j$ is occlusion level. The purpose of $W^S$ is to evaluate the input shape with non-occluded landmarks as accurately as possible, and $\odot$ selects the most important k-nearest exemplar shapes from the shape dictionary. $\left(W^S S \odot W^S D_S\right)$ is used to select the nearest exemplar shapes of $W^S S$ from the dynamic exemplar shape dictionary $W^S D_S$. In other words, $W^S$ and $\odot$ impose different weights on the row and column respectively of the exemplar shape dictionary, which makes the shape constraint more flexible, robust and efficient.

Compared to time-consuming Lasso optimization, the proposed adaptive exemplar-based shape model is more efficient. Given the shape dictionary $\mathrm{D_S} \in \mathbb{R}^{2\mathrm{N}\times\mathrm{M}}$, the computational complexity of the interior-point convex optimization solver for $l_1$ optimization problem is $O\left(N^2 M\right)$. The computational complexity of our method lies only in the selection of the K-Nearest Neighbors, and the reconstruction coefficients are computed directly by the least squares method. In Lasso optimization, the sparse coefficients tend to be local and items with larger coefficients are more similar to the input sample [46], [47]. In the proposed adaptive shape model, we select the k-Nearest exemplar shapes of the input shape as the shape bases for reconstructing the input shape directly and set the coefficients corresponding to the remaining exemplars in the dictionary as zero. This means that the input face shape can be reconstructed by the nearest exemplar shapes [22],

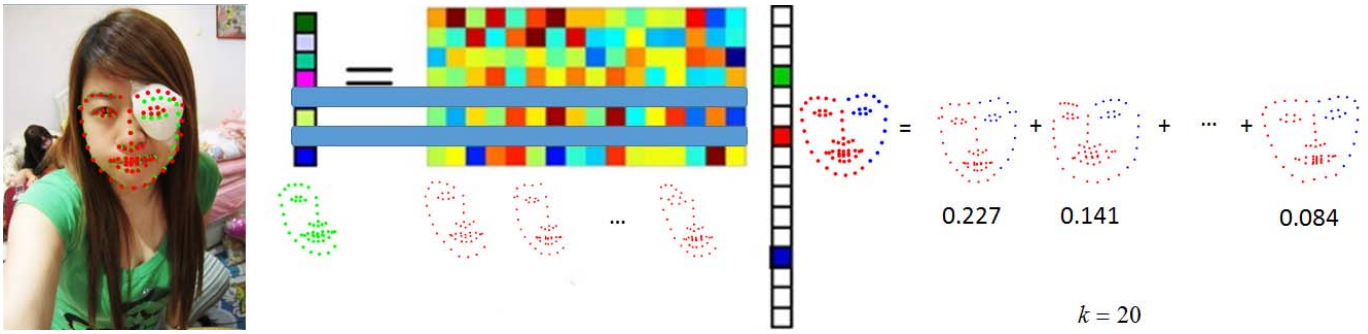



Fig. 2. Sparse shape constraint by adaptive exemplar based shape model.

which is feasible both in empirical observation and theory analysis [22], [48].

As shown in Fig. 2, we can only take account of the non-occluded landmarks when the occlusion levels of the landmarks are given, and we use the nearest exemplar-based shape model to reconstruct the shape by Eq 4. Surprisingly, even though we still reconstruct the face shape using only 20 exemplar face shapes, as in Fig. 1, the result is very satisfying. The gross error on the occluded area is almost removed, and the landmarks on the non-occluded area are as accurate.

### B. Occlusion Inference Model

It is well known that it is difficult to detect occluded landmarks if only the local features around the landmarks are used. However, occlusion is generally a continuous local region with an irregular size and becomes obvious in the shape-normalized appearance [15]. Inspired by [34], the facial shape and appearance tend to be consistent on the exemplar dictionary. We construct the shape-normalized appearance dictionary $D_A$, which is directly derived from the exemplars $D_S$. The shape-indexed appearance is then constructed by $D_A\beta, \beta \in \mathbb{R}^{M\times 1}$ on the exemplar appearance dictionary, and the reconstruction discrepancy $\|A - D_A\beta\|_2^2$ is taken to estimate the occlusion levels $w_j, j = 1, \cdots, N$ of the landmarks and detect the occluded landmarks.

To effectively calculate the appearance reconstruction coefficients $\beta$, we utilize Canonical Correlation Analysis (CCA) [49] to build the relationship between the appearance reconstruction coefficients $\beta$ and the shape reconstruction coefficients $\alpha$. We can then estimate $\beta$ directly from $\alpha$.

$$X = \begin{bmatrix} D_S \\ D_A \end{bmatrix}, E(X) = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, Var(X) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad (5)$$

where $D_S$ is the exemplar shape dictionary, $D_A$ is the exemplar appearance dictionary, $E(X)$ is the mean value of $X$, and $Var(X)$ is the covariance of $X$. CCA aims to find $a^T$ and $b^T$ to obtain the maximum correlation between $u = a^T D_S$ and $v = b^T D_A$.

$$Var(u) = a^T \Sigma_{11} a, Var(v) = b^T \Sigma_{22} b, Cov(u, v) = a^T \Sigma_{12} b, \quad (6)$$

where $Var(u)$ is the variance of $u$, $Var(v)$ is the variance of $v$, and $Cov(u, v)$ is the covariance between $u$ and $v$. The correlation between $u$ and $v$ is

$$Corr(u, v) = \frac{a^T \Sigma_{12} b}{\sqrt{a^T \Sigma_{11} a}\sqrt{b^T \Sigma_{22} b}}. \quad (7)$$

The optimization objective is

$$\begin{aligned} &\max \quad a^T \Sigma_{12} b \\ &s.t. \quad a^T \Sigma_{11} a = 1, b^T \Sigma_{22} b = 1. \end{aligned} \quad (8)$$

We construct a Lagrangian equation

$$\ell = a^T \Sigma_{12} b - \frac{\lambda_1}{2}\left(a^T \Sigma_{11} a - 1\right) - \frac{\lambda_2}{2}\left(b^T \Sigma_{22} b - 1\right), \quad (9)$$

and finally obtain $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} a = \lambda_1^2 a$. $\lambda_1$ and $a$ can be easily calculated by calculating eigenvalues and eigenvectors. In the same way, we can obtain $\lambda_2$ and $b$. We build a correlation relationship on the face shape and appearance by the canonical variables $a$ and $b$:

$$b^T D_A = C a^T D_S, \quad (10)$$

where $C$ is the correlation constant calculated by Eq. 7. As a result, we can efficiently estimate the face appearance parameter $\beta$ from the face shape parameter $\alpha$ by utilizing the correlation relationship between the face appearance and shape:

$$b^T D_A \beta = C a^T D_S \alpha. \quad (11)$$

The occlusion level $w_j, j = 1, \ldots, N$ for each landmark is calculated as follows:

$$E = \|A - D_A\beta\|_2^2, \quad (12)$$

$$w_j = \frac{\sum_{j=1}^{N} \bar{E}_j}{\bar{E}_j}, \quad (13)$$

where $E$ is the discrepancy between the observed appearance $A$ and the synthesized shape-normalized appearance $D_A\beta$, and $\bar{E}_j$ is the discrepancy on the local area around the *j-th* landmark.

Fig. 3 illustrates the flow of occlusion inference. We first calculate the shape-normalized appearance $A$ from the input face image at the current shape, and then we can obtain the discrepancy $E$ between the observed appearance $A$ and the

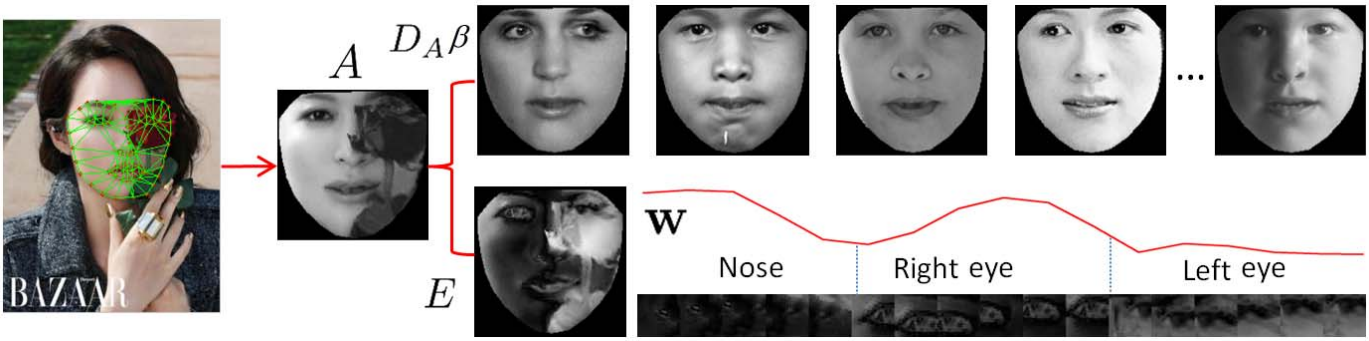



Fig. 3. Adaptive weights by occlusion inference.

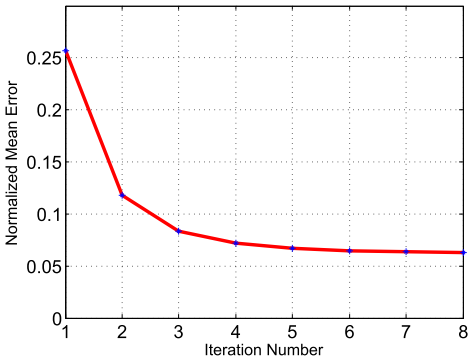



Fig. 4. The decrease of normalized mean error on the IBUG dataset.

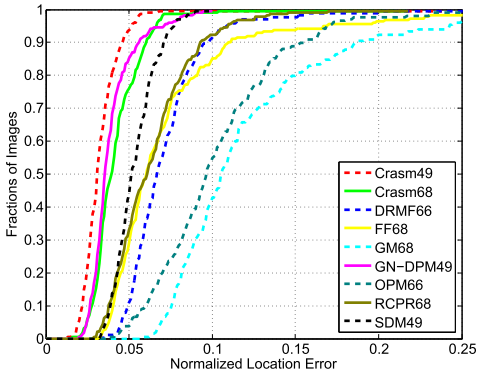



Fig. 5. CED curves over the LFPW dataset.

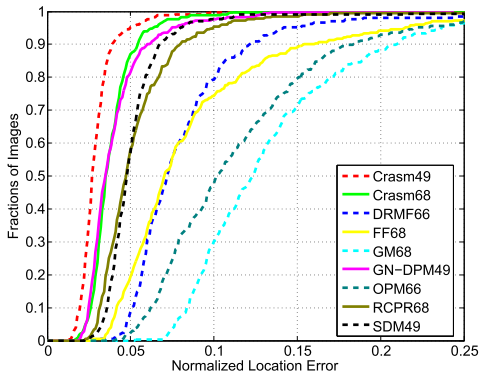



Fig. 6. CED curves over the HELEN dataset.

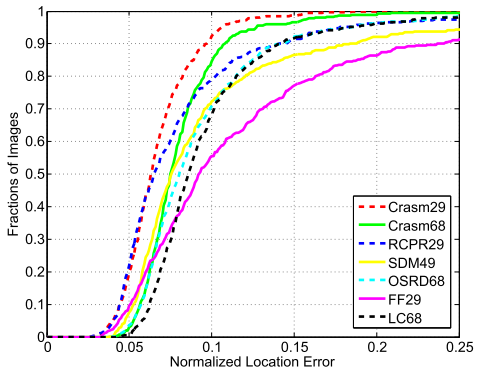



Fig. 7. CED curves over the COFW dataset.

synthesized shape-normalized appearance $D_A\beta$ by Eq. (12). Lastly, based on the error map $E$, we calculate the occlusion level in the local area around each landmark and infer the weight $w_j$ by Eq. (13). There is a flower above the left eye, so the weights around the left eye are smaller than those around the non-occluded right eye. By occlusion inference from the appearance discrepancy, we can obtain an adaptive weight for each landmark. $w_j$ indicates the occlusion status for each landmark, which can also be used for occlusion detection directly by thresholding.

### *C. Cascade Regression With Adaptive Exemplar-Based Shape Model*

We denote the proposed Cascade Regression with Adaptive Shape Model as CRASM for simplicity, and the CRASM learning algorithm is summarized in **Algorithm** 1.

## IV. Experiments

### *A. Datasets*

There are a number of databases for evaluating face alignment algorithms [51]. Different databases present variations in face pose, expression, illumination and occlusion, and the landmark configurations are also different. In [13], the IBUG team created a large 300-W database for face alignment competition, which contains 11147 face images from six datasets: LFPW [22], AFW [20], HELEN [45], XM2VTS [52], FRGC [53] and their IBUG dataset. All the face images are re-labeled by 68 landmarks. We train the proposed method on AFW, the training set of LFPW, and the training set of Helen, with 3148 face images in total. We report the experimental results on the test set of LFPW (224), the test set of HELEN (330), and IBUG dataset (135) respectively to deeply integrate the performance as in [35].

**Algorithm 1** The CRASM Learning Algorithm

**Input:** Face image $I_i$, face box $R_i$(DPM [50]), landmark label $S_i^*$.

1) Compute the normalized mean shape $\bar{S}$, and construct the shape dictionary $D_S$ from non-occluded exemplars by K-SVD.
2) Construct the corresponding shape-normalized appearance dictionary $D_A$.
3) Calculate the correlation relationship parameters $a$, $b$, and $C$ by Eq. 7 8 10.
4) Generate initial shape $S_i^0$ for each training data according to face box $R_i$ with DPM parts.
5) Initialize weights $w_{ij}$, $W_i^S$, and $W_i^F$.
6) **for** $t = 0$ to $T$
7) Shape reconstruction $\arg\min_{\alpha_i^t} \left\| W_i^S S_i^t - \left( W_i^S S_i^t \odot W_i^S D_S \right) \alpha_i^t \right\|_2^2$.
8) Generate constrained shape $S_i^t = D_S \alpha_i^t$.
9) Estimate appearance reconstruction coefficients $b^T D_A \beta_i^t = C a^T D_S \alpha_i^t$.
10) Calculate corresponding appearance error $E_i = \|A_i - D_A \beta_i^t\|_2^2$.
11) Update weights $w_{ij} = \frac{\sum_{j=1}^{N} \bar{E}_{ij}}{\bar{E}_{ij}}$, $W_i^S$ and $W_i^F$.
12) Extract multi-scale HoG features at current shape $\Phi(I_i, S_i^t)$.
13) Compute the regression matrix $R^t$, $\arg\min_{R^t} \sum_{i=1}^{M} \left\| (S_i^* - S_i^t) - R^t W_i^F \Phi(I_i, S_i^t) \right\|_2^2$.
14) Update the shape $S_i^{t+1} = S_i^t + R^t W_i^F \Phi(I_i, S_i^t)$.
15) **end**

**Output:**Crasm model $R^t, t = 1, \ldots, T$.

TABLE I
FACE DETECTION RESULTS ON FOUR DATASETS

| Dataset | LFPW | HELEN | IBUG | COFW |
|---|---|---|---|---|
| Face DR(%) | 100% | 100% | 99.26% | 99.41% |
| Failed Number | 0 | 0 | 1 | 3 |

We also test the proposed CRASM method on the **COFW** database [41], which is widely used to evaluate robustness to occlusion and occlusion detection. It has 1852 face images in total, of which the training set has 1345 face images. The face images are captured in real-world conditions and have large variations in shape and occlusion due to differences in pose, expression, use of accessories such as sunglasses and hats, and interactions with objects (e.g. food, hands, microphones). The faces are occluded to different degrees and COFW has an average occlusion of over 23%, with large variations in the type of occlusion. All the images are labeled with 29 landmarks. Note that there are two incorrect annotations in the training set which we have re-annotated.

### *B. Experimental Setting*

An automatic face alignment system should integrate face detection and face alignment, and face detection provides the shape initialization for face alignment. Occlusion usually affects face detection too. Following [50], [54], we develop a robust DPM-based face detector and utilize the deformable parts to estimate the face pose. Our face detection results on four datasets are shown in Table I, where the face detection rate (Face DR) is the proportion of the successfully detected faces on the dataset. The DPM-based face detector only fails to detect one face on the IBUG dataset, and three faces on the COFW dataset.

TABLE II
RESULTS OF DIFFERENT DICTIONARY SIZES

| M | 1000 | 1500 | 2000 | 2500 | 3000 |
|---|---|---|---|---|---|
| NME(%) | 4.03 | 3.65 | 3.38 | 3.39 | 3.44 |
| Time(s) | 0.47 | 0.62 | 0.72 | 0.85 | 0.97 |

TABLE III
RESULTS OF DIFFERENT K

| k | 10 | 20 | 30 | 40 | 50 |
|---|---|---|---|---|---|
| NME(%) | 7.34 | 5.26 | 3.78 | 3.56 | 3.43 |
| Time(s) | 0.68 | 0.69 | 0.71 | 0.73 | 0.75 |

As in [32], we use multi-scale HOG as the local feature descriptor for cascade regression. For the adaptive shape model, we use K-SVD [55] to select the representative exemplar shapes to construct the shape dictionary from the non-occluded training data. Given the shape dictionary, we can generate the corresponding shape-normalized exemplar appearance dictionary. We use the shape dictionary to reconstruct the face images on IBUG (135), to evaluate the representation ability of the exemplar shape dictionary. Table II shows the reconstruction error and computation time, where M is the dictionary size. We set M = 2000 in the following experiments, taking account of efficiency and accuracy. The number of nearest neighbors *k* is tuned in Table III. We finally set *k* as 30, which means that the input shape is reconstructed by 30 example shapes. As is shown in Figure 4, the normalised mean error on the IBUG dataset gradually decreases when we increase iteration number from 1 to 8. Finally, we set the number of cascade regression iterations as 7, because there is no obvious decrease in the normalised mean error after that.

Given the ground-truth, the localization performance can be evaluated by normalized mean error (NME) [13], and normalization is typically performed with respect to Inter-Ocular

TABLE IV
NORMALIZED MEAN ERROR ON LFPW DATASET

| Algorithm | CRASM | CRASM | CR | SSC-CR | RCPR | SDM | LBF | GM | FF | DRMF | OPM | GN-DPM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| landmarks | 68 | 49 | 68 | 68 | 68 | 49 | 68 | 68 | 68 | 66 | 66 | 49 |
| Face DR(%) | 100 | 100 | 100 | 100 | Label | 87.95 | Label | 89.29 | Label | 73.21 | 92.86 | Label |
| NME(%) | 4.23 | 3.32 | 4.84 | 4.46 | 6.44 | 5.36 | 5.12 | 14.11 | 7.39 | 7.22 | 10.31 | 3.92 |

TABLE V
NORMALIZED MEAN ERROR ON HELEN DATASET

| Algorithm | CRASM | CRASM | CR | SSC-CR | RCPR | SDM | LBF | GM | FF | DRMF | OPM | GN-DPM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| landmarks | 68 | 49 | 68 | 68 | 68 | 49 | 68 | 68 | 68 | 66 | 66 | 49 |
| Face DR(%) | 100 | 100 | 100 | 100 | Label | 93.64 | Label | 92.42 | Label | 63.03 | 89.39 | Label |
| NME(%) | 3.90 | 3.00 | 4.43 | 4.24 | 5.46 | 5.84 | 4.70 | 13.42 | 8.93 | 8.29 | 11.59 | 4.06 |

TABLE VI
NORMALIZED MEAN ERROR ON IBUG DATASET

| Algorithm | ESR | SDM | LBF | TCDCN | CR | SSC-CR | Crasm |
|---|---|---|---|---|---|---|---|
| landmarks | 68 | 68 | 68 | 68 | 68 | 68 | 68 |
| Face DR(%) | Label | Label | Label | Label | 99.26 | 99.26 | 99.26 |
| NME(%) | 17.00 | 15.40 | 11.98 | 9.15 | 8.59 | 7.36 | 6.29 |

TABLE VII
NORMALIZED MEAN ERROR ON COFW DATASET

| Algorithm | CRASM | CRASM | CR | SSC-CR | RCPR | SDM | OSRD | LBF | GM | FF | LC | HPM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| landmarks | 29 | 68 | 29 | 29 | 29 | 49 | 68 | 68 | 68 | 29 | 68 | 68 |
| Face DR(%) | 99.41 | 99.41 | 99.41 | 99.41 | Label | 71.40 | Label | 89.15 | 79.68 | Label | Label | Label |
| NME(%) | 6.68 | 8.02 | 8.34 | 7.46 | 8.38 | 6.99 | 9.27 | 13.7 | 11.82 | 12.24 | 9.82 | 7.46 |

Distance (IOD). Based on NME, we plot the cumulative error distribution (CED) curves, calculated from the normalized mean errors over each image. The allowed error (localization threshold) is taken to be a percentage of the inter-ocular distance IOD, typically 10% or less of IOD.

$$err = \frac{1}{M}\sum_{i=1}^{M}\frac{\frac{1}{N}\sum_{j=1}^{N}\left|p_{i,j}-g_{i,j}\right|_2}{|l_i - r_i|_2}, \tag{14}$$

where $M$ is the number of images in the test set, $N$ is the number of landmarks, $p_{i,j}$ is the predicted location of the *j-th* landmark of the *i-th* image, $g_{i,j}$ is the ground truth of the *j-th* landmark of the *i-th* image, $l_i$ and $r_i$ are the positions of the left and left eye center of the the *i-th* image.

There are three kinds of landmark configuration (29, 49 and 68 landmarks) in this paper. For the models tested on the datasets of LFPW, HELEN and IBUG, the proposed methods are named Crasm68 (with all landmarks) and Crasm49 (without the landmarks on the face contour and the inner mouth corner). On the dataset of COFW, We train CRASM with the landmark configuration of 68 and 29 respectively, and we obtain two CRASM models: Crasm68 and Crasm29. The Crasm68 model is trained on the 300-W datasets, and Crasm29 is trained on the COFW training set (1345) and Helen (348 faces) [21].[1] To evaluate the pre-trained models (68 landmarks) in relation to other methods on the COFW (29 landmarks) dataset, we use only 19 intersected landmarks for comparison. For a fair comparison, we report the normalized mean error (NME) taking the face detection rate into consideration. "Label" in each table means that no face detection is integrated and the face boxes labelled in the databases are used directly. To confirm the effectiveness of the proposed adaptive exemplar-based shape model, we set up the baseline methods: cascade regression (CR) and sparse shape constraint cascade regression (SSC-CR) [24]. Both CR and SSC-CR are trained with the initialization from the DPM-based face detector [50], [54] and the multi-scale HOG features [32].

[1] http://www.f-zhou.com/fa.html

### C. Evaluation on Face Alignment

The **LFPW** test subset contains 224 face images which vary significantly in pose, illumination and occlusion. The **HELEN** test subset consists of 330 face images under all kinds of natural conditions, both indoor and outdoor, and most of the images in HELEN are of high resolution. As in the latest literature [1], we compare our algorithm with ten state-of-the-art methods including robust cascaded pose regression (RCPR) [41], supervised descent model (SDM) [56], local binary feature (LBF) [35], graphical models (GM) [20], fast AAM fitting (FF) [57], discriminative response map fitting (DRMF) [58], optimized part mixtures (OPM) [59], and Gauss-newton deformable part model (GN-DPM) [60].

Fig. 8. Example alignment results by CRASM on LFPW, HELEN, and IBUG.

The results of Face DR and NME on **LFPW** and **HELEN** are listed in Table IV and V respectively, and the CED curves are shown in Figure 5 and 6 respectively. It can be seen that CRASM outperforms all other methods due to the better initialization from the DPM-based face detector, the more expressive multi-scale HoG features, and the occlusion-adaptive shape constraint. Compared to the direct sparse shape constraint without the occlusion inference (SSC-CR) citeliu2015dual, the proposed adaptive exemplar-based shape model decreases the NME from 4.46% to 4.23% on LFPW test set and decreases the NME from 4.24% to 3.90% on HELEN test set respectively. The proposed adaptive exemplar-based shape model not only improves the landmark localization accuracy under occlusions but also improves the performance by smoothing the face shapes. The performance of CRASM on LFPW and HELEN is almost equivalent to the human performances of 0.0328 on LFPW and 0.033 on Helen reported in [41].

The **IBUG** test subset, which has 135 face images with large variations in pose, expression and illumination, is more challenging than **LFPW** and **HELEN**. The latest works report the results obtained by explicit shape regression (ESR) [26], SDM, LBF, and deep-based TCDCN [61] on IBUG. Table VI reports our results compared with these four methods. We can see that only one face was missed by our DPM-based face detector, and the normalized mean error on this missing face is 64.62%. CRASM obtains a NME of 6.72% with the missing face, which is better than the most recent TCDCN method which has NME of 9.15%. The failure rate ($\geqslant 0.1 * IOD$)

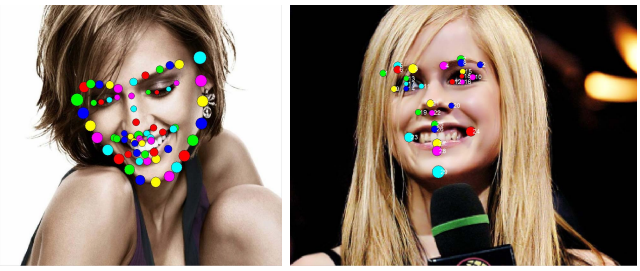


Fig. 9. Normalized Statistical Mean Error of each landmark by CRASM on IBUG and COFW.

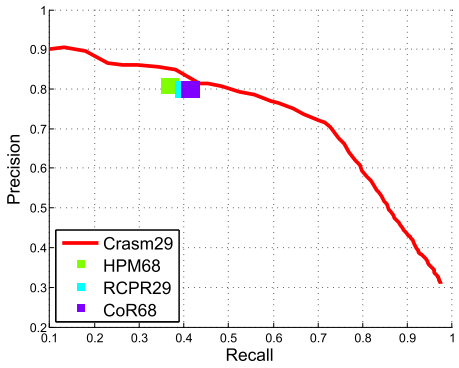



Fig. 10. Occlusion detection result on COFW.

of the proposed CRASM method is 12.59% on the IBUG test set. Compared to the baseline method CR, SSC-CR decreases the NME by 14.32%. By contrast, CRASM decreases the

Fig. 11. Face alignment and occlusion detection by CRASM on COFW.

NME by 26.78%, which indicates that the proposed adaptive exemplar-based shape model is more flexible and effective than the sparse shape constraint. In Figure 8, some examples exhibit superior ability in handling difficult cases with pose variation, occlusion and exaggerated expression, thanks to the coupling of DPM-based face detector, multi-scale HoG features, and adaptive shape-constrained cascade regression. It can also be seen that CRASM is robust to illumination because the weighting scheme reduces the corrupting noises, including illumination.

Many alignment methods have failed to locate facial landmarks accurately on the **COFW** database due to the large variation in occlusion. State-of-the-art results were achieved on this database by RCPR [41]. Apart from RCPR, we add the other three occlusion-robust baseline methods for comparison, that is, occlusion stagewise relational dictionary (OSRD) [34], learn to combine model (LC) [32], and hierarchical deformable part model (HPM) [42]. The NME comparisons are reported in Table VII, and the CED curves are shown in Figure 7.

It can be seen that Crasm68 achieves accurate performance compared to the other methods. Compared to RCPR, the proportion of highly accurate faces aligned by Crasm68 is smaller because there is a bias between these two landmark configurations. However, the failure rate of Crasm68 is 15.38%, and the failure rate of RCPR 20.71%, which indicates that the proposed adaptive shape model is effective under occlusion. HPM68 is slightly better than Crasm68, the failure rate of HPM68 being reported in [42] as 13.24%. This is because we use our DPM-based detector for automatic initialization, while HPM68 uses the face boxes directly attached to the database. In addition, there is a fine-tuning procedure for HPM68 to transfer the model trained on Helen to COFW, whereas we do not use the COFW training set for Crasm68. However, crasm29 outperforms all of the other methods with a failure rate of 7.69%. Thus, it can be said that CRASM obtains competitive results compared to HPM. Compared to the baseline method CR and SSC-CR, the proposed Crasm29 decreases the NME by 16.71% and 10.46% respectively, which confirms the effectiveness of the proposed method under occlusions.

Figure 9 shows the statistically normalized errors for each landmark aligned by CRASM on IBUG and COFW. It is clear that the landmarks on the face contour are most difficult to locate accurately, and this is because the outline is easily affected by pose variation and occlusion. In contrast, the inner and outer corners of the eyes and the nose tips are relatively easy to locate, since these landmarks are barely affected by facial expression, while the landmarks around the mouth are heavily dependent on facial expression.

### *D. Evaluation on Occlusion Detection*

Since only COFW provides the ground truth of occlusion, we only evaluate the occlusion detection on COFW, and we compare our method with RCPR [41], CoR [30] and HPM [42], because they have also reported the occlusion detection results on COFW. Figure 10 shows the experimental results of occlusion detection. Quantitatively, we set the false alarm at the same level(80%), our method achieves

48.45% accuracy while RCPR is 40%, CoR is 41.44%, and HPM is 37%. CRASM improves the detection precision by about 10%. Figure 11 gives example results of occlusion detection. Occlusions caused by continuous objects such as sunglass and microphone are easy to detect. However, occlusions caused by complexional objects (hand) and discontinuous objects (a few strands of hair) are hard to detect.

## V. Conclusions

In this paper, we propose a cascade regression method with adaptive exemplar-based shape model for robust face alignment. During each cascade regression, the shape-indexed appearance feature is introduced to estimate the occlusion level of each landmark, and each landmark is then assigned a weight according to its estimated occlusion level. The occlusion levels of the landmarks act as adaptive weights on the shape-indexed features to decrease the noise on the shape-indexed features. At the same time, an adaptive exemplar-based shape prior is designed to suppress the influence of local image corruption. Extensive experiments are conducted on challenging benchmarks LFPW, HELEN, IBUG, and COFW. The experimental results demonstrate that the proposed adaptive exemplar-based shape model is able to effectively rectify the facial landmark locations under occlusions, and the proposed CRASM method achieves state-of-the-art results for facial landmark localization and occlusion detection. Our future research will focus on joint face alignment and occlusion parsing based on the shape-indexed appearance.

## References


[1] N. Wang, X. Gao, D. Tao, and X. Li. (2014). "Facial feature point detection: A comprehensive survey." [Online]. Available: https://arxiv.org/abs/1410.1037

[2] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[3] R. Weng, J. Lu, and Y.-P. Tan, "Robust point set matching for partial face recognition," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1163–1176, Mar. 2016.

[4] D. Li, H. Zhou, and K. M. Lam, "High-resolution face verification using pore-scale facial features," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2317–2327, Aug. 2015.

[5] Y. Tai, J. Yang, Y. Zhang, L. Luo, J. Qian, and Y. Chen, "Face recognition with pose variations and misalignment via orthogonal Procrustes regression," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2673–2683, Jun. 2016.

[6] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognit.*, vol. 36, no. 1, pp. 259–275, Jan. 2003.

[7] I. Mpiperis, S. Malassiotis, and M. G. Strintzis, "Bilinear models for 3-D face and facial expression recognition," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 3, pp. 498–511, Sep. 2008.

[8] F. I. Parke and K. Waters, *Computer Facial Animation*, vol. 289. Wellesley, MA, USA: AK Peters Wellesley, 1996.

[9] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.

[10] X. Lu and A. Jain, "Deformation modeling for robust 3D face matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1346–1357, Aug. 2008.

[11] S. Berretti, A. D. Bimbo, and P. Pala, "Sparse matching of salient facial curves for recognition of 3-D faces with missing parts," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 2, pp. 374–389, Feb. 2013.

[12] O. Çeliktutan, S. Ulukaya, and B. Sankur, "A comparative study of face landmarking techniques," *EURASIP J. Image Video Process.*, vol. 13, no. 1, pp. 1–27, Mar. 2013.

[13] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 397–403.

[14] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, 1995.

[15] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proc. Eur. Conf. Comput. Vis.*, 1998, pp. 484–498.

[16] G. Dedeoğlu, S. Baker, and T. Kanade, "Resolution-aware fitting of active appearance models to low resolution images," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 83–97.

[17] E. Antonakos, J. Alabort-i-Medina, G. Tzimiropoulos, and S. P. Zafeiriou, "Feature-based Lucas–Kanade and active appearance models," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2617–2632, Sep. 2015.

[18] L. Liang, F. Wen, Y.-Q. Xu, X. Tang, and H.-Y. Shum, "Accurate face alignment using shape constrained Markov network," in *Proc. Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2006, pp. 1313–1319.

[19] Y. Huang, Q. Liu, and D. Metaxas, "A component based deformable model for generalized face alignment," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[20] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.

[21] F. Zhou, J. Brandt, and Z. Lin, "Exemplar-based graph matching for robust facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1025–1032.

[22] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 545–552.

[23] S. Zhang, Y. Zhan, M. Dewan, J. Huang, D. N. Metaxas, and X. S. Zhou, "Towards robust and effective shape modeling: Sparse shape composition," *Med. Image Anal.*, vol. 16, no. 1, pp. 265–277, Jan. 2012.

[24] Q. Liu, J. Deng, and D. Tao, "Dual sparse constrained cascade regression for robust face alignment," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 700–712, Feb. 2016.

[25] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

[26] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *Proc. Comput. Vis. Pattern Recognit.*, 2012, pp. 2887–2894.

[27] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.

[28] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. Comput. Vis. Pattern Recognit.*, 2013, pp. 3476–3483.

[29] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *Proc. Int. Conf. Comput. Vis. Workshops*, 2013, pp. 386–391.

[30] X. Yu, Z. Lin, J. Brandt, and D. N. Metaxas, "Consensus of regression for occlusion-robust facial feature localization," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 105–118.

[31] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1078–1085.

[32] J. Yan, Z. Lei, D. Yi, and S. Li, "Learn to combine multiple hypotheses for accurate face alignment," in *Proc. Int. Conf. Comput. Vis. Workshops*, 2013, pp. 392–396.

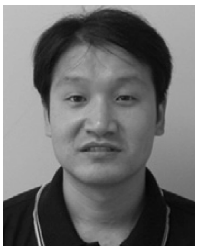
[33] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1859–1866.

[34] J. Xing, Z. Niu, J. Huang, W. Hu, and S. Yan, "Towards multi-view and partially-occluded face alignment," in *Proc. Comput. Vis. Pattern Recognit.*, 2013, pp. 1829–1836.

[35] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 FPS via regressing local binary features," in *Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 1685–1692.

[36] D. Cristinacce and T. F. Cootes, "Boosted regression active shape models," in *Proc. Brit. Mach. Vis. Conf.*, 2007, pp. 1–10.

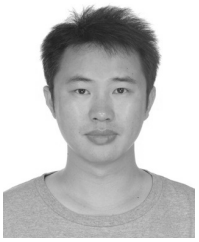
[37] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer, "Robust and accurate shape model fitting using random forest regression voting," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 278–291.
[38] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment via regressing local binary features," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1233–1245, Mar. 2016.
[39] M.-C. Roh, T. Oguri, and T. Kanade, "Face alignment robust to occlusion," in *Proc. Autom. Face Gesture Recognit.*, Mar. 2011, pp. 239–244.
[40] F. Yang, J. Huang, and D. Metaxas, "Sparse shape registration for occluded facial feature localization," in *Proc. Autom. Face Gesture Recognit.*, Mar. 2011, pp. 272–277.
[41] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 1513–1520.
[42] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 2385–2392.
[43] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, Mar. 1975.
[44] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, Jul. 2009.
[45] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 679–692.
[46] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2223–2231.
[47] X. Zhao, S. Shan, X. Chai, and X. Chen, "Locality-constrained active appearance model," in *Proc. Asian Conf. Comput. Vis.*, 2013, pp. 636–647.
[48] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.
[49] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
[50] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 720–735.
[51] D. Rathod, A. Vinay, S. S. Shylaja, and S. Natarajan, "Facial landmark localization—A literature survey," *Int. J. Current Eng. Technol.*, vol. 4, no. 3, pp. 1901–1907, Jun. 2014
[52] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. Audio Video-Based Biometric Pers. Authentication*, vol. 964. 1999, pp. 965–966.
[53] P. J. Phillips *et al.*, "Overview of the face recognition grand challenge," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Jun. 2005, pp. 947–954.
[54] J. Yan, Z. Lei, L. Wen, and S. Li, "The fastest deformable part model for object detection," in *Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 2497–2504.
[55] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[56] X. Xiong and F. De la Torre. (2014). "Supervised descent method for solving nonlinear least squares problems in computer vision." [Online]. Available: https://arxiv.org/abs/1405.0601
[57] G. Tzimiropoulos and M. Pantic, "Optimization problems for fast AAM fitting in-the-wild," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 593–600.
[58] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. Comput. Vis. Pattern Recognit.*, 2013, pp. 3444–3451.
[59] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 1944–1951.
[60] G. Tzimiropoulos and M. Pantic, "Gauss-Newton deformable part models for face alignment in-the-wild," in *Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 1851–1858.
[61] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. (2014). "Learning deep representation for face alignment with auxiliary attributes." [Online]. Available: https://arxiv.org/abs/1408.3967



**Qingshan Liu** (SM'07) received the M.S. degree from the Department of Auto Control, South-East University, in 2000, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Chinese Academic of Science, in 2003. He was as an Associate Professor with the National Laboratory of Pattern Recognition, Chinese Academic of Science, and an Associate Researcher with the Multimedia Laboratory, The Chinese University of Hong Kong, from 2004 to 2005. He was an Assistant Research Professor with the Computational Biomedicine Imaging and Modeling Center, Department of Computer Science, Rutgers The State University of New Jersey, from 2010 to 2011. He is currently a Professor with the School of Information and Control Engineering, Nanjing University of Information Science and Technology, China. His research interests are Image and Vision Analysis, including face image analysis, graph and hypergraph-based image and video understanding, medical image analysis, and event-based video analysis. He received the President Scholarship of the Chinese Academy of Sciences in 2003.

**Jiankang Deng** received the bachelor's and master's degrees from the Nanjing University of Information Science and Technology in 2012 and 2015, respectively. His research interest is face analysis.

**Jing Yang** received the bachelor's degree from the School of Information and Control, Nanjing University of Information Science and Technology, Nanjing, China, in 2014, where she is currently pursuing the master's degree. Her current research interests include face alignment and visual tracking.



**Guangcan Liu** (M'13) received the bachelor's degree in mathematics and the Ph.D. degree in computer science and engineering from Shanghai Jiao Tong University, Shanghai, China, in 2004 and 2010, respectively. He was a Post-Doctoral Researcher with the National University of Singapore, Singapore, from 2011 to 2012, the University of Illinois at Urbana–Champaign, Champaign, IL, USA, from 2012 to 2013, Cornell University, Ithaca, NY, USA, from 2013 to 2014, and Rutgers University, Piscataway, NJ, USA, in 2014. Since 2014, he has been a Professor with the School of Information and Control, Nanjing University of Information Science and Technology, Nanjing, China. His research interests touch on the areas of machine learning, computer vision, and image processing.

**Dacheng Tao** (F'15) is a Professor of Computer Science and the Director of the Centre for Artificial Intelligence, and the Faculty of Engineering and Information Technology, University of Technology Sydney. He mainly applies statistics and mathematics to artificial intelligence and data science. His research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and over 200 publications at prestigious journals and prominent conferences, such as the IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, ICML, CVPR, ICCV, ECCV, AISTATS, ICDM, and ACM SIGKDD, with several best paper awards, such as the Best Theory and Algorithm Paper Runner Up Award in the IEEE ICDM07, the Best Student Paper Award in the IEEE ICDM13, and the 2014 ICDM 10-Year Highest-Impact Paper Award. He received the 2015 Australian ScopusEureka Prize, the 2015 ACS Gold Disruptor Award, and the 2015 UTS ViceChancellors Medal for Exceptional Research. He is a Fellow of OSA, IAPR, and SPIE.