Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2017, Article ID 4579398, 11 pages
http://dx.doi.org/10.1155/2017/4579398

Research Article

A Robust Shape Reconstruction Method for Facial Feature Point Detection

Shuqiu Tan, Dongyi Chen, Chenggang Guo, and Zhiqi Huang

School of Automation Engineering, University of Electronic Science and Technology of China, No. 2006, Xiyuan Ave, West Hi-Tech Zone, Chengdu 611731, China

Correspondence should be addressed to Shuqiu Tan; tanshuqiu123136@hotmail.com and Dongyi Chen; dychen@uestc.edu.cn

Received 24 October 2016; Revised 18 January 2017; Accepted 30 January 2017; Published 19 February 2017

Academic Editor: Ezequiel López-Rubio

Copyright © 2017 Shuqiu Tan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Facial feature point detection has been receiving great research advances in recent years. Numerous methods have been developed and applied in practical face analysis systems. However, it is still a quite challenging task because of the large variability in expression and gestures and the existence of occlusions in real-world photo shoot. In this paper, we present a robust sparse reconstruction method for the face alignment problems. Instead of a direct regression between the feature space and the shape space, the concept of shape increment reconstruction is introduced. Moreover, a set of coupled overcomplete dictionaries termed the shape increment dictionary and the local appearance dictionary are learned in a regressive manner to select robust features and fit shape increments. Additionally, to make the learned model more generalized, we select the best matched parameter set through extensive validation tests. Experimental results on three public datasets demonstrate that the proposed method achieves a better robustness over the state-of-the-art methods.

1. Introduction

In most literatures, facial feature points are also referred to facial landmarks or facial fiducial points. These points mainly locate around edges or corners of facial components such as eyebrows, eyes, mouth, nose, and jaw (see Figure 1). Existing databases for method comparison are labeled with different number of feature points, varying from the minimum 5-point configuration [1] to the maximal 194-point configuration [2]. Generally facial feature point detection is a supervised or semisupervised learning process that trains model on a large number of labeled facial images. It starts from a face detection process and then predicts facial landmarks inside the detected face bounding box. The localized facial feature points can be utilized for various face analysis tasks, for example, face recognition [3], facial animation [4], facial expression detection [5], and head pose tracking [6].

In recent years, regression-based methods have gained increasing attention for robust facial feature point detection. Among these methods, a cascade framework is adopted to recursively estimate the face shape S of an input image, which is the concatenation of facial feature point coordinates. Beginning with an initial shape $S^{(1)}$, S is updated by inferring a shape increment ΔS from the previous shape:

$$\Delta S^{(t)} = W^{(t)} \Phi^{(t)} \left(I, S^{(t)} \right), \tag{1}$$

where $\Delta S^{(t)}$ and $W^{(t)}$ are the shape increment and linear regression matrix after t iterations, respectively. As the input variable of the mapping function $\Phi^{(t)}$, I denotes the image appearance and $S^{(t)}$ denotes the corresponding face shape. The regression goes to the next iteration by the additive formula:

$$S^{(t)} = S^{(t-1)} + \Delta S^{(t-1)}. \tag{2}$$

In this paper, we propose a sparse reconstruction method that embeds sparse coding in the reconstruction of shape increment. As a very popular signal coding algorithm, sparse coding has been recently successfully applied to the fields of computer vision and machine learning, such as feature selection and clustering analysis, image classification, and face recognition [7–11]. In our method, sparse overcomplete dictionaries are learned to encode various facial poses and local textures considering the complex nature of imaging

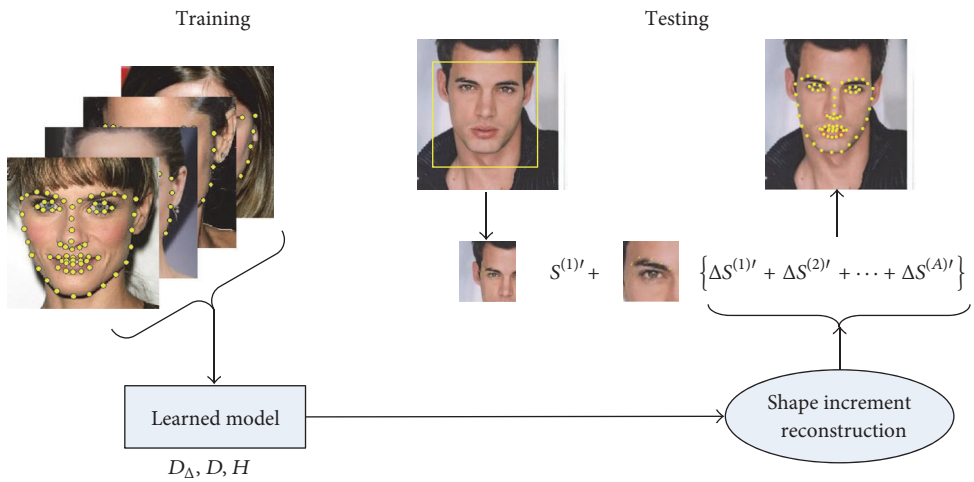


Figure 1: Schematic diagram of our robust sparse reconstruction method for facial feature point detection.

conditions. The schematic diagram of the proposed shape increment reconstruction method is illustrated in Figure 1. In the training stage, two kinds of overcomplete dictionaries need to be learned. The first kind of dictionary is termed shape increment dictionary since the atoms consist of typical shape increments in each iteration. The other kind of dictionary is termed local appearance dictionary because of the atoms abstracting the complex facial feature appearance. In the testing stage, local features are extracted around the shape points of current iteration and then encoded into feature coefficients using the local appearance dictionary. Thus shape increments can be reconstructed by the shape increment dictionary and the shape coefficients transformed from the feature coefficients. Considering the holistic performance, we adopt a way of alternate verification and local enumeration to get the best parameter set in a large number of experiments. Comparison with three previous methods is evaluated on three publicly available face datasets. Experimental results show that the proposed sparse reconstruction method achieves a superior detection robustness comparing with other methods.

The following contents of this paper are organized as follows: related work is introduced in Section 2. The proposed sparse reconstruction method is described in detail in Section 3 and experimental results are compared in Section 4. Finally we conclude the whole paper in Section 5.

2. Related Work

During the past two decades, a large number of methods have been proposed for facial feature point detection. Among the early methods, Active Appearance Model (AAM) [12] is a representative parametric model that aims to minimize the difference between the texture sampled from the testing image and the texture synthesized by the model. Later many improvements and extensions of AAM are proposed [6, 13–20]. To improve the efficiency in real-time system, Tzimiropoulos and Pantic [16] proposed a model to efficiently solve the AAM fitting problem. Tresadern et al. [14] used Haar-like features to reduce computation, which can help the mobile device to perform real-time tracking. Nguyen et al. [17–19] thought AAMs are easily converged to local minima. And to overcome this problem, they designed a new model that learns a cost function having local minima only at desired places. In terms of improving robustness, Huang et al. [6] combined view-based AAM [20] and Kalman filter to perform pose tracking and use shape parameters to rebuild the view space. Hansen et al. [13] introduced a nonlinear shape model that based on Riemannian elasticity model to handle the problem of poor pose initialization.

Generally, Constrained Local Model- (CLM-) based methods [21–25] are to learn a group of local experts and then take various shape prior for refinement. Vogler et al. [21] used Active Shape Model (ASM) [23] to build a 3D deformable model for real-time tracking. Yu et al. [22] used the mean-shift method [25] to rapidly approach the global optimum. And Liang et al. [24] constrained the structure of facial feature points using the component locations.

The aforementioned methods share the same characteristic which controls face shape variations through some certain parameters. But different from those methods, the regression-based methods [26–28] directly learn the regression function from image appearance to target output. Gao et al. [26] adopt a two-level cascaded boosted regression [27] structure to obtain a vectorial output for all points. To solve the problems of large shape variations and occlusions, Burgos-Artizzu et al. [28] improved their method in three aspects: firstly, they first reference pixels by linear interpolation between two landmarks. Secondly, the regression model directly embeds the occlusion information for robustness. Thirdly, they designed a smart initialization restart scheme to avoid unsuitable random initializations.

Our method belongs to the regression-based method, like [28–31]. However, our work is different from previous methods in several aspects. Firstly, existing methods, like [31], acquire the descent direction in a supervised learning manner. But in our proposed method, the information of descent direction is included in the sparse dictionaries for reconstructing the shape increment. And then unlike the method in [28], our method has no usage of the occlusion

information of each feature point. Finally the method from [30] designed a two-level boosted regression model to infer the holistic face shape. In our regression model, we refine the face shape by the learned two coupled dictionaries stage by stage.

3. Regression-Based Sparse Reconstruction Method

3.1. Problem Formulation. In this paper, the alignment target of all methods is assessed through the following formula:

$$\left\| S^{(\text{final})} - S^{(\text{ground-truth})} \right\|_2, \tag{3}$$

where $S^{(\text{final})}$ and $S^{(\text{ground-truth})}$ denote the final estimated shape and the corresponding ground-truth shape of an image, respectively. In the regression-based methods, the iterative equation is formulated as the following:

$$S^{(k)} = S^{(k-1)} + \Delta S^{(k-1)*}; \tag{4}$$

here $\Delta S^{(k-1)*}$ is a variable of shape increment after $k-1$ iterations, and its value should approximate the ground-truth shape increment $\Delta S^{(k-1)}$, where $\Delta S^{(k-1)} = S^{(\text{ground-truth})} - S^{(k-1)}$.

3.2. Multi-Initialization and Multiparameter Strategies. Multi-initialization means diversification of initial iteration shape which can improve robustness of the reconstruction model. Specifically, we randomly select multiple ground-truth face shapes from the training set to form a group of initial shapes for the current image. Obviously, the multi-initialization strategy is able to enlarge the training sample size and enrich extracted feature information that makes each regression model more robust, while, during the testing stage, multi-initialization can create more chances to step out of potential local minima that may lead to inaccurate feature point localization.

In our method, there are four key parameters that are the size of feature dictionary, the size of shape increment dictionary, and their corresponding sparsity. The selection of the four parameters has a direct influence on the learned reconstruction model. Therefore we do a large number of validation tests to find the best matched parameters. Then according to the validation results, we decide to adopt three sets of parameters to train the model.

3.3. The Learning of Sparse Coding. We use the Orthogonal Matching Pursuit (OMP) [32] algorithm and the K-Singular Value Decomposition (K-SVD) [33] algorithm to find the overcomplete dictionary by minimizing the overall reconstruction error:

$$\begin{aligned} \min_{D,\gamma} \quad & \|S - D\gamma\|_F^2 \\ \text{subject to} \quad & \|\gamma_i\|_0 \le T_0, \quad \forall i, \end{aligned} \tag{5}$$

where S is the input data and D and γ denote sparse dictionary and sparse coefficient, respectively. T_0 defines the number of nonzero values in a coefficient vector and is termed the sparsity.

3.4. The Learning of Shape Increment. In Supervised Descend Method (SDM [31]), authors adopt a linear regression equation to approximate shape increments:

$$\Delta S^{(k)} = R^{(k)} \phi^{(k)} + b^{(k)}; \tag{6}$$

here $\phi^{(k)}$ denotes Histograms of Oriented Gradients (HoG) features extracted from the shapes $S^{(k)}$ of previous stage. $R^{(k)}$ and $b^{(k)}$ are got from the training set by minimizing

$$\arg\min_{R^{(k)}, b^{(k)}} \sum_i \left\| \Delta S^{(ki)*} - R^{(k)} \phi^{(ki)} - b^{(k)} \right\|_2^2. \tag{7}$$

Different from the idea of linear approximation proposed in SDM, we introduce the concept of direct sparse reconstruction for reconstructing shape increments:

$$\Delta S^{(k)*} = D_\Delta^{(k)} \gamma_\Delta^{(k)}. \tag{8}$$

Here $D_\Delta^{(k)}$ and $\gamma_\Delta^{(k)}$ represent the shape increment dictionary and its corresponding sparse coefficient in the kth iteration, respectively. From another perspective the generic descent directions are embedded into the sparse dictionary $D_\Delta^{(k)}$ which can be more robust in facing large shape variations.

3.5. The Shape Regression Framework. To better represent local appearances around facial feature points, the extracted HoG features are also encoded into sparse coefficients:

$$\phi^{(k)} = D^{(k)} \gamma^{(k)}, \tag{9}$$

where $D^{(k)}$ and $\gamma^{(k)}$ are the local appearance dictionary and the local appearance sparse coefficient, respectively. Instead of a direct mapping from the whole feature space to the shape increment space, we propose to perform regression only in the sparse coefficient space. Since both coefficient matrixes are sufficient sparse, the regression matrix can be quickly solved. The equation is formulated as follows:

$$\gamma_\Delta^{(k)} = H^{(k)} \gamma^{(k)}. \tag{10}$$

Now we describe the shape regression framework in detail (see Pseudocode 1). During the training stage, we can get the shape prediction of the next stage using (4). By iterative learning shape increment $\Delta S^{(k)*}$, we can obtain the final face shape. Combining (10) and (8) $\Delta S^{(k)*}$ is computed from $\Delta S^{(k)*} = D_\Delta^{(k)} H^{(k)} \gamma^{(k)}$, where $D_\Delta^{(k)}$ and $\gamma^{(k)}$ are variables that can be acquired by the following sparse reconstruction formulas:

$$\begin{aligned} \arg\min_{D^{(k)}, \gamma^{(k)}} \quad & \left\| \phi^{(k)} - D^{(k)} \gamma^{(k)} \right\|_2^2, \\ \text{s.t} \quad & \left\| \gamma^{(k)} \right\|_0 \le T_\phi \\ \arg\min_{D_\Delta^{(k)}, \gamma_\Delta^{(k)}} \quad & \left\| \Delta S^{(k)} - D_\Delta^{(k)} \gamma_\Delta^{(k)} \right\|_2^2, \\ \text{s.t} \quad & \left\| \gamma_\Delta^{(k)} \right\|_0 \le T_\Delta. \end{aligned} \tag{11}$$

Input: Training set images and corresponding shapes: $I_{\text{train}} = [I_1, I_2, \ldots, I_N]$ and $S_{\text{train}} = [S_1, S_2, \ldots, S_N]$. Testing set images and corresponding initial shapes: $I_{\text{test}} = [I_1, I_2, \ldots, I_{N'}]$ and $S_{\text{initial}} = [S_1^{(1)\prime}, S_2^{(1)\prime}, \ldots, S_{N'}^{(1)\prime}]$. Sparse coding parameters set: T_Δ, T_ϕ, size of D_Δ and size of D. Total iterations Q.
Output: Final face shapes $S_{\text{test}} = [S_1^{(Q+1)\prime}, S_2^{(Q+1)\prime}, \ldots, S_{N'}^{(Q+1)\prime}]$.
Training Stage:
  for k **from** 1 **to** Q**, do**
    step 1: Given S^* and $S^{(k)}$, obtain $\Delta S^{(k)}$. Then extract $\phi^{(k)}$ from $S^{(k)}$.
    step 2: Sequentially get $\gamma^{(k)}$, $D_\Delta^{(k)}$ and $\gamma_\Delta^{(k)}$ using Equations (11).
    step 3: Get $H^{(k)}$ using Equation (12).
  end for
Training Model: $[D_\Delta^{(1)}, D_\Delta^{(2)}, \ldots, D_\Delta^{(Q)}]$, $[D^{(1)}, D^{(2)}, \ldots, D^{(Q)}]$ and $[H^{(1)}, H^{(2)}, \ldots, H^{(Q)}]$.
Testing Stage:
  for k **from** 1 **to** Q**, do**
    step 1: Similarly, extract $\phi^{(k)\prime}$ on $S^{(k)\prime}$.
    step 2: Given $D_\Delta^{(k)}$, $D^{(k)}$ and $H^{(k)}$, calculate $\Delta S^{(k)\prime}$ using Equation (14).
    step 3: Obtain $S^{(k+1)}$ using Equation (13).
  end for

Pseudocode 1: Pseudocode of our proposed regression-based method.

T_Δ and T_ϕ represent the shape increment sparsity and the local appearance sparsity, respectively. Given $\gamma^{(k)}$ and $\gamma_\Delta^{(k)}$ we can get $H^{(k)}$ by

$$\arg\min_{H^{(k)}} \left\| \gamma_\Delta^{(k)} - H^{(k)} \gamma^{(k)} \right\|_2^2. \tag{12}$$

Finally, we can generate a set of $[D_\Delta^{(1)}, D_\Delta^{(2)}, \ldots, D_\Delta^{(Q)}]$, $[D^{(1)}, D^{(2)}, \ldots, D^{(Q)}]$ and $[H^{(1)}, H^{(2)}, \ldots, H^{(Q)}]$ after Q iterations. Here Q is the number of iterations and $k = 1, 2, \ldots, Q$.

During the testing stage, we can get the local appearance coefficients $\gamma^{(k-1)\prime}$ using the already learned $D^{(k-1)}$. Then the final face shape is estimated using (16) and (17) after Q iterations.

$$S^{(k+1)\prime} = S^{(k)\prime} + \Delta S^{(k)\prime}, \tag{13}$$

$$\Delta S^{(k+1)\prime} = D_\Delta^{(k)} H^{(k)} \gamma^{(k)\prime}. \tag{14}$$

3.6. Major Contributions of the Proposed Method. In this section, we summarize the following three contributions of the proposed method:

(1) Sparse coding is utilized to learn a set of coupled dictionaries, named the shape increment dictionary and the local appearance dictionary. The solved corresponding sparse coefficients are embedded in a regression framework for approximating the ground-truth shape increments.

(2) A way of alternate verification and local enumeration is applied for selecting the best parameter set in extensive experiments. Moreover, it is shown in experimental results that the proposed method has a strong stability under different parameter settings.

(3) We also rebuild testing conditions that the top 5%, 10%, 15%, 20%, and 25% of the testing images are removed according to the descending order sorted by the normalized alignment error. And then the proposed method is compared with three classical methods on three publicly available face datasets. Results support that the proposed method achieves better detection accuracy and robustness than the other three methods.

4. Experiments

4.1. Face Datasets. In this section, three publicly available face datasets are selected for performance comparison: Labeled Face Parts in the Wild (LFPW-68 points and LFPW-29 points [16]) and Caltech Occluded Faces in the Wild 2013 (COFW) [34]. In the downloaded LFPW dataset, 811 training images and 224 testing images are collected. Both the 68 points' configuration and the 29 points' configuration labeled for the LFPW dataset are evaluated. The COFW dataset includes 1,345 training images and 507 testing images, and each image is labeled with 29 facial feature points and related binary occlusion information. Particularly, collected images in this dataset show a variety of occlusions and large shape variations.

4.2. Implementation Details

4.2.1. Codes. The implementation codes of SDM [31], Explicit Shape Regression (ESR) [30], and Robust Cascaded Pose Regression (RCPR) [28] are got from the Internet. Except that the codes of RCPR and ESR are released on the personal websites by at least one of the authors, we get the code of SDM from Github.

4.2.2. Parameter Settings. Generally, the size of shape increment dictionary and local appearance dictionary in our method depends on the dimensionality of the HoG descriptor. And in the following validation experiments, we will

Table 1: Comparison of different parameter sets on LFPW (68 points) dataset. Here T_Δ is fixed to 2.

T_Δ	T_ϕ	Size of D_Δ	Size of D	Q	K (tr & ts)	Mean errors
2	2	256	256	5	1 & 1	0.079381
		256	512	5	1 & 1	0.08572
		512	256	5	1 & 1	0.085932
		512	512	5	1 & 1	0.086731
2	4	256	256	5	1 & 1	0.081958
		256	512	5	1 & 1	0.083715
		512	256	5	1 & 1	0.086187
		512	512	5	1 & 1	0.087157
2	6	256	256	5	1 & 1	0.07937
		256	512	5	1 & 1	0.07986
		512	256	5	1 & 1	0.075987
		512	512	5	1 & 1	0.084429
2	8	256	256	5	1 & 1	0.075863
		256	512	5	1 & 1	0.082588
		512	256	5	1 & 1	0.077048
		512	512	5	1 & 1	0.082644
2	10	256	256	5	1 & 1	0.076178
		256	512	5	1 & 1	0.076865
		512	256	5	1 & 1	0.080907
		512	512	5	1 & 1	0.088414

introduce how to select the best combination of parameters. Parameters settings of SDM, ESR, and RCPR are consistent with the original settings reported in the papers. In SDM, the regression runs 5 stages. In ESR, the number of features in a fern and candidate pixel features are 5 and 400, respectively. To build the model, the method uses 10 and 500 stages to train a two-level boosted framework. And in RCPR, 15 iterations, 5 restarts, 400 features, and 100 random fern regressors are adopted.

4.2.3. Assessment Criteria. In our experiments, we use the following equation to calculate and normalize the alignment errors. Firstly, we calculate the localization errors between the ground-truth point coordinates and the detected point coordinates, that is, the Euclidean distance between two vectors. Then it is further normalized by the interocular distance as follows:

$$d_{\text{error}} = \frac{\|P - G\|_2}{\left\|G_{\text{leye}} - G_{\text{reye}}\right\|_2}. \tag{15}$$

In (15), P denotes the detected facial point coordinates and G denotes the ground-truth point coordinates. G_{leye} and G_{reye} denote the ground-truth center coordinates of left eye and right eye, respectively.

4.3. Experiments

4.3.1. Parameter Validation. In this section, we will introduce how to use the way of alternate verification and local enumeration to find the final values of parameters. As described above, there are six variables T_Δ, T_ϕ, size of D_Δ, size of D, Q, and K that need to be fixed; here K is the initialization number during the process of training and testing. Depending on the requirements of sparsity, the candidate values of T_Δ and T_ϕ are selected from the following set:

$$T_\Delta, T_\phi \in \{2, 4, 6, 8, 10\}. \tag{16}$$

Similarly, the candidate sizes of D_Δ, sizes of D, Q, and K form the following sets:

$$\begin{aligned} \text{size of } D_\Delta &\in \{256, 512\}, \\ \text{size of } D &\in \{256, 512\}, \\ Q &\in \{1, 2, 3, 4, 5\}, \\ K &\in \{1, 6, 8, 10, 12, 14\}. \end{aligned} \tag{17}$$

Firstly the values of Q and K are set to 5 and 1, respectively. Note that the value of K in the testing stage should be equal to the value of K in the training stage. Then we set the value of T_ϕ to 2, 4, 6, 8, and 10 sequentially. The size of D_Δ and D is selected in random combination. For different values of T_Δ we can get five groups of results. And Table 1 gives the detailed results when T_Δ is fixed to 2. From Table 1 we may find that the parameter set {2, 8, 256, 256, 5, 1 & 1} achieves the lowest alignment error. Similarly we conduct the rest experiments and find the best parameter sets. The corresponding sparsity is also fixed and therefore we get three sets of parameters that are {6, 10, 512, 256}, {8, 8, 512, 256}, and {10, 10, 512, 256}. In Table 2, we test the multi-initialization and multiparameter strategies while the regression runs 4 iterations and 10 initializations with different parameter settings. In the final, all point localizations are averaged to get the fusion result.

Table 2: Comparison of multi-initialization and multiparameter strategies on LFPW (68 points) dataset. Here Q and K are set to 4 and 10, respectively.

T_Δ	T_ϕ	Size of D_Δ	Size of D	Q	K (tr & ts)	Mean errors	Fusion errors
6	10	512	256	4	10 & 10	0.062189	0.055179
10	10	512	256	4	10 & 10	0.06075	
8	8	512	256	4	10 & 10	0.061787	

Table 3: Mean error of each facial component on LFPW (68 points), LFPW (29 points), and COFW (29 points) datasets. The top 5% maximal mean errors of the testing facial images in each dataset are removed.

(a)

	Method	Contour	Eyebrow	Mouth	Nose	Eye
LFPW (68 points)	SDM	0.0829	0.0619	0.0478	0.0395	0.0369
	ESR	0.0862	0.0750	0.0651	0.0596	0.0527
	RCPR	0.0948	0.0690	0.0562	0.0493	0.0433
	Our method	0.0747	0.0587	0.0455	0.0405	0.0392

(b)

	Method	Jaw	Eyebrow	Mouth	Nose	Eye
LFPW (29 points)	SDM	0.0422	0.0422	0.0410	0.0422	0.0328
	ESR	0.0570	0.0459	0.0531	0.0502	0.0400
	RCPR	0.0748	0.0507	0.0568	0.0509	0.0357
	Our method	0.0382	0.0403	0.0401	0.0406	0.0323

(c)

	Method	Jaw	Eyebrow	Mouth	Nose	Eye
COFW (29 points)	SDM	0.0713	0.0714	0.0709	0.0600	0.0519
	ESR	0.1507	0.1022	0.1082	0.0952	0.0801
	RCPR	0.1209	0.0810	0.0781	0.0655	0.0539
	Our method	0.0668	0.0642	0.0702	0.0567	0.0497

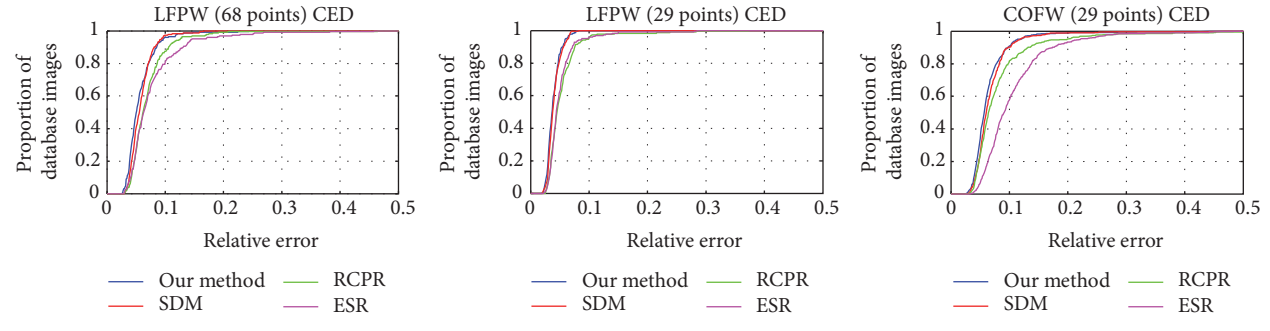
4.3.2. Comparison with Previous Methods. Due to the existence of a small number of facial images having large shape variations and severe occlusions, it challenges the random multi-initialization strategy which fails to generate an appropriate starting shape. Therefore we compare our method with three classic methods on rebuilt datasets. These datasets still include most of the images coming from LFPW (68 points), LFPW (29 points), and COFW (29 points). We just remove the top 5%, 10%, 15%, 20%, and 25% of the testing facial images in each dataset by sorting the alignment errors in a descending order (see Figure 2).

In Figure 2, all curves of COFW show a more dispersive distribution than the other two datasets. Since this dataset consists of many more facial images with large shape variations and occlusions, it may affect the detection accuracy more or less. Meanwhile, the irregular textural features around facial feature points are challenging for learning of structural model during the training. Obviously, in Figure 2, the curves of our proposed method are superior to the others. Additionally the LFPW (68 points) and LFPW (29 points) share the same facial images but different face shapes, so we may find some useful information about the performance of methods through these datasets.

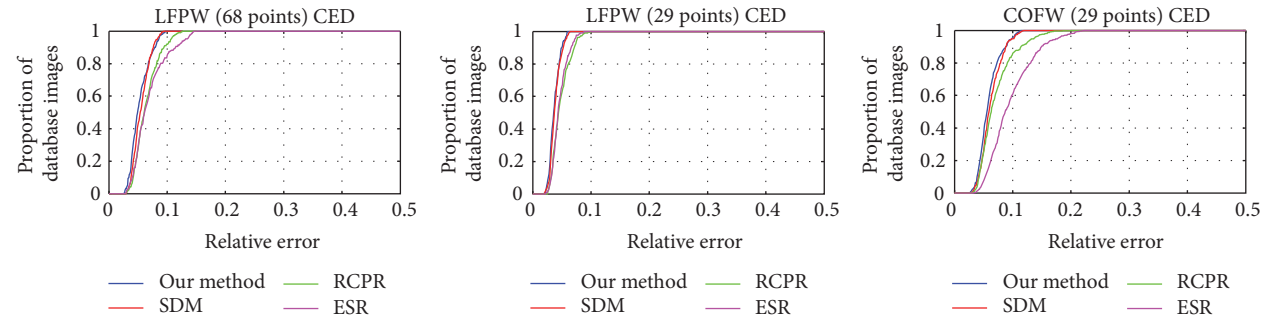
In general, the more facial feature points are, the more difficult they are to detect. By comparing among five facial components, the mean errors of nose and eyes given in Tables 3 and 4 do not change obviously across three datasets, because the vicinal textural information of eyes is easy to recognize and the textural information around nose has a less possibility to be occluded. Moreover, the facial feature points located in the regions of nose and eyes are denser than the points of contour, which is also benefit to the regressive searching process.

Figure 3 shows the alignment errors of four methods tested on LFPW (68 points), LFPW (29 points), and COFW (29 points) datasets. In Figure 3 we may find that the mean error curves show a rapid descending trend when the most difficult 5% of testing images are removed. It indicates that the statistical average can be biased by a few challenge images. Then as the removal proportion increases, all the curves become smoother. It shows in Figure 3 that our proposed method is more stable than other methods, which means our training model has a robustness in dealing with occlusions and large shape variations.

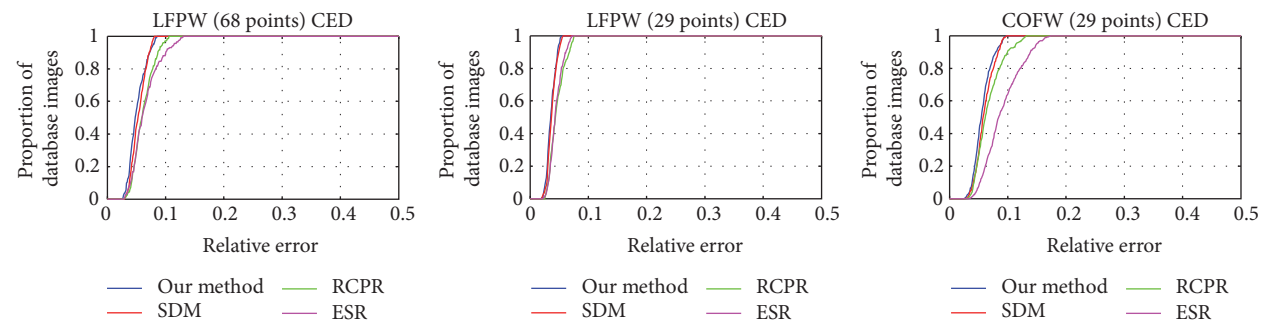
Specifically, we plot the detection curves of five facial components in Figure 4. It is obvious in Figure 4 that ESR and RCPR has a less competitive performance for localizing each facial components. And our method shows better robustness in localizing feature points that belong to eyebrows and contour, since these two facial components are very likely to



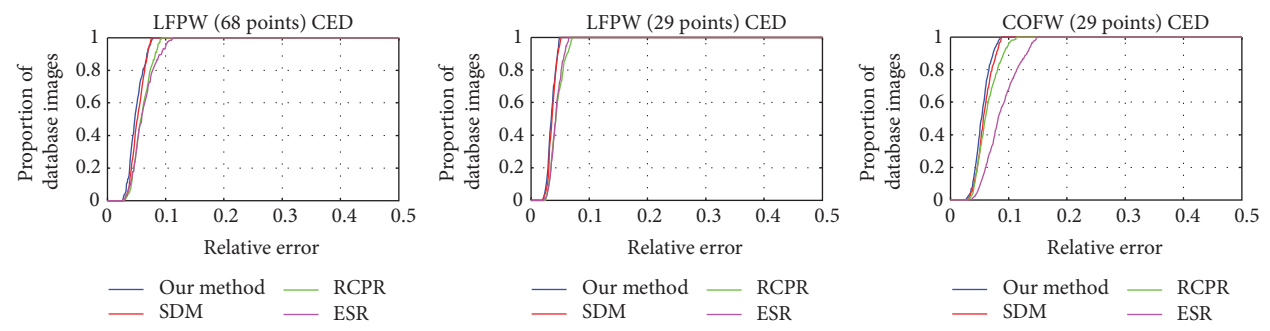
(a)



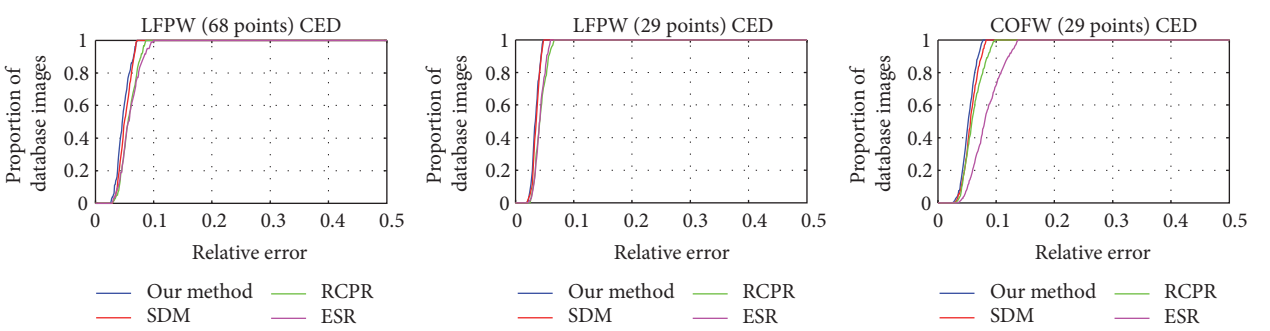
(b)



(c)

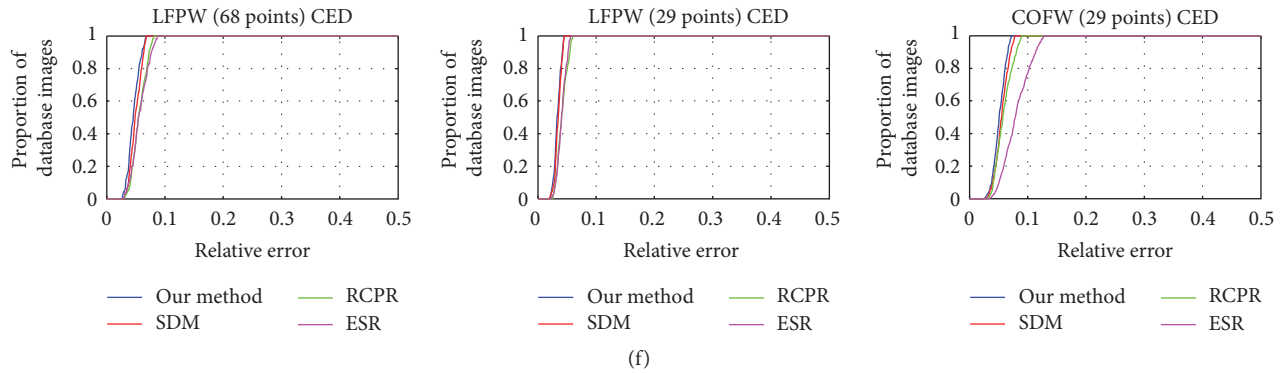


(d)



(e)

Figure 2: Continued.



(f)

Figure 2: Cumulative Error Distribution (CED) curves of four methods tested on LFPW (68 points), LFPW (29 points), and COFW (29 points) datasets. The top (a) 0%, (b) 5%, (c) 10%, (d) 15%, (e) 20%, and (f) 25% of the testing images are removed according to the descending order sorted by the normalized alignment errors.

Table 4: Mean error of each facial component on LFPW (68 points), LFPW (29 points), and COFW (29 points) datasets. The top 25% maximal mean errors of the testing facial images in each dataset are removed.

(a)

	Method	Contour	Eyebrow	Mouth	Nose	Eye
LFPW (68 points)	SDM	0.0718	0.0581	0.0433	0.0363	0.0337
	ESR	0.0746	0.0634	0.0535	0.0435	0.0408
	RCPR	0.0830	0.0615	0.0484	0.0414	0.0367
	Our method	0.0634	0.0518	0.0406	0.0364	0.0332

(b)

	Method	Jaw	Eyebrow	Mouth	Nose	Eye
LFPW (29 points)	SDM	0.0385	0.0381	0.0376	0.0389	0.0295
	ESR	0.0498	0.0419	0.0461	0.0435	0.0350
	RCPR	0.0637	0.0460	0.0471	0.0433	0.0309
	Our method	0.0336	0.0360	0.0365	0.0362	0.0292

(c)

	Method	Jaw	Eyebrow	Mouth	Nose	Eye
COFW (29 points)	SDM	0.0607	0.0643	0.0614	0.0525	0.0457
	ESR	0.1255	0.0873	0.0882	0.0781	0.0664
	RCPR	0.1055	0.0679	0.0633	0.0533	0.0440
	Our method	0.0561	0.0569	0.0603	0.0497	0.0435

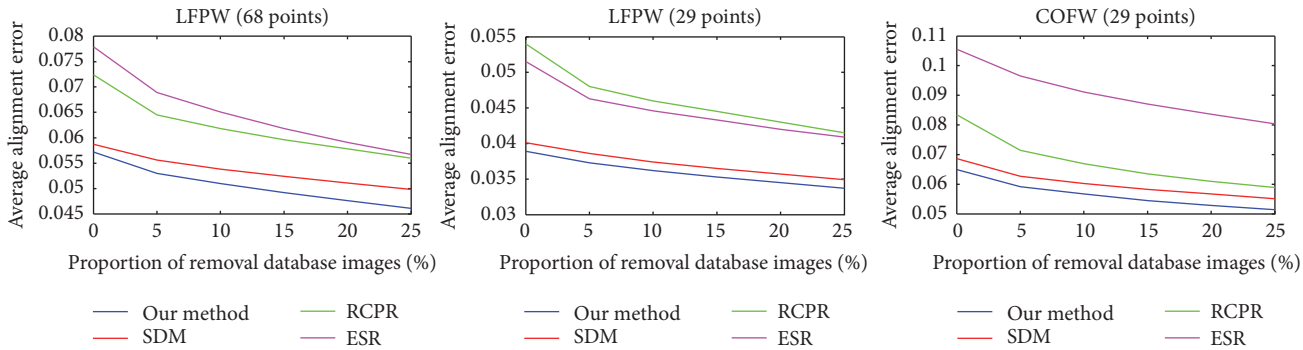


Figure 3: Mean errors of four methods tested on LFPW (68 points), LFPW (29 points), and COFW (29 points) datasets.

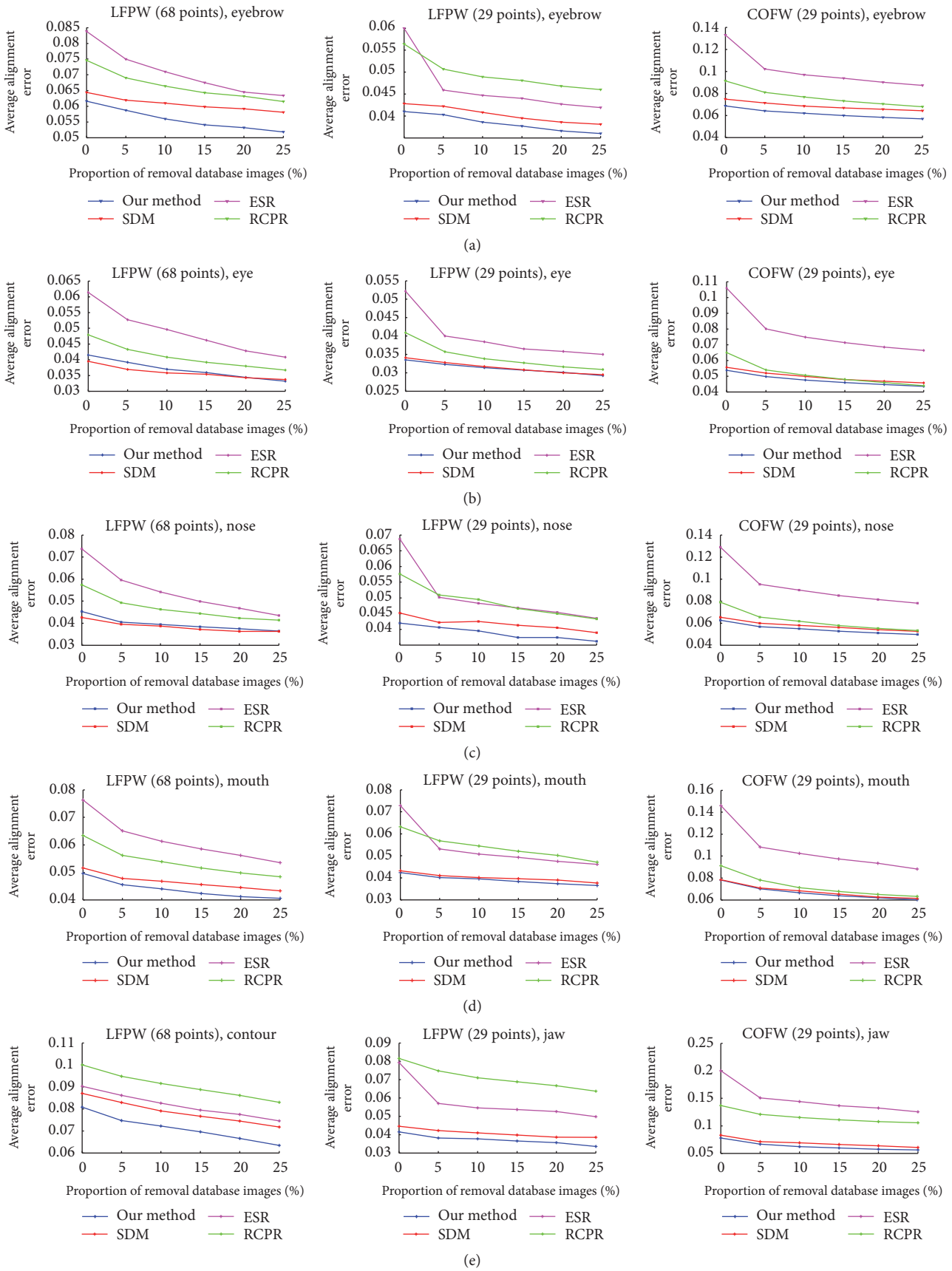


Figure 4: Facial feature point detection curves of four methods for each facial component on LFPW (68 points), LFPW (29 points), and COFW (29 points) datasets. (a) Eyebrow. (b) Eye. (c) Nose. (d) Mouth. (e) Contour or jaw.

be occluded by hair or objects and have a more separated distribution pattern. Experimental results demonstrates that our proposed method can estimate facial feature points with high accuracy and is able to deal with the task of face alignment on complex occlusions and large shape variations.

5. Conclusion

A robust sparse reconstruction method for facial feature point detection is proposed in this paper. In the method, we build the regressive training model by learning a coupled set of shape increment dictionaries and local appearance dictionaries which are learned to encode various facial poses and rich local textures. And then we apply the sparse model to infer the final face shape locations of an input image by continuous reconstruction of shape increments. Moreover, in order to find the best matched parameters, we perform extensive validation tests by using the way of alternate verification and local enumeration. It shows in the comparison results that our sparse coding based reconstruction model has a strong stability. In the later experiments, we compare our proposed method with three classic methods on three publicly available face datasets when removing the top 0%, 5%, 10%, 15%, 20%, and 25% of the testing facial images according to the descending order of alignment errors. The experimental results also support that our method is superior to the others in detection accuracy and robustness.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work is supported by the National Key Research & Development Plan of China (no. 2016YFB1001401) and National Natural Science Foundation of China (no. 61572110).

References

[1] Z. Zhang, P. Luo, C. Change Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proceedings of the European Conference on Computer Vision (ECCV '14)*, pp. 94–108, Springer International, Zurich, Switzerland, 2014.

[2] L. Vuong, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proceedings of the European Conference on Computer Vision*, pp. 679–692, Springer, Florence, Italy, October 2012.

[3] H.-S. Lee and D. Kim, "Tensor-based AAM with continuous variation estimation: application to variation-robust face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 1102–1116, 2009.

[4] T. Weise, S. Bouaziz, L. Hao, and M. Pauly, "Realtime performance-based facial animation," *ACM Transactions on Graphics*, vol. 30, no. 4, p. 77, 2011.

[5] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, and S. Sridharan, "Person-independent facial expression detection using constrained local models," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG '11)*, pp. 915–920, IEEE, Santa Barbara, Calif, USA, March 2011.

[6] C. Huang, X. Ding, and C. Fang, "Pose robust face tracking by combining view-based AAMs and temporal filters," *Computer Vision and Image Understanding*, vol. 116, no. 7, pp. 777–792, 2012.

[7] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1794–1801, IEEE, Miami, Fla, USA, June 2009.

[8] S. Gao, I. W.-H. Tsang, L.-T. Chia, and P. Zhao, "Local features are not lonely—Laplacian sparse coding for image classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 3555–3561, IEEE, June 2010.

[9] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[10] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 625–632, IEEE, Providence, RI, USA, June 2011.

[11] D.-S. Pham, O. Arandjelovic, and S. Venkatesh, "Achieving stable subspace clustering by post-processing generic clustering results," in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN '16)*, pp. 2390–2396, IEEE, Vancouver, Canada, July 2016.

[12] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[13] M. F. Hansen, J. Fagertun, and R. Larsen, "Elastic appearance models," in *Proceedings of the 22nd British Machine Vision Conference (BMVC '11)*, September 2011.

[14] P. A. Tresadern, M. C. Ionita, and T. F. Cootes, "Real-time facial feature tracking on a mobile device," *International Journal of Computer Vision*, vol. 96, no. 3, pp. 280–289, 2012.

[15] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Robust and efficient parametric face alignment," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 1847–1854, IEEE, Barcelona, Spain, November 2011.

[16] G. Tzimiropoulos and M. Pantic, "Optimization problems for fast AAM fitting in-the-wild," in *Proceedings of the 14th IEEE International Conference on Computer Vision (ICCV '13)*, pp. 593–600, IEEE, Sydney, Australia, December 2013.

[17] M. H. Nguyen and F. De la Torre, "Local minima free parameterized appearance models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, IEEE, Anchorage, Alaska, USA, June 2008.

[18] M. H. Nguyen and F. De La Torre, "Learning image alignment without local minima for face detection and tracking," in *Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG '08)*, pp. 1–7, IEEE, Amsterdam, Netherlands, September 2008.

[19] N. M. Hoai and F. De la Torre, "Metric learning for image alignment," *International Journal of Computer Vision*, vol. 88, no. 1, pp. 69–84, 2010.

[20] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor, "View-based active appearance models," *Image and Vision Computing*, vol. 20, no. 9-10, pp. 657–664, 2002.

[21] C. Vogler, Z. Li, A. Kanaujia, S. Goldenstein, and D. Metaxas, "The best of both worlds: combining 3D deformable models with active shape models," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–7, IEEE, Rio de Janeiro, Brazil, October 2007.

[22] 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 *Proceedings of the 14th IEEE International Conference on Computer Vision (ICCV '13)*, pp. 1944–1951, IEEE, Sydney, Australia, December 2013.

[23] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," in *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[24] L. Liang, R. Xiao, F. Wen, and J. Sun, "Face alignment via component-based discriminative search," in *Computer Vision—ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II*, vol. 5303 of *Lecture Notes in Computer Science*, pp. 72–85, Springer, Berlin, Germany, 2008.

[25] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.

[26] H. Gao, H. K. Ekenel, and R. Stiefelhagen, "Face alignment using a ranking model based on regression trees," in *Proceedings of the British Machine Vision Conference (BMVC '12)*, pp. 1–11, London, UK, September 2012.

[27] N. Duffy and D. Helmbold, "Boosting methods for regression," *Machine Learning*, vol. 47, no. 2-3, pp. 153–200, 2002.

[28] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '13)*, pp. 1513–1520, Sydney, Australia, 2013.

[29] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1078–1085, IEEE, San Francisco, Calif, USA, June 2010.

[30] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[31] X. Xiong and F. De La Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 532–539, Portland, Ore, USA, June 2013.

[32] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proceedings of the 27th Asilomar Conference on Signals, Systems & Computers*, pp. 40–44, Pacific Grove, Calif, USA, November 1993.

[33] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[34] X. Gao, Y. Su, X. Li, and D. Tao, "A review of active appearance models," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 40, no. 2, pp. 145–158, 2010.