Towards Multi-view and Partially-occluded Face Alignment

Junliang Xing[1]*, Zhiheng Niu[2], Junshi Huang[2], Weiming Hu[1], Shuicheng Yan[2]
[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, P. R. China
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576
{jlxing,wmhu}@nlpr.ia.ac.cn {niuzhiheng,junshi.huang,eleyans}@nus.edu.sg

Abstract

We present a robust model to locate facial landmarks under different views and possibly severe occlusions. To build reliable relationships between face appearance and shape with large view variations, we propose to formulate face alignment as an ℓ_1-induced Stagewise Relational Dictionary (SRD) learning problem. During each training stage, the SRD model learns a relational dictionary to capture consistent relationships between face appearance and shape, which are respectively modeled by the pose-indexed image features and the shape displacements for current estimated landmarks. During testing, the SRD model automatically selects a sparse set of the most related shape displacements for the testing face and uses them to refine its shape iteratively. To locate facial landmarks under occlusions, we further propose to learn an occlusion dictionary to model different kinds of partial face occlusions. By deploying the occlusion dictionary into the SRD model, the alignment performance for occluded faces can be further improved. Our algorithm is simple, effective, and easy to implement. Extensive experiments on two benchmark datasets and two newly built datasets have demonstrated its superior performances over the state-of-the-art methods, especially for faces with large view variations and/or occlusions.

1. Introduction

Face alignment, *i.e*., locating the facial landmark points of a face image, is an important computer vision task and essential for many other applications, *e.g*., face recognition [29], face synthesis [24], and 3D face modeling [13]. Although many efforts are devoted in solving this task and great progress has been made during the past decades [5, 28, 9, 3, 4, 30, 27], face alignment still remains a very challenging task, especially when the face images are taken from different views and/or undergo severe occlusions.

Traditional approaches addressing the face alignment problem employ parameterized models to describe the face appearance and shape. The Active Shape Model (ASM) [7] represents the face shapes by conducting principal component analysis on the manually labeled training samples and iteratively fitting the face instance in a test image using the learned face shape. The Active Appearance Model (AAM) [5, 15] further reconstructs the entire face using an appearance model and estimates the face shape by minimizing the texture residual. The AAM approach, together with ASM, provides a general framework for solving the face alignment problem. Following studies [28, 6, 4, 27], however, have found that the classic AAM approach is computationally expensive and sensitive to the initialization due to the involved gradient descent based optimization.

To deal with these problems, there are two main kinds of models to improve the classic ASM and AAM framework. The first kind is the part based models [9, 19, 3, 30, 6]. These models perform face alignment by maximizing a posterior probability of part locations given the image and then fuse the probabilities of all the parts together enforced by a global shape model, *e.g*. enhanced ASM [9, 19] or pictorial structures [30], to generate the final result. Unlike AAM which tries to approximate the raw image pixels directly, the constrained local models [9] employ an extended appearance model to generate the feature templates of the parts, which obtains improved robustness and accuracy.

The other kind of models is the regression based face alignment approaches [8, 18, 11, 22, 4, 6, 27], which directly learn a mapping function from the image appearance to the face shape. The distinctions among these methods mainly lie in the employed learning algorithm (*e.g*. boosting [8], random forest [6], or non-linear least squares [27]) and the adopted features (*e.g*. Haar wavelets [8], random ferns [11], or SIFT [27]). The pose-indexed features [11], which are obtained by re-computing the features every time when a new face landmark estimation is updated, proves to be important for learning a robust alignment model [11, 4, 27]. Moreover, with an initial shape provided by the face detector [23], mapping from pose-indexed features to face landmark displacements provides a natural and effective way to iteratively update the estimated face landmarks towards their true positions. Since the mapping functions are usu-

*This work was mainly performed when J. Xing was staying in NUS.

ally non-linear [8, 18, 11, 22, 6], training them needs many annotated samples and usually takes hours of time to learn the complex mapping relationships.

For nearly frontal faces, most of the existing face alignment algorithms can work considerably well. For faces with large view variations, however, their performances often degrade significantly. This is mainly due to the complex appearance-shape relations exhibited in multi-view faces. For faces with severe occlusions, the situation becomes even worse, since most of existing face alignment algorithms do not explicitly model the occlusions. Even though some algorithms are claimed to be robust to occlusions, the underlying mechanism on how to model occlusions and why it works is not clear. Therefore, if a face alignment model can inherently or explicitly address the face view variation and occlusion problems simultaneously, it will be very useful to improve alignment performance.

To obtain such a model for multi-view and partially-occluded face alignment, we propose an ℓ_1-induced Stagewise Relational Dictionary (SRD) model to learn consistent and coherent relationships between face appearance and shape for face images with large view variations. The SRD model jointly learns at each training stage two relational dictionaries, one for face shape using landmark displacement and one for object appearance using pose-indexed features. The learned dictionaries automatically capture distinct modes of face shape and related modes in appearance, which directly characterize faces from different views and thus form a multi-view face model. The relations between shape and appearance are naturally embedded in the relational dictionaries and can be obtained quite efficiently. Given a test face image, the SRD model iteratively selects a small subset of related appearance modes from the dictionary via sparse representation, and then predicts the shape displacement towards the true face shape.

To perform robust face alignment under occlusions, we further propose to learn an occlusion dictionary, whose elements form different elemental occlusion patterns and a sparse combination of them simulates different kinds of face occlusions. By deploying the occlusion dictionary into the original SRD model with a modified joint learning method, its robustness is further improved, especially for faces with occlusions. We conduct extensive experiments to evaluate and analyze the proposed SRD model for face alignment under different experimental settings and over several benchmark datasets. The results demonstrate state-of-the-art performance of the proposed algorithm, especially on multi-view and partially-occluded face alignment tasks.

2. Stagewise Relational Dictionary Learning

Face alignment is essentially an appearance-shape modeling process. One of the main challenges comes from large face view variations which cause very complex appearance-shape relationships. To learn robust face appearance and shape models for multi-view faces and to capture consistent appearance-shape relations, we propose to formulate face alignment as an ℓ_1-induced relational dictionary learning problem and develop a stagewise optimization procedure to learn multiple relational dictionaries.

2.1. Model Formulation

Denote a training set with N training samples as $\mathcal{X} = \{(\mathbf{x}_i, \mathbf{p}_i)\}_{i=1}^N$, where each sample $(\mathbf{x}_i, \mathbf{p}_i)$ contains one face image $\mathbf{x}_i$ and the labeled landmark positions $\mathbf{p}_i$. A face alignment model $\mathcal{M}$ need capture the relationships between face appearance and shape by abstracting $\mathcal{X}$ into a compact representation Θ, *e.g.*, a set of model parameters. This can be achieved by minimizing a loss function over the training set. Generally, the loss function can be represented as $l(\mathbf{a}, \mathbf{s}, \Theta)$, where $\mathbf{a} \in \mathbb{R}^{n_a}$ and $\mathbf{s} \in \mathbb{R}^{n_s}$ express the face appearance and shape features from one training sample respectively. The loss function measures the incompatibilities between them. Note that here the face appearance $\mathbf{a}$ is not necessarily to be the face image $\mathbf{x}$, more expressive image features can be used to represent face appearance. Similarly, the face shape $\mathbf{s}$ is not restricted to be landmark locations $\mathbf{p}$, and can be other more effective face shape representations, *e.g.*, landmark displacement [4, 27]. With the loss function, learning the model is equivalent to solving

$$\hat{\Theta} = \operatorname*{argmin}_{\Theta} \sum_{i=1}^{N} l(\mathbf{s}_i, \mathbf{a}_i, \Theta). \tag{1}$$

The definition of the loss function, therefore, has a fundamental impact on the final face alignment model. Previous methods define this loss function either solely based on face appearance [5, 9, 6, 27] or on face shape [7, 8, 11, 3, 4]. For multi-view face alignment, since both the face appearance and shape exhibit huge variations, the appearance-shape relationship becomes very complex. It therefore requires the loss function not only to guide the learning process to automatically find reliable modes of the face appearance and shape, but also to ensure the learned model captures consistent face shape-appearance relationships. To this end, we propose to define the loss function as

$$l(\mathbf{s}, \mathbf{a}, \mathbf{D}) \triangleq \min_{\mathbf{c} \in \mathbb{R}^m} \|[\mathbf{s}; \mathbf{a}] - \mathbf{D}\mathbf{c}\|_2^2 + \lambda \|\mathbf{c}\|_1. \tag{2}$$

As an instantiation of Θ in Eqn. (1), $\mathbf{D} \in \mathbb{R}^{n \times m}$ is a dictionary used to simultaneously represent the face shape and appearance, $n = n_a + n_s$, and m is the dictionary size. The underlying shape-appearance relationship is enforced to be consistent by sharing the same representation coefficients $\mathbf{c}$. Note that $\mathbf{c}$ is encouraged to be sparse (controlled by λ). The sparsity ensures a face appearance-shape instance to be represented by only a few elements in the dictionary. The sparsity assumption has been proved to be very effective for many vision problems [20, 25]. In our problem, this sparsity regularization will benefit both the model training and

testing. For model training, since the dictionary can represent one sample using only a few of its elements, it will be forced to learn distinct shape-appearance modes so as to well represent all the samples with different shapes and appearances. For model testing, since a testing face is represented using only a few modes, it provides a mechanism to automatically select the most related modes to the testing face and thus generate more robust and accurate estimation.

In practice, it usually constrains $\mathbf{D}$'s columns $\mathbf{d}_1,\ldots,\mathbf{d}_m$ to have ℓ_2-norms less than or equal to 1 to avoid trivial solutions. The constraint set of $\mathbf{D}$ is thus denoted as

$$\mathcal{D}^{n\times m} \triangleq \{\mathbf{D}\in\mathbb{R}^{n\times m}, \text{s.t. } \forall j\in\{1,\ldots,m\}, \|\mathbf{d}_j\|_2\leq 1\}. \quad (3)$$

Now, by minimizing the loss function in Eqn. (2) over the training set $\mathcal{X}$, the dictionary model can be learned by

$$\hat{\mathbf{D}} = \underset{\mathbf{D}\in\mathcal{D}^{n\times m}}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N}\left\{\min_{\mathbf{c}\in\mathbb{R}^m} \|[\mathbf{s}_i;\mathbf{a}_i]-\mathbf{D}\mathbf{c}\|_2^2+\lambda\|\mathbf{c}\|_1\right\}. \quad (4)$$

2.2. Model Learning

Before learning the model in Eqn. (4), we provide some discussions on how to represent the face shape and appearance. For the face shape, since the groundtruth landmarks are not accessible during testing, we cannot directly use the groundtruth landmark positions as the shape feature in training, but instead use the displacements between the true landmark locations and the estimated landmark locations, *e.g.*, those provided by a face detector [23]. For the face appearance, we can use the pose-indexed features to build a robust representation [11, 4, 27]. The original training set, therefore, can be transformed to a new one which can be directly used for learning the model. Formally, this new training set can be denoted as $\mathcal{X}^0 = \{(\mathbf{s}_i^0, \mathbf{a}_i^0)\}_{i=1}^N$, where $\mathbf{s}_i^0 = \mathbf{p}_i - \mathbf{p}_i^0$, and $\mathbf{a}_i^0 = \mathbf{h}(\mathbf{x}_i, \mathbf{p}_i^0)$. Here $\mathbf{p}_i^0$ denotes the estimated landmark locations and $\mathbf{h}(\mathbf{x}_i, \mathbf{p}_i^0)$ denotes a pose-indexed feature extraction function.

With the new training set, a local minimum solution of the problem in Eqn. (4) can be obtained using a widely used two-step optimization process [2, 14, 21]: in the first step $\mathbf{D}$ is fixed to minimize the loss function with respect to the $\mathbf{c}$; and in the second step $\mathbf{c}$ is fixed to perform gradient descent method to minimize the loss function with respect to $\mathbf{D}$. The obtained dictionary in this way, however, is not compatible to the testing setting of the model, because during testing we can only obtain the representation coefficients from the appearance part of dictionary $\mathbf{D}$. To fit the testing settings of the model, we propose a stagewise rational dictionary learning process to learn the model.

Mathematically, the loss function in Eqn. (2) can be equivalently written as

$$l(\mathbf{a},\mathbf{s},\mathbf{D}) = \min_{\mathbf{c}\in\mathbb{R}^m} \|\mathbf{s}-\mathbf{D}_s\mathbf{c}\|_2^2+\|\mathbf{a}-\mathbf{D}_a\mathbf{c}\|_2^2+\lambda\|\mathbf{c}\|_1, \quad (5)$$

where we split the original dictionary $\mathbf{D}$ into two parts, *i.e.*, $\mathbf{D}=[\mathbf{D}_s;\mathbf{D}_a]$, $\mathbf{D}_s\in\mathbb{R}^{n_s\times m}$ and $\mathbf{D}_a\in\mathbb{R}^{n_a\times m}$ corresponding to the shape and appearance dictionary respectively. Note that these two dictionaries are related to each other and their underlying relationships are controlled by the coefficients $\mathbf{c}$, which ensures the two dictionaries to represent a face shape-appearance instance consistently. We refer to these two dictionaries together as a *relational* dictionary. With this substitution, the problem in Eqn. (4) becomes

$$\begin{aligned}\{\hat{\mathbf{D}}_s,\hat{\mathbf{D}}_a\} &= \underset{\mathbf{D}_s,\mathbf{D}_a}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} l(\mathbf{s}_i,\mathbf{a}_i,\mathbf{D}_s,\mathbf{D}_a)\\ \text{s.t. } l(\mathbf{s}_i,\mathbf{a}_i,\mathbf{D}_s,\mathbf{D}_a) &= \min_{\mathbf{c}} \|\mathbf{s}_i-\mathbf{D}_s\mathbf{c}\|_2^2 \\ &\quad +\|\mathbf{a}_i-\mathbf{D}_a\mathbf{c}\|_2^2+\lambda\|\mathbf{c}\|_1 \\ \mathbf{c}\in\mathbb{R}^m, \mathbf{D}_s\in\mathcal{D}^{n_s\times m}&, \mathbf{D}_a\in\mathcal{D}^{n_a\times m}.\end{aligned} \quad (6)$$

To fit model testing settings, we propose a four-step iterative procedure to solve the above problem: in Step 1, we fix the $\mathbf{D}_a$ to learn $\mathbf{D}_s$ and $\mathbf{c}$; in Step 2 we update $\mathbf{D}_a$ using $\mathbf{D}_s$ and $\mathbf{c}$; in Step 3 we fix $\mathbf{D}_s$ and learn $\mathbf{D}_a$ and $\mathbf{c}$, and in Step 4 we updates $\mathbf{D}_s$ using $\mathbf{D}_a$ and $\mathbf{c}$. These four steps are summarized in Line 3-10 in Algorithm 1, where a batch training mode is used for speeding up convergence.

The underlying oracles for this optimization procedure hide behind the starting steps (*i.e.* Step 1 and Step 2) and exiting steps (*i.e.* Step 3 and Step 4) in the iteration. In Step 1, in order to capture different modes from multi-view face shapes, the face shape dictionary is firstly learned and then used to initialize the face appearance dictionary in Step 2. In Step 4, since we have no access to the face shapes during testing, the representation coefficients can only be obtained from the face appearance dictionary. Therefore, the coefficients used to update the face shape dictionary are updated only from the face appearance dictionary in Step 3. Like the two-step optimization procedure in [2, 14, 21], the proposed four-step optimization procedure may also not find the global optimal solution of the problem in Eqn. (4), but it can guarantee a compatible setting for the testing of the model and thus produce a more effective model.

2.3. Stagewise Optimization

Denoting the learned relational dictionary model using the four-step optimization procedure from training set $\mathcal{X}^0$ as $\mathbf{D}^0=[\mathbf{D}_s^0;\mathbf{D}_a^0]$, we can use this relational dictionary to update the initially estimated face shape and appearance in the training set. For example, given the i-th training sample, we first represent its initial appearance $\mathbf{a}_i^0$ sparsely using the learned appearance dictionary $\mathbf{D}_a^0$, *i.e.*,

$$\hat{\mathbf{c}}_i = \underset{\mathbf{c}\in\mathbb{R}^m}{\operatorname{argmin}} \|\mathbf{a}_i^0-\mathbf{D}_a^0\mathbf{c}\|_2^2+\lambda\|\mathbf{c}\|_1. \quad (7)$$

Then the estimated face shape of the i-th training sample can be updated by a linear combination of the learned shape displacements indicated by representation coefficients $\hat{\mathbf{c}}$:

$$\mathbf{s}_i^1 = \mathbf{s}_i^0 + \mathbf{D}_s^0\hat{\mathbf{c}}_i. \quad (8)$$

Based on the updated face shapes $\{\mathbf{s}_i^1\}_{i=1}^N$, we can continue to rebuild a new training set $\mathcal{X}^1=\{(\mathbf{s}_i^1,\mathbf{a}_i^1)\}_{i=1}^N$ from the original training set $\mathcal{X}=\{(\mathbf{x}_i,\mathbf{p}_i)\}_{i=1}^N$, by extracting the

Algorithm 1 SRD model learning

Input: training set $\mathcal{X} = \{(\mathbf{x}_i, \mathbf{p}_i)\}_{i=1}^N$, m (dictionary size), λ (regularization parameter), T (maximal stage number).
Output: learned SRD model $\mathcal{M}$.

Initialization: Initialize elements of $\mathcal{M}$ randomly.
for $t = 0 \rightarrow T$ **do**
  Build training set $\mathcal{X}^t = \{(\mathbf{s}_i^t, \mathbf{a}_i^t)\}_{i=1}^N$ from $\mathcal{X}$.
  $\mathbf{A}^t \leftarrow [\mathbf{a}_1^t, \ldots, \mathbf{a}_N^t]$, $\mathbf{S}^t \leftarrow [\mathbf{s}_1^t, \ldots, \mathbf{s}_N^t]$.
  while not converged **do**
    Step 1: fix $\mathbf{D}_a^t$ to learn $\mathbf{D}_s^t$ and $\mathbf{C}$ on $\mathcal{X}^t$:

$$\underset{\mathbf{D}_s^t, \mathbf{C}}{\operatorname{argmin}} \left\{ \min_{\mathbf{C}} \|\mathbf{S}^t - \mathbf{D}_s^t \mathbf{C}\|_2^2 + \lambda \|\mathbf{C}\|_1 \right\}.$$

    Step 2: update $\mathbf{D}_a^t$ using $\mathbf{A}^t$ and $\mathbf{C}$: $\mathbf{D}_a^t = \mathbf{A}^t / \mathbf{C}$.
    Step 3: fix $\mathbf{D}_s^t$ to learn $\mathbf{D}_a^t$ and $\mathbf{C}$ on $\mathcal{X}^t$:

$$\underset{\mathbf{D}_a^t, \mathbf{C}}{\operatorname{argmin}} \left\{ \min_{\mathbf{C}} \|\mathbf{A}^t - \mathbf{D}_a^t \mathbf{C}\|_2^2 + \lambda \|\mathbf{C}\|_1 \right\}.$$

    Step 4: update $\mathbf{D}_s^t$ using $\mathbf{D}_a^t$ and $\mathbf{C}$: $\mathbf{D}_s^t = \mathbf{S}^t / \mathbf{C}$.
  end while
  Update $(\mathbf{s}_i, \mathbf{a}_i)$: $\mathbf{s}_i^{t+1} = \mathbf{s}_i^t + \mathbf{D}_s^t \mathbf{c}_i$, get $\mathbf{a}_i^{t+1}$ from $\mathbf{s}_i^{t+1}$ in $\mathcal{X}$.
end for
Generate the learned SRD model $\mathcal{M} = \{[\mathbf{D}_s^t; \mathbf{D}_a^t]\}_{t=0}^T$.

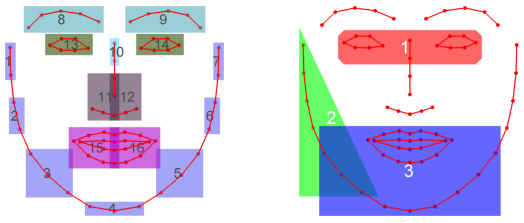
pose-indexed features of the training samples from the updated face landmarks, *i.e.* $\mathbf{a}_i^1 = \mathbf{h}(\mathbf{x}_i, \mathbf{p}_i + \mathbf{s}_i^1)$. Then again we can use this new training set $\mathcal{X}^1$ to learn a new relational dictionary $\mathbf{D}^1 = [\mathbf{D}_s^1; \mathbf{D}_a^1]$ at a new stage. This process can be repeated until the rebuilt dataset converges. The final model will contain multiple relational dictionaries trained at different stages, *i.e.* $\mathcal{M} = \{[\mathbf{D}_s^t; \mathbf{D}_a^t]\}_{t=0}^T$, which we refer to as *Stagewise Relational Dictionary* (SRD) model. In Algorithm 1, we summarize the complete learning process of the SRD model. In all our experiments, the rebuilt training set quickly converges in only 2 or 3 stages.

Given a test image, we first estimate an initial shape of the landmarks from the face detection result and extract the pose-indexed features to form the initial face appearance around the estimated landmarks. Then we use the learned SRD model to iteratively update the estimated face shape and re-extract the pose-indexed features to build the face appearance. This process is exactly the same as the face shape and appearance updating process during training the model, which will guide the updated face shape and appearance towards the true value in the test image.

3. SRD with Occlusion Learning

The SRD model proposed in Section 2 can naturally deal with face view variations via simultaneously appearance-shape modeling. To deal with face occlusions, we further propose to simulate the occlusion via jointly learning an occlusion dictionary within the rational dictionary. It enables the SRD to model explicitly the occlusion.

The most popular approach of modeling occlusion in sparse representation is to add an identity matrix in the learned dictionary [26, 16]. The identity matrix explains the occluded image pixels which cannot be well represented solely by the learned dictionary. This occlusion modeling approach, however, is very computationally expensive for high dimensional identity matrix. For example, when adopting a 100-dimensional appearance feature for one landmark, the dimension of the identity matrix to account for occlusions of 100 landmarks will be $10,000 \times 10,000$. This extremely high dimension prohibits the ℓ_1 minimization process in the SRD model for face alignment.



(a) (b)

Figure 1. Definitions of the Elemental Occlusion Patterns (EOPs) and the Partial Occlusion Patterns (POPs). (a). A set of 16 EOPs defined for the 68 landmarks drawn on a mean face shape. (b). Three different POPs by combining a few EOPs.

To address the occlusion problem effectively and efficiently, we propose to learn a more compact and representative occlusion dictionary. In our SRD model, the occlusion dictionary can be added to the appearance dictionary and, consequently, the loss function in Eqn. (5) becomes

$$\begin{aligned} l(\mathbf{a}, \mathbf{s}, \mathbf{D}) = \min_{\mathbf{c}, \mathbf{e}} \{ & \|\mathbf{s} - \mathbf{D}_s \mathbf{c}\|_2^2 + \\ & \|\mathbf{a} - [\mathbf{D}_a, \mathbf{D}_o][\mathbf{c}; \mathbf{e}]\|_2^2 + \lambda \|[\mathbf{c}; \mathbf{e}]\|_1 \}, \end{aligned} \tag{9}$$

where $\mathbf{D}_o \in \mathbb{R}^{n_a \times k}$ denotes the occlusion dictionary with k columns, $\mathbf{e} \in \mathbb{R}^k$ is the representation coefficients of $\mathbf{D}_o$, and $\mathbf{D} = [\mathbf{D}_s; \mathbf{D}_a, \mathbf{D}_o]$. We denote this occlusion handling SRD model as OSRD. One main problem is how to build suitable training set to guide the model learning procedure to achieve desired occlusion modeling effects: $\mathbf{D}_o$ models the appearance of the occluded face part, $\mathbf{D}_a$ models face appearance without occlusions, and $\mathbf{D}_s$ models the true face shape regardless of occlusions.

The most direct way to build such a training set is to collect many occluded face images and manually label the occlusion groundtruth in pixel level. This is obviously very difficult and too time-consuming. We therefore design an automatic procedure to build a training set which provides similar effects. We do not try to restrict the explicit form of the face occluder (*e.g.* glasses or hair), but let it be any object in the wild. What we restrict is that the occluder appearance must satisfy the face shape constraint when building face appearance from it, which can thus well simulate the appearance of occluded faces when performing alignment on it. We also restrict the occluder appearance to follow the common patterns in occluded faces from nature images.

To generate occluded training samples satisfying these restrictions, we first build a training set that simulates the original training with full occlusions. Based on the assumption that occlusions can be any form of natural objects, we build a full occlusion training set by copying all the im-

Algorithm 2 OSRD model learning

Input: training set $\mathcal{X} = \{(\mathbf{x}_i, \mathbf{p}_i)\}_{i=1}^N$, m (dictionary size), k (full occlusion dictionary size), λ (regularization parameter), T (maximal stage number), EOPs and POPs.
Output: learned OSRD model $\mathcal{M}$.

Initialization: Initialize elements of $\mathcal{M}$ randomly.
for $t = 0 \rightarrow T$ **do**
  Build training set $\mathcal{X}^t$, $\mathcal{Y}^t$, $\mathcal{Z}^t$ from $\mathcal{X}$.
  $\mathbf{A}^t \leftarrow [\mathbf{a}_1^t, \ldots, \mathbf{a}_N^t]$, $\mathbf{S}^t \leftarrow [\mathbf{s}_1^t, \ldots, \mathbf{s}_N^t]$, $\mathbf{B}^t \leftarrow [\mathbf{b}_1^t, \ldots, \mathbf{b}_N^t]$, $\tilde{\mathbf{A}}^t \leftarrow [\tilde{\mathbf{a}}_1^t, \ldots, \tilde{\mathbf{a}}_N^t]$.
  while not converged **do**
    Step 1: fix $\mathbf{D}_a^t$ to learn $\mathbf{D}_s^t$ and $\mathbf{C}$ on $\mathcal{X}^t$:
    $$\underset{\mathbf{D}_s^t, \mathbf{C}}{\operatorname{argmin}} \left\{ \min_{\mathbf{C}} \|\mathbf{S}^t - \mathbf{D}_s^t \mathbf{C}\|_2^2 + \lambda \|\mathbf{C}\|_1 \right\}.$$
    Step 2: update $\mathbf{D}_a^t$ using $\mathbf{A}^t$ and $\mathbf{C}$: $\mathbf{D}_a^t = \mathbf{A}^t / \mathbf{C}$.
    Step 3: fix $\mathbf{D}_s^t$ to learn $\mathbf{D}_a^t$ and $\mathbf{C}$ on $\mathcal{X}^t$:
    $$\underset{\mathbf{D}_a^t, \mathbf{C}}{\operatorname{argmin}} \left\{ \min_{\mathbf{C}} \|\mathbf{A}^t - \mathbf{D}_a^t \mathbf{C}\|_2^2 + \lambda \|\mathbf{C}\|_1 \right\}.$$
    Step 4: fix $\mathbf{D}_a^t$ to learn $\mathbf{D}_b^t$ on $\mathcal{Y}^t$:
    $$\underset{\mathbf{D}_b^t}{\operatorname{argmin}} \left\{ \min_{\mathbf{C}, \mathbf{E}} \|\mathbf{B}^t - [\mathbf{D}_a^t, \mathbf{D}_b^t][\mathbf{C}; \mathbf{E}]\|_2^2 + \lambda \|[\mathbf{C}; \mathbf{E}]\|_1 \right\}.$$
    Step 5: Expend $\mathbf{D}_b^t$ to the block dialog form to get $\mathbf{D}_o^t$.
    Step 6: fix $\mathbf{D}_a^t$ and $\mathbf{D}_o^t$ on $\mathcal{Z}$ to find best $\mathbf{C}$ and $\mathbf{E}$:
    $$\min_{\mathbf{C}, \mathbf{E}} \|\tilde{\mathbf{A}}^t - [\mathbf{D}_a^t, \mathbf{D}_o^t][\mathbf{C}; \mathbf{E}]\|_2^2 + \lambda \|[\mathbf{C}; \mathbf{E}]\|_1.$$
    Step 7: fix $\mathbf{D}_a^t$ and $\mathbf{C}$ to update $\mathbf{D}_s^t$: $\mathbf{D}_s^t = \mathbf{S}^t / \mathbf{C}$.
  end while
  Update $(\mathbf{s}_i, \mathbf{a}_i)$: $\mathbf{s}_i^{t+1} = \mathbf{s}_i^t + \mathbf{D}_s^t \mathbf{c}_i$, extract $\mathbf{a}_i^{t+1}$ from $\mathbf{s}_i^{t+1}$ in $\mathcal{X}$, extract $\mathbf{b}_i^{t+1}$ from $\mathbf{s}_i^{t+1}$ in $\mathcal{Y}$.
end for
Generate learned OSRD model $\mathcal{M} = \left\{ [\mathbf{D}_s^t; \mathbf{D}_a^t, \mathbf{D}_o^t] \right\}_{t=0}^T$.

ages from the original training set but shifting the face annotations and corresponding face detection bounding box to another random place in the same image where no face exists. Note that during the shifting process, the face shape constraint of the landmark annotations and detection bounding box stay unchanged. Then we extract training samples for the occlusion dictionary from this new dataset just the same as in the original training set. These samples can well approximate the shapes and appearances of fully occluded face images. This training set simulates full face occlusions, therefore we call it the full occlusion training set and denote it as $\mathcal{Y}^t = \{(\mathbf{s}_i^t, \mathbf{b}_i^t)\}_{i=1}^N$. From this training set, we can extract appearances features of possible face occluders and later use them to train an initial occlusion dictionary.

Compared with face alignment under full face occlusions, face alignment for partially-occluded faces is more important in practice. Therefore, we further generate another training set to simulate partial face occlusions. To this end, we first define a set of Elemental Occlusion Patterns (EOPs) based on observations of real world face occlusion patterns. Take a face with 68 labeled landmarks as an example, 16 EOPs are defined in this work as shown in Figure 1(a). Each EOP covers several landmarks and the combinations of these EOPs can approximate almost all kinds of face occlusion patterns in realistic ways. Based on the defined EOPs, we further generate a set of Partial Occlusion Patterns (POPs) by different sparse combinations of the EOPs (Figure 1(b)). By randomly applying one POP on each corresponding sample pair in the training set $\mathcal{X}^0$ and full occlusion training set $\mathcal{Y}$, we can generate the third partial occlusion training set and denote it as $\mathcal{Z}^t = \{(\mathbf{s}_i^t, \tilde{\mathbf{a}}_i^t)\}_{i=1}^N$. This dataset can be used to simulate the appearance from partially-occluded faces.

Based on the three training sets $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$, now we can learn the OSRD model. We propose a similar procedure as Algorithm 1 to solve the OSRD learning problem. The main idea is as follows. First, we learn the appearance dictionary $\mathbf{D}_a$ on $\mathcal{X}^t$ (Step 1-3) as in Algorithm 1. Then, fixing $\mathbf{D}_a$, we learn a full occlusion dictionary $\mathbf{D}_b$ using $\mathbf{D}_a$ on $\mathcal{Y}$. Thirdly, we extend $\mathbf{D}_b$ to a block dialog form based on the EOPs to generate the partial occlusion dictionary $\mathbf{D}_o$. Formally, denote l EOPs together as $\mathbf{O}_e = [\mathbf{o}_1^e, \ldots, \mathbf{o}_l^e]$, and $\mathbf{D}_b = [\mathbf{d}_1^b, \ldots, \mathbf{d}_k^b]$, where $\mathbf{o}_l^e \in \{0, 1\}^{n_a}$ is the l-th EOP, whose j-th entry indicates whether the corresponding landmark is occluded, k is the size of $\mathbf{D}_b$. Then the partial occlusion dictionary can be represented as $\mathbf{D}_o = [\mathbf{d}_1^o, \ldots, \mathbf{d}_{k \times l}^o]$, where $\mathbf{d}_{i \times l + j}^o = \mathbf{d}_i^b \otimes \mathbf{o}_j^e$, and $\otimes$ is the element-wise multiplication. In the next step, upon fixing $\mathbf{D}_a$ and $\mathbf{D}_o$, it minimizes the representation errors on $\mathcal{Z}$. With the best representation coefficients, the shape dictionary $\mathbf{D}_s$ is updated by least square fitting. Algorithm 2, summarizes the complete OSRD learning process.

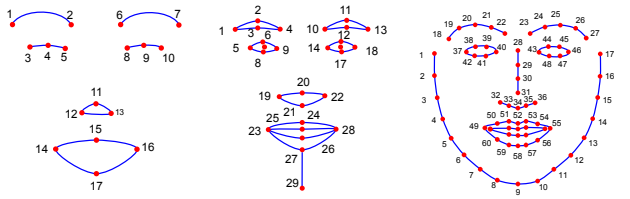
4. Experiments

The main objective of this work is for multi-view and partially-occluded face alignment. We therefore design experiments particularly for this objective to perform evaluations and analyses. For completeness, we also conduct experiments for general face alignment and compare with previous reported results. In all the conducted experiments, our algorithm achieves the best performance, although the improvement in general face alignment is not so obvious due to its performance saturation. We first introduce the implementation details and then conduct experiments for general face alignment, multi-view face alignment, and partially-occluded face alignment tasks, respectively.

4.1. Implementation Details

Features. We explore two feature descriptors, HoG [10] and SIFT [12], around the landmarks to represent the face appearance. It is observed that HoG is more computationally efficient and performs comparably or even better, especially for landmarks located around edges. However, SIFT is more robust for locating a relative small number of inner face landmarks. To make a fair comparison with the SDM method [27], the performances reported in the experiments are all based on the 128-dimensional SIFT descriptor [27].

Parameters. The testing process of SRD is parameter-free, as it only requires the same parameter settings as in the



(a) (b) (c)

Figure 2. Different definitions of facial landmark annotations and their corresponding mean shapes estimated from a face detector: (a) 17 points, (b) 29 points, and (c) 68 points.

model training. In particular, we tune the regularizer λ via cross validation over the training set, and throughout experiments we find that setting its value within $[0.01, 0.1]$ generates stably good performances. Therefore, we set $\lambda = 0.05$ throughout the experiments. As for the dictionary size, we find that a larger size will consistently improve the performance, but will also increase the computationally cost. We take a trade-off and set the size of the shape and appearance dictionary size as $m = 100$ and the size of the full occlusion dictionary $k = 5$ in all the experiments.

Speed. In current unoptimized MATLAB implementation, it takes about 30 minutes to train the SRD model from 2000 samples. In testing, provided with the face detection results, it only costs about 0.1 seconds to align one face. Since about half of the testing time is cost by feature extraction, the testing speed of the SRD model can be further improved if more efficient features are adopted.

4.2. General Face Alignment Evaluation

We evaluate the proposed SRD model on two widely used benchmark datasets, BioID [1] and LFPW [3], and compare it with the state-of-the-art algorithms ever reported on them. The BioID dataset contains 1521 face images annotated with 17 landmarks (Figure 2(a)). The LFPW dataset contains 1132 training images and 300 testing images, annotated with 29 landmarks (Figure 2(b)). Since its original version is no longer available from its URLs, we use its augmented version provided by [4] in the experiments. For fair comparisons, we follow the same settings as in [3, 4] and measure the alignment error by the average Euclidean distance between predicted landmarks and labeled landmarks, which is normalized by the inter-ocular distance.

In Figure 3, we plot the cumulative error distribution (CED) curves of our methods on the BioID dataset as in [4], which are calculated from the normalized mean errors over each point. The six baseline methods are the explicit shape regression method (ESR) [4], the supervised descent model (SDM) [27], the consensus exemplar method (CEM) [3], the constrained local models (CLE) [9], the extended ASM method (EASM) [17], and the boosted regression method (BRM) [22]. The SDM method is a very simple yet effective face alignment algorithm developed recently. Because its CED curve on BioID is not reported in [27], we implement it based on the partially released testing code in [27]

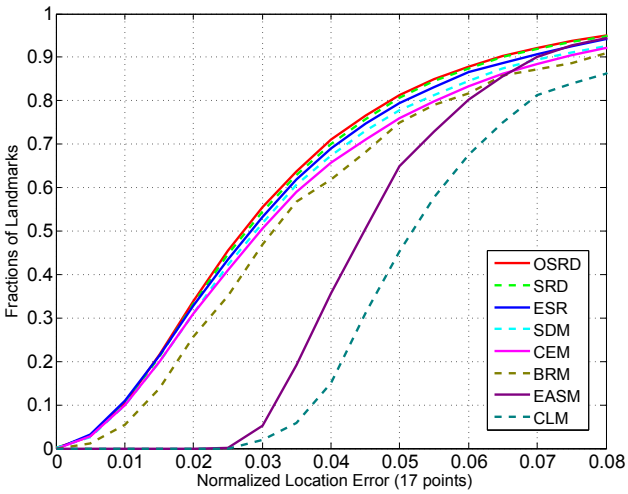


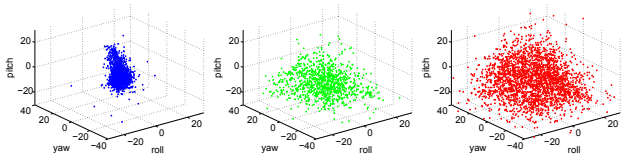
Figure 3. CED curves over the BioID dataset.

and train the model with the same settings as for our model. The CED curves for other five methods are all directly obtained from their authors. As shown in Figure 3, our OSRD and SRD models are the two best ones among these algorithms. However, since this dataset is quite easy, the performance gaps among different algorithms are not so obvious, especially for the most recent methods.

Table 1. Mean alignment errors on the LFPW dataset.

Algorithm	CEM[3]	OSDM[27]	SDM	ESR[4]	SRD	OSRD
ME ($\times 10^{-2}$)	3.99	3.47	3.49	3.43	3.24	**3.19**

In Table 1, we report the performance of our algorithm on LFPW in terms of mean alignment error, and compare it with other state-of-the-art methods on this dataset. The mean error of OSDM is from the reported results in [27]. Again, our proposed two methods achieve the best performances on this more challenging dataset, with about 7% and 5% improvements respectively over the best baseline. From Table 1, it can be observed that the SDM model implemented by ourselves has comparable performance to the one reported in [27]. The slight performance difference may be caused by the different constitutions of the training sets.

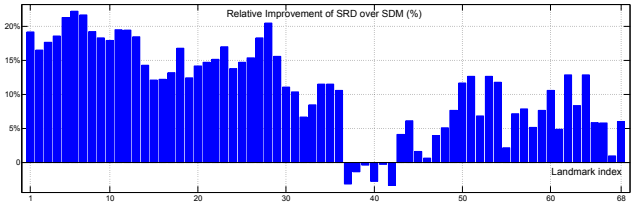


(a) (b) (c)

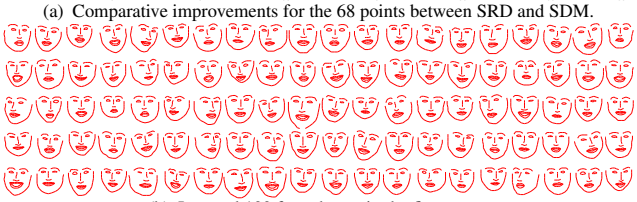
Figure 4. Face view distributions of (a) the BioID dataset, (b) the LFPW dataset, and (c) our MVFW dataset. MVFW has much larger view variations and is more well-proportioned.

4.3. Multi-view Face Alignment Evaluation

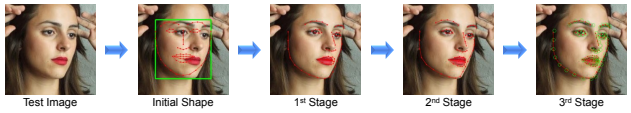
As we can see from previous experiments, most existing algorithms perform quite well on the BioID dataset, since all the faces in this dataset are captured from near-frontal views. The LFPW dataset, although more challenging than BioID, is still dominated by faces with small view varia-



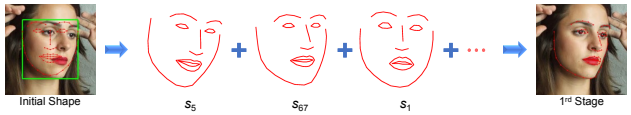
(a) Comparative improvements for the 68 points between SRD and SDM.



(b) Learned 100 face shapes in the first stage.



(c) Stagewise optimization process on one testing image.



(d) Automatically selected face shapes in the 1st stage for the testing image in (c).

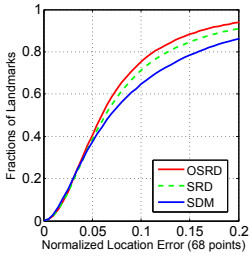
Figure 5. Evaluation and analysis of multi-view face alignment on our MVFW dataset.

tions. In Figure 4, we verify these claims by plotting the view distributions of these datasets. Here we represent the face view using its three angles of in-plane roll, out-plane pitch, and out-plane yaw, which is obtained from a face gesture estimator using the labeled landmarks as inputs. Figure 4 shows that the face views from both BioID and LFPW distribute in a very small range and very few samples have large view angles. These two datasets, therefore, are not suitable for evaluating multi-view face alignment.

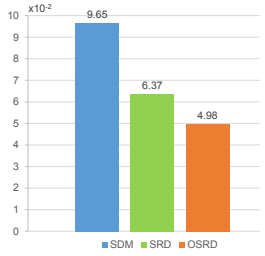
To build a suitable dataset for multi-view face alignment, we assemble a large "face in the wild" dataset from many existing datasets collected by 300-W[1] and select samples with different view angles uniformly to construct a new multi-view face dataset in the wild. The obtained dataset, denoted as MVFW, contains 2050 training samples and 450 testing samples, annotated with 68 landmarks (Figure 2(c)). Figure 4(c) plots its view distribution, from which we can see that it is well-proportioned for different view angles.

We train the SDM model (implemented by ourselves) and our SRD model on MVFW. We then calculate the individual alignment errors for each landmark. Figure 5(a) plots the relative improvement brought by SRD over SDM on the testing set. The SRD model, on average, gives an over 10% improvement for multi-view face alignment. This owes to its ability of simultaneously appearance and shape modeling. Figure 5(b) shows the 100 learned face shapes from the first stage of SRD, which exhibits very typical face view modes from in-plane roll, out-plane pitch and yaw.

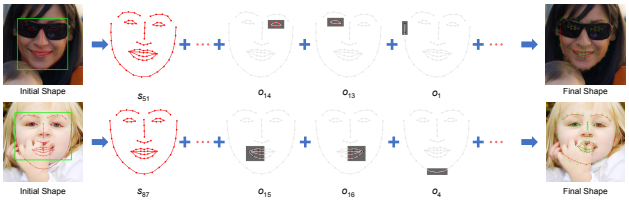
[1] http://ibug.doc.ic.ac.uk/resources



(a) CED curves over OCFW.



(b) Mean errors on OCFW.



(c) Automatically selected shapes and EOPs for two testing images with occlusions.

Figure 6. Evaluation and analysis of partially-occluded face alignment on our OCFW dataset.

To better understand the effectiveness of the stagewise optimization process of SRD, we plot in Figure 5(c) the intermediate results of three stages for one testing sample. The SRD quickly converges to the groundtruth landmarks from a mean face shape provided by face detector. In each stage, the SRD model automatically selects a sparse combination of the most related shape bases to estimate the displacement of the testing face. In Figure 5(d), we visualize the sparse representation in the first stage by drawing the three shape bases with the largest coefficients. They are very relevant to the shape of the testing face.

4.4. Partially-occluded Face Alignment Evaluation

Since the testing set of LFPW contains only a few occluded faces, it is also not suitable for evaluating partially-occluded face alignment. We therefore design a similar procedure to build a new dataset to evaluate the ability of handling occlusion for a face alignment algorithm. The objective of building the dataset is to evaluate the generalization ability of a face alignment model to deal with occlusions when trained on samples without occlusions. To this end, we manually label the occlusion state of all the collected samples and split them into two parts: samples without occlusions and with occlusions. The former one is used for training and the latter is used for testing. The obtained dataset, denoted as OCFW, contains 2591 training samples and 1246 testing samples. To facilitate further studies on multi-view and occluded face alignment, the MVFW and OCFW datasets will be made publicly available online[2].

In Figure 6(a) and 6(b), we respectively plot the CED curves and mean errors of different models over the OCFW dataset. Our OSRD model obtains the best performance among the three models, which demonstrates its good generalization ability to deal with occlusions. To illustrate how

[2] https://sites.google.com/site/junliangxing/codes

Figure 7. Exemplar face alignment results. Top row: general face alignment on the BioID dataset and the LFPW dataset. Middle row: multi-view face alignment on the MVFW dataset. Bottom row: partially-occluded face alignment on the OCFW dataset.

occlusion is handled by the OSRD model, Figure 6(c) plots the automatically selected face shape bases and the corresponding EOPs in the occlusion dictionary for two typical occluded face images. It can be seen that most of the occluded landmarks are successfully detected by the occlusion dictionary using its bases. In Figure 7, some alignment results obtained by the OSRD model are plotted on challenging examples with large variations in face view, expression, illumination, scale, and occlusion.

5. Conclusions and Future Work

We present a novel algorithm for multi-view and partially-occluded face alignment. By simultaneously modeling face shape and appearance, the proposed SRD model captures consistent shape-appearance relationships for faces with huge view variations. By a learning based occlusion modeling procedure, the OSRD model deals with partial face occlusions quite robustly. Extensive experiments have demonstrated the state-of-the-art alignment performance, especially for multi-view and partially-occluded faces. In future work, we plan to perform deeper analyses on the SRD and OSRD models, *e.g.*, its theoretical convergence properties and comparison of different optimization procedures. We also plan to apply the proposed model to other vision problems like human pose estimation and object tracking.

Acknowledgement: This work is partially supported by NSFC (Grant No. 61303178 and 60935002), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), Beijing NSF (Grant No. 4121003), Guangdong NSF (Grant No. S2012020011081), and the Singapore National Research Foundation under its International Research Centre @Singapore Funding Initiative and administered by the IDM Programme Office.

References

[1] BioID dataset. `http://www.bioid.com`.

[2] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *TSP*, 54(11):4311–4322, 2006.

[3] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.

[4] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*, 2012.

[5] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *TPAMI*, 23(6):681–685, 2001.

[6] T. Cootes, M. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In *ECCV*, 2012.

[7] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *TPAMI*, 61(1):38–59, 1995.

[8] D. Cristinacce and T. Cootes. Boosted regression active shape models. In *BMVC*, 2007.

[9] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *PR*, 41(10):3054–3067, 2008.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[11] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *CVPR*, 2010.

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[13] X. Lu and A. Jain. Deformation modeling for robust 3d face matching. *TPAMI*, 30(8):1346–1357, 2008.

[14] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *JMLR*, 11:19–60, 2010.

[15] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 60(2):135–164, 2004.

[16] X. Mei and H. Ling. Robust visual tracking using ℓ_1 minimization. In *ICCV*, 2009.

[17] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *ECCV*, 2008.

[18] J. Saragih and R. Goecke. A nonlinear discriminative approach to AAM fitting. In *ICCV*, 2007.

[19] J. Saragih, S. Lucey, and J. Cohn. Deformable model fitting by regularized landmark mean-shift. *IJCV*, 91(2):200–215, 2011.

[20] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 58(1):267–288, 1996.

[21] I. Tosic and P. Frossard. Dictionary learning. *IEEE Signal Process. Mag.*, 28(2):27–38, 2011.

[22] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *CVPR*, 2010.

[23] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[24] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *TPAMI*, 31(11):1955–1967, 2009.

[25] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proc. IEEE*, 98(6):1031–1044, 2010.

[26] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *TPAMI*, 31(2):210–227, 2009.

[27] X. Xiong and F. Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[28] S. Yan, X. Hou, S. Li, H. Zhang, and Q. Cheng. Face alignment using view-based direct appearance models. *Int. J. Image. Sys. Tech.*, 13(1):106–112, 2003.

[29] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–459, 2003.

[30] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.