Towards Robust and Accurate Multi-View and Partially-Occluded Face Alignment

Junliang Xing, *Member, IEEE*, Zhiheng Niu, Junshi Huang, Weiming Hu, *Senior Member, IEEE*, Xi Zhou, *Senior Member, IEEE*, and Shuicheng Yan, *Fellow, IEEE*

Abstract—Face alignment acts as an important task in computer vision. Regression-based methods currently dominate the approach to solving this problem, which generally employ a series of mapping functions from the face appearance to iteratively update the face shape hypothesis. One keypoint here is thus how to perform the regression procedure. In this work, we formulate this regression procedure as a sparse coding problem. We learn two relational dictionaries, one for the face appearance and the other one for the face shape, with coupled reconstruction coefficient to capture their underlying relationships. To deploy this model for face alignment, we derive the relational dictionaries in a stage-wised manner to perform close-loop refinement of themselves, i.e., the face appearance dictionary is first learned from the face shape dictionary and then used to update the face shape hypothesis, and the updated face shape dictionary from the shape hypothesis is in return used to refine the face appearance dictionary. To improve the model accuracy, we extend this model hierarchically from the whole face shape to face part shapes, thus both the global and local view variations of a face are captured. To locate facial landmarks under occlusions, we further introduce an occlusion dictionary into the face appearance dictionary to recover face shape from partially occluded face appearance. The occlusion dictionary is learned in a data driven manner from background images to represent a set of elemental occlusion patterns, a sparse combination of which models various practical partial face occlusions. By integrating all these technical innovations, we obtain a robust and accurate approach to locate facial landmarks under different face views and possibly severe occlusions for face images in the wild. Extensive experimental analyses and evaluations on different benchmark datasets, as well as two new datasets built by ourselves, have demonstrated the robustness and accuracy of our proposed model, especially for face images with large view variations and/or severe occlusions.

Index Terms—Face alignment, dictionary learning, sparse representation, appearance-shape modeling

1 Introduction

Face alignment, i.e., locating the positions of some predefined facial landmarks from a face image, is an important problem in computer vision. For vision tasks related to human faces, it usually acts as an essential middle-level bridge between the low-level tasks like face detection [1], face tracking [2], and high-level tasks like facial traits classification [3], face synthesis [4], and face recognition [5]. Although many efforts have been devoted in solving this task from computer version researchers and great progresses have been made in the past decades [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], face alignment still remains a very challenging task, especially when the face images are taken from different views and/or undergo severe occlusions.

Regression-based face alignment methods [10], [12], [16], [17], [18], [19], [20], [21] have demonstrated very good performance for near frontal face images. A general pipeline of these methods during testing is to update the face shape hypothesis from the face appearance iteratively using learned regression models. For face images with large view variations, the relationship between the face appearance and shape becomes very complex and thus poses great challenges to these methods. 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 demonstrated to be robust to occlusions, the underlying mechanism on how to model occlusions and why it works is not clear. In the Robust Cascaded Pose Regression (RCPR) method [22], the occlusion problem is explicitly addressed and very good alignment results are obtained. The training of the RCPR model, however, needs to annotate the occlusion state of all the landmarks in the training set, which requires a

- *J. Xing is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P.R. China. E-mail: jlxing@nlpr.ia.ac.cn.*
- *Z. Niu is with the Advanced Engineering Electronics & Safety, Delphi Deutschland GMBH, Delphiplatz 1, Wuppertal, North Rhine-Westfalia 42119, Germany. E-mail: niuzhiheng@gmail.com.*
- *J. Huang is with the AI Institute of Qihoo/360 Company, Jiuxianqiao Road, Chaoyang District, Beijing 100015, P.R. China. E-mail: huangjunshi@360.cn.*
- *W. Hu is with the CAS Center for Excellence in Brain Science and Intelligence Technology, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing 100190, P.R. China. E-mail: wmhu@nlpr.ia.ac.cn.*
- *X. Zhou is with the Intelligent Media Technique Research Center, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, P.R. China. E-mail: zhouxi@cigit.ac.cn.*
- *S. Yan is with the AI Institute, Qihoo/360 Company, Beijing 102206, China, and the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576. E-mail: eleyans@nus.edu.sg.*

Manuscript received 11 Apr. 2015; revised 10 Mar. 2017; accepted 23 Mar. 2017. Date of publication 24 Apr. 2017; date of current version 13 Mar. 2018. Recommended for acceptance by M. Tistarelli.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2017.2697958

considerable amount of time. If a face alignment model can inherently and explicitly address the face view variation and occlusion problems simultaneously and automatically, it will be very important to current face alignment research and also useful for practical vision applications.

To obtain such a model, we view the face alignment task as an appearance-shape modeling procedure and the learning objective is to minimize the incompatibility between the face appearance and shape. We formulate the regression procedure in face alignment as a sparse coding problem, which is designed to model both the face view variations and partial occlusions. We learn two relational dictionaries to represent the face appearance and face shape, respectively. These two dictionaries share the same reconstruction coefficient when representing a face sample, thus capturing the underlying relationship between the face appearance and shape. During the testing phase, the regression model learns to select only a subset of related appearance modes from the appearance dictionary via sparse representation, and then predicts the shape displacement towards the true face shape to refine the shape hypothesis.

For the appearance-shape modeling under large face view variations, both the face appearance dictionary and shape dictionary need to model the face patterns exhibited from different views. In the general setting of the face alignment problem, since the true face shape is not available and we are instead given only a face shape hypothesis, we cannot directly learn the relational dictionaries from the ground-truth face appearance and shape in the training set. Therefore, we derive a stage-wised version of the relational dictionaries which allows them to refine from each other using the training samples in a closed loop. At first, the shape dictionary is initialized from the ground-truth face shape to generate a set of shape hypotheses. Then the face appearance dictionary is learned from this shape dictionary and later further updates the shape dictionary. The updated shape dictionary is then in-return used to refine the appearance dictionary. This stage-wise relational dictionary learning procedure produces regression model deployable in the face alignment problem.

For the modeling of face appearances with partial occlusions, we introduce an occlusion dictionary in the appearance dictionary to represent occluded face appearance. The occlusion modeling stage-wise relational dictionary is then used to recover the face shape from partially occluded face appearance. We design a set of elemental occlusion patterns and let each element in the occlusion dictionary represent only one elemental occlusion pattern. A sparse combination of these elemental occlusion patterns thus represents different kinds of partial occlusion patterns. To learn the occlusion dictionary, we collect a set of background images without faces to train the occlusion dictionary jointly with the appearance and shape dictionaries in a data driven manner, which removes the need of labeling the occlusion state of face landmarks. During testing, the elements in the occlusion dictionary are automatically selected to represent the occluded face appearances, thus making the full alignment model robust to partial occlusions.

To summarize, we in this work have made the following contributions to the face alignment problem.

- We propose to formulate face alignment as an ℓ_1-induced dictionary learning problem which performs simultaneously face appearance and shape modeling (Section 3).
- We derive a stage-wise relational dictionary model (SRD) to learn compatible models for the appearance-shape modeling problem in multi-view face alignment (Section 4).
- We extend the SRD model to a hierarchical version (HSRD) from the whole face shape to face part shapes, which further improves the alignment accuracy by addressing both global and local view variations (Section 5).
- We introduce an occlusion dictionary in the SRD model (OSRD) to recover face shape from partially occluded face appearance, and develop a data driven learning procedure to train the model, which is more effective and efficient than previous occlusion modeling methods in sparse representation (Section 6).

Based on these contributions, we have presented a robust and accurate face alignment system. Extensive experiments under different experimental settings and over several benchmark datasets demonstrate significant improvements of the face alignment performance over many state-of-the-art algorithms, especially for faces with large view variations and serious occlusions.

2 Related Work

Face alignment has been studied for many years. Traditional approaches employ parameterized models to describe the face appearance and shape. The Active Shape Model (ASM) [23] represents the face shapes by conducting principal component analysis on the manually labeled training samples and progressively fitting the face instance in a test image using the learned face shape. The Active Appearance Model (AAM) [6], [24] 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 [7], [10], [12], [25], however, find 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 [8], [9], [11], [25], [26]. 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 [8], [27] or pictorial structures [11]. Unlike AAM which approximates the raw image pixels directly, the constrained local models [8] 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 approaches [10], [12], [16], [17], [25], [27], [29], [30], which directly learn a mapping function from the image features to the face shape. The distinctions among these methods mainly lie in the employed learning algorithm (e.g., boosting

[27], random forest [25], or non-linear least squares [12]) and the adopted features (e.g., Haar wavelets [27], random ferns [29], or SIFT [12]). The pose-indexed features [29], which are obtained by re-computing the features every time when a new face shape hypothesis is updated, are demonstrated to be very important for learning a robust alignment model [10], [12], [29]. Moreover, with an initial shape provided by the face detector [1], mapping from the pose-indexed features to the face landmark displacements provides a natural and effective way to iteratively update the locations face landmarks. Since the mapping functions are often highly non-linear [15], [25], [27], [28], [29], [30], training them usually needs many annotated samples and takes considerable time to learn the complex relationships.

There are also some work on joint face alignment from multiple images [31], [32] and facial landmark tracking in video sequence [33]. One crucial part for a joint face alignment algorithm is how to model the global shape constraint to guide the face alignment in each face image. To track facial landmarks in the video, the prior on the temporal coherency is useful to refine and speed up the alignment process. In this work, the exploration of information among multiple images and the temporal coherency prior are beyond the scope of our model.

3 Face Alignment as Dictionary Learning

Essentially, face alignment can be viewed as an appearance-shape modeling problem. During the training phase, the representations of the face appearance and shape are respectively extracted from a set of detected face images with the corresponding landmark annotations. A face alignment model needs learning these training appearance-shape pairs into a compact model to capture their underlying relationships. During the testing phase, given a face image, the face shape is supposed to be estimated from the face appearance using the learned model.

3.1 Model Formulation

Denote a dataset 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 the 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, a 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, respectively. And the loss function here measures their incompatibility. Note that here the face appearance $\mathbf{a}$ is not necessarily to be the face image $\mathbf{x}$, more robust and expressive image features like the pose-indexed features extracted around all the landmarks [12], [29] can be used to represent face appearance. Similarly, the face shape $\mathbf{s}$ is not restricted directly to be the landmark locations $\mathbf{p}$, and can be some more suitable and effective representations. With this formulation of the loss function, learning the model is equivalent to solving

$$\Theta^* = \arg\min_{\Theta} \sum_{i=1}^{N} l(\mathbf{s}_i, \mathbf{a}_i, \Theta). \tag{1}$$

The definition of the loss function, therefore, has a fundamental impact on the final face alignment model. For multi-view face alignment, since both the face appearance and the face shape exhibit huge variations, the appearance-shape relationship becomes very complex. It thus 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 appearance-shape relationships. To this end, we propose to employ the sparse representation model [34] to formulate this loss function. Denote $\mathbf{D}$ as a dictionary, each column of which is an element to represent a data point $\mathbf{z}$, i.e., $\mathbf{z} = \mathbf{Dc}$, where $\mathbf{c}$ is the representation coefficient. In our case, the data point is the composition of the face appearance $\mathbf{a}$ and face shape $\mathbf{s}$, i.e., $\mathbf{z} \triangleq [\mathbf{s}; \mathbf{a}]$, where $[\cdot;\cdot]$ denotes the row based concatenation of vectors or matrices.

With the above denotations, the sparse representation based loss function is thus formulated as

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

Compared to Eqn. (1), the dictionary $\mathbf{D} \in \mathbb{R}^{n \times m}$ here is an instantiation of the parameter Θ, which is used to simultaneously represent the face appearance $\mathbf{a}$ and shape $\mathbf{s}$. Here $n = n_a + n_s$, the dimension of the dictionary basis, and m is the dictionary size. The underlying appearance-shape relationship is enforced to be consistent by sharing the same representation coefficient $\mathbf{c}$. Note that the coefficient $\mathbf{c}$ here is encouraged to be sparse, the sparsity of which is controlled by the regularization parameter λ.

The sparsity assumption has been proved to be very effective for many computer vision problems [35], [36]. In our problem for face alignment, this sparsity regularization ensures a face appearance-shape instance to be represented by only a few bases in the dictionary, which will benefit both the model training and testing. For model training, the regularization term drives the dictionary to learn distinct appearance-shape bases, a sparse combination of which can well represent the face samples with different appearances and shapes. For model testing, since a face sample is regularized to get represented using only a few bases in the dictionary, it provides a mechanism to automatically select the most related bases to the testing face and thus generates more robust estimation.

In practice, to prevent the dictionary from being too large and causing numerical issues, 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. The constraint set is thus denoted as

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

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

$$\mathbf{D}^* = \arg\min_{\mathbf{D} \in \mathcal{D}^{n \times m}} \frac{1}{N} \sum_{i=1}^{N} \left\{ \min_{\mathbf{c} \in \mathbb{R}^m} \|[\mathbf{s}_i; \mathbf{a}_i] - \mathbf{Dc}\|_2^2 + \lambda \|\mathbf{c}\|_1 \right\}. \tag{4}$$

During the testing phase, given a face image, the model needs to estimate the landmark locations from the face appearance $\mathbf{a}$. To guarantee robustness, the face appearance $\mathbf{a}$ can be extracted based on an initial estimation of the face shape provided by a face detector. Then the representation coefficient $\mathbf{c}^*$ is obtained from the appearance part of the dictionary via sparse coding. And finally the face shape $\mathbf{s}^*$ is constructed from the shape part of the dictionary using the representation coefficient $\mathbf{c}^*$, i.e.,

$$\begin{aligned}\mathbf{s}^* &= \mathbf{D}_s\mathbf{c}^*,\\ \mathbf{c}^* &= \arg\min_{\mathbf{c}\in\mathbb{R}^m} \|\mathbf{a}-\mathbf{D}_a\mathbf{c}\|_2^2 + \lambda\|\mathbf{c}\|_1,\end{aligned} \tag{5}$$

where $\mathbf{D}_s$ and $\mathbf{D}_a$ correspond to the shape part and the appearance part of the dictionary $\mathbf{D}$, i.e., $\mathbf{D} = [\mathbf{D}_s; \mathbf{D}_a]$. Hereafter, we refer them as the shape dictionary and the appearance dictionary, respectively.

3.2 Model Analyses

The ℓ_1-induced optimization problem in Eqns. (4) and (5) provides a new formulation for the face alignment problem which simultaneously models the face appearance and face shape, as well as their underlying relationships. The proposed formulation has some fundamental differences from previous approaches for face alignment. Previous algorithms design loss functions based on either appearance [6], [8], [12], [25] or shape [9], [10], [23], [27], [29]. Our proposed loss function minimizes the incompatibilities between the face appearance and shape cooperatively. Moreover, due to the ℓ_1-induced regularization term in Eqn. (2), each iteration of its optimization learns different descent directions from different sparse combinations of the dictionary bases to update the landmark positions, thus making the refinement process adapt to the specific situation of the current estimated face shape and appearance. This also is quite different from the steepest decent direction [6] or the supervised descent direction [12] commonly adopted by previous methods [16], [18], [22], [24], [26], [28], [29], during each iteration of which only one descent direction is estimated to optimize the loss function.

The optimization problem in Eqn. (4) is often refereed to as dictionary learning [34], [37], [38] and can be optimized using a two-step iterated procedure: the first step fixes $\mathbf{D}$ and minimizes the cost function with respect to the coefficient $\mathbf{c}$; and the second step fixes the coefficient $\mathbf{c}$ and performs gradient descent like methods to minimize the lost function with respect to $\mathbf{D}$. By using the true face landmarks from the training set to model the face shape $\mathbf{s}$, and the features extracted from the landmarks to model the face appearance $\mathbf{a}$, solving the optimization problem in Eqn. (4) will generate a very effective face appearance-shape model to express the training data. The trained appearance-shape dictionary in this way, however, are not feasible to the testing phase in face alignment. The core reason for this is due to the fact that we do not have access to the true face shape during testing, which will cause serious problems for the dictionary model in Eqn. (4) if trained using the two-step optimization procedure.

One main problem for the two-step optimization procedure is that it will cause the *incompatibility* between the model training phase and the model testing phase. When testing a face image, we need to estimate the face shape $\mathbf{s}$ based on the face appearance $\mathbf{a}$. Therefore, during the testing phase, as indicated in Eqn. (5), the representation coefficient $\mathbf{c}^*$ is obtained only from the appearance dictionary $\mathbf{D}_a$ by solving the ℓ_1-regularized minimization problem. The two-step optimization procedure, however, updates the representation coefficient $\mathbf{c}^*$ from the full dictionary $\mathbf{D}$ in the first step, which actually uses a different model from testing. The model learned with the two-step optimization procedure, therefore, is incompatible to the testing setting.

Another problem of the two-step optimization procedure is even more serious when considering the alignment performance. Since the learned model in Eqn. (4) is inevitably to have reconstruction errors and more importantly, due to the unavailability of the true face shape $\mathbf{s}$, the face appearance $\mathbf{a}$ can only be roughly extracted from the face region provided by a face detector. The accuracy of the representation coefficient $\mathbf{c}^*$ as well as the face shape $\mathbf{s}^*$ obtained by Eqn. (5) are thus very likely to be insufficient for landmark location estimations. Considering that most face alignment algorithms use an iterated optimization strategy, the dictionary learning formulation should also have a similar mechanism to guarantee alignment accuracy.

To deal with these problems, we propose to learn a new relational dictionary in a stage-wise manner based on the formulation in Section 3.1. The details of this new model are elaborated in the following section.

4 SRD Model Learning

As noted previously, the core reason for the incompatibility and inaccuracy problems in the dictionary learned with the two-step procedure comes from the fact that the ground-truth landmark locations is unavailable during the testing phase. To deal with these problems, we need a dictionary learning algorithm for the problem in Eqn. (4) which does not directly use the ground-truth landmark locations to learn the dictionary and provides a mechanism to iteratively update the face shapes during the testing phase. To this end, we model the face shape using the displacements between the estimated landmark positions and the ground truths.

In this way, the locations of the ground-truth landmarks are only implicitly and indirectly used to guide the training process. For the face appearance, we extract some robust image features around the estimated landmark locations, which is more distinctive than the raw image pixels for face appearance modeling [10], [29]. The original training set $\mathcal{X} = \{(\mathbf{x}_i, \mathbf{p}_i)\}_{i=1}^N$, therefore, can be transformed in a new form which can be denoted as $\mathcal{X}^0 = \{(\mathbf{s}_i^0, \mathbf{a}_i^0)\}_{i=1}^N$. Here the face shape representation $\mathbf{s}_i^0$ is the displacements between the true landmark locations $\mathbf{p}_i$ and the current estimated landmark locations $\mathbf{p}_i^0$, i.e., $\mathbf{s}_i^0 = \mathbf{p}_i - \mathbf{p}_i^0$, and the face appearance representation $\mathbf{a}_i^0$ comes from a pose-indexed feature extraction function $\mathbf{h}(\mathbf{x}_i, \mathbf{p}_i^0)$, i.e., $\mathbf{a}_i^0 = \mathbf{h}(\mathbf{x}_i, \mathbf{p}_i^0)$.

4.1 Relational Dictionary Decomposition

The loss function in Eqn. (2) can be equivalently written as

$$l(\mathbf{a}, \mathbf{s}, \mathbf{D}) = \min_{\mathbf{c}\in\mathbb{R}^m} \|\mathbf{s}-\mathbf{D}_s\mathbf{c}\|_2^2 + \|\mathbf{a}-\mathbf{D}_a\mathbf{c}\|_2^2 + \lambda\|\mathbf{c}\|_1. \tag{6}$$

As discussed before, here we decompose the original dictionary $\mathbf{D}$ into two parts, the shape dictionary $\mathbf{D}_s \in \mathbb{R}^{n_s\times m}$ and the appearance dictionary $\mathbf{D}_a \in \mathbb{R}^{n_a\times m}$. Note that these two dictionaries are related to each other and their underlying relationships are controlled by the coefficient $\mathbf{c}$, which ensures the two dictionaries to represent a face appearance-shape instance consistently. We therefore refer to these two dictionaries together as a *relational dictionary*. With this substitution, the problem in Eqn. (4) becomes

$$\begin{aligned}\{\mathbf{D}_s^*, \mathbf{D}_a^*\} &= \mathop{\arg\min}_{\mathbf{D}_s,\mathbf{D}_a} \frac{1}{N}\sum_{i=1}^{N} l(\mathbf{s}_i, \mathbf{a}_i, \mathbf{D}_s, \mathbf{D}_a) \\ l(\mathbf{s}_i, \mathbf{a}_i, \mathbf{D}_s, \mathbf{D}_a) &= \min_{\mathbf{c}} \|\mathbf{s}_i - \mathbf{D}_s\mathbf{c}\|_2^2 \\ &\quad + \|\mathbf{a}_i - \mathbf{D}_a\mathbf{c}\|_2^2 + \lambda\|\mathbf{c}\|_1 \\ \mathbf{c} \in \mathbb{R}^m,\ \ \mathbf{D}_s &\in \mathcal{D}^{n_s\times m},\ \ \mathbf{D}_a \in \mathcal{D}^{n_a\times m}.\end{aligned} \tag{7}$$

To get compliable to the testing settings of the model in Eqn. (5), 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 update $\mathbf{D}_s$ using $\mathbf{D}_a$ and $\mathbf{c}$. These four steps are summarized from Lines 6 to 9 in Algorithm 1, where a batch training mode is employed to speed up the model convergence.

The underlying oracles for this optimization procedure hide behind the starting steps (i.e., Steps 1 and 2) and exiting steps (i.e., Steps 3 and 4) of the iteration. In Step 1, in order to capture different modes from multi-view face shapes, the face shape dictionary is first 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 the testing phase, the representation coefficient can only be obtained from the face appearance dictionary. Therefore, the coefficient used to update the face shape dictionary in Step 4 are only from the face appearance dictionary in Step 3. Like the two-step optimization procedure in [34], [37], [38], 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 phase and thus learns a more effective model.

Algorithm 1. SRD Model Learning

```
Input: training set $\mathcal{X} = \{(\mathbf{x}_i, \mathbf{p}_i)\}_{i=1}^N$, $m$ (dictionary size), $\lambda$ (regularization parameter), $T$ (maximal stage number).
Output: learned SRD model $\mathcal{M}$.
1:  Initialization: generate $\mathbf{P}^0 = [\mathbf{p}_1^0; \ldots; \mathbf{p}_N^0]$ using the face detector, and fill $\mathcal{M}$ with $m$ training samples.
2:  for $t = 0 \rightarrow T$ do
3:    Build training set $\mathcal{X}^t = \{(\mathbf{s}_i^t, \mathbf{a}_i^t)\}_{i=1}^N$ from $\mathcal{X}$ and $\mathbf{P}^t$.
4:    $\mathbf{A}^t \leftarrow [\mathbf{a}_1^t, \ldots, \mathbf{a}_N^t]$, $\mathbf{S}^t \leftarrow [\mathbf{s}_1^t, \ldots, \mathbf{s}_N^t]$.
5:    while not converged do
6:      Step 1: fix $\mathbf{D}_a^t$ to learn $\mathbf{D}_s^t$ and $\mathbf{C}$ on $\mathcal{X}^t$:
          $\mathop{\arg\min}_{\mathbf{D}_s^t,\mathbf{C}} \left\{ \min_{\mathbf{C}} \|\mathbf{S}^t - \mathbf{D}_s^t\mathbf{C}\|_2^2 + \lambda\|\mathbf{C}\|_1 \right\}.$
7:      Step 2: update $\mathbf{D}_a^t$ using $\mathbf{A}^t$ and $\mathbf{C}$: $\mathbf{D}_a^t = \mathbf{A}^t/\mathbf{C}$.
8:      Step 3: fix $\mathbf{D}_s^t$ to learn $\mathbf{D}_a^t$ and $\mathbf{C}$ on $\mathcal{X}^t$:
          $\mathop{\arg\min}_{\mathbf{D}_a^t,\mathbf{C}} \left\{ \min_{\mathbf{C}} \|\mathbf{A}^t - \mathbf{D}_a^t\mathbf{C}\|_2^2 + \lambda\|\mathbf{C}\|_1 \right\}.$
9:      Step 4: update $\mathbf{D}_s^t$ using $\mathbf{D}_a^t$ and $\mathbf{C}$: $\mathbf{D}_s^t = \mathbf{S}^t/\mathbf{C}$.
10:   end while
11:   Update $\mathbf{P}^{t+1} = [\mathbf{p}_1^{t+1}; \ldots; \mathbf{p}_N^{t+1}]$: $\mathbf{P}^{t+1} = \mathbf{P}^t + \mathbf{D}_s^t\mathbf{C}$.
12: end for
13: Generate the learned SRD model $\mathcal{M} = \{[\mathbf{D}_s^t; \mathbf{D}_a^t]\}_{t=0}^T$.
```

4.2 Stage-Wise Optimization

One crucial point for the regression-based face alignment algorithms is an iterative mechanism to refine the face shape hypothesis towards the true face shape. Take the supervised descent model (SDM) [12] as an example, the face shape hypothesis is progressively refined up to four times from the SIFT descriptors extracted around landmarks in the current estimation. Inspired by this observation, we propose to learn the relational dictionary in a stage-wised manner to provide an iterative mechanism for the ℓ_1-induced face alignment model.

Denoting the learned relational dictionary using the four-step optimization procedure from the 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 landmark positions in the training set. Specifically, for the ith training sample, we first represent its initial appearance $\mathbf{a}_i^0$ sparsely using the learned appearance dictionary $\mathbf{D}_a^0$, i.e.,

$$\mathbf{c}_i^* = \mathop{\arg\min}_{\mathbf{c}\in\mathbb{R}^m} \|\mathbf{a}_i^0 - \mathbf{D}_a^0\mathbf{c}\|_2^2 + \lambda\|\mathbf{c}\|_1. \tag{8}$$

Based on the learned relational dictionary, the representation coefficient $\mathbf{c}_i^*$ can be used to give an estimation of the displacement between its current estimated landmark positions $\mathbf{p}_i^0$ and the true landmark positions $\mathbf{p}_i$. Using a linear combination of the corresponding shape displacements indicated by the representation coefficient $\mathbf{c}_i^*$, the estimated landmark locations $\mathbf{p}_i^0$ for the ith training sample can be updated by adding this displacement

$$\mathbf{p}_i^1 = \mathbf{p}_i^0 + \mathbf{D}_s^0\mathbf{c}_i^*. \tag{9}$$

Then the face shape representation for the ith training sample can be updated correspondingly

$$\mathbf{s}_i^1 = \mathbf{p}_i - \mathbf{p}_i^1 = \mathbf{s}_i^0 - \mathbf{D}_s^0\mathbf{c}_i^*. \tag{10}$$

With the updated face shape estimations $\{\mathbf{s}_i^1\}_{i=1}^N$, we can continue to build 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 features of the training samples from the updated face landmarks, i.e., $\mathbf{a}_i^1 = \mathbf{h}(\mathbf{x}_i, \mathbf{p}_i^1)$. 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 is repeated until the rebuilt dataset converges, i.e., the differences between the estimated landmark locations and labeled ones are small enough. The final model will contain multiple relational dictionaries trained at different stages, i.e., $\mathcal{M} = \{[\mathbf{D}_s^t; \mathbf{D}_a^t]\}_{t=0}^T$, where T is the stage number of the stage-wised learning process. Considering the characteristics of these learned dictionaries, we refer to them as *Stage-wise Relational Dictionary* (SRD) model. In Algorithm 1, we summarize the learning process of the SRD model. Note that in Algorithm 1, the dictionary learning problem and sparse representation problem are all performed in a batch model. The matrix $\mathbf{C} = [\mathbf{c}_1; \ldots; \mathbf{c}_N]$ is defined as the representation coefficients for all the samples. In our experiments, we observe that the rebuilt training set quickly converges in only 2 or 3 stages.

During the testing phase, given a face image $\mathbf{x}$, we first produce the initial estimation of landmark locations $\mathbf{p}^0$ from face detection, e.g., the mean positions of the face landmarks learned from the training set. Then, based on this initial estimation, we employ the learned SRD model to iteratively update the representations of the face appearance $\mathbf{a}$ and shape $\mathbf{s}$ using Eqns. (8) and (9). In Algorithm 2, we summarize the overall testing procedure of face alignment using the learned SRD model. Here we can clearly see that this process is exactly the same as the face appearance and shape updating process performed in training the model, thus making the model testing phase compatible to the model training phase. The stage-wised updating process will progressively guide the updated face appearance and shape towards the true value in the test image.

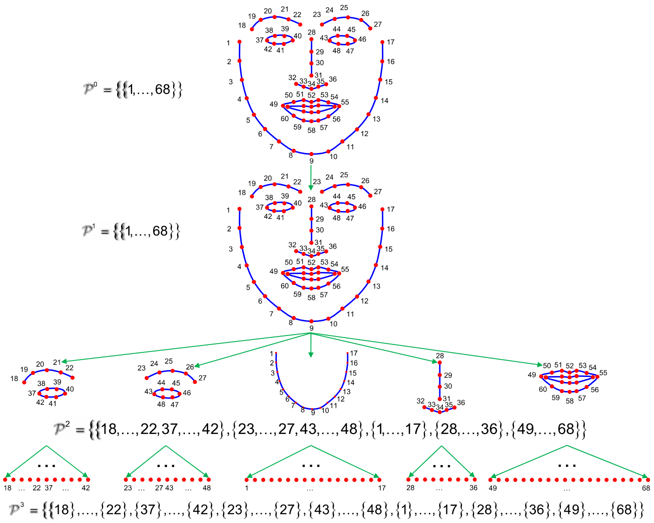
Algorithm 2. Face Alignment Using the SRD Model

```
Input: the testing face image $\mathbf{x}$, learned SRD model $\mathcal{M} = \{[\mathbf{D}_s^t; \mathbf{D}_a^t]\}_{t=0}^T$.
Output: estimated landmark locations $\hat{\mathbf{p}}$.
Initialization: get an initial estimation of landmark locations $\mathbf{p}^0$ in $\mathbf{x}$ by face detection.
for $t = 0 \to T$ do
    Extract face appearance $\mathbf{a}^t$ around $\mathbf{p}^t$ in $\mathbf{x}$:
        $\mathbf{a}^t = \mathbf{h}(\mathbf{x}, \mathbf{p}^t).$
    Sparse representation of face appearance $\mathbf{a}^t$ using $\mathbf{D}_a^t$:
        $\mathbf{c}^* = \arg\min_{\mathbf{c} \in \mathbb{R}^m} \|\mathbf{a}^t - \mathbf{D}_a^t \mathbf{c}\|_2^2 + \lambda \|\mathbf{c}\|.$
    Update the estimation of the landmark locations $\mathbf{p}^{t+1}$:
        $\mathbf{p}^{t+1} = \mathbf{p}^t + \mathbf{D}_s^t \mathbf{c}^*.$
end for
Generate final landmark locations $\hat{\mathbf{p}} = \mathbf{p}^{T+1}$.
```

5 Hierarchical SRD Model Learning

A very natural way to apply the SRD model to face alignment is to extract the face appearance and shape representations from all the facial landmarks—that is, by training and testing the SRD model on the whole face to perform alignment globally. Supposing that the training set $\mathcal{X} = \{(\mathbf{x}_i, \mathbf{p}_i)\}_{i=1}^N$ is labeled with L facial landmarks, the landmark annotation for one sample $(\mathbf{x}, \mathbf{p})$ can be denoted as $\mathbf{p} = [p_1, \ldots, p_L]^\top$, where $p_l = [x_l, y_l]^\top$, $l = 1, \ldots, L$, is the 2D coordinates of the lth landmark. Based on the formulations in Sections 3 and 4, and denoting $\mathbf{p}^0 = [p_1^0, \ldots, p_L^0]^\top$ as the initial landmark estimation, the face shape representation for the first stage training can thus be represented as $\mathbf{s}^0 = [s_1^0, \ldots, s_L^0]^\top = [p_1 - p_1^0, \ldots, p_L - p_L^0]^\top$, and the corresponding appearance representation is extracted around all these L landmarks accordingly.

This kind of global representations for the face appearance and shape has obtained very promising alignment results [39]. One key reason for its success is that the global shape constraint guarantees very robust alignment results. The main problem for the global representations, however, is that sometimes its alignment results may not be so accurate, especially when the testing face undergoes exaggerated local deformations in some of its regional parts, such as blinking eyes and pursed lips. To simultaneously ensure the robustness and accuracy of the alignment results, we propose to learn a hierarchical SRD model, which we refer to as the HSRD model. The basic idea for the HSRD model is to learn coarse-to-fine face representations for different training stages and progressively adapt them from the whole face to individual landmarks. We first introduce some denotations for the formulation of the HSRD model.



$\mathcal{P}^0 = \{\{1,\ldots,68\}\}$
$\mathcal{P}^1 = \{\{1,\ldots,68\}\}$
$\mathcal{P}^2 = \{\{18,\ldots,22,37,\ldots,42\},\{23,\ldots,27,43,\ldots,48\},\{1,\ldots,17\},\{28,\ldots,36\},\{49,\ldots,68\}\}$
$\mathcal{P}^3 = \{\{18\},\ldots,\{22\},\{37\},\ldots,\{42\},\{23\},\ldots,\{27\},\{43\},\ldots,\{48\},\{1\},\ldots,\{17\},\{28\},\ldots,\{36\},\{49\},\ldots,\{68\}\}$

Fig. 1. A four-level hierarchy in the face partition sets for a face annotated with 68 landmarks.

For a training set labeled with L landmarks, let $\mathcal{I}$ be the *index set* of all the landmarks, i.e., $\mathcal{I} = \{1, \ldots, L\}$. A *partition set* of the index set $\mathcal{I}$, denoted as $\mathcal{P}$, is a set corresponding to an arbitrary splitting of the $\mathcal{I}$ into some nonempty disjoint sets. For any subset $p \in \mathcal{P}$, we use $\mathbf{s}_p$ and $\mathbf{a}_p$ to respectively denote the face shape representation and the appearance representation built from its corresponding landmarks. Let $\mathbf{s}_\mathcal{P} = \{\mathbf{s}_p\}_{p\in\mathcal{P}}$ and $\mathbf{a}_\mathcal{P} = \{\mathbf{a}_p\}_{p\in\mathcal{P}}$ denote the compositional representation sets respectively built for the face appearance and shape based on $\mathcal{P}$. If the partition set $\mathcal{P}$ contains only one element, i.e., $|\mathcal{P}| = 1$, $\mathbf{s}_p$ and $\mathbf{a}_p$ correspond to the appearance and shape representations built from the whole face. If $|\mathcal{P}| = L$, $\mathbf{s}_p$ and $\mathbf{a}_p$ correspond to the appearance and shape representations built from each individual landmark.

Based on above definitions and denotations, we get a general way to represent the face appearance and shape which can be built from any partitions of the whole face, thus making it feasible to learn coarse-to-fine face representations. During the early training stages, the HSRD model learns the appearance and shape representations from the whole face to refine the alignment globally. As the training goes on, it learns finer representations for regional face parts and refines their corresponding alignments on the local part level. At the last training stage, it learns the representations from each landmark to refine the alignment individually.

To realize these ideas, appropriate partition sets should be employed for different stages to learn their specific training objectives. Specifically, denoting the partition sets as $\{\mathcal{P}_t\}_{t=0}^T$, it should employ coarser partition set $\mathcal{P}_t$ when t is small and finer one when t is large. In Fig. 1, we plot the partition sets employed for aligning a face with 68 landmarks. The designation of the four-layer hierarchy for the partition sets in Fig. 1 has very obvious semantic meanings and provides a straightforward way to perform coarse-to-fine alignment. In the partition sets for the first two stages, all the 68 landmarks are taken as a whole to perform alignment globally, which ensures the model to find a robust primary alignment result. In the following stage, all the regional semantic face parts, including left eye, right eye, face contour, nose, and mouth, are respectively grouped together to adjust the alignment in the part level. At the last stage, each landmark is isolated in a partition set and refined individually to finally improve the alignment accuracy.

Algorithm 3. HSRD Model Learning

```
Input: training set $\mathcal{X} = \{(\mathbf{x}_i, \mathbf{p}_i)\}_{i=1}^N$, $m$ (dictionary size), $\lambda$ (regularization parameter), $T$ (maximal stage number).
Output: learned HSRD model $\mathcal{M}$.
1: Initialization: generate $\mathbf{P}^0 = [\mathbf{p}_1^0; \ldots; \mathbf{p}_N^0]$ using the face detector, and fill $\mathcal{M}$ with $m$ training samples.
2: Create the partition sets $\{\mathcal{P}^t\}_{t=0}^T$ like in Fig. 1.
3: for $t = 0 \to T$ do
4:     Build training set $\mathcal{X}^t$ from $\mathcal{X}$, $\mathbf{P}^t$ and $\mathcal{P}^t$.
5:     for every $p$ in $\mathcal{P}^t$ do
6:         Learn $\mathbf{D}_{a,p}^t$, $\mathbf{D}_{s,p}^t$ and $\mathbf{C}_p$ from $\{(\mathbf{s}_{i,p}^t, \mathbf{a}_{i,p}^t)\}_{i=1}^N$ in $\mathcal{X}^t$ using the four-step procedure in Algorithm 1.
7:         Update $\mathbf{P}^{t+1}$ partially: $\mathbf{P}_p^{t+1} = \mathbf{P}_p^t + \mathbf{D}_{s,p}^t \mathbf{C}_p$.
8:     end for
9: end for
10: Generate the model $\mathcal{M} = \{\{[\mathbf{D}_{s,p}^t; \mathbf{D}_{a,p}^t]\}_{p \in \mathcal{P}^t}\}_{t=0}^T$.
```

Based on the partition sets, the HSRD model learns, at each stage, a set of appearance and shape dictionaries according to the corresponding partition set. Take stage t for example, for each element p in the partition set $\mathcal{P}^t$, the HSRD model learns a shape dictionary $\mathbf{D}_{s,p}^t$ and an appearance dictionary $\mathbf{D}_{a,p}^t$ from the corresponding training samples using the four-step optimization procedure described in Algorithm 1. It then updates the estimations of the corresponding landmarks indicated by p. Algorithm 3 summarizes the overall learning process of the HSRD model. During the testing phase, the deployment of the learned HSRD model for face alignment is very similar to Algorithm 2. The only difference is that each relational dictionary only needs to update the locations of its corresponding landmarks and the whole hierarchical and stage-wise optimization process therefore naturally performs alignment from a coarser level to a finer one.

6 Occlusion Learning SRD Model

The SRD model proposed in Section 4, as well as its hierarchical extension HSRD model described in Section 5, can intrinsically deal with face view variations via simultaneously appearance-shape modeling. To deal with partial face occlusions, we further propose to perform occlusion modeling via jointly learning an occlusion dictionary within the rational dictionary. Hereafter, we refer to this occlusion handling SRD model as OSRD for short.

6.1 OSRD Model Formulation

In our OSRD model, an occlusion dictionary is added along with the appearance dictionary in the SRD model, and, consequently, the loss function in Eqn. (6) becomes

$$l(\mathbf{a}, \mathbf{s}, \mathbf{D}) = \min_{\mathbf{c},\mathbf{e}} \{\|\mathbf{s} - \mathbf{D}_s\mathbf{c}\|_2^2 + \|\mathbf{a} - [\mathbf{D}_a, \mathbf{D}_o][\mathbf{c}; \mathbf{e}]\|_2^2 + \lambda\|[\mathbf{c}; \mathbf{e}]\|_1\}, \tag{11}$$

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 coefficient of the occlusion dictionary $\mathbf{D}_o$, denotation $[\cdot, \cdot]$ is the column based concatenation of vectors or matrices, and $\mathbf{D} = [\mathbf{D}_s; \mathbf{D}_a, \mathbf{D}_o]$. The testing procedure of this model is similar to that of the SRD model (see Algorithm 2). The difference is that the occlusion dictionary $\mathbf{D}_o$ is expected to represent the occluded part of the face appearance, while the visible part of the face appearance is represented by the dictionary $\mathbf{D}_a$. If the OSRD model is trained to work in this way, it will be very robust to partial face occlusions.

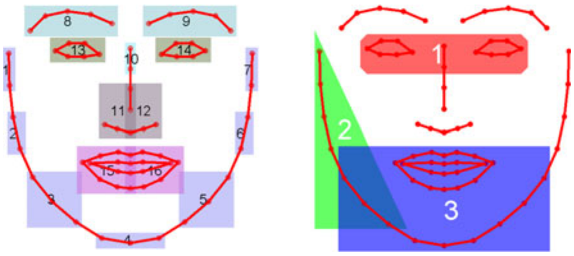
6.2 OSRD Model Analyses

One of the most widely adopted approaches to occlusion modeling in the sparse representation is using an identity matrix [41], [42], i.e., let $\mathbf{D}_o = \mathbf{I}_{n_a \times n_a}$. This kind of occlusion dictionary, however, has several serious problems in our setting for the face alignment problem. One practical problem is that the size of the identity matrix equals to the dimension of the appearance dictionary, which is often of very high dimensionality. This prohibits the ℓ_1 minimization process involved in the training and testing phases. Another more serious problem is that an identity matrix itself can theoretically represent anything, even with the sparsity constraint, it still represents much more than occlusions. This is likely to spoil the learned dictionary. Last but not least, since the identity matrix is directly added for the testing phase but not simultaneously learned with the relational dictionary during model training, it also causes the incompatibility problem between the model training and testing. To address all these problems, we propose to learn a more compact and expressive occlusion dictionary in a data driven manner to represent different kinds of face occlusions.

6.3 OSRD Model Learning

Our principle is to learn the occlusion dictionary jointly with the appearance and shape dictionary in a data driven manner to model face occlusions efficiently and effectively. To this end, there are three issues that need to be resolved. The first issue is how to automatically build a training set with enough samples of occluded faces for the occlusion dictionary, since manually labeling the occlusion states of the face images in the pixel level or at least in the landmark level is very difficult and time-consuming for a large dataset. The second one is how to model the occlusion dictionary in a compact form that can be learned efficiently and effectively. The last one is how to jointly learn the occlusion dictionary with the appearance and shape dictionaries to obtain compatible models. In the following, we will elaborate on the solving of these three issues respectively.

First, to automatically build a training set with different kinds of partial face occlusions, we employ the appearances of non-face image regions to simulate the appearances of the face occluders. Supposing that the original training set $\mathcal{X} = \{(\mathbf{x}_i, \mathbf{p}_i)\}_{i=1}^N$ has no face occlusion, we first build a new training set $\mathcal{Y}$ by copying the images in $\mathcal{X}$ but shifting the



(a) EOPs (b) POPs

Fig. 2. 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 plotted by combining some EOPs.

landmark annotations and corresponding face detection bounding box to another place in the same image where no face exists. Note that the face shape constraint between the landmark annotations and detection bounding box stays unchanged during the shifting process. Based on $\mathcal{Y}$, we extract the appearance and shape features from it to build a training set for the full occlusion dictionary learning just as the same to that from $\mathcal{X}$ (cf. Section 4). We refer the extracted training set as the *full occlusion training set* and denote it as $\mathcal{Y}^t = \{(\mathbf{s}_i^t, \hat{\mathbf{a}}_i^t)\}_{i=1}^N$ in the tth stage, where $\hat{\mathbf{a}}_i^t$ denotes the face appearance representation under full occlusions. The training set $\mathcal{X}^t = \{(\mathbf{s}_i^t, \mathbf{a}_i^t)\}_{i=1}^N$ extracted from the original training set $\mathcal{X}$ is referred to as the *no occlusion training set*. Based on the training set $\mathcal{Y}^t$ and $\mathcal{X}^t$, we can simulate all kinds of partial face occlusions by randomly replacing the face appearance at some landmarks in $\mathcal{X}^t$ using the corresponding ones in $\mathcal{Y}^t$. We denote this *partial occlusion training set* as $\mathcal{Z}^t = \{(\mathbf{s}_i^t, \tilde{\mathbf{a}}_i^t)\}_{i=1}^N$. Here $\tilde{\mathbf{a}}_i^t$ is face appearance representation under partial occlusions.

Second, to learn the occlusion dictionary D_o in a compact form, we restrict our occlusion dictionary to deal with only partial face occlusions. This restriction is reasonable, since due to the lack of observations, face alignment under full occlusion is an infeasible problem. To deploy this restriction, we first define a set of Elemental Occlusion Patterns (EOPs) based on observations of the face occlusions in the wild. Take a face image labeled with 68 landmarks as an example, 16 EOPs are defined as shown in Fig. 2a, each of which covers several neighboring landmarks. With this kind of definition, each element (one column vector) in the occlusion dictionary D_o represents one specific EOP and its size k (column number) can thus be much smaller than the size of appearance representation n_a. Since the combinations of these EOPs can approximate almost all kinds of face occlusion patterns in realistic ways, we can further generate a set of Partial Occlusion Patterns (POPs) by different sparse combinations of the EOPs. Fig. 2b shows three examples of the POPs. When generating the partial occlusion training set $\mathcal{Z}^t$, we restrict it to follow these POP constraints, thus drive the model to deal with partial occlusions.

The last issue now is how to jointly learn the occlusion dictionary with the appearance and shape dictionaries from the training sets. We propose an extended procedure as in Algorithm 1 to solve the OSRD learning problem, which is described in Algorithm 4. The matrix $\mathbf{E} = [\mathbf{e}_1; \ldots; \mathbf{e}_N]$ is defined as the representation coefficient for the partial occlusion dictionary for all the samples. Its main optimization process is a seven-step iteration which proceeds 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 $\mathbf{D}_O$ using $\mathbf{D}_a$ on $\mathcal{Y}^t$ (Step 4) and then extend $\mathbf{D}_O$ to a block dialog form based on the EOPs to generate the partial occlusion dictionary $\mathbf{D}_o$ (Step 5). Next, upon fixing $\mathbf{D}_a$ and $\mathbf{D}_o$, it minimizes the representation errors on $\mathcal{Z}^t$ (Step 6). With the best representation coefficients, the shape dictionary $\mathbf{D}_s$ is updated by least square fitting (Step 7). These seven steps are iterated several times to converge.

Algorithm 4. 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), $\lambda$ (regularization parameter),
       $T$ (maximal stage number), EOPs and POPs.
Output: learned OSRD model $\mathcal{M}$.
Initialization: generate $\mathbf{P}^0 = [\mathbf{p}_1^0; \ldots; \mathbf{p}_N^0]$ using the face
  detector, generate $\mathcal{Y}$ from $\mathcal{X}$, and fill $\mathcal{M}$ with $m$ training
  samples in $\mathcal{X}$ and $k$ training samples in $\mathcal{Y}$.
for $t = 0 \rightarrow T$ do
    Build training set $\mathcal{X}^t$, $\mathcal{Y}^t$, $\mathcal{Z}^t$ from $\mathcal{X}$ and $\mathcal{Y}$.
    $\mathbf{A}^t \leftarrow [\mathbf{a}_1^t, \ldots, \mathbf{a}_N^t]$, $\mathbf{S}^t \leftarrow [\mathbf{s}_1^t, \ldots, \mathbf{s}_N^t]$,
    $\hat{\mathbf{A}}^t \leftarrow [\hat{\mathbf{a}}_1^t, \ldots, \hat{\mathbf{a}}_N^t]$, $\tilde{\mathbf{A}}^t \leftarrow [\tilde{\mathbf{a}}_1^t, \ldots, \tilde{\mathbf{a}}_N^t]$.
    while not converged do
        Step 1: fix $\mathbf{D}_a^t$ to learn $\mathbf{D}_s^t$ and $\mathbf{C}$ on $\mathcal{X}^t$:
            $\arg\min_{\mathbf{D}_s^t, \mathbf{C}} \left\{ \min_{\mathbf{C}} \|\mathbf{S}^t - \mathbf{D}_s^t \mathbf{C}\|_2^2 + \lambda \|\mathbf{C}\|_1 \right\}.$
        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 $\mathcal{X}^t$:
            $\arg\min_{\mathbf{D}_a^t, \mathbf{C}} \left\{ \min_{\mathbf{C}} \|\mathbf{A}^t - \mathbf{D}_a^t \mathbf{C}\|_2^2 + \lambda \|\mathbf{C}\|_1 \right\}.$
        Step 4: fix $\mathbf{D}_a^t$ to learn $\mathbf{D}_O^t$ on $\mathcal{Y}^t$:
            $\arg\min_{\mathbf{D}_O^t} \left\{ \min_{\mathbf{C}, \mathbf{E}} \|\hat{\mathbf{A}}^t - [\mathbf{D}_a^t, \mathbf{D}_O^t][\mathbf{C}; \mathbf{E}]\|_2^2 + \lambda \|[\mathbf{C}; \mathbf{E}]\|_1 \right\}.$
        Step 5: Expend $\mathbf{D}_O^t$ to block dialog form to get $\mathbf{D}_o^t$.
        Step 6: fix $\mathbf{D}_a^t$ and $\mathbf{D}_o^t$ on $\mathcal{Z}^t$ 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 $\mathbf{P}^{t+1} = [\mathbf{p}_1^{t+1}; \ldots; \mathbf{p}_N^{t+1}]$: $\mathbf{P}^{t+1} = \mathbf{P}^t + \mathbf{D}_s^t \mathbf{C}$.
end for
Generate learned OSRD model $\mathcal{M} = \left\{[\mathbf{D}_s^t; \mathbf{D}_a^t, \mathbf{D}_o^t]\right\}_{t=0}^T$.
```

As a complement to the last issue, here we describe in detail on how to extend $\mathbf{D}_O$ to a block diagonal matrix to form $\mathbf{D}_o$ in Step 5. Denote the 16 EOPs illustrated in Fig. 2a using a matrix form as $\mathbf{O}_e = [\mathbf{o}_1^e, \ldots, \mathbf{o}_{16}^e]$, where $\mathbf{o}_i^e \in \{0, 1\}^{n_a}$ is the indicator vector of the ith EOP, whose entry to each of its n_a dimensions indicates whether the appearance representation is simulated with occlusion based on its corresponding landmark. Denote the full occlusion dictionary as $\mathbf{D}_O = [\mathbf{d}_1^O, \ldots, \mathbf{d}_r^O]$ and the partial occlusion dictionary as $\mathbf{D}_o = [\mathbf{d}_1^o, \ldots, \mathbf{d}_k^o]$, where $r = k/16$ is the size of $\mathbf{D}_O$. This extension process is illustrated in Fig. 3. Then, using the ith element in $\mathbf{O}_e$ and jth element in $\mathbf{D}_O$,

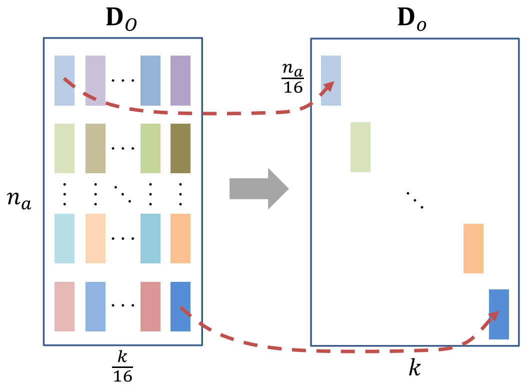


Fig. 3. Illustration of extending the full occlusion dictionary $\mathbf{D}_O$ to a block diagonal form for the partial occlusion dictionary $\mathbf{D}_o$.

the $(j \times 16 + i)$th element in the partial occlusion dictionary $\mathbf{D}_o$ can be obtained by $\mathbf{d}^o_{j\times16+i} = \mathbf{d}^O_j \otimes \mathbf{o}^e_i$, where $\otimes$ is the element-wise multiplication.

6.4 OSRD Model Short Summary

Our OSRD model provides an explicit mechanism for occlusion modeling in the face alignment problem, and is designed to represent partial face occlusions from different patterns in a generative and a data-driven manner. To train the OSRD model, there is no need to manually annotate the occlusion state of the face in the pixel level or in the landmark level, which saves lots of time in practice. It is remarkable that the OSRD model can be trained using only simple face samples without occlusions and be used to test samples with different kinds of partial occlusions. With the testing results of the model, the representation coefficient for the occlusion dictionary can give some evidence for the occurrences of specific occlusions. All these properties make the OSRD model distinctive from other face alignment models for face images with partial occlusions.

7 Experiments

We have conducted extensive experiments to analyze and verify the performances of the proposed approach on different benchmark datasets, including the BioID dataset [42], the LFPW dataset [9], the 300-W dataset [43] (a combination of the AFW dataset [11], Helen dataset [44], the XM2VTS [45] dataset and iBug dataset [43]), the COFW dataset [22], the MVFW and OCFW datasets [39]. To make the experimental evaluations feasible, we have also employed various kinds of ground-truth landmark annotations, i.e., 19, 29, 51, and 68 points. The definitions of these landmark annotations are shown in Fig. 4.

7.1 Model Analyses

To gain some insights of the proposed model and facilitate the reproduction of this work, we provide some deeper analyses of the proposed model, as well as some implementation details. All the results in this section is performed on a small dataset we build from the Helen dataset for analysis purpose. This small analyzing dataset contains 1,000 training samples and 200 testing samples, with the annotations of 68 landmarks as shown in Fig. 4d.

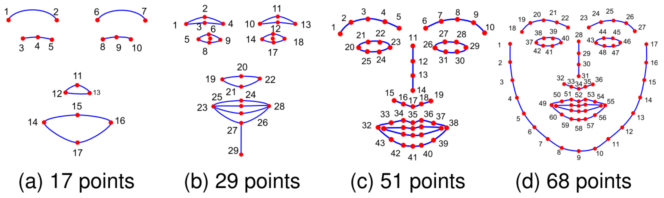

(a) 17 points (b) 29 points (c) 51 points (d) 68 points

Fig. 4. Different definitions of facial landmark annotations and their corresponding mean shapes estimated from a face detector: (a) 17 points, (b) 29 points, (c) 51 points, and (d) 68 points.

7.1.1 Feature Descriptors

We explore two different types of feature descriptors, HoG [46] and SIFT [47], around the landmarks to represent the face appearance. By training and testing the SRD model using these two types of feature descriptors respectively on the analyzing dataset, it is observed that the HoG descriptors is more computationally efficient and performs comparably or even slightly better than SIFT, especially for landmarks located around the face contours. The SIFT descriptor, on the other side, is more robust for locating the inner facial landmarks (e.g., the 51 points in Fig. 4c). To make a fair comparison with the SDM [12] in the following experiments, the performances reported are all based on the same 128-dimensional SIFT descriptor adopted in [12].

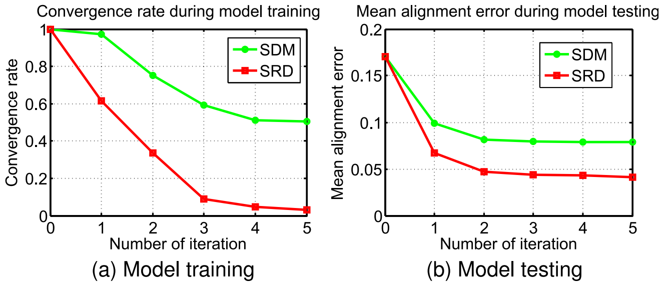
7.1.2 Convergence Rate

To investigate the convergence properties of the proposed model, we first calculate the convergence rate after each training stage of the SRD model on the analyzing dataset. Fig. 5a plots the convergence line chart of the SRM model, with comparisons to the SDM model. The SDM method is a quite effective face alignment algorithm proposed recently [12]. We implement it based on the partially released code in [12]. Since the SDM model and our SRD model use different training objectives, to compare their convergency in one figure, we normalize the convergence rate in each stage using the objective value in the previous stage.

Besides model training, we further examine the change of the mean alignment errors after each testing stage of the SRD model using the metric widely adopted in previous works [9], [10], [12]. This metric measures the alignment error using the average euclidean distance between the predicted landmark locations and the labeled landmark locations, normalized by the inter-ocular distance, which is defined as the euclidean distance between the outer corners of the eyes [43]. The line charts for the mean alignment errors of our SRD model and the SDM model are plotted in Fig. 5b. As can be observed from Fig. 5, during the training phase, the convergence rate of the SRD model decreases quickly and approaches to zero after only 3 iterations, while the convergence rate of the SDM model remains at a very high value, i.e., 0.5, which indicates not only more iterations to converge, but also a tendency to over-fitting. During the testing phase, similar results can also be observed for the alignment error. All these results demonstrate good convergency of the SRD model.

7.1.3 Learning Algorithms

One contribution of this work is the proposal of the four-step iterative procedure to solve the relational dictionary



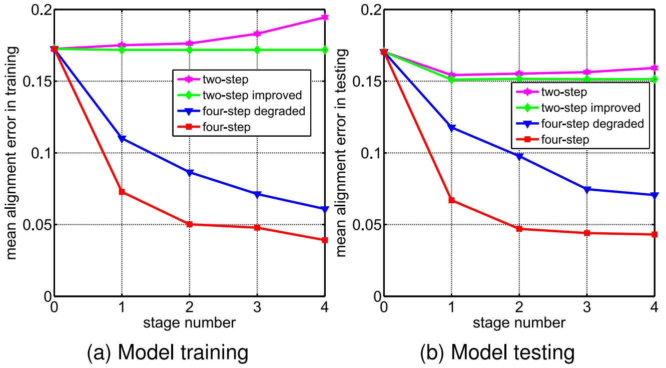
(a) Model training (b) Model testing

Fig. 5. The convergence properties of the SRD model during training phase (a) and testing phase (b), with comparisons to the SDM model.

learning problem formulated in Eqn. (7). Here we examine the effectiveness of this optimization algorithm and compare it with the widely used two-step iterative procedure [34], [38] discussed in Section 3.2.

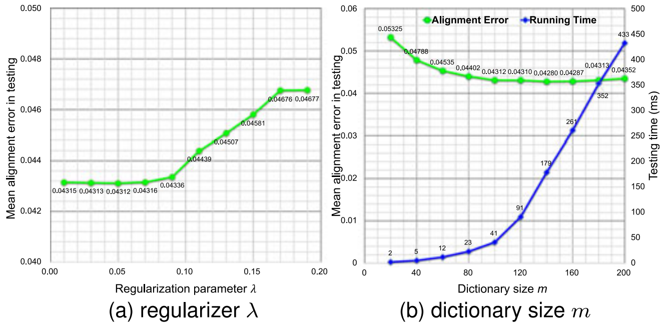
To gain a deeper insight into the learning procedure, we also present two other baselines for learning the relational dictionary. One baseline is an improved version of the two-step iterative procedure to make it comparable to the testing settings of the SRD model. This baseline first learns the appearance dictionary $\mathbf{D}_a$ using the two-step iterative procedure, and then updates the shape dictionary $\mathbf{D}_s$ via the least square fitting. The other baseline is a degraded version of our four-step iterative procedure, which only performs one iteration of four steps in Algorithm 1. We train the SRD model using these four optimization procedures on the analyzing dataset respectively and record the training errors and the testing errors at each stage.

Fig. 6 plots the training and testing errors in each stage of the SRD model using the four procedures. We can observe that the original two-step iterative procedure even do not converge during the training stage. This is because the updating of the landmark estimations (i.e., Line 11 in Algorithm 1) is actually a testing process of the learned model. The learned relational dictionary model using the two-step iterative procedure is incompatible to the testing setting of the model, thus cannot reduce the alignment error for the training of the relational dictionary in the next stage. By making the training setting of the relational dictionary comparable to its testing setting, the improved version of the two-step iterative procedure starts to converge in the training stage. The degraded version of the four-step iterative procedure first learns the shape dictionary and uses it to initialize the appearance dictionary. This makes it more effective to learn the relational dictionary and obtains reasonable results in the testing phase. By iterative learning and updating, the four-step iterative procedure obtains the best performance within the four learning algorithms.



(a) Model training (b) Model testing

Fig. 6. The effectiveness of different optimization procedures for learning the relational dictionary: (a) mean alignment errors in different training stages; (b) mean alignment errors in different testing stages.



(a) regularizer λ (b) dictionary size m

Fig. 7. Investigating the effects of the regularizer λ (a) and dictionary size m (b) on the alignment errors.

7.1.4 Model Parameters

The testing phase of the proposed model does not need to set additional parameters, as it only requires the same parameter settings as in the model training phase. During the training phase, two parameters may affect the performances of the alignment result, i.e., the regularization parameter λ and the dictionary size m. They are set through cross-validation during training as follows.

Regularization Parameter λ. The regularization parameter λ in Eqn. (7) controls the sparsity of the representation coefficient $\mathbf{c}$. It is the most common parameter in the sparse representation framework. In particular, we tune this parameter via cross validation over the analyzing dataset. Through the experiments we find that setting its value within the range $[0.01, 0.1]$ generates stably good performances (see Fig. 7a). Therefore, we set $\lambda = 0.05$ throughout the experiments.

Dictionary Size m. For the dictionary size m, we find that a larger size usually improve the performance consistently, but will also increase the computationally cost (See Fig. 7b). We take a trade-off and set the size of the appearance and shape dictionary size as $m = 100$. Similar observations can also be found for the full occlusion dictionary size k, which is set to $k = 80$ in all the experiments. With this setting, averagely five bases are learned for each of the 16 EOPs defined in Fig. 2a.

7.1.5 Running Speed

In our current unoptimized MATLAB implementation, the ℓ_1 minimization problem is solved via the LASSO algorithm [35] based on the implementation from [34]. For the model training, it takes about 30 minutes to train the SRD model from 2,000 samples. There are two main procedures in the model training, feature extraction and dictionary learning. Both of them have the linear time complexity with the number of the training samples. In the testing phase, provided the face detection results, it costs about 100 ms to align one face. Since about half of the computational cost comes from the feature extraction procedure, the testing process of the SRD model is thus very efficient and the overall alignment speed can be further improved if more efficient feature representations are adopted.

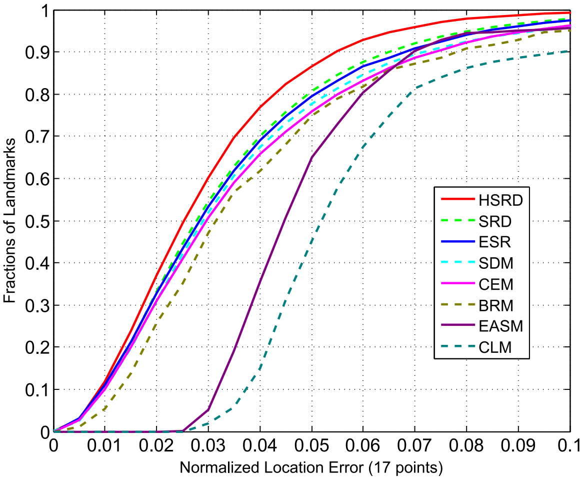


Fig. 8. Cumulative error distribution curves on the BioID dataset.

7.2 General Face Alignment Evaluation

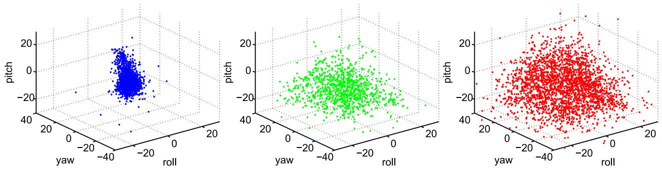
To compare with other face alignment algorithms, we first evaluate our proposed model on two widely used benchmark datasets, BioID [42] and LFPW [9]. The BioID dataset contains 1521 testing face images annotated with 17 landmarks (Fig. 4a). Most of the faces in this dataset are frontal or near frontal. The LFPW dataset contains 1,132 training images and 300 testing images, annotated with 29 landmarks (Fig. 4b). Since the original version of the LFPW datset is no longer available from its provided URLs, we use its augmented version provided by [10] in the experiments. For fair comparisons, we follow the same experimental settings and evaluation metrics as adopted in [9], [10] and [39].

In Fig. 8, we plot the cumulative error distribution (CED) curves of our approach on the BioID dataset as in [10]. The six baseline methods are the explicit shape regression method (ESR) [10], the SDM method [12], the consensus exemplar method (CEM) [9], the boosted regression method (BRM) [30], the extended ASM method (EASM) [48], and the constrained local models (CLM) [8]. Note that the CED curves here are calculated from the normalized mean alignment errors over each landmark, but not over the whole 17 landmarks. The CED curve of the SDM method on the BioID dataset is not reported in [12], we include it here based our implementation. The CED curves for the other five methods are all from the results reported in [10].

As shown in Fig. 8, most of the methods performs quite well in this relatively easy dataset. For many applications based on face alignment, the acceptable normalized location error should bellow 0.1. In this sense, all of these methods obtain more 90 percent success rate. The SRD model is among the best methods and obtains slightly better results than previous best method ESR [10] in this dataset. By training the SRD model hierarchically, our newly proposed HSRD model obtains more than 7 percent improvement over other methods at the 0.05 point for normalized location error, which will be very useful for applications that require high alignment accuracy, e.g., face synthesis.

TABLE 1
Mean Alignment Errors on the LFPW Dataset

Algorithm	CEM	OSDM	SDM	ESR	SRD	HSRD
ME ($\times 10^{-2}$)	3.99	3.47	3.49	3.43	3.24	**2.95**



(a) BioID dataset (b) LFPW dataset (c) MVFW dataset

Fig. 9. 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.

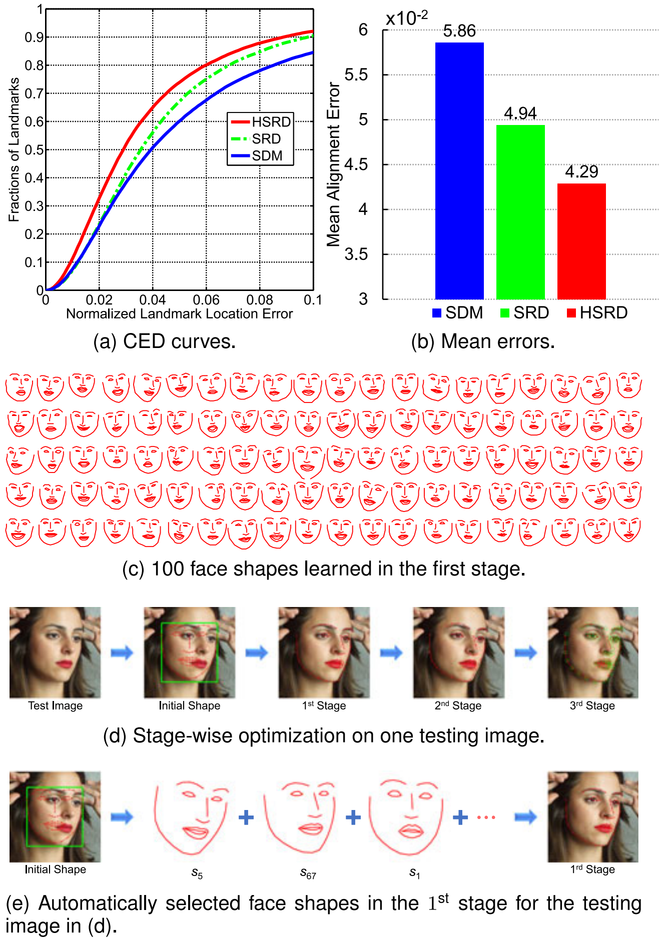
In Table 1, we report the performance of our algorithm on the LFPW dataset in terms of mean alignment error, and compare it with other state-of-the-art methods reported on this dataset. Here we use the OSDM acronym to denote the corresponding mean alignment error obtained from the reported result in the original SDM paper [12]. Again, our proposed two models, SRD and HSRD, achieve the best performances on this more challenging dataset, with about 14 and 6 percent improvements over the best baseline model ESR, respectively. From Table 1, it is observed that the SDM model implemented by ourselves has comparable performance to the one reported in [12]. The slight difference on the alignment performance may be caused by the different constitutions of the training dataset.

7.3 Multi-View Face Alignment Evaluation

As we can see from the previous experiments, most existing methods perform quite well on the BioID dataset, since all the faces in it are captured from near-frontal views. The LFPW dataset, though more challenging than the BioID dataset, is still dominated by faces with small view variations. In Fig. 9, we verify these claims by plotting the view distributions of these datasets. 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. The face views from both BioID (Fig. 9a) and LFPW (Fig. 9b) 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 better evaluate multi-view face alignment, a dataset with face views distributed evenly in a large range is needed. To this end, we have build a new dataset collected from multi-view faces in the wild [39]. This new dataset, denoted as MVFW, contains 2050 training samples and 450 testing samples, with annotations of 68 landmarks (Fig. 4d). Fig. 9c plots its view distribution, from which we can see that it is well-proportioned in all the three face view angles.

We train the SDM model [12], our SRD and HSRD models on the training set of the MVFW dataset under the same settings. Then we evaluate these models on the testing set of MVFW. In Figs. 10a and 10b, we plot the CED curves and the mean alignment errors of these models on the testing set. The CED curves in Fig. 10a show that the SRD model produces more robust alignment results over the SDM model (cf. the alignment error between 0.02 and 0.1). The HSRD further improves the alignment accuracy by a very large margin over the other two models. In respect to the mean alignment



(a) CED curves.

(b) Mean errors.

(c) 100 face shapes learned in the first stage.

(d) Stage-wise optimization on one testing image.

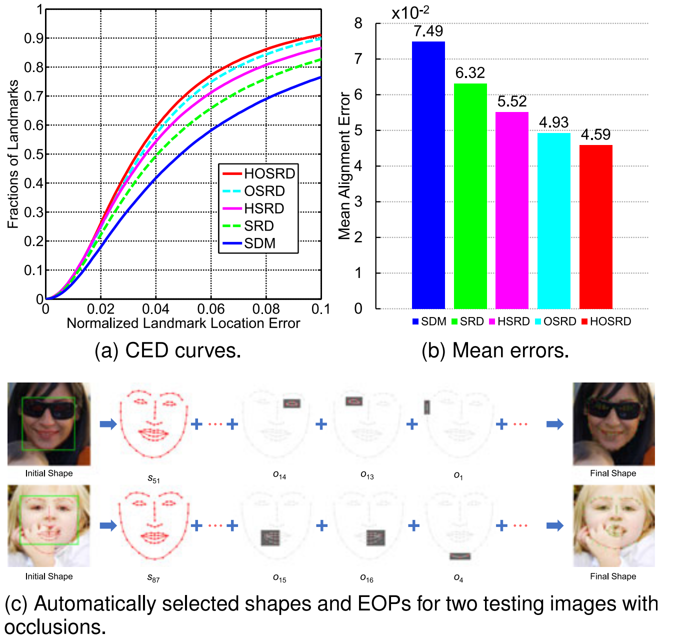
(e) Automatically selected face shapes in the 1st stage for the testing image in (d).

Fig. 10. Evaluation and analyses of multi-view face alignment on our MVFW dataset.

error, the HSRD model, on average, obtains an over 25 percent improvement for the multi-view face alignment over the SDM model. It therefore has great potential for practical applications built on multi-view face alignment.

Fig. 10c shows the 100 learned face shapes from the first stage in the relational dictionary. These face shapes are plotted by adding the shape bases to the mean face shape. These face shapes exhibit very typical face view modes from in-plane roll, out-plane pitch and yaw. The proposed model successfully captures these shape variations and simultaneously models the complex relationships between these face shapes and their appearances. By performing sparse representation using these face shapes and their corresponding appearances, it can well model face images undergo very large view variations.

To better understand the effectiveness of the stage-wise optimization process, we plot in Fig. 10d the intermediate results of the three stages when testing one sample. The alignment result quickly converges to the ground-truth landmarks from a mean face shape provided by a face detector in a stage-wised manner. 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 Fig. 10e, we visualize the sparse representation in the first stage by drawing the shape bases with the three largest coefficient. Those selected face shape bases are very relevant to the shape of the testing face.



(a) CED curves.

(b) Mean errors.

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

Fig. 11. Evaluation and analyses of partially-occluded face alignment on our OCFW dataset.

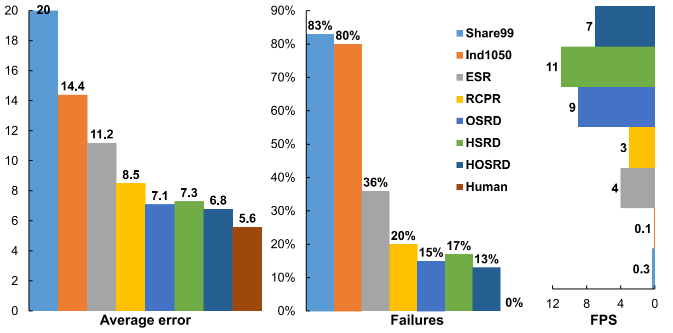


Fig. 12. Evaluation of occluded face alignment on the COFW dataset.

7.4 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 have also built a new dataset to evaluate the occlusion handling ability of a face alignment algorithm for occluded faces in the wild [39]. This new dataset, denoted as OCFW, contains 2,591 training samples and 1,246 testing samples. The objective of this dataset is to evaluate the generalization ability of an alignment model to deal with occlusions when trained on samples without occlusions. Its training set, therefore, only contains face samples without occlusions and its testing set, on the contrary, contains face samples with obvious occlusions. This dataset composition poses great challenges to a face alignment algorithm. To facilitate further studies on multi-view and occluded face alignment, we have made the MVFW and OCFW datasets publicly available.[1]

In Figs. 11a and 11b, we respectively plot the CED curves and mean alignment errors of different models over the OCFW dataset. The OSRD model and its hierarchical version HOSRD obtain much better results than other three

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

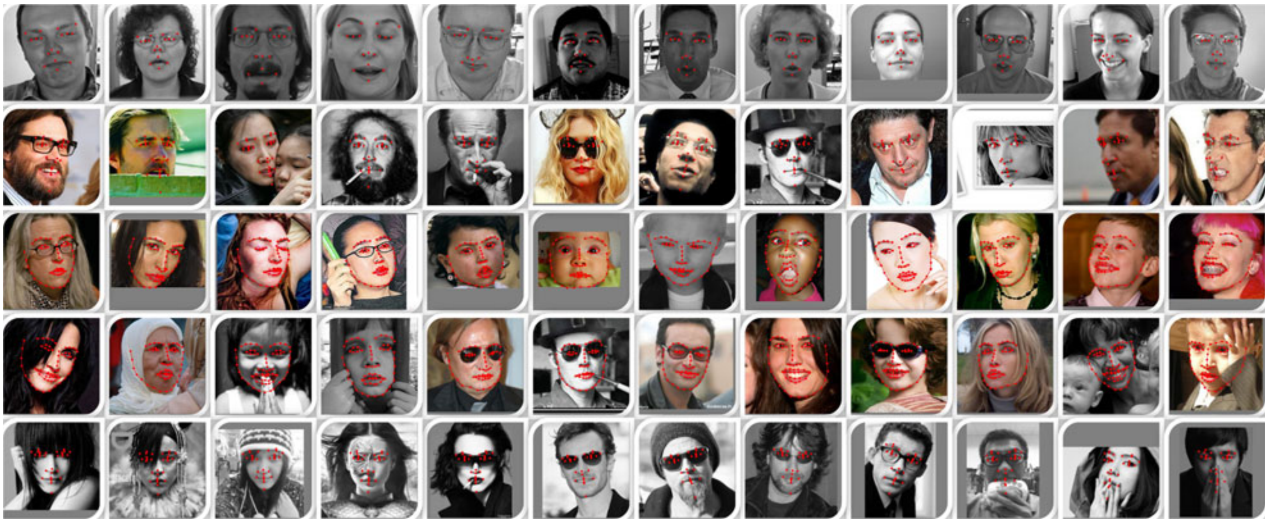


Fig. 13. Exemplar face alignment results. Top row: General face alignment on the BioID dataset. Second row: General face alignment on the LFPW dataset. Middle row: Multi-view face alignment on the MVFW dataset. Fourth row: partially-occluded face alignment on the OCFW dataset. Bottom row: partially-occluded face alignment on the COFW dataset.

models and reduces the mean alignment error by more than 30 percent compared with the SDM model. These results demonstrate their good generalization ability to deal with occlusions. To illustrate how occlusion is handled by the OSRD model, Fig. 11c plots the automatically selected face shape bases and the corresponding EOPs in the occlusion dictionary for two typical occluded face images. It shows that most of the occluded landmarks are successfully detected by the occlusion dictionary using its bases learned from the EOPs in natural face images.

To further verify the occlusion handling ability of our model, we also evaluate it on the challenging COFW dataset [22]. This dataset is labeled with 29 landmarks similar to Fig. 4b as well as the occlusion state of each landmark. The RCPR model [22] explicitly uses these annotations of occlusion information and produces very good results for occluded face alignment. Since both of the training set (1,345 samples) and the testing set (507 samples) contain severe occlusions, it is not feasible to train our OSRD model, which is designed to train on face samples without occlusions. We therefore train the OSRD model and its hierarchical version HOSRD model on the training set of OCFW and test it on the testing set of COFW. To make a comprehensive comparison, we also train our HSRD model on the COFW dataset.

In Fig. 12, we report the evaluation results in terms of average error, failures and speed, as defined in [22]. We also include the results of some state-of-the-art alignment algorithms, i.e., Share09 [11], Ind1050 [11], and ESR [10], reported by [22]. Fig. 12 shows that, without using the annotations of occlusion information, the HSRD model obtains better results than other competing models. The OSRD model and the HOSRD model further improve the performance trained even using non-occluded samples. The HSRD model, the OSRD model, and the HOSRD model are also the three fastest algorithms among the seven algorithms.

Finally, to give a qualitative view of the results obtained by our alignment model, in Fig. 13, we plot some alignment results using our model on challenging examples with large variations in face view, expression, illumination, scale, and occlusion.

8 Conclusion and Future Work

We have presented a novel model for multi-view and partially-occluded face alignment. By formulating face alignment as a dictionary learning and sparse coding problem, it provides a new framework to perform simultaneous appearance-shape modeling for face alignment. By learning the dictionary in a stage-wise and hierarchical manner, the proposed HSRD model guarantees robust and accurate face alignment results for faces with both exaggerated global and local view variations. By modeling the partial face occlusions in a data driven manner, the proposed OSRD model deals with partial face occlusions quite efficiently and effectively. Extensive experiments have demonstrated the state-of-the-art alignment performance of our model, especially for multi-view and partially-occluded faces.

Currently, the hierarchical SRD model is learned by semantically defined hierarchy of facial landmarks. In future work, we intend to further improve its performance by automatically learning this hierarchy. We also plan to perform some formal proofs and deeper analyses on the proposed model, e.g., its theoretical convergence properties and comparison of different optimization procedures. Finally, we will try to apply the proposed model to other vision problems like human pose estimation and object tracking.

Acknowledgments

The authors would like to thank the associate editor and the reviewers for their constructive comments and suggestions. This work is partly supported by the Natural Science Foundation of China (Grant No. 61672519, 61303178, 61472421, and U1636218), the 973 basic research program of China (Grant No. 2014CB349303), and the Strategic Priority Research Program of the CAS (Grant No. XDB02070003).

References

[1] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[2] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade, "Tracking in low frame rate video: A cascade particle filter with discriminative observers of different life spans," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1728–1740, Oct. 2008.

[3] S. Li, J. Xing, Z. Niu, S. Shan, and S. Yan, "Shape driven kernel adaptation in convolutional neural network for robust facial trait recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 222–230.

[4] 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.

[5] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surveys*, vol. 35, no. 4, pp. 399–459, Dec. 2003.

[6] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[7] S. Yan, X. Hou, S. Z. Li, H. Zhang, and Q. Cheng, "Face alignment using view-based direct appearance models," *Int. J. Imag. Syst. Tech.*, vol. 13, no. 1, pp. 106–112, Jun. 2003.

[8] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognit.*, vol. 41, no. 10, pp. 3054–3067, Oct. 2008.

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

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

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

[12] X. Xiong and F. Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 532–539.

[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*, 2013, pp. 397–403.

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

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

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

[17] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1867–1874.

[18] G. Tzimiropoulos and M. Pantic, "Gauss-Newton constrained local models for face alignment in-the-wild," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1851–1858.

[19] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1859–1866.

[20] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 109–122.

[21] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine autoencoder networks (CFAN) for real-time face alignment," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 1–16.

[22] X. Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1513–1520.

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

[24] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 135–164, Nov. 2004.

[25] T. Cootes, M. 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.

[26] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. J. Comput. Vis.*, vol. 91, no. 2, pp. 200–215, Jan. 2011.

[27] D. Cristinacce and T. Cootes, "Boosted regression active shape models," in *Proc. British Mach. Vis. Conf.*, 2007, pp. 880–889.

[28] J. Saragih and R. Goecke, "A nonlinear discriminative approach to AAM fitting," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

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

[30] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2729–2736.

[31] B. Smith and L. Zhang, "Joint face alignment with non-parametric shape models," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 43–56.

[32] X. Zhao, X. Chai, and S. Shan, "Joint face alignment: Rescue bad alignments with good ones by regularized re-fitting," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 616–630.

[33] J. Shen, S. Zafeiriou, G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 1003–1011.

[34] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Mar. 2010.

[35] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. Series B-Statist. Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[36] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vis. and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010.

[37] 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.

[38] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, Mar. 2011.

[39] J. Xing, Z. Niu, J. Huang, W. Hu, and S. Yan, "Towards multi-view and partially-occluded face alignment," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1829–1836.

[40] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[41] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 33, no. 11, pp. 2259–2272, Nov. 2011.

[42] BioID dataset. (2011). [Online]. Available: https://www.bioid.com/About/BioID-Face-Database

[43] 300 Faces In-The-Wild Challenge. (2013). [Online]. Available: https://ibug.doc.ic.ac.uk/resources/300-W/

[44] V. Le, J. Brandt, Z. Lin, L. Boudev, and T. Huang, "Interactive facial feature localization," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 679–692.

[45] K. Messer, J. Matas, J. Kittler, J. Luttin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. Int. Conf. Audio Video-Based Biometric Person Authentication*, 1999, pp. 72–77.

[46] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[47] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

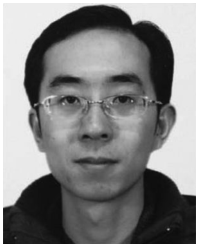
[48] S. Milborrow and F. Nicolls, "Locating facial features with an extended active shape model," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 504–513.

Junliang Xing received the dual BS degrees in computer science and mathematics from Xi'an Jiaotong University, Shaanxi, China, in 2007, and the PhD degree in computer science from Tsinghua University, Beijing, China, in 2012. He is currently an associate professor in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision problems related to faces and humans. He is a member of the IEEE.

Zhiheng Niu received the bachelor's, master's, and doctor degrees from the Harbin Institute of Technology, in 2003, 2005, and 2009, respectively. He was a senior R&D engineer in the Panasonic Research and Development Center Singapore from 2009 to 2013. He was a senior research fellow in the ECE Department, National University of Singapore from 2013 to 2015. He is currently a senior engineer in the Delphi Deutschland GMBH, Germany. His research interests include machine learning and pattern recognition.

Junshi Huang received the bachelor's degree from the Beijing Institute of Technology, in 2011 and the PhD degree from the National University of Singapore, in 2015. He is currently a researcher of the AI Institute, Qihoo/360 Company. His research interests include object detection, image annotation, and image retrieval.

Weiming Hu received the PhD degree from the Department of Computer Science and Engineering, Zhejiang University, Zhejiang, China. Since 1998, he has been in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, where he is currently a professor. He has published more than 200 papers on peer reviewed international conferences and journals. His current research interests include visual motion analysis and recognition of harmful Internet multimedia. He is a senior member of the IEEE.



Xi Zhou received the BS and MS degrees from the University of Science and Technology of China, Hefei, China, and the PhD degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana, in 2010. He is currently a full professor in the Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, China. His research interests include computer vision and multimedia. He is a senior member of the IEEE.

Shuicheng Yan is the chief scientist of Qihoo/360, the director of 360 AI Institute, and the former director of Learning and Vision Research Group, National University of Singapore. His major research areas include computer vision, machine learning, and multimedia analysis, with more than 25,000 citations and H-index 70. He is TR Highly-cited researcher 2014, 2015, and 2016, and his team won seven times of winner or honorable-mention prizes in PASCAL VOC and ILSVRC challenges, as well as more than 10 times of best (student) paper awards. He is a fellow of the IEEE, the IAPR, and the ACM Distinguished Scientist.