Noname manuscript No.
(will be inserted by the editor)

Discriminative Deep Face Shape Model for Facial Point Detection

Yue Wu · Qiang Ji

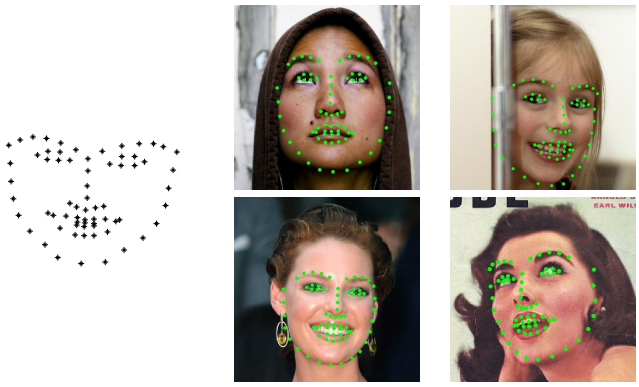
Received: 9 Feb. 2014 / Accepted: Oct. 2014

Abstract Facial point detection is an active area in computer vision due to its relevance to many applications. It is a nontrivial task, since facial shapes vary significantly with facial expression, pose or occlusion. In this paper, we address this problem by proposing a discriminative deep face shape model that is constructed based on an augmented factorized three-way Restricted Boltzmann Machines model (RBM). Specifically, the discriminative deep model combines the top-down information from the embedded face shape patterns and the bottom up measurements from local point detectors in a unified framework. In addition, along with the model, effective algorithms are proposed to perform model learning and to infer the true facial point locations from their measurements. Based on the discriminative deep face shape model, 68 facial points are detected on facial images in both controlled and "in-the-wild" conditions. Experiments on benchmark data sets show the effectiveness of the proposed facial point detection algorithm against state-of-the-art methods.

Keywords Facial point detection · Restricted Boltzmann Machine · Deep learning

Yue Wu
Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180-3590
E-mail: wuy9@rpi.edu

Qiang Ji
Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180-3590
E-mail: jiq@rpi.edu



(a) (b)

Fig. 1 Facial point detection. (a) Facial points that define the face shapes. (b) Facial images with detected facial poinits.

1 Introduction

The facial points refer to a few salient landmark points around facial components, such as eyes, mouth, and the face contour (Figure 1). They serve as the anchors on face and provide important information for face analysis. The detection of the facial point is crucial, due to its relevance to many vision applications like human head pose estimation, facial expression recognition, and face recognition.

Facial point detection algorithms usually rely on the distinct human face shape (Figure 1(a)). The face shape refers to the unique patters of the spatial relationships among facial points, which are usually used to either constrain the searching areas or correct the estimations from the image appearance. However, modeling the face shapes is challenging, since they can vary significantly with subjects, poses and facial expressions.

The existing works (Cristinacce and Cootes, 2008; Zhu and Ramanan, 2012; Saragih et al, 2011; Martinez

et al, 2013; Valstar et al, 2010) usually construct generative model to capture the face shape variations as prior and combine it with the measurements for facial point detection. However, there are some shortcomings about the existing works. First, combining the generative face shape prior model with the measurements requires extra efforts and it may rely on some strong assumptions (e.g. iid assumptions about the parameters of active shape model). Second, the existing models are only powerful in capturing the local spatial relationship among sets of facial points (Martinez et al, 2013; Valstar et al, 2010; Zhu and Ramanan, 2012), but the spatial relationship among all facial points is high order and global. Third, even though there are some works (Cristinacce and Cootes, 2008; Saragih et al, 2011) that capture the global spatial relationship among all points, their models are linear which may not be effective when the face shape variation is large due to different facial expressions and poses. To overcome the limitations of the existing methods, in this paper, we propose the novel discriminative deep face shape model which is non-linear, higher order, and directly combines the prior and measurement in a unified model.

The discriminative deep face shape model proposed in this paper captures the significant face shape variations under varying facial expressions and poses for different subjects, and it is constructed based on the factorized three-way Restricted Boltzmann Machines (RBM) model (Memisevic and Hinton, 2010). It combines top-down information from the face shape patterns, and the bottom-up measurements from the local facial point detectors in a unified framework. Along with the model, we propose the effective algorithms to perform model learning and to infer the true facial point locations given their measurements. Finally, we demonstrate that the facial point detection algorithm based on the deep face shape model performs well on images in both controlled and "in-the-wild" conditions against the state-of-the-art approaches.

The remainder of this work is organized as follows. In section 2, we review the related work. In section 3, we will introduce the proposed discriminative deep face shape model. We will discuss how to use the face shape model for facial point detection in section 4. A comprehensive evaluation of the proposed method, including the comparison with state-of-the-art works is provided in section 5. We then conclude the paper in section 6.

2 Related work

2.1 Facial point detection

The facial point detection methods can be classified into two major categories: the holistic methods, and the constrained local methods. The holistic methods predict the facial point locations from the whole facial images. The constrained local methods usually have two separate parts, including the local point detectors that generate the initial measurements of the facial point locations from the images, and a face shape model that constrains the results based on the embedded face shape patterns. This research follows the framework of the Constrained Local methods.

Holistic methods: the classic holistic method is the Active Appearance Model (AAM) (Cootes et al, 2001). It is a statistical model that fits the facial images with a small number of parameters, controlling the appearance and shape variations. Over the past decades, researches focused on estimating the model parameters, either with a regression method or in a least-square sense. Typical methods include the project-out inverse compositional algorithm (POIC) (Matthews and Baker, 2004), the simultaneous inverse compositional algorithm (SIC) (Baker et al, 2002), and the fast SIC algorithm (Tzimiropoulos and Pantic, 2013).

Besides the AAM methods, recently, more sophisticated models are proposed. For example, in (Sun et al, 2013) and (Zhou et al, 2013), a set of deep convolutional networks are constructed in a cascade manner to detect the facial points. In (Xiong and De la Torre Frade, 2013), face alignment is formulated as a nonlinear least-square problem and robust supervised descent method is proposed to fit the facial images.

Constrained local methods: the constrained local methods decouple the information from the facial appearance and shape by explicitly building the response maps from the independent local patch-based point detectors, and the shape model, which are then combined together to predict the locations of the facial points.

There are some works that focus on improving the face shape model. The early work is the Active Shape Model (ASM) (Cootes et al, 1995; Cristinacce and Cootes, 2008; Saragih et al, 2011), which represents the shape variations in a linear subspace based on the Principle Component Analysis. More recently, in (Valstar et al, 2010; Martinez et al, 2013), a generative shape model embedded in the Markov Random Field is proposed, which captures the spatial relationships among sets of points and then combines them together. In (Zhu and Ramanan, 2012), Zhu and Ramanan proposed a method

that simultaneously performs face detection, pose estimation, and landmark localization (FPLL). It builds a shape prior model based on the tree structure graphical model for each face pose. In (Le et al, 2012), two levels of ASM are constructed, one for the shape variations of each components and the other for the joint spatial relationship among facial components. In (Belhumeur et al, 2011, 2013), instead of using the parametric model, the works of Belhumeur et al. represent the face shape in a non-parametric manner with optimization strategy to fit facial images.

In this paper, our facial point detection algorithm is within the framework of the Constrained Local methods and we focus on face shape model. Our face shape model differs from the existing face shape models. First,the existing methods (Valstar et al, 2010; Martinez et al, 2013; Zhu and Ramanan, 2012) model the relationship among sets of points and then combine them together (e.g. in a tree-structure), while the proposed method jointly captures the global spatial relationship among all the facial points in a unified non-linear deep face model. Second, the tree structure face shape models in (Zhu and Ramanan, 2012) are constructed for each head pose/facial expression, and therefor rely on the good estimation of the poses/expresions. In contrast, our model directly captures the face shapes under varying facial expressions and poses. Third, while the existing face shape models are usually generative prior models (Valstar et al, 2010; Martinez et al, 2013; Cristinacce and Cootes, 2008), our model direct combines the prior and the measurements in a discriminative manner.

2.2 Restricted Boltzmann Machines based shape model

Recent works have shown the effectiveness of Restricted Boltzmann Machines and their variants in terms of representing objects' shapes. Due to the nonlinear and global nature embedded in these models, they are more suitable for capturing the variations of objects' shape, compared with the linear models like the Active Shape Model (ASM) (Cootes et al, 1995). In (Eslami et al, 2012), Eslami et al. proposed a strong method based on Boltzmann Machines to model the binary masks of objects, instead of the vertices of the key points. Specifically, they build a Deep Belief Networks(DBNs)-like model but with only locally shared weights in the first hidden layer to represent the shape of horse and motorbikes. The sampling results from the model look realistic and have a good generalization. RBM also has been applied to model human body pose. In (Taylor et al, 2010), Taylor et al. proposed to use a new prior model called implicit mixture of conditional Restricted Boltzmann Machines to capture the human poses and motions (imRBM). The mixture nature of imRBM makes it possible to learn a single model to represent the human poses and motions under different activities such as walking, running, etc. In (Kae et al, 2013), Kae et al. proposed an augmented CRF with Boltzmann machine to capture shape priors for image labeling. In (Wu et al, 2013), we proposed a generative model to capture the face shape variations as prior. It is then combined with the measurements for facial point tracking, even when subject is with significant facial expressions and poses.

There are some major difference between our deep face model and the previous shape model that are based on the Restricted Boltzmann Machine. First, the proposed shape model has a different deep structure from the previous methods and it is specifically constructed to model the face shapes under varying facial expressions and poses. Second, in contrast to our previous work (Wu et al, 2013), in which Restricted Boltzmann Machine is used to build a generative model to capture the face shape as prior, the proposed model in this paper is discriminative and it directly combines the prior with the measurements. The generative model needs more data to train and is problematic in combining with image measurements for discriminative facial landmark detection task. Third, we introduce learning algorithm to ensure the model learning with both complete and incomplete data. The whole model is learned jointly, which is in contrast to our previous work (Wu et al, 2013) where parts of the model are learned independently.

3 Deep face shape model

3.1 Problem formulation

Given the initial locations of the 68 facial points $\mathbf{m} = [m_{1,x}, m_{1,y}, ...m_{68,x}, m_{68,y}]^T$ generated using the individual facial point detectors, our goal is to infer their true locations $\mathbf{x} = [x_{1,x}, x_{1,y}, ..., x_{68,x}, x_{68,y}]^T$. This can be formed in a probabilistic formulation:

$$\mathbf{x}^* = arg\max_{\mathbf{x}} P(\mathbf{x}|\mathbf{m}) \quad (1)$$

The probability in Equation 1 captures the conditional joint probability of $\mathbf{x}$ given their measurements $\mathbf{m}$. It also embeds the face shape patterns as constraints. If we can construct a face shape model to effectively represent the probability, we can find the true facial point locations by maximizing the conditional probability. However, due to cross-subject variations, different poses and expressions, the probability is multimodal and difficult

to model. To alleviate the problem, we propose the following deep face shape model.

3.2 Model

As shown in Figure 2, the deep face shape model consists of three layers of nodes, where $\mathbf{x}$ denotes the ground truth facial point locations that we want to infer, and $\mathbf{m}$ is their measurements from the local point detectors. In the middle layer, node $\mathbf{y}$ represents the corresponding frontal face shapes of $\mathbf{x}$ with the same facial expression. Figure 4 shows a pair of corresponding images as an example. $\mathbf{y}$ can be regarded as additional data that is available for some facial images during training. In the top layer, there are two sets of binary hidden nodes, including $\mathbf{h^1}$ and $\mathbf{h^2}$.

The model captures two levels of information. The first level of information refers to the face shape patterns captured with the nodes in two top layers, including $\mathbf{x}, \mathbf{y}, \mathbf{h^1}$ and $\mathbf{h^2}$. The second level of information is the input from the measurements $\mathbf{m}$. In the model, it jointly combines the top-down information from face shape patterns and the bottom-up information from the measurements. In the following, we explain each part separately.

Top two layers: The nodes in the top two layers capture the joint spatial relationship among all facial points, as represented by $\mathbf{x}$, under varying facial expressions, poses, and for different subjects. It consists of two parts. In the right part, $\mathbf{h^2}$ represents the hidden nodes that are connected to $\mathbf{x}$, following a standard Restricted Boltzmann Machine (RBM) connection as shown in Figure 3 (c). In the left part, to better model the shape variations, we explicitly add the nodes $\mathbf{y}$ in the middle layer, which represents the corresponding frontal face shape of $\mathbf{x}$ with the same facial expression. The idea is that similar frontal face shape relates to similar non-frontal face shape through the hidden nodes $\mathbf{h^1}$, which captures the pose information. To model the joint compatibility among face shape $\mathbf{x}$ with varying poses, its corresponding frontal shape $\mathbf{y}$, and their relationships embedded in $\mathbf{h^1}$, we use the factored three-way Restricted Boltzmann Machine model (Memisevic and Hinton, 2010; Ranzato et al, 2010). Figure 3 (a)(b) show their connections in details. Each node x_i, y_j, and ${h^1}_k$ are connected to a factor f with parameters $\{W^x_{i,f}, W^y_{j,f}, W^{h^1}_{k,f}\}$. With multiple factors, the model captures the high order global relationship among $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{h^1}$. The left ($\mathbf{h^1}$ and $\mathbf{y}$) and right ($\mathbf{h^2}$) parts are complementary with each other. $\mathbf{h^1}$ would focus on the variations due to poses and $\mathbf{h^2}$ may focus on the other variations that are not related to poses (e.g. facial expressions in frontal pose).

Bottom layer: The nodes in the bottom layer model the joint compatibility among the facial feature point locations $\mathbf{x}$ and their measurements $\mathbf{m}$. The connection of $\mathbf{x}$ and $\mathbf{m}$ is shown in Figure 3 (d). We model their compatibility with the standard RBM model.

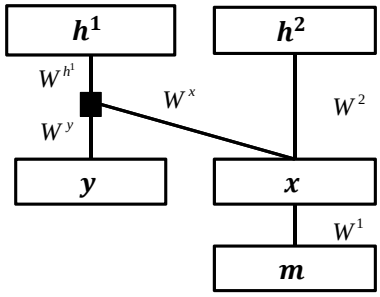


Fig. 2 The proposed discriminative deep face shape model. It consists of a factorized three-way RBM connecting nodes $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{h^1}$. It also includes two RBMs that model the connections among $\mathbf{x}$, $\mathbf{h^2}$ and $\mathbf{m}$, $\mathbf{x}$, respectively.

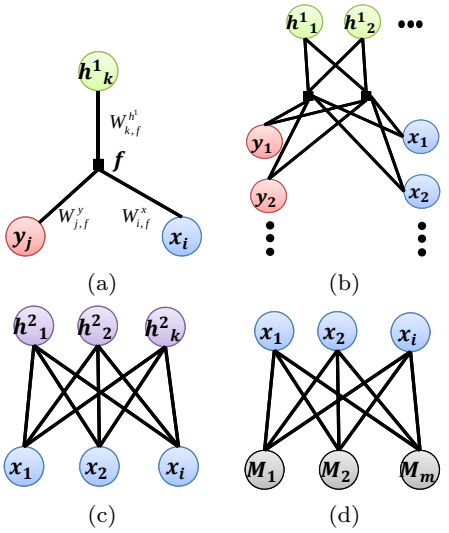
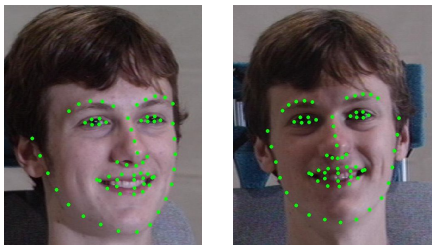


Fig. 3 Graphical depiction about different parts of the model. (a)(b) Factored three way RBM model that captures the joint compatibility among nodes $\mathbf{x}, \mathbf{y}, \mathbf{h^1}$. (c) Model connecting $\mathbf{x}$ and $\mathbf{h^2}$ (d) Model connecting $\mathbf{m}$ and $\mathbf{x}$.

Given the model structure, the joint energy function is defined in Equation 2 with model parameter $\theta = \{W^x, W^y, W^{h^1}, W^1, W^2, c^x, c^y, b^{h^1}, b^{h^2}\}$. Specifically, $W^x \in \Re^{d_x \times d_f}$, $W^y \in \Re^{d_y \times d_f}$, $W^{h^1} \in \Re^{d_{h^1} \times d_f}$ are parameters that model the compatibility of nodes $\mathbf{x}, \mathbf{y}, \mathbf{h^1}$ in the 3-way connection. Here, d_* represents the dimension of the corresponding variable. Similarly, $W^1 \in \Re^{d_m \times d_x}$ and $W^2 \in \Re^{d_x \times d_{h^2}}$ are parameters for the connections of $\mathbf{m}, \mathbf{x}$ and $\mathbf{x}, \mathbf{h^2}$, respectively. $c^x \in$



(a) (b)

Fig. 4 Corresponding images. (a)Images with poses and expressions, indicated as $\mathbf{x}$. (b) Corresponding frontal face, represented as $\mathbf{y}$.

$\Re^{d_x}$, $c^y \in \Re^{d_y}$, $b^{h^1} \in \Re^{d_{h^1}}$, and $b^{h^2} \in \Re^{d_{h^2}}$ represent the bias terms for corresponding variables.

$$
\begin{aligned}
&- E(\mathbf{x}, \mathbf{y}, \mathbf{m}, \mathbf{h^1}, \mathbf{h^2}; \theta) \\
&= \sum_f (\sum_i x_i W^x_{i,f})(\sum_j y_j W^y_{j,f})(\sum_k h^1_k W^{h^1}_{k,f}) \\
&+ \sum_{m,i} m_m W^1_{m,i} x_i + \sum_{i,l} x_i W^2_{i,l} h^2_l \\
&+ \sum_i x_i c^x_i + \sum_j y_j c^y_j + \sum_k h^1_k b^{h^1}_k + \sum_l h^2_l b^{h^2}_l
\end{aligned} \tag{2}
$$

As illustrated in Equation 1, our goal is to capture the conditional joint probability of the true facial point locations given their measurements. Therefore, instead of building a generative model, we directly model the conditional probability with a discriminative model. Specifically, the discriminative model defines the conditional joint probability $P(\mathbf{x}, \mathbf{y}|\mathbf{m}; \theta)$ and conditional probability $P(\mathbf{x}|\mathbf{m}; \theta)$ in Equation 3 and 4. $Z_m(\theta)$ is the partition function defines in Equation 5.

$$
P(\mathbf{x}, \mathbf{y}|\mathbf{m}; \theta) = \frac{\sum_{\mathbf{h^1}, \mathbf{h^2}} exp(-E(\mathbf{x}, \mathbf{y}, \mathbf{m}, \mathbf{h^1}, \mathbf{h^2}; \theta))}{Z_m(\theta)}, \tag{3}
$$

$$
P(\mathbf{x}|\mathbf{m}; \theta) = \frac{\sum_{\mathbf{y}, \mathbf{h^1}, \mathbf{h^2}} exp(-E(\mathbf{x}, \mathbf{y}, \mathbf{m}, \mathbf{h^1}, \mathbf{h^2}; \theta))}{Z_m(\theta)}, \tag{4}
$$

$$
Z_m(\theta) = \sum_{\mathbf{x}, \mathbf{y}, \mathbf{h^1}, \mathbf{h^2}} exp(-E(\mathbf{x}, \mathbf{y}, \mathbf{m}, \mathbf{h^1}, \mathbf{h^2}; \theta)) \tag{5}
$$

The conditional probabilities of one variable given the other variables and $\mathbf{m}$ are shown in Equation 6, 7, 8, and 9.

$$
\begin{aligned}
&P(x_i = 1|\mathbf{y}, \mathbf{h^1}, \mathbf{h^2}, \mathbf{m}; \theta) = \\
&\quad \sigma(\sum_f W^x_{i,f}(\sum_j y_j W^y_{j,f})(\sum_k h^1_k W^{h^1}_{k,f}) \\
&\quad + \sum_m m_m W^1_{m,i} + \sum_l W^2_{i,l} h^2_l + c^x_i)
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
&P(y_j = 1|\mathbf{x}, \mathbf{h^1}; \theta) = \\
&\quad \sigma(\sum_f W^y_{j,f}(\sum_i x_i W^x_{i,f})(\sum_k h^1_k W^{h^1}_{k,f}) + c^y_j)
\end{aligned} \tag{7}
$$

$$
\begin{aligned}
&P(h^1_k = 1|\mathbf{x}, \mathbf{y}; \theta) = \\
&\quad \sigma(\sum_f W^{h^1}_{k,f}(\sum_i x_i W^x_{i,f})(\sum_j y_j W^y_{j,f}) + b^{h^1}_k)
\end{aligned} \tag{8}
$$

$$
P(h^2_l = 1|\mathbf{x}; \theta) = \sigma(\sum_{i,l} x_i W^2_{i,l} + b^{h^2}_l) \tag{9}
$$

3.3 Learning the face shape model

We refer model learning as to jointly learn the model parameters θ given the training data. We have two sets of training data. The first set contains complete data $Data^C = \{\mathbf{x}_c, \mathbf{y}_c, \mathbf{m}_c\}_{c=1}^{N^C}$, including the face shape $\mathbf{x}$ in varying facial expressions and poses, its corresponding frontal face shape $\mathbf{y}$, and the measurements $\mathbf{m}$. The second set contains the incomplete data $Data^I = \{\mathbf{x}_i, \mathbf{m}_i\}_{i=1}^{N^I}$, in which frontal face shape $\mathbf{y}$ is not available. For the incomplete data, $\mathbf{y}$ is also hidden variable, therefore makes the proposed model (3-way connection part) a deep model. One observation from (Welling and Hinton, 2002; Hinton, 2002) is that in order for Contrastive Divergence (CD) algorithm (Hinton, 2002) to work well, it requires the exact samples from the probability of the hidden nodes given the visible data. Unfortunately this probability is intractable in deep model including the proposed model. To perform model learning, we use the following algorithm.

Given the data, the model parameters are estimated by maximizing the log conditional likelihood defined in Equation 11 and 12 with gradient ascent method.

$$
\begin{aligned}
\theta^* &= arg\max_\theta L(\theta) \\
&= arg\max_\theta L(\theta; Data^C) + L(\theta; Data^I)
\end{aligned} \tag{10}
$$

$$
L(\theta; Data^C) = \frac{1}{N^C} \sum_{c=1}^{N^C} log(P(\mathbf{x}_c, \mathbf{y}_c|\mathbf{m}_c; \theta)) \tag{11}
$$

$$
L(\theta; Data^I) = \frac{1}{N^I} \sum_{i=1}^{N^I} log(P(\mathbf{x}_i|\mathbf{m}_i; \theta)) \tag{12}
$$

The gradient of model parameters are calculated as:

$$
\frac{\partial L(\theta)}{\partial \theta} = -\langle \frac{\partial E}{\partial \theta} \rangle_{P^C_{data}} - \langle \frac{\partial E}{\partial \theta} \rangle_{P^I_{data}} + \langle \frac{\partial E}{\partial \theta} \rangle_{P_{model}}. \tag{13}
$$

It contains three terms. The first and second terms indicate the expectations over the complete data $P^C_{data} = p(\mathbf{h^1}, \mathbf{h^2}|\mathbf{x}_c, \mathbf{y}_c, \mathbf{m}_c; \theta)$, and the incomplete data $P^I_{data} = p(\mathbf{h^1}, \mathbf{h^2}, \hat{y}|\mathbf{x}_i, \mathbf{m}_i; \theta)$, respectively. The third term represents the expectation over the model for all the data $P_{model} = p(\mathbf{h^1}, \mathbf{h^2}, \tilde{\mathbf{y}}, \tilde{\mathbf{x}}|\mathbf{m}_q; \theta)$, $q \in Data^C \bigcup Data^I$. In the following, we explain how to estimate each term.

The estimation of the data dependent probability for the complete data in the first term of Equation 13 could be done through Equation 8 and 9, since hidden nodes are independent given $\mathbf{x}$ and $\mathbf{y}$. For the second term, directly estimation of the data dependent probability for the incomplete data is intractable, since only $\mathbf{x}$ is given and $\mathbf{y}, \mathbf{h^1}, \mathbf{h^2}$ are all hidden nodes. To tackle this problem, we follow (Salakhutdinov and Hinton, 2009) and pursue the variational learning method, which results in the mean-field fixed point equations as shown in Equation 14, 15, 16. In this method, we iteratively estimate and update the values of $\mathbf{y}$, $\mathbf{h^1}$ and $\mathbf{h^2}$ with their mean-field estimation results. To estimate the model expectation in the third term, we use the Persistent Markov Chains for each data based on Equation 6, 7, 8, 9. The overall model learning algorithm is shown in Algorithm 1.

$$y_j \leftarrow \sigma(\sum_f W^y_{j,f}(\sum_i x_i W^x_{i,f})(\sum_k h^1_k W^{h_1}_{k,f}) + c^y_j) \quad (14)$$

$$h^1_k \leftarrow \sigma(\sum_f W^h_{k,f}(\sum_i x_i W^x_{i,f})(\sum_j y_j W^y_{j,f}) + b^{h_1}_k) \quad (15)$$

$$h^2_l \leftarrow \sigma(\sum_{i,l} x_i W^2_{i,l} + b^{h^2}_l) \quad (16)$$

Algorithm 1 Learning the deep face shape model.

Input: Complete data $\{\mathbf{x}_c, \mathbf{y}_c, \mathbf{m}_c\}_{c=1}^{N^C}$, including the face shape in arbitrary pose and expression, its measurement, and its corresponding frontal shape with same expression. Incomplete data $\{\mathbf{x}_i, \mathbf{m}_i\}_{i=1}^{N^I}$ without the frontal face shape.
Output: Model parameters $\theta = \{W^x, W^y, W^{h^1}, W^1, W^2, c^x, c^y, b^{h^1}, b^{h^2}\}$.
Randomly initialize the parameters θ^0, and the chains for each data.
for iteration t=0, to T **do**
(a) For each complete training data $\{\mathbf{x}_c, \mathbf{y}_c, \mathbf{m}_c\} \in Data^C$
Calculate the data dependent probability through Equation 8 and 9.
Sample $\mathbf{h^1}_c$ and $\mathbf{h^2}_c$ from these probabilities.
(b) For each incomplete training data $\{\mathbf{x}_i, \mathbf{m}_i\} \in Data^I$
Randomly initialize $\hat{\mathbf{y}}_i, \mathbf{h^1}_i, \mathbf{h^2}_i$ and run mean-field updates using Equation 14, 15, 16 for a few updates until convergence.
Set $\hat{\mathbf{y}}_i, \mathbf{h^1}_i, \mathbf{h^2}_i$ as their mean-filed results
(c) For each training data $q \in Data^C \bigcup Data^I$, and its chains k=1...K.
Run the Gibbs sampler for a few steps, and get the updated states $\{\tilde{\mathbf{x}}_q^{t,k}, \tilde{\mathbf{x}}_q^{t,k}, \tilde{\mathbf{h}}_q^{1,t,k}, \tilde{\mathbf{h}}_q^{2,t,k}\}$ through the Chains.
(d) Update.
$\theta^{t+1} = \theta^t + \alpha^t(\frac{\partial L(\theta)}{\partial \theta})$. For example, $W^{2,t+1} = W^{2,t} + \alpha^t[(\sum_c^{N^C} \mathbf{x}_c^T \mathbf{h^2}_c) + (\sum_i^{N^I} \mathbf{x}_i^T \mathbf{h^2}_i) - \frac{1}{K}(\sum_q^{N^C+N^I} \sum_k (\tilde{\mathbf{x}}_q^{t,k})^T \tilde{\mathbf{h}}_q^{2,t,k})]/(N^C + N^I)$.
decrease α^t
end for

3.4 Inference of the facial point locations given their measurements in the face shape model

During testing, we infer the true facial feature locations $\mathbf{x}$ given their measurements $\mathbf{m}$ using Equation 1 and 4. In Equation 4, calculating the conditional probability involves the estimation of the partition function that sums over all the variables. The estimation is intractable if the dimensions of the variables are high. To alleviate the problem, we pursue the Gibbs Sampling method which relies on the conditional probabilities of one variable given all the other variables which are tractable in the deep model. Those conditional probabilities are defined in Equation 6, 7, 8, 9.

The overall algorithm is shown in Algorithm 2. The input is the measurements of facial point locations indicated as $\mathbf{m}$, and the model parameters θ that defines the conditional probability $P(\mathbf{x}|\mathbf{m};\theta)$. The output is the estimated true facial point locations $\mathbf{x}^*$. The algorithm starts by randomly initializing all the hidden variables in the multiple chains $c = 1, ...C$, including $\mathbf{x}$, $\mathbf{y}$, $\mathbf{h^1}$, and $\mathbf{h^2}$. Within each iteration, the Gibbs Sampler samples each variable from their conditional probability, assuming other variables are given. This involves the sampling of $\mathbf{x}$, $\mathbf{y}$, $\mathbf{h^1}$, and $\mathbf{h^2}$ in a sequential manner using Equation 6, 7, 8, 9. After that, we collect the samples to estimate $P(\mathbf{x}|\mathbf{m})$. Then, the facial point locations $\mathbf{x}^*$ are set as the values that can maximize the conditional probability.

Algorithm 2 Infer the facial point locations $\mathbf{x}$ given their measurements in the deep face shape model using Gibbs Sampling.

Input: The measurements $\mathbf{m}$ from independent point detectors, and model parameters θ that defines $P(\mathbf{x}|\mathbf{m};\theta)$
Output: The inferred facial point locations $\mathbf{x}^*$
for chain c=0, to C **do**
Randomly initialize the values of all hidden variables $\mathbf{x}, \mathbf{y}, \mathbf{h^1}$, and $\mathbf{h^2}$.
for iteration t=0, to T **do**
sample $\mathbf{x}$ given $\mathbf{m}, \mathbf{y}, \mathbf{h^1}$ and $\mathbf{h^2}$ using Equation 6.
sample $\mathbf{y}$ given $\mathbf{x}$ and $\mathbf{h^1}$ using Equation 7.
sample $\mathbf{h^1}$ and $\mathbf{h^2}$ given $\mathbf{x}$ and $\mathbf{y}$ using Equation 8 and 9.
end for
end for
For each chain, collect the last K samples of $\mathbf{x}$ and estimate $P(\mathbf{x}|\mathbf{m})$ from the samples.
Estimate $\mathbf{x}^* = arg\max_{\mathbf{x}} P(\mathbf{x}|\mathbf{m})$.

3.5 Major contributions of the proposed model

There are several important properties and benefits of the proposed model:

1. The discriminative deep face shape model combines top-down and bottom-up information in a unified framework. The top-down information refers to the face shape patterns embedded in the top two layers through multiple hidden nodes. The bottom-up information refers to the input from the measurements. More importantly, the proposed model combines the top-down and bottom-up information in a discriminative manner, and it directly model the conditional distribution $P(\mathbf{x}|\mathbf{m})$. In contrast, our previous work (Wu et al, 2013), the Active Shape Model (Cootes et al, 1995; Cristinacce and Cootes, 2008; Le et al, 2012), and some state-of-the-art works (Martinez et al, 2013; Valstar et al, 2010; Zhu and Ramanan, 2012) usually model $P(\mathbf{x})$ and learn the face shape prior during training. This generative face shape prior is then combined with the measurements during testing, although some of them may train these generative models discriminatively.
2. The non-linear deep model directly captures the high order dependencies among all 68 facial feature points, while the previous works (Martinez et al, 2013; Valstar et al, 2010; Zhu and Ramanan, 2012) usually capture the relationship among sets of points and then combine them together. In addition, the proposed unified model captures the face shape variations due to different kinds of facial expressions and poses, while the tree-based shape model in (Zhu and Ramanan, 2012) is constructed for each pose/facial expression. The negative consequence is that they need to search for the facial landmark locations under each pose (or expression) and search over poses (or expressions).
3. Besides the standard connection through RBM (connection through nodes $\mathbf{h^2}$), the model explicitly leverages the additional frontal face shapes to help the learning (nodes $\mathbf{h^1}$ and $\mathbf{y}$). The intuition is that similar frontal face $\mathbf{y}$ with similar facial expressions would lead to similar face shapes $\mathbf{x}$ with face poses, where the transition is captured through different hidden nodes in $\mathbf{h^1}$. In this case, the proposed model is specifically designed to model the face shape patterns under varying facial expressions and poses, and it is different from the existing RBM based shape model.
4. The model is learned with both complete and incomplete data which is different from our previous work (Wu et al, 2013). Due to the existent of the incomplete data, the model is deep. Therefor, Contrastive Divergence (CD) (Hinton, 2002) is not effective, since it requires the exact estimate of the probability of hidden nodes (Welling and Hinton, 2002; Hinton, 2002). To tackle this problem, we propose an effective learning method. We jointly learn the whole model in this work, while our previous work (Wu et al, 2013) learns sperate parts of the model and combines them together.

4 Facial point detection using the face shape model

4.1 Facial point detection algorithm

Our facial point detection algorithm is illustrated in Figure 5. The algorithm starts with face detection. Within the bounding box, the algorithm generates the measurements from the independent local point detectors, and constrains the results through the discriminative deep face shape model iteratively. In each iteration, the local point detectors (will be discussed in Section 4.2) search each facial point independently (e.g. eye corner as in (b)). As shown in (c), the results from independent point detectors are treated as measurements which are then constrained and refined through the face shape model (discussed in Section 3.4). Finally, the algorithm outputs the locations of the facial points as shown in (d).

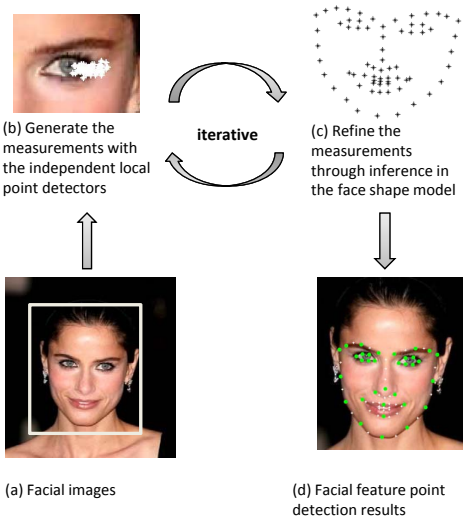


Fig. 5 Diagram illustration of the facial point detection algorithm. (a)The algorithm starts with face detection. The local point detection and face shape constrains are performed iteratively in (b)(c). (d) The detected facial points.

4.2 Local point detector

An integral part of our facial point detection algorithm is the facial point detectors. It detects each facial point independently based on the local image patches. The detector results are served as the input to the face shape model, and used to infer the true facial point locations. In section 5.3, we evaluate different local point detectors and the combinations of those local point detectors with the proposed discriminative deep face shape model.

5 Experimental results

5.1 Data sets

In this experiments, we used four benchmark databases and some sample images can be found in Figure 13.

CMU Multi-PIE face (MultiPIE) database: The MultiPIE database (Gross et al, 2010) is collected in controlled conditions with lab environment. It contains 337 subjects, imaged under 15 view points and 19 illumination conditions with six expressions, including neutral, smile, surprise, squint, disgust, and scream.

Helen dataset: The Helen dataset (Le et al, 2012) contains high quality "in-the-wild" facial images searched from the website with a variety of keywords which leads to facial images with arbitrary expressions, poses, illuminations, etc. There are 2000 images in the training set and 330 in the testing set.

Labeled Face Parts in the Wild (LFPW): The LFPW dataset (Belhumeur et al, 2011) contains "in-the-wild" facial images collected from the website with significant variations. Due to broken URLs, we are only able to download 800 images as training set, and 224 images as testing set.

Annotated Face in the Wild (AFW) database: The AFW database (Zhu and Ramanan, 2012) contains 205 "in-the-wild" images with 468 faces collected from Flickr. Images contain cluttered backgrounds with arbitrary viewpoints, expressions and appearances.

5.2 Implementation details

Training and testing data: For training, we used the training sets from Helen and LFPW datasets and the images from the first 200 subjects in MultiPIE database (2500 images). There are four testing sets, including the provided testing sets from Helen and LFPW, the AFW database, and our collected testing set from MultiPIE. The testing set from MultiPIE includes images of the remaining subjects (id 201-337) not used during training (1054 images). To ensure fair comparison with state-of-the-art approach (Martinez et al, 2013), we also choose images with no larger than 30 degree pose angles and arbitrary illuminations and expressions. The facial point annotations are provided either by the database or by the iBug group (Sagonas et al, 2013b,a). Since the facial images are with great scale change, we normalize the images based on the size of the face bounding box.

Deep face shape model: In the deep face shape model, the number of hidden nodes $\mathbf{h^1}$, factors and hidden nodes $\mathbf{h^2}$ are 100, 64 and 300 respectively, determined from the training data. Following (Salakhutdinov and Hinton, 2009), we preprocess the continuous data with the Gaussian-binary RBM (GRBM) model (Hinton and Salakhutdinov, 2006; Mohamed et al, 2011) and then treat the values of the hidden layers as the preprocessed data to speed up the learning process. Before the input to the GRBM, the training data is first normalized so that the standard deviation equals to 1. In this case, the sigma of GRBM is fixed and we only need to learn the mean.

Evaluation criteria: We measure the error using the displacement w.r.t the ground truth normalized by the inter-ocular distance in Equation 17.

$$error = \frac{||D - L||_2}{||L_{leye} - L_{reye}||_2}, \tag{17}$$

where D and L represent the detected and labeled facial point locations. L_{leye} and L_{reye} indicate the locations of left and right eyes, respectively. Through out the experiments, we detect 68 facial landmarks, but there are two sets of points for evaluations. The first set only includes 17 interior points defined in (Cristinacce and Cootes, 2008). The second set includes all 68 points. We indicate the error measurements for these two sets as me_{17} and me_{68}, respectively.

5.3 Results

The proposed algorithm consists of the local point detectors and the discriminative deep face shape model. In this section, we evaluate the varying local point detectors, the proposed face shape model, their different combinations, and the comparison of the proposed method with the state-of-the-art works in terms of accuracy and computational efficiency.

5.3.1 Local point detectors

For each facial landmark, the local point detector scans a region and classifies the patch at each pixel location. It consists of a feature descriptor to encode the information of each image patch and a classifier. In our experiments, we evaluate a few feature descriptors and

classifiers. They are (1) Scale Invariant Feature Transform (SIFT) features (Lowe, 2004) + L2 regularized Logistic Regression (LR) (Fan et al, 2008) classifier, (2) Histograms of Oriented Gradients (HOG) (Dalal and Triggs, 2005) features + LR, (3) learned features using Deep Boltzmann Machine (DBM) (Salakhutdinov and Hinton, 2009) + LR, (4) learned features using DBM + Neutral Network as described in (Salakhutdinov and Hinton, 2009), and (5) the reported local detectors and their results in (Belhumeur et al, 2011, 2013) which use two scale SIFT features with Support Vector Regressor (SVR) (Smola and Schölkopf, 2004). We cut the local image patch centered at the ground truth point locations as positive data and the negative patches are at least 1/4 the inter-ocular distance away from the ground truth. The patch height and width are about 1/4 the inter-ocular distance and all the patches are then normalized to 25*25. To learn the features, we use two layers DBM with 1000 and 800 hidden nodes in the first and second layers. Following (Salakhutdinov and Hinton, 2009), we preprocess the continuous data using the Gaussian-binary RBM model (Hinton and Salakhutdinov, 2006; Mohamed et al, 2011) with 1000 hidden nodes and then treat the values of the hidden layers as the preprocessed data to speed up the learning process. We test all the local point detectors on the LFPW database and measure the errors on the 17 points.

The experimental results are shown in Figure 6. Comparing (1)(2)(3), we see that SIFT, HOG, and the learned features perform similarly, where HOG is slightly better. Comparing (3)(4), we see that Logistic Regression and Neutral Network perform similarly with the same learned features using DBM. Our implementations including (1)-(4) achieve similar performance as the reported results (5) on the same testing database with same evaluation criteria.

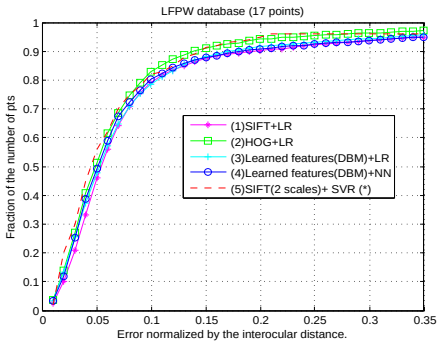
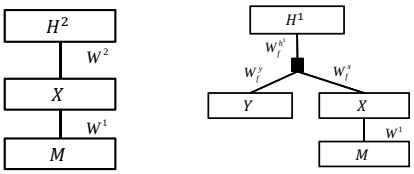


Fig. 6 Comparing local point detectors with different feature descriptors and classifiers. They are (1) SIFT + Logistic Regression (LR), (2) HOG + LR, (3) Learned features using Deep Boltzmann Machine (DBM) + LR, (4) Learned features using DBM + Neutral Network (NN), and (5) the reported local detectors (Belhumeur et al, 2011, 2013) which use two-scales SIFT features with Support Vector Regressor (SVR).



(a) CRBM (b) Conditional 3-way RBM

Fig. 7 Different variations of the proposed model.(a) Model only with $\mathbf{h^2}$ (CRBM). (b) Model only with $\mathbf{y}$ and $\mathbf{h^1}$ (Conditional 3-way RBM).

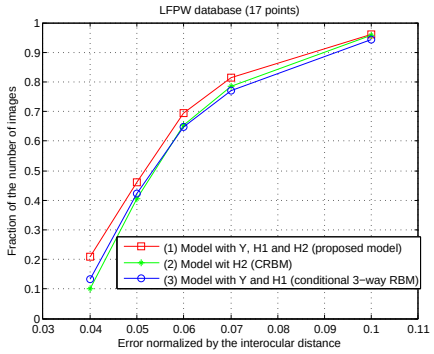


Fig. 8 Comparing different variations of the proposed model.

5.3.2 Face shape model

In this section, we evaluate the face shape model. First, we compare the proposed discriminative deep face shape model as described in section 3 and Figure 2, with its variations shown in Figure 7 using the same local point detectors (HOG+LR) on the LFPW database. Specifically, by excluding the 3-way connection in the proposed model, we get the model shown in Figure 7 (a), which can be regarded as a conditional RBM model. By excluding the hidden nodes $\mathbf{h^2}$ in the original proposed model, we get the model shown in Figure 7 (b), which can be considered as a conditional 3-way RBM model. The connections and parameterizations are the same as the proposed model. There are 500 hidden nodes in the CRBM model, 300 hidden nodes and 128 factors in the conditional 3-way RBM model. Both models learn the conditional probability as the proposed model. To train the conditional 3-way RBM model, we use both complete and incomplete data with the proposed learning method. For the CRBM training, the proposed learning algorithm becomes similar as the persistent CD algorithm (Tieleman, 2008). As shown in Figure 8, the proposed model with both parts performs better than CRBM and conditional 3-way RBM models.

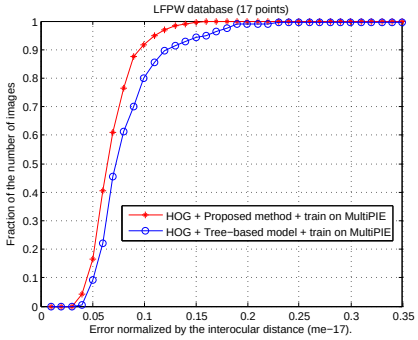


Fig. 9 Comparing the proposed discriminative deep face shape model with the tree-based model of FPLL method (Zhu and Ramanan, 2012) on the LFPW database.

Second, we compare the proposed method to the FPLL method in (Zhu and Ramanan, 2012) with tree-based face shape model. In this experiment, to ensure fair comparison and focus on the face shape model, both methods are utilizing the same HOG feature descriptor, retrained on the same MultiPIE database and tested on the LFPW database. We use the public available code from the authors. It is important to notice that this setting is biased in favor of the FPLL method. Specifically, we restrict the training database to MultiPIE, since FPLL requires the fully supervised data with facial landmark, facial expression and head pose annotations, while our model does not have this requirement. In addition, in the original work, tree-based face shape models are constructed for each head pose and facial expression while our model encodes the face shape variations of varying head poses and facial expressions. Even with this advantage, the tree-based method performs worse than the proposed method as shown in Figure 9.

5.3.3 Facial point detection algorithm

In this section, we evaluate the overall facial point detection algorithm with the proposed discriminative deep face shape model.

First, we visualize the performance of the facial point detection algorithm in different iterations on one sample image. As shown in Figure 10 (a), the initial estimations of the points from the local point detectors are poor, especially for the contour points. Given the poor initial measurements, the algorithm constrains their locations through the deep face shape model, leading to the results shown in Figure 10 (b). As shown in Figure 6 (c)(d)(e), the algorithm iteratively achieves reasonable results even at iteration 3, and the error converges quickly after 8 iterations. In practice, we stop the algorithm if the facial landmark locations in two consecutive iterations do not change.

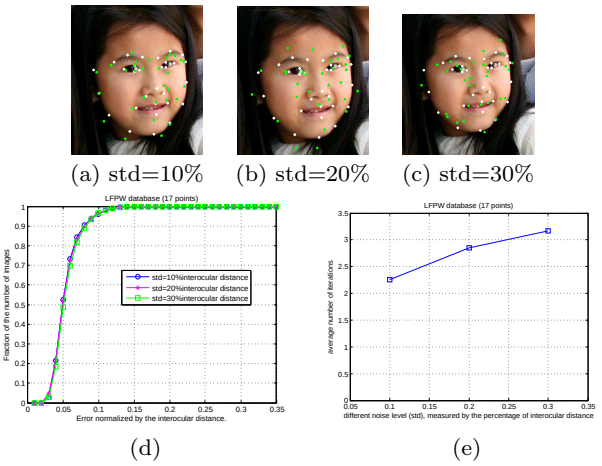
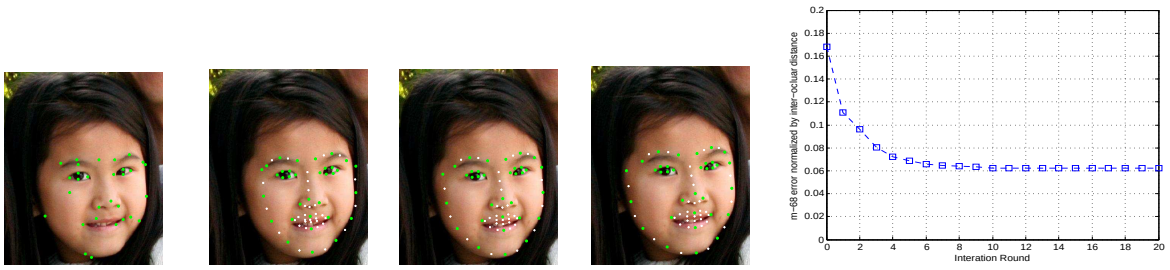


Fig. 11 Evaluate the robustness of the algorithm with different initial detection accuracies on the LFPW database (Better see in color). (a)(b)(c) show the synthesized initial detections results (green dots) by perturbing the ground truth (white stars) with different Gaussian noises (std=10%, 20%, 30% of the inter-ocular distance). (d)The facial point detection results on LFPW with different noise levels. (e) Number of iterations required for convergence with different noise levels.

Second, we evaluate the robustness of the algorithm on different initial detection accuracies on the LFPW database using the proposed method (HOG+LR as local point detectors). Specifically, we generate synthetic initial detection results by adding different levels of Gaussian noise to the ground truth locations. The standard derivations of the Gaussian noise equal to 10%, 20%, and 30% of the inter-ocular distance, respectively. Figure 11(a)-(c) show different initializations on one sample image. As shown in Figure 11(d)(e), the algorithm achieves almost identical results with a slightly increase of the number of iterations for large noise level.

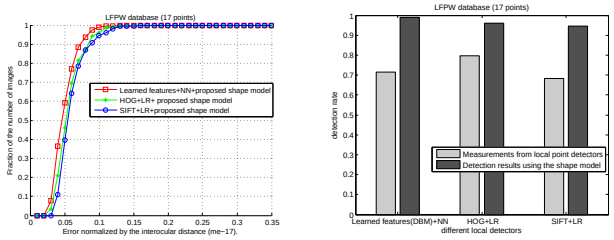
Third, we evaluate different local point detectors described in section 5.3.1 with the proposed discriminative deep face shape model on LFPW database. As can be seen from Figure 12 (a), the combination of learned features (DBM), Neural Network as local point detectors, and the proposed shape model achieves the best results. As in Figure 12 (b), comparing to the measurements from local point detectors, the detection rates of the algorithm with the proposed discriminative deep face shape model increase dramatically for all different combinations.

Forth, we evaluate the algorithm on all four benchmark databases, including the MultPIE, Helen, LFPW and AFW databases. In this experiment and the experiments below we use the learned features and the Neural Network as local point detectors since they perform the best on the LFPW database as shown in the last experiment. Figure 13 shows our detection results on



(a) Measurements (b) Iteration 1 (c) Iteration 3 (d) Iteration 20 (e) results

Fig. 10 Facial point detection errors for different iterations on one sample image. (a) measurements from local point detectors. (b)(c)(d) show results for different iterations. (e)average errors across iterations.



(a) (b)

Fig. 12 (a) Facial point detection results on LFPW using different combinations of local point detectors and the proposed discriminative deep face shape model. (b) Comparing the detection rates (percentage of images on which error $<$ 10% of inter-ocular distance) of the local point detectors and the algorithms with the shape model.

some representative images. We can see from the figure that the facial point detection algorithm is robust on images with varying illuminations, poses, facial expressions, occlusions (by eyeglasses, sunglasses, hair, and objects), resolutions, etc. The results demonstrate that our method can accurately detection facial points on image in both controlled and "in-the-wild" conditions. Figure 14 (a) shows the average results for each facial point across four testing databases, including MultiPIE, Helen, LFPW, and AFW, and the point index can be found in Figure 14 (b). As can be seen, points around eyes can be detected robustly, while contour points are difficult to detect. The is due to the fact that contour points are less distinctive and tend to be occluded by hair, and other objects.

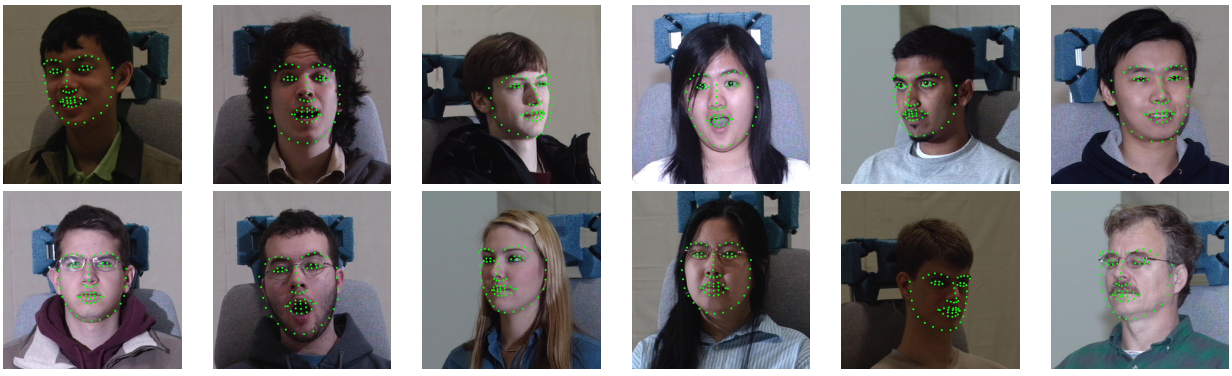
5.3.4 Comparison with the state-of-the-art works

In this section, we compare our work with 7 state-of-the-art works, including the FPLL (Zhu and Ramanan, 2012), the LEAR (Martinez et al, 2013), the RLMS (Saragih et al, 2011), the Consensus of exemplars (Belhumeur et al, 2011, 2013), the interactive method (Le et al, 2012), the AAM fitting in the wild (Tzimiropoulos and Pantic, 2013), and the Supervised descent method (Xiong and De la Torre Frade, 2013). The first five works follow the framework of the Constrained Local Methods and their algorithms are based on sophisticated face shape models. The later two approaches are the holistic methods. Here, we use learned features and NN as local point detector. For the state-of-the-art works besides FPLL method, we show the reported results in the original papers for fair comparison. Those results are indicated using "*". For FPLL method, since the evaluation criteria in the original paper is different from ours and the other state-of-the-art works, we use the public available code provided by the authors to generate the results for comparison.

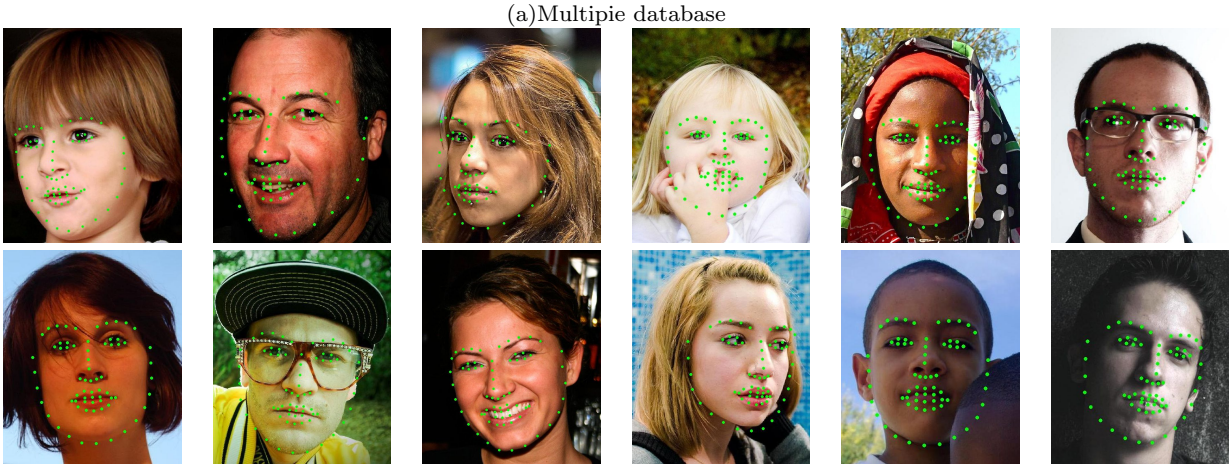
The comparison on four testing databases is shown in Figure 15 and Table 1. In Figure 15, we produce the cumulative curve corresponding to the percentage of testing images for which the error are less than specific values. On the left column, we show the popular me_{17} error that averages over 17 interior points (Cristinacce and Cootes, 2008). On the right column, we show me_{68} that averages over all 68 points. In Table 1, we show the percentage of images that are correctly detected (average error less than 0.1) using m_{17}, and m_{68} measurements, respectively.

MultiPIE testing set: The testing results on MultiPIE dataset shows that our algorithm performs robustly even if images are with varying poses, expressions and illuminations. As shown in Figure 15 (a1) and Table 1, our algorithm performs better than LEAR (Martinez et al, 2013), especially in large poses angles. As shown in Figure 15 (a2) and Table 1, our algorithm outperforms RLMS (Saragih et al, 2011).

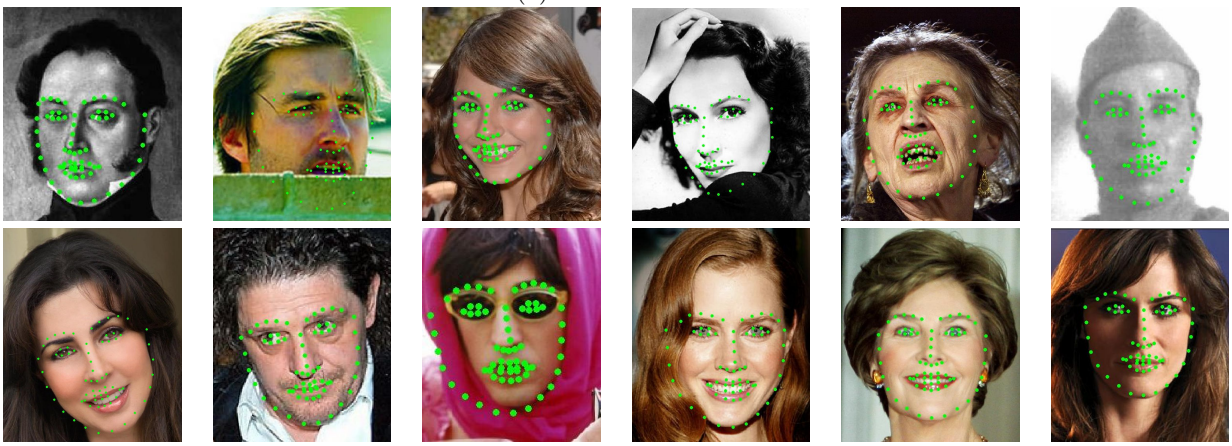
Helen testing sets: The effectiveness of the proposed facial point detection algorithm on "in-the-wild" images is supported by its performance on the Helen databases. Figure 15 (b1)(b2), and Table 1 show that our method outperforms the iterative method (Le et al, 2012) and FPLL method (Zhu and Ramanan, 2012) on this testing set.



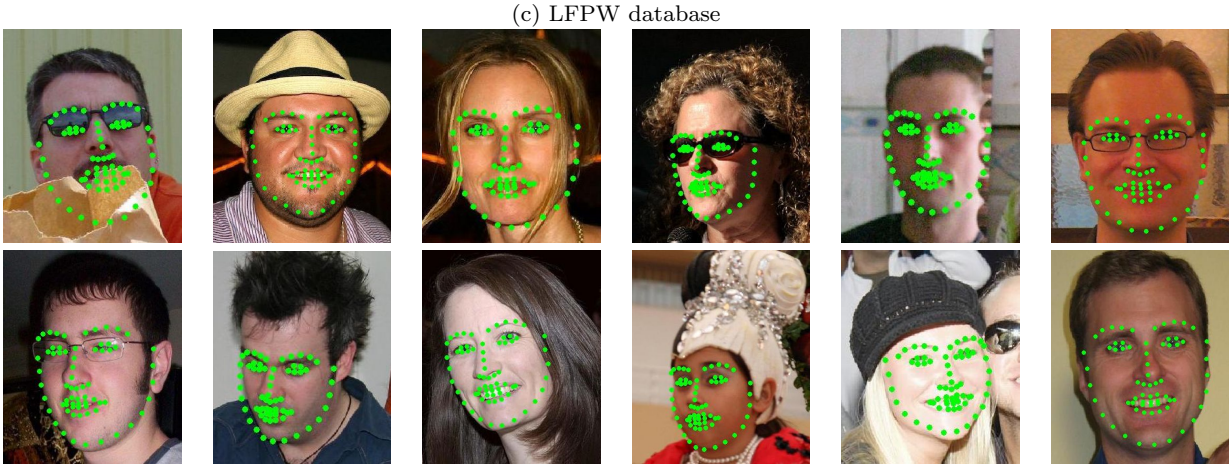
(a)Multipie database



(b) Helen database

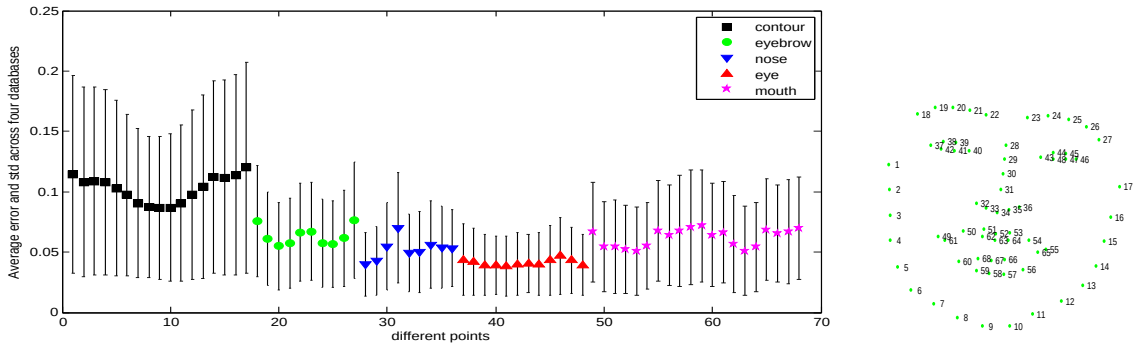


(c) LFPW database



(d) AFW database

Fig. 13 Detection results on sample images from four databases. (Better view in color)



(a) (b)

Fig. 14 Detection error (mean and std) for each point across four testing databases. (a) Detection results. (b) Point index.

LFPW testing sets: From Figure 15 (c1) and Table 1, we see that out method outperforms the AAM fitting in the wild (Tzimiropoulos and Pantic, 2013), the Consensus of exemplars (Belhumeur et al, 2011, 2013), and the FPLL method (Zhu and Ramanan, 2012). It is slightly worse than the Supervised descent method (Xiong and De la Torre Frade, 2013). But, Supervised descent method detect 29 points, while our method detect all 68 points. From Figure 15 (c2) and Table 1, we see that our method outperforms FPLL (Zhu and Ramanan, 2012) counting all the 68 points.

AFW testing sets: We perform cross-database testing on the AFW database. Our method performs better than FPLL (Zhu and Ramanan, 2012), as shown in Figure 15 and Table 1. Both methods show decreased accuracies on AFW. This indicates that AFW is a even more challenging database than Helen and LFPW database. We also observed in the experiments that the image quality from AFW is poorer (e.g. low resolution, larger pose angles, etc.). But, our facial point detection algorithm manages to generate good detection results on this challenging database.

Table 1 The comparison of detection rates (error<0.1) against state-of-the-art approaches on four benchmark databases using the me_{17} and me_{68} error measurements. Those algorithms include LEAR (Martinez et al, 2013), RLMS (Saragih et al, 2011), FPLL (Zhu and Ramanan, 2012), the interactive method (Le et al, 2012), AAM wild (Tzimiropoulos and Pantic, 2013), Consensus of exemplars (Belhumeur et al, 2011, 2013), and Supervised descent method (Xiong and De la Torre Frade, 2013). The reported results in the original paper are indicated as (*)

Database		method	accuracy(me_{17})	accuracy(me_{68})
MultiPIE	0^o	LEAR (20pts*)	98.0%	-
		RLMS (*)	-	75%
		Our method	**98.60%**	**96.08%**
	15^o	LEAR (20pts*)	96%	-
		Our method	**97.56%**	95.93%
	30^o	LEAR (20pts*)	65%	-
		Our method	**75.36%**	79.71%
Helen		FPLL	84.15%	61.33%
		Interactive method(194pts*)	-	76%
		Our method	**95.52%**	**89.91%**
LFPW		FPLL	79.90%	56.70%
		AAM wild(*)	93%	-
		Consensus exemplars (29pts*)	95%	-
		Supervised descent (29pts*)	**100%**	-
		Our method	99.04%	**96.39%**
AFW		FPLL	67.18%	51.15%
		Our method	**90.48%**	**75.82%**

The experimental comparison above shows that our facial point detection algorithm based on the novel deep face shape model outperforms the state-of-the-art approaches. It also shows that our method performs well on images in both controlled and "in-the-wild" conditions even with cross-database setting.

Efficiency: During facial landmark detection, the inference (section 3.4) through the proposed discriminative deep face shape model is efficient and it only takes about 0.12 second for one iteration (about 3 iterations to process one image). The major computational cost of the proposed method comes from the dense feature extractions of the image patches within the face region for local point detection. Specifically, in our current implementation, it takes about 2.83, 8.85, and 77.52 seconds to calculate the feature descriptors using SIFT, HOG and learned features (DBM) for one image. In total, it needs about 3.18, 9.2 and 78.0 seconds to detect the 68 facial landmarks on one image. The algorithm is implemented mainly in Matlab and tested on Intel Core 2 Duo processor E8400.

Among all the 7 state-of-the-art works for comparison, only two algorithms report their computational complexity to detect 68 facial landmark points. Specifically, the FPLL (Zhu and Ramanan, 2012) with the tree-based face shape model requires 40 second to process one image. Using their fast and less accurate version, it requires about 4 second to process one image. It is reported in (Saragih et al, 2011) that the RLMS algorithm requires about 0.12 second to process one image. Generally speaking, the computational efficiency of the proposed algorithm is comparable with the state-of-

the-art works. The efficiency can be further increased if we speed up the feature extraction step or use other strategy to generate the measurements (e.g. regression based local point detection). This work is beyond the scope of the paper and we will leave it as the future work.

6 Conclusion

This paper presents a facial point detection method that is based on a novel discriminative deep face shape model. The discriminative deep face shape model captures the joint spatial relationship among all facial points under varying facial expressions and poses for different subjects. In addition, throughout the discriminative modeling, it combines the top-down information from the embedded face shape patterns and the bottom-up measurements from the local point detectors in one unified model. Along with the model, we propose effective algorithms to perform model learning and to infer the the facial point locations given their measurements. Based on the discriminative deep face shape model, the proposed facial point detection algorithm outperforms the state-of-the-art approaches on both controlled and "in-the-wild" challenging images.

As the future work, we plan to improve the accuracy and efficiency of the local point detector to further boost the performance of our facial point detection algorithm. Specifically, the detection of the contour points is still nontrivial and challenging for existing methods including ours. The dense feature extraction step is inefficient in our current implementation. Another working direction is to learn the local point detectors with the discriminative deep face shape model jointly, which would further improve the performance of the proposed facial landmark detection algorithm.

Acknowledgements This work is supported in part by a grant from US Army Research office (W911NF- 12-C- 0017).

References

Baker S, Gross R, Matthews I (2002) Lucas-kanade 20 years on: A unifying framework: Part 3. International Journal of Computer Vision 56:221–255

Belhumeur P, Jacobs D, Kriegman D, Kumar N (2013) Localizing parts of faces using a consensus of exemplars. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(12):2930–2940

Belhumeur PN, Jacobs DW, Kriegman DJ, Kumar N (2011) Localizing parts of faces using a consensus of exemplars. In: The 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Cootes TF, Taylor CJ, Cooper DH, Graham J (1995) Active shape models their training and application. Comput Vis Image Underst 61(1):38–59

Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6):681–685

Cristinacce D, Cootes T (2008) Automatic feature localisation with constrained local models. Pattern Recognition 41(10):3054 – 3067

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: International Conference on Computer Vision and Pattern Recognition, vol 2, pp 886–893

Eslami S, Heess N, Winn J (2012) The shape boltzmann machine: a strong model of object shape. In: CVPR, pp 406–413

Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9:1871–1874

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

Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural Comput 14(8):1771–1800

Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507

Kae A, Sohn K, Lee H, Learned-Miller EG (2013) Augmenting crfs with boltzmann machine shape priors for image labeling. In: CVPR, IEEE, pp 2019–2026

Le V, Brandt J, Lin Z, Bourdev L, Huang TS (2012) Interactive facial feature localization. In: Proceedings of the 12th European Conference on Computer Vision - Volume Part III, ECCV'12, pp 679–692

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. International Journal on Computer Vision 60(2):91–110

Martinez B, Valstar MF, Binefa X, Pantic M (2013) Local evidence aggregation for regression-based facial point detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(5):1149–1163

Matthews I, Baker S (2004) Active appearance models revisited. International Journal of Computer Vision 60(2):135–164

Memisevic R, Hinton GE (2010) Learning to represent spatial transformations with factored higher–order boltzmann machines. Neural Computation 22(6):1473–1492

Mohamed A, Dahl G, Hinton G (2011) Acoustic modeling using deep belief networks. IEEE Transactions on Audio, Speech, and Language Processing PP(99):1

Ranzato M, Krizhevsky A, Hinton GE (2010) Factored 3-way restricted boltzmann machines for modeling natural images. Journal of Machine Learning Research - Proceedings Track 9:621–628

Sagonas C, Tzimiropoulos G, Zafeiriou S, Pantic M (2013a) 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: Proceedings of IEEE Int. Conf. on Computer Vision (ICCV-W 2013), 300 Faces in-the-Wild Challenge (300-W), Sydney, Australia

Sagonas C, Tzimiropoulos G, Zafeiriou S, Pantic M (2013b) A semi-automatic methodology for facial landmark annotation. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on, pp 896–903

Salakhutdinov R, Hinton G (2009) Deep boltzmann machines. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, vol 5, pp 448–455

Saragih JM, Lucey S, Cohn JF (2011) Deformable model fitting by regularized landmark mean-shift. International Journal of Computer Vision 91(2):200–215

Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Statistics and Computing 14(3):199–222

Sun Y, Wang X, Tang X (2013) Deep convolutional network cascade for facial point detection. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp 3476–3483

Taylor G, Sigal L, Fleet D, Hinton G (2010) Dynamical binary latent variable models for 3d human pose tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 631–638

Tieleman T (2008) Training restricted boltzmann machines using approximations to the likelihood gradient. In: Proceedings of the 25th international conference on Machine learning, pp 1064–1071

Tzimiropoulos G, Pantic M (2013) Optimization problems for fast aam fitting in-the-wild. In: Proceedings of IEEE International conference on Computer Vision, pp 593–600

Valstar M, Martinez B, Binefa V, Pantic M (2010) Facial point detection using boosted regression and graph models. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp 13–18

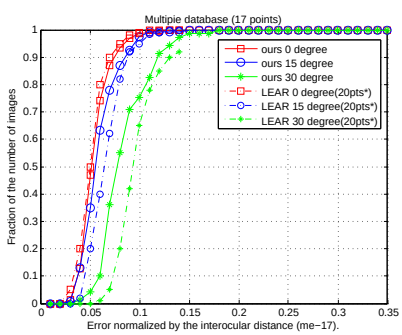
Welling M, Hinton GE (2002) A new learning algorithm for mean field boltzmann machines. In: Proceedings of the International Conference on Artificial Neural Networks, Springer-Verlag, London, UK, UK, ICANN '02, pp 351–357

Wu Y, Wang Z, Ji Q (2013) Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp 3452–3459

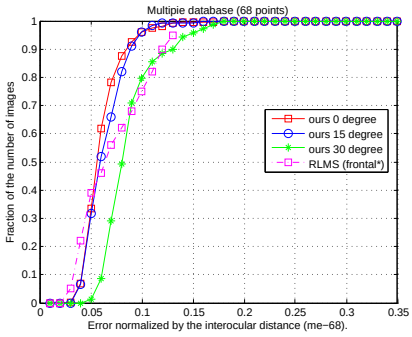
Xiong X, De la Torre Frade F (2013) Supervised descent method and its applications to face alignment. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)

Zhou E, Fan H, Cao Z, Jiang Y, Yin Q (2013) Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: IEEE International Conference on Computer Vision Workshops, pp 386–391

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

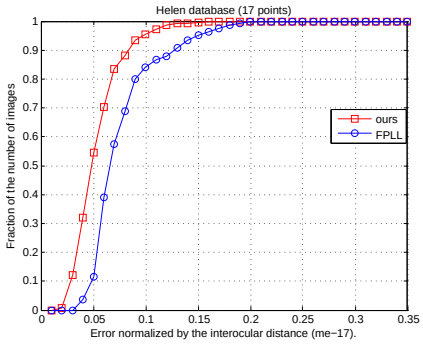


(a1) Average over 17 points

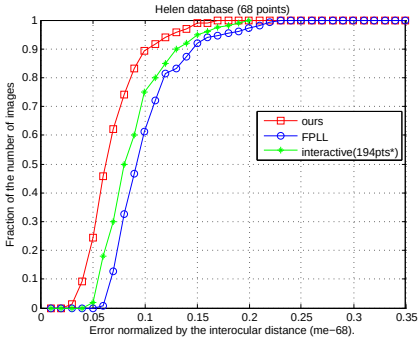


(a2) Average over 68 points

(a) MultiPIE database

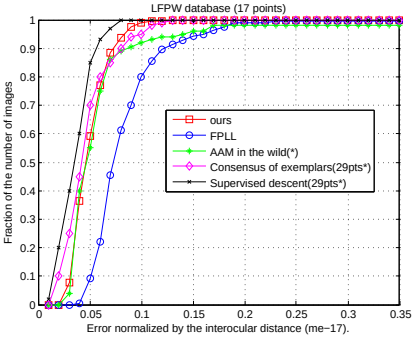


(b1) Average over 17 points

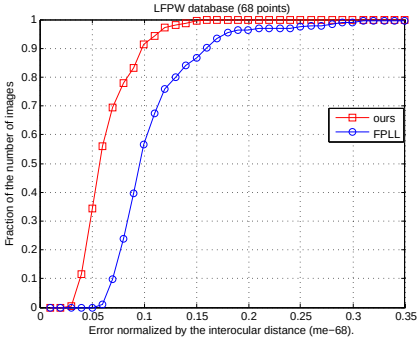


(b2) Average over 68 points

(b) Helen database

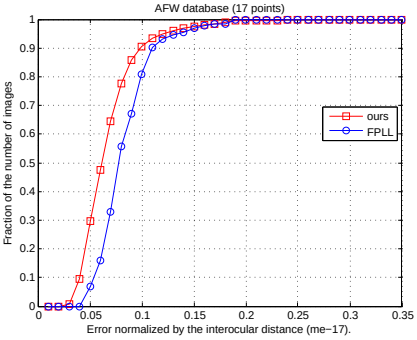


(c1) Average over 17 points

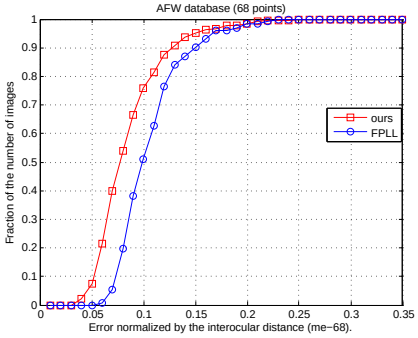


(c2) Average over 68 points

(c) LFPW database



(d1) Average over 17 points



(d2) Average over 68 points

(d) AFW database

Fig. 15 Experimental comparison with state-of-the-art works on four benchmark databases, including the MultiPIE (Gross et al, 2010), the Helen (Le et al, 2012), the LFPW (Belhumeur et al, 2011), and the AFW (Zhu and Ramanan, 2012) databases. Those state-of-the-art approaches include the LEAR (Martinez et al, 2013), the RLMS (Saragih et al, 2011), the FPLL (Zhu and Ramanan, 2012), the interactive method (Le et al, 2012), AAM fitting in the wild (Tzimiropoulos and Pantic, 2013), Consensus of exemplars (Belhumeur et al, 2011, 2013), and the Supervised descent method (Xiong and De la Torre Frade, 2013). For each row, we show the results for one database. On the left and right columns, we show the comparison based on the average error over 17 (Cristinacce and Cootes, 2008) and 68 points, respectively. The reported results in each paper are indicated as (*).