Head Pose Estimation on Low-Quality Images

Kang Wang[1], Yue Wu[2] and Qiang Ji[1]
[1] Rensselaer Polytechnic Institute {wangk10, jiq@rpi.edu} , [2] Nvidia {wuyuesophia@gmail.com}

***Abstract*— Head pose estimation methods can be broadly classified into learning-based methods and model-based methods. The learning based methods use machine learning techniques to directly predict the pose from image appearance, while the model-based methods link the 2D observation (e.g. facial landmarks) and 3D model through the projection model for pose estimation. However, both methods may have difficulty on images with very low quality (e.g. low resolution, occlusion, and noisy images). For example, there would be limited appearance information to generate accurate landmark detection on low-quality images for reliable face pose estimation. To tackle pose estimation on low-quality images, we propose to combine the learning and model based methods. Specifically, we first build the relationship between facial landmark locations and image appearance using the Restricted Boltzmann Machine (RBM) model. Then, we link the landmark locations and 3D model analytically using the projection model. By combining the RBM model with the projection model, without explicit landmark detection, we predict the head pose with a KL-divergence based method and a gradient-based method. Experimental results demonstrate the effectiveness of the proposed method.**

I. Introduction

Head pose estimation aims to identify the relative orientation and location of the head w.r.t the camera coordinate frame. Head pose has been an major component in face-related research areas, including gaze estimation ([37], [38], [39]), which reflects human intent and focus of attention, face frontalization [42], [26] which is helpful for face recognition and security. There are also many applications related to head pose, such as social attention [33], [27], [28], human-and-robot interactions [10], human-behavior analysis [32], [8], security surveillance [4], driver behavior analysis [23]. All those applications require the head pose estimation algorithms to achieve good pose estimation accuracy in challenging conditions, even if the face is far away from the camera.

In computer vision, head pose estimation based on facial images or videos have been studied extensively and intensively. The major methods can be broadly classified into learning-based approaches and model-based approaches (Fig. 1). The learning-based approaches aim to learn the relationship between facial image and the head pose angles with machine learning techniques, while the model-based methods try to link the 2D observation (e.g. facial landmark locations) and 3D model through the projection models. However, both methods have limitations on images with low-quality as shown in Fig. 2. In particular, for low-quality images, the image may contain limited facial appearance information. Similarly, low-quality image also poses challenge for the model-based approach since accurate landmark detection on the low quality images is difficult. Therefore, both methods would suffer and lead to low performances due to the noisy and incomplete observations from low-quality image.

978-1-5386-2335-0/18/$31.00 ©2018 IEEE

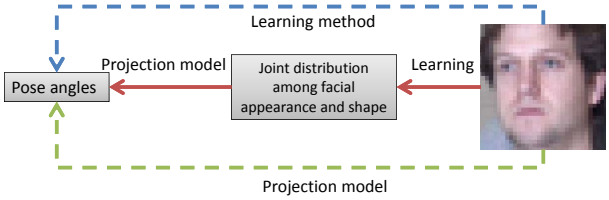


Fig. 1. Proposed method and general head pose estimation methods in different categories. The learning-based methods (blue dotted line) directly learn the mapping between image appearance and pose. The model based methods (green dotted line) link 2D facial shape and 3D face pose through the projection model. The proposed method (red line) combines the learning and model based methods.

Fig. 2. Low quality images (e.g. low-resolution, occlusion and noisy images).

To tackle those problems, in this paper, we propose a method that combines the learning based with model based methods for pose estimation on low-quality images, as shown in Fig. 1. The general intuition of the proposed method is to combine both methods, so that they may compensate each other and boost the performances of pose estimation. In particular, we first learn the joint relationship between facial landmark locations and facial appearance using the Restricted Boltzmann Machine (RBM) model. We then combine the RBM model with the projection model to perform face pose estimation directly from facial images, without explicit landmark detection. RBM has been widely used in computer vision [24], [31] to relate different variables. In this work, we link the 2D shape information and 3D facial shape model using the projection model, and link the face image appearance with image face shape through RBM statistically. Then, without explicit facial landmark detection, we can predict the head pose with a KL-divergence based method and a gradient-based method directly from face images.

The main novelty of the proposed approach lies in combining model-based approach with learning-based approach in order to take advantage of their respective strength. It allows to seamlessly combine model-based pose estimation

with learning-based pose estimation. Through this combination, we can perform head pose estimation without explicit landmark detection. The idea hence can apply to applications which may be difficult to explicit accurate landmark detection with low quality images. In summary, the major contributions of the proposed work are as follows:

- The proposed method combines learning techniques and projection model for pose estimation. This is novel compared to most of the existing methods that either follow data-driven learning approaches or purely projection model based approaches.
- Different from the existing methods, the proposed method does not explicitly detect the facial landmarks, which are difficult to perform on the low-quality images. Instead, we propose the KL-divergence and gradient based methods to estimate the face pose.
- We propose to use RBM model to capture the joint distribution for face appearance and face shape to allow head pose inference directly from the face appearance without explicitly performing facial landmark detection.
- The experimental results on images with low-quality images demonstrate the effectiveness of the proposed method on these very challenging conditions.

II. Related Work

Many methods have been proposed to estimate the head pose. These methods can be classified into two categories: learning-based methods and model-based methods.

The learning-based approaches learn the classifiers or regression models to map the appearance features to head pose. There are methods using K-nearest-neighbor [7] or using quantization [43] to improve efficiency. In [3], the authors propose to classify head pose into eight directions specific for individual scenes, exploiting the tracking information of walking direction. In [13], the authors propose to use image abstraction and local directional quaternary patterns for head pose estimation. In [14], the authors propose the K-clusters regression forests for head pose estimation. They introduce more flexible node splitting by clustering in target space instead of binary splitting. In [6], the authors propose a two-layer framework to model the globally nonlinear manifold by local linear functions. In [45], the authors propose a novel supervised descriptor learning algorithm formulated as generalized low rank approximations of matrices with a supervised manifold regularization. In [17], the authors extract dense SIFT features on face region and reduce the dimensions of features by random projection method, followed by SVR to do the regression for head pose estimation. In [1], the authors propose to classify the driver's head pose by feature selection and fusion of SURF, HOG, Haar and SF (steerable filters) features. Experiments demonstrate that the fusion of features boost the performance on head pose classification. The method in [5] maps the handcrafted HOG features to head pose and bounding box shifts by a mixture of Gaussian models.

The model based methods try to find the projection matrix that can be used to project the 3D model to a 2D image. The head pose can be extracted from the projection matrix. The method in [21] estimates head pose by applying the POSIT algorithm on the mapping of 2D projected points and 3D head model. The method in [34] fits the cylinder head models to get the projected facial points and model parameters. By further combining with AAM, they can extend the pose coverage for tracking. In [40], the authors propose a landmark-free approach to fit a 3D model to 2D image based on the sparse coding and texture of annotated face model based regression. In [18], the authors propose to fit a 3D morphable model to a 2D image using local image features (SIFT) based on the cascade regression method. During cascade regression, it iteratively updates the model parameters based on appearance features to minimize the landmark position projection error. The head pose can be decoded from the model parameters after it converges. Another cascade regression based method is from [36], and it estimates the pose from the predefined face basis vector which is used for 3D facial shape estimation. The method in [35] first performs landmark detection based on appearance features, followed by fitting the 3D model to estimate pose. In [20], the authors propose to estimate head pose from 2D face image using a 3D model morphed from a reference 3D model, which refers to a 3D face of a person of the same ethnicity and gender as the query subject. The head pose is predicted by depth parameters which are estimated by minimizing the mismatch between the feature of the query face image and related morphed 3D model projected onto 2D image plane.

There are limited works that perform head pose on low quality images. In [11], the authors address the problem of head pose estimation in low quality images. They train one particular one-layer linear neural network called auto-associative network for each pose based on the Widrow-Hoff learning rule. The head pose estimation is achieved by choosing the highest score from the related auto-associative network. Experimental results show that it outperforms humans in pan for head pose estimation. In [30], the authors propose to classify low quality head images with size from 20 to 40 segmented from the wild images. It combines the color histograms of skin and hair region for classification. In [25], the authors propose a KL distance based facial appearance descriptor, which indexes each pixel to mean appearance templates. They use the the proposed descriptor to train a multi-class SVM for head pose estimation on low quality images.

III. Approach

In this work, we propose a method that combines learning and projection model for head pose estimation on low-quality images. The proposed method contains two parts. We first capture the joint probability distribution of facial landmark locations and facial appearances using the Restricted Boltzmann Machine (RBM) model [22][29]. Then, we use the learned joint distribution and the projection model for head pose estimation. In the following, we first discuss how to learn the model, the 3D deformable model and then the proposed pose estimation algorithm.

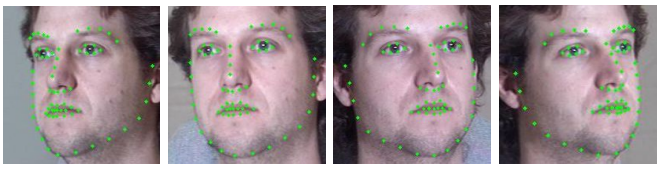


Fig. 3. Facial images [12] with marked facial landmarks.

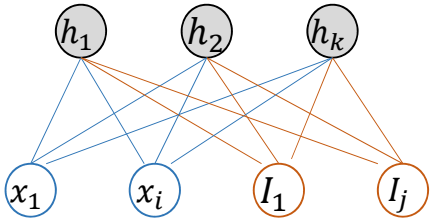


Fig. 4. Restricted Boltzmann machine (RBM) model that learns the joint distribution for shape and facial appearance.

A. Learning the joint probability distribution for facial shape and appearance

It is clear from Fig. 3 that there is a direct relationship between the facial landmark locations that determines the facial shape and facial appearance. Even though the relationship may vary from subject to subject, there are clearly certain patterns. In addition, we also can observe that both the facial shape and facial appearance can reflect the pose angles. Therefore, we would like to learn the relationship between facial landmark locations and facial appearance and use the relationship for pose estimation. In particular, we can denote the facial landmark locations as $\boldsymbol{x} = \{x_1, x_2, ..., x_D\}$, where D is the number of landmarks. We also can denote the vectorized image as $\boldsymbol{I}$. Our goal is to learn their joint probability distribution $p(\boldsymbol{x}, \boldsymbol{I})$. In particular, we used the Restricted Boltzmann Machine (RBM) [22][29].

The Restricted Boltzmann Machine [22][29] is an undirected probabilistic graphical model as shown in Fig. 4. It consists of one layer of binary hidden nodes, denoted as $\boldsymbol{h}$, and one layer of visible variables at the bottom layer. RBM captures the joint distribution of the variables in the bottom layer with the multiple hidden nodes. In our application, the bottom layer variables include the image appearance $\boldsymbol{I}$ and the facial landmark locations $\boldsymbol{x}$. Since they are all continuous variables, we use the Gaussian Bernoulli RBM [16]. Overall, the RBM captures the joint distribution for the facial shape and appearance as follows:

$$p(\boldsymbol{x}, \boldsymbol{I}; \Theta) = \frac{1}{Z} \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{x}, \boldsymbol{I}, \boldsymbol{h}; \Theta)), \tag{1}$$

where Z is the partition function to ensure a valid probability distribution. $E(\boldsymbol{x}, \boldsymbol{I}, \boldsymbol{h}; \Theta)$ is the energy function with model parameters $\Theta = \{\boldsymbol{a}^x, \boldsymbol{a}^I, \boldsymbol{b}, W^x, W^I\}$,

$$\begin{aligned} E(\boldsymbol{x}, \boldsymbol{I}, \boldsymbol{h}; \Theta) = &\sum_i \frac{(x_i - a_i^x)^2}{2} - \sum_{i,k} x_i W_{i,k}^x h_k \\ &+ \sum_j \frac{(I_j - a_j^I)^2}{2} - \sum_{j,k} I_j W_{j,k}^I h_k - \sum_k b_k h_k . \end{aligned} \tag{2}$$

The bottom layer variables $\boldsymbol{x}$ and $\boldsymbol{I}$ are normalized so that their standard derivation is 1. The RBM parameters are usually learned through the contrastive Divergence (CD) algorithm [15], given the training data $\{\boldsymbol{x}_m, \boldsymbol{I}_m\}_m$.

B. 3D deformable model and head pose

Before we introduce the proposed pose estimation method, we also need to discuss the 3D deformable model. To capture the 3D facial shape variations, we build a 3D deformable shape model. Given the N 3D facial shape training data from different subjects, denoted as $\{S_n\}_{n=1}^N \in \Re^{3*D}$, consisting of 3D coordinates of the facial landmark points, we can build the 3D deformable model using the principal component analysis technique. In particular, each new shape S can be represented using the deformable model coefficients $\boldsymbol{q} = \{q_1, q_2, ..., q_k\}$:

$$S = \bar{S} + \sum_k \Phi_k q_k, \tag{3}$$

where $\bar{S}$ represents the average 3D shape and Φ_k represents the basis.

Given the 3D deformable model, we can link the 3D face shape (represented using the deformable coefficients) and the 2D facial landmark locations using the projection model (we use weak perspective projection in this work). In particular, if we denote the column and row coordinates of the dth 2D point on image as c_d and r_d for corresponding 3D points in shape S, we have:

$$\begin{bmatrix} c_1, c_2, ..., c_D \\ r_1, r_2, ..., r_D \end{bmatrix} = M(\bar{S} + \sum_k \Phi_k q_k) + \begin{bmatrix} t_1, t_1, ..., t_1 \\ t_2, t_2, ..., t_2 \end{bmatrix}. \tag{4}$$

Here, M is a $2*3$ matrix corresponding to the scaled first and second rows of the rotation matrix. t is the translation vector on the image plane that is applied to all the points. Overall, for all the facial landmarks $\boldsymbol{x} = \{c_1, r_1, c_2, r_2, ...c_D, r_D\}$, we can write the projected points as $\boldsymbol{x}_{proj}$:

$$\boldsymbol{x}_{proj} = g(M, t, \boldsymbol{q}), \tag{5}$$

where $g(.)$ denotes the projection model. The detected 2D facial landmarks are usually the noisy observations of the projection results. Therefore, we can denote the conditional distribution of the 2D shape given the pose and deformable coefficients as follow:

$$p(\boldsymbol{x}|M, t, \boldsymbol{q}) \sim \mathcal{N}(g(M, t, \boldsymbol{q}), \alpha\Lambda). \tag{6}$$

Here, we assume the noise is Gaussian noise with diagonal covariance matrix, and Λ represents the identity matrix.

For general model-based pose estimation methods, given the detected 2D facial landmarks on images, we can estimate the pose parameters by minimizing the projection errors:

$$M^*, t^*, \boldsymbol{q}^* = arg \min_{M,t,\boldsymbol{q}} \|\boldsymbol{x} - g(M, t, \boldsymbol{q})\|^2. \tag{7}$$

Here, $\boldsymbol{x}$ denotes the detected landmarks. To solve the optimization problem, we can alternatively update the pose matrix M, t and the deformable coefficients $\boldsymbol{q}$.

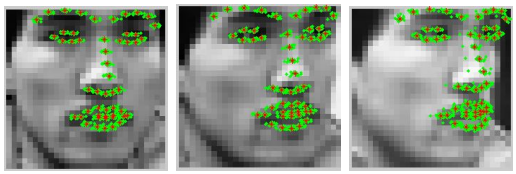


Fig. 5. Generate the samples of facial landmark locations given the facial images [9]. For each images, we generate multiple samples (clusters of green dots). The red dots indicate their mean.

C. Pose estimation

Given the learned joint distribution for facial shape and facial appearance $p(\boldsymbol{x}, \boldsymbol{I}; \Theta)$, the 3D deformable model and projection equations, we can perform head pose estimation. In particular, we propose two methods: the KL-divergence based method and the gradient-based method.

1) KL-divergence based method: The intuition of the KL-based approach is that on low-quality images, facial landmark detection may not be accurate. Instead of relying on the detected landmarks, we calculate the probabilistic landmark locations, which would give us more information. In particular, given the facial appearance information of the testing image $\boldsymbol{I}_{test}$ and the learned joint distribution for shape and the appearance $p(\boldsymbol{x}, \boldsymbol{I}; \Theta)$, we can calculate the conditional distribution of the facial landmark locations given the appearance $p(\boldsymbol{x}|\boldsymbol{I}_{test}; \Theta)$. In addition, we can relate the 2D facial landmark locations and the 3D pose and deformable coefficients as in Eq. 6. Therefore, we can formulate the head pose estimation problem by minimize the KL-divergence of those two distributions of $\boldsymbol{x}$:

$$\begin{aligned} M^*, t^*, \boldsymbol{q}^* &= arg \min_{M,t,\boldsymbol{q}} \mathcal{KL}(p(\boldsymbol{x}|\boldsymbol{I}_{test}; \Theta)||p(\boldsymbol{x}|M, t, \boldsymbol{q})) \\ &= arg \max_{M,t,\boldsymbol{q}} \int_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{I}_{test}; \Theta) log p(\boldsymbol{x}|M, t, \boldsymbol{q}). \end{aligned} \tag{8}$$

Since the integral is intractable, we can use importance sampling method and approximate the integral through the samples:

$$M^*, t^*, \boldsymbol{q}^* = arg \max_{M,t,\boldsymbol{q}} \sum_s \frac{1}{N_s} log p(\boldsymbol{x}^s|M, t, \boldsymbol{q}), \tag{9}$$

where $\boldsymbol{x}^s$ denotes the samples generated from $p(\boldsymbol{x}|\boldsymbol{I}_{test}; \Theta)$ and N_s denotes the number of sample. Note that, for RBM model, to generate the samples of $\boldsymbol{x}^s$ from $p(\boldsymbol{x}, \boldsymbol{I}_{test}; \Theta)$, we can use the Gibbs sampling method that iteratively calls the following function:

$$p(x_i|\boldsymbol{h}; \Theta) \sim \mathcal{N}\left(\sum_k W_{i,k} h_k + a_i^x, 1\right), \tag{10}$$

$$p(h_k = 1|\boldsymbol{x}, \boldsymbol{I}_{test}; \Theta) = \sigma\left(\sum_i x_i W_{i,k} + \sum_j I_j W_{l,k} + b_k\right), \tag{11}$$

where $\mathcal{N}(.,.)$ denotes the Gaussian distribution and $\sigma(.)$ denotes the sigmoid function. Fig. 5 shows some sampling outputs given different facial images.

By applying Eq. 6, the problem in Eq. 9 can be further simplified as:

$$M^*, t^*, \boldsymbol{q}^* = arg \min_{M,t,\boldsymbol{q}} \sum_s \frac{1}{N_s} \|\boldsymbol{x}^s - g(M, t, \boldsymbol{q})\|. \tag{12}$$

The above equation indicates that different from the existing model-based head pose estimation methods that estimate the pose by fitting the detected facial landmarks on the testing image (Eq. 7), we can perform head pose estimation by fitting multiple samples of the 2D facial shape. This is especially important for images with low-quality, since it's difficult to have very accurate landmark detection results, but the collection of samples may provide more information. To solve the optimization problem, similar to the general model-based pose estimation methods (Eq. 7), we can alternatively update the pose matrix M, t and the deformable coefficients $\boldsymbol{q}$. But, this time, the pose parameters are shared by all the samples.

2) Gradient-based method: Another way to estimate the head pose with the learned joint distribution is to directly optimize the pose parameters that maximize the joint distribution given the image appearance. In particular, we want to find the pose parameters that maximize the conditional distribution of pose given the image appearance $p(M, t, \boldsymbol{q}|\boldsymbol{I}_{test})$. Then, pose estimation problem can be formulated as follows:

$$\begin{aligned} M^*, t^*, \boldsymbol{q}^* &= arg \max_{M,t,\boldsymbol{q}} p(M, t, \boldsymbol{q}|\boldsymbol{I}_{test}) \\ &= arg \max_{M,t,\boldsymbol{q}} \log p(M, t, \boldsymbol{q}|\boldsymbol{I}_{test}) \end{aligned} \tag{13}$$

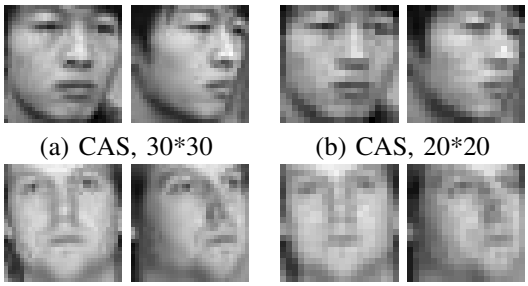
Furthermore, since we know the landmark locations can be represented in terms of the pose matrix M, t, and deformable coefficient $\boldsymbol{q}$, as in Eq. 5, the problem can be equally converted to find the pose parameters that maximize the joint distribution among the 2D shape (determined by the pose parameters) and the given image appearance (detailed proof can be found in the supplemental material). Assume M, t and $\boldsymbol{q}$ are independent, then the gradient can be calculated w.r.t $M, t, \boldsymbol{q}$ iteratively. Let's take M as example,

$$\begin{aligned} M^* &= arg \max_M \log p(\boldsymbol{x}|\boldsymbol{I}_{test}) \cdot \left| \frac{\partial \boldsymbol{x}}{\partial M} \right| \\ &= arg \max_M \log p(\boldsymbol{x}, \boldsymbol{I}_{test}) \cdot \left| \frac{\partial \boldsymbol{x}}{\partial M} \right| \end{aligned} \tag{14}$$

In order to solve the optimization problem, we use gradient ascent method. Denote the objective function as $f(M, \boldsymbol{I}_{test})$, the gradient w.r.t M can be calculated as:

$$\begin{aligned} \frac{\partial f(M, \boldsymbol{I}_{test})}{\partial M} &= \frac{\partial \log p(\boldsymbol{x}, \boldsymbol{I}_{test})}{\partial M} + \frac{\partial \log \left| \frac{\partial \boldsymbol{x}}{\partial M} \right|}{\partial M} \\ &= \frac{\partial \log p(\boldsymbol{x}, \boldsymbol{I}_{test})}{\partial \boldsymbol{x}} \cdot \frac{\partial \boldsymbol{x}}{\partial M} \end{aligned} \tag{15}$$

The second term is zero since $\boldsymbol{x}$ is the linear function of M. As for the first term, taking advantage of the chain rule, we can take the gradient of the log joint probability w.r.t the shape $\boldsymbol{x}$, and then take the gradient of the shape w.r.t the pose parameters or deformable coefficients. "$\cdot$" represents the dot product between two vectors. The gradient of $\boldsymbol{x} = g(M, t, \boldsymbol{q})$



(a) CAS, 30*30 (b) CAS, 20*20

(c) MultiPie, 30*30 (d) MultiPie, 20*20

Fig. 6. Sample testing images from CAS-PEAL face database and MultiPie database (30*30 pixels).

w.r.t pose parameters or the deformable coefficients can be calculated based on Eq. 4. The following equation can be used to calculate the gradient of $\log p(\boldsymbol{x}, \boldsymbol{I}_{test})$ w.r.t $\boldsymbol{x}$ for the RBM model:

$$\frac{\partial \log p(\boldsymbol{x}, \boldsymbol{I}_{test}; \Theta)}{\partial x_t}|_{\boldsymbol{x}^0} = -(x_t^0 - a_t^x) + \sum_k W_{t,k}\, p(h_k = 1|\boldsymbol{x}^0, \boldsymbol{I}_{test}). \tag{16}$$

The beauty of this approach is that it combines analytic model (projection model) with the learned RBM model to jointly compute the pose gradient update instead of computing directly from face image or from projection model.

IV. Experimental Results

A. Implementation details

1) databases: In the experiments, we use four different databases: the CAS-PEAL face database [9], the MultiPie database [12], the Caltech Occluded Face in the Wild (COFW) database [2] and the Annotated Facial Landmarks in the Wild (AFLW) database[19]. The CAS-PEAL face database contains images of 1040 subjects with varying poses changing in pitch and yaw angles. There are 6000 images for training and 520 images for testing (cross-subjects). For the MultiPie database, we use 1333 images from the first 150 subjects for training and 908 images from the remaining subjects for testing. For both CAS-PEAL and MultiPie databases, we only use the images with yaw variations for pose estimation (0, +-15, +-30, and +-45 degree). We also resize them to generate two sets of images with different resolutions. For the first set, the face region is about 30*30, while the face region is about 20*20 for the second set of images. The COFW database contains "in-the-wild" facial images with significant facial occlusion. There are 1345 images in the training set and 507 images in the testing set. There are 29 facial landmark annotations for each image. For this database, we use the model-based pose estimation method (Eq. 7) to generate the head pose labels based on the ground truth landmarks. The AFLW database contains "in-the-wild" facial images with significant facial occlusion as well as different light conditions. There are 1500 images in the training set and 500 images in the testing set and 21 annotated facial landmarks for each image. In our experiment, we normalize the image from both the COFW and the AFLW so that the face region is about 30*30 pixels.

2) Model details: For the RBM, we use 600 hidden nodes and the model is trained for 1000 epochs. During inference, 20 samples are generated for the KL-divergence based method. For the gradient based method, we use learning rate as $2.5 * 10^{-5}$. We will stop the gradient ascent iteration if the pose parameters do not change in two consecutive iteration. We calculate the mean absolute difference between the estimated pose and the ground truth pose to evaluate the pose estimation accuracy.

B. Low-resolution image

1) CAS-PEAL face database: We first perform experiments on the CAS-PEAL face database [9]. Some sample images can be found in Fig. 6 (a)(b). There are two sets of baseline methods. The first set of baseline methods follow the model-based approaches, which first perform facial landmark detection, followed by the model-based pose estimation. We used two state-of-the-art algorithms for facial landmark detection, including the Supervised Descent Method (SDM) [41], which is a general cascade regression model and the TCDCN method [44], which is a CNN and multi-task learning based method. Given the detected facial landmarks, we used the weak perspective projection and deformable model for head pose estimation by minimizing the projection errors (Eq. 7). The second set of baseline method is the learning-based method. In particular, we use the same Restricted Boltzmann Machine model to directly learn the joint distribution for the yaw pose angle and the image appearance and infer the pose directly for pose estimation. For the proposed method, we can combine the RBM with the KL-divergence based method and the gradient based method. We denote them as RBM + KL, and RBM + GD, respectively.

The overall experimental results on the CAS-PEAL face database can be found in Tab. I. There are a few observations. First, for the proposed methods, the gradient-based method is better than the KL-divergence based method. Second, overall, the proposed method with the combination of RBM and gradient-based method achieves the best performance on both images of 30*30 pixel and 20*20 pixel, especially for images with 20*20 pixel.

The performances of the proposed method with the combination of RBM and gradient-based method (section III-C.2) across different gradient ascent iterations are shown in Fig. 8. Here, we show the projection of the 3D landmark points using the estimated pose parameters and the deformable coefficients to the 2D image plane with Eq. 4 for better visualization. As we can see, the gradient based method will gradually find the good pose parameters to fit the testing image. For the proposed method with the combination of RBM and KL-divergence based method, we have shown some intermediate shape samples in Fig. 5.

The detailed pose estimation results for different poses are shown in Fig. 7. For pose with 0 degree, the landmark detection based methods, such as SDM and TCDCN are slightly

TABLE I

HEAD POSE ESTIMATION ACCURACY ON CAS-PEAL

methods	30* 30				20* 20			
	pitch	yaw	roll	average	pitch	yaw	roll	average
SDM [41] + model	4.41	3.29	2.38	3.35	4.43	3.58	2.43	3.48
TCDCN [44] + model	4.73	4.45	3.61	4.25	5.83	5.49	3.67	4.99
Learning-based (RBM)	-	3.51	-	-	-	3.18	-	-
proposed (RBM+KL)	4.05	**3.25**	**2.21**	3.17	3.82	3.25	**2.11**	3.06
proposed (RBM+Gradient)	**3.64**	3.29	2.51	**3.15**	**3.49**	**3.14**	2.77	**3.13**

TABLE II

HEAD POSE ESTIMATION ACCURACY ON MULTIPIE

methods	30* 30				20* 20			
	pitch	yaw	roll	average	pitch	yaw	roll	average
SDM [41] + model	5.92	**3.68**	2.16	4.02	6.47	4.68	2.21	4.53
TCDCN [44] + model	5.40	4.86	2.54	4.19	6.03	6.19	2.83	4.91
Learning-based (RBM)	-	4.02	-	-	-	4.30	-	-
proposed (RBM+ KL)	6.43	3.81	**2.01**	4.15	4.62	4.55	**1.74**	3.64
proposed (RBM+ Gradient)	**4.26**	4.33	2.38	**4.01**	**3.91**	**4.29**	2.17	**3.46**

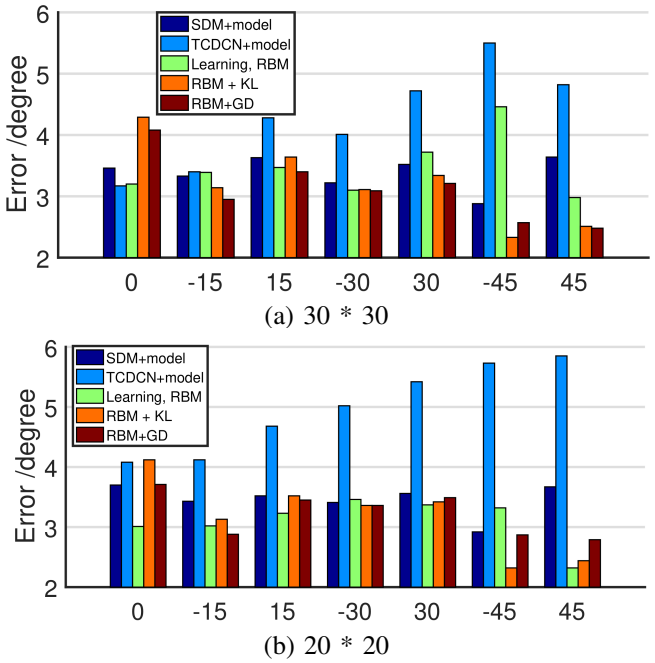


Fig. 7. Comparison of head pose estimation errors on different yaw angles of CAS-PEAL face database with low resolution images.

better than the proposed method. But, with the increase of the pose angles, the proposed method outperforms the landmark detection based methods.

2) MultiPie database: We also performed experiments on the MultiPie database [12]. Some sample images can be found in Fig. 6 (c)(d). The results can be found in Tab. II. Overall, the performance of the proposed method is similar to that on the CAS-PEAL face database. For all methods, the performances on 20*20 images are generally worse than that on 30*30 images. For the proposed method, the gradient based method is always better than the Kl-divergence based method. Compared to other baseline methods that combine

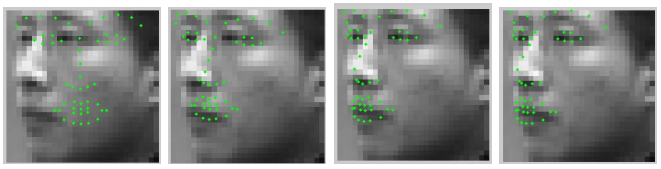


Fig. 8. Performance (projection of 3D points using the estimated pose parameters onto image plane) of the RBM + Gradient based method across different iterations on one sample image (30*30) from CAS-PEAL face database [9]. The first image shows the initialization and the remaining images show the consecutive results in gradient ascent iterations. The last image show the final result after convergence.

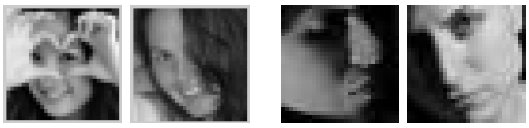
landmark detection methods such as SDM and TCDCN and model-based pose estimation, the proposed method achieves better average errors. The proposed method is also better than the learning-based method.

C. Image with facial occlusion

We further performed experiments (See Tab. III) on low-resolution images with facial occlusion. In particular, we use the CAS-PEAL face database and normalize images so that the face region is about 30*30. In addition, we add random patch with size 5*5 to cover part of the face.

Fig. 9. Images (30*30) with facial occlusion (5*5).

First, compared to the results on the same set without occlusion as shown in Tab. I, the learning-based methods has significant performance drop, while the proposed methods only show minor performance drop. The result demonstrates



(a) (b)

Fig. 10. "In-the-wild" facial images with facial occlusions and different light conditions from COFW and AFLW database .

that by incorporating the model into the learning framework, the proposed method can better handle facial occlusion compared to the data-driven learning method. Second, the proposed method is better than the other baseline methods.

TABLE III
POSE ESTIMATION ACCURACY ON CAS-PEAL WITH OCCLUSION

methods	pitch	yaw	roll	average
SDM [41] + model	4.45	**3.61**	2.49	3.62
TCDCN [44] + model	5.36	5.81	3.71	4.96
Learning-based (RBM)	-	5.12	-	-
proposed (RBM+KL)	4.06	4.78	**1.83**	**3.56**
proposed (RBM+Gradient)	**3.86**	4.51	2.52	3.63

D. "In-the-wild" facial image

Finally, we perform experiments on "in-the-wild" facial images from COFW and AFLW. Faces from COFW include self-occlusion and object-occlusion while faces from AFLW focus more on different lighting conditions as well as occlusions. All images are resized to 30*30 (Fig. 10). The experimental results are shown in Tab. IV and Tab. V respectively.

As we can see, the performances of all the methods drop noticeably. In addition, since the COFW database only contain 29 points and the AFLW database with 21 annotated points even contain less, with smaller number of points, the model-based methods and the proposed methods would suffer more than the learning based methods, that doesn't include the projection model. But, the proposed methods still can achieve comparable results.

E. Contribution of model-based and learning-based methods

To further demonstrate the contribution of each step, we perform two experiments on MultiPie and AFLW. Here, we mainly explore the advantage of proposed RBM+KL method without explicit landmark detection. For comparison, we also include results from the Max-out method which provides the most possible landmark locations to the projection model and a baseline method that uses the ground truth landmark points

TABLE IV
POSE ESTIMATION ACCURACY ON COFW OCCLUSION

methods	pitch	yaw	roll	average
SDM + model	6.96	8.38	5.45	6.93
TCDCN [44] + model	7.78	**6.77**	5.19	6.58
Learning-based (RBM)	6.95	7.54	**3.99**	**6.16**
proposed (RBM+ KL)	7.23	6.86	5.61	6.56
proposed (RBM + Gradient)	**6.89**	6.84	5.16	6.30

TABLE V
POSE ESTIMATION ACCURACY ON AFLW WITH DIFFERENT LIGHT CONDITION

methods	pitch	yaw	roll	average
SDM + model	11.15	8.66	**3.44**	7.75
proposed (RBM+ KL)	**8.80**	**6.29**	4.04	**6.38**
proposed (RBM + Gradient)	10.62	6.49	4.75	7.28

TABLE VI
CONTRIBUTION OF MODEL-BASED AND LEARNING-BASED METHODS

methods	MultiPie				AFLW			
	pitch	yaw	roll	avg	pitch	yaw	roll	avg
Max-out	6.4	5.8	2.4	4.9	10.4	7.7	5.2	7.8
Baseline	5.8	3.4	2.3	3.8	6.6	4.6	1.7	4.3
RBM+ KL	6.4	3.8	2.0	4.2	4.8	6.3	4.0	6.4

to feed in the projection model. We first evaluate the landmark estimation accuracy of Max-out method. The landmark estimation accuracy is calculated by comparing the detection landmark locations to the groundtruth facial landmark locations normalized by inter-ocular distance. The results are 10.41% and 7.65% on MultiPie and AFLW respectively. To evaluate the head pose accuracy, the mean absolute difference between the estimated pose and the ground truth pose are calculated. The experimental results are shown in Tab. VI. As we can see from the result, the proposed method takes advantage of uncertainty and improves the performance a lot comparing to only trusting on exactly detected landmarks. And, it can achieve acceptable performance comparing to the baseline method which estimate head pose from ground truth landmarks.

V. CONCLUSION

In conclusion, in this paper we proposed methods for head pose estimation on low-quality images (e.g. low-resolution, occlusion, and noisy images). Different from the existing methods that either perform learning-based approaches or model-based approaches, we combine learning and model based methods. We learn the joint probabilistic distribution of the facial image and facial shape, and use it in combination with a KL-divergence or a gradient based method for pose estimation without explicit facial landmark detection. The experiments on benchmark databases demonstrate the effectiveness of the proposed methods on low-quality images including the low-resolution images, images with facial occlusion, and "in-the-wild" images with facial occlusion. In the future, we would exploit the other model to learn the joint probabilistic distribution of facial shape and facial appearance. In addition, we would further evaluate the proposed method on other challenging conditions (e.g. illumination variations).

Acknowledgements: This work is partially supported by Cognitive Immersive Systems Laboratory (CISL), a collaboration between IBM and RPI, and also a center in IBM's AI Horizon Network.

REFERENCES

[1] N. Alioua, A. Amine, A. Rogozan, A. Bensrhair, and M. Rziza. Driver head pose estimation using efficient descriptor fusion. *EURASIP Journal on Image and Video Processing*, 2016(1):1–14, 2016.

[2] X. Burgos-Artizzu, P. Perona, and P. Dollr. Robust face landmark estimation under occlusion. In *International Conference on Computer Vision (ICCV)*. IEEE, 2013.

[3] I. Chamveha, Y. Sugano, D. Sugimura, T. Siriteerakul, T. Okabe, Y. Sato, and A. Sugimoto. Appearance-based head pose estimation with scene-specific adaptation. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1713–1720. IEEE, 2011.

[4] C. Chen and J.-M. Odobez. We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1544–1551. IEEE, 2012.

[5] V. Drouard, R. Horaud, A. Deleforge, S. Ba, and G. Evangelidis. Robust head-pose estimation based on partially-latent mixture of linear regression. *arXiv preprint arXiv:1603.09732*, 2016.

[6] J. Foytik and V. K. Asari. A two-layer framework for piecewise linear manifold-based head pose estimation. *International journal of computer vision*, 101(2):270–287, 2013.

[7] Y. Fu and T. S. Huang. Graph embedded analysis for head pose estimation. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 6–pp. IEEE, 2006.

[8] Q. Gan, S. Nie, S. Wang, and Q. Ji. Differentiating between posed and spontaneous expressions with latent regression bayesian network. 2017.

[9] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao. The cas-peal large-scale chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 38(1):149–161, Jan 2008.

[10] A. Gaschler, S. Jentzsch, M. Giuliani, K. Huth, J. de Ruiter, and A. Knoll. Social behavior recognition using body posture and head pose for human-robot interaction. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 2128–2133. IEEE, 2012.

[11] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley. Head pose estimation on low resolution images. In *International Evaluation Workshop on Classification of Events, Activities and Relationships*, pages 270–280. Springer, 2006.

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

[13] B. Han, S. Lee, and H. S. Yang. Head pose estimation using image abstraction and local directional quaternary patterns for multiclass classification. *Pattern Recognition Letters*, 45:145–153, 2014.

[14] K. Hara and R. Chellappa. Growing regression forests by classification: Applications to object pose estimation. In *Computer Vision–ECCV 2014*, pages 552–567. Springer, 2014.

[15] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computing*, 14(8):1771–1800, Aug. 2002.

[16] G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, Jul. 2006.

[17] H. T. Ho and R. Chellappa. Automatic head pose estimation using randomly projected dense sift descriptors. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 153–156. IEEE, 2012.

[18] P. Huber, Z.-H. Feng, W. Christmas, J. Kittler, and M. Rätsch. Fitting 3d morphable face models using local features. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 1195–1199. IEEE, 2015.

[19] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[20] S. G. Kong and R. O. Mbouna. Head pose estimation from a 2d face image using 3d face morphing with depth parameters. *IEEE Transactions on Image Processing*, 24(6):1801–1808, 2015.

[21] R. O. Mbouna, S. G. Kong, and M.-G. Chun. Visual analysis of eye state and head pose for driver alertness monitoring. *IEEE transactions on intelligent transportation systems*, 14(3):1462–1469, 2013.

[22] R. Memisevic and G. E. Hinton. Learning to represent spatial transformations with factored higher–order boltzmann machines. *Neural Computation*, 22(6):1473–1492, Jun. 2010.

[23] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):607–626, 2009.

[24] S. Nie, Z. Wang, and Q. Ji. A generative restricted boltzmann machine based method for high-dimensional motion data modeling. *Computer Vision and Image Understanding*, 136:14–22, 2015.

[25] J. Orozco, S. Gong, and T. Xiang. Head pose classification in crowded scenes. In *BMVC*, volume 1, page 3, 2009.

[26] G.-J. Qi. Loss-sensitive generative adversarial networks on lipschitz densities. *arXiv preprint arXiv:1701.06264*, 2017.

[27] G.-J. Qi, C. C. Aggarwal, and T. Huang. Link prediction across networks by biased cross-network sampling. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pages 793–804. IEEE, 2013.

[28] G.-J. Qi, C. C. Aggarwal, and T. S. Huang. On clustering heterogeneous social media objects with outlier links. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 553–562. ACM, 2012.

[29] M. Ranzato, A. Krizhevsky, and G. E. Hinton. Factored 3-way restricted boltzmann machines for modeling natural images. In *Int. Conf. Artificial Intell. and Stat.*, pages 621–628, Sardinia, Italy, 2010.

[30] N. Robertson and I. Reid. Estimating gaze direction from low-resolution faces in video. In *European Conference on Computer Vision*, pages 402–415. Springer, 2006.

[31] R. Salakhutdinov, A. Mnih, and G. Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798. ACM, 2007.

[32] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 53–58. IEEE, 2002.

[33] R. Subramanian, Y. Yan, J. Staiano, O. Lanz, and N. Sebe. On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 3–10. ACM, 2013.

[34] J. Sung, T. Kanade, and D. Kim. Pose robust face tracking by combining active appearance models and cylinder head models. *International Journal of Computer Vision*, 80(2):260–274, 2008.

[35] Y. Tong, Y. Wang, Z. Zhu, and Q. Ji. Robust facial feature tracking under varying face pose and facial expression. *Pattern Recognition*, 40(11):3195–3208, 2007.

[36] S. Tulyakov and N. Sebe. Regressing a 3d face shape from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3748–3755, 2015.

[37] R. Valenti, N. Sebe, and T. Gevers. Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing*, 21(2):802–815, 2012.

[38] K. Wang and Q. Ji. Real time eye gaze tracking with kinect. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 2752–2757. IEEE, 2016.

[39] K. Wang and Q. Ji. Real time eye gaze tracking with 3d deformable eye-face model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1003–1011, 2017.

[40] Y. Wu, X. Xu, S. K. Shah, and I. A. Kakadiaris. Towards fitting a 3d dense facial model to a 2d image: A landmark-free approach. In *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*, pages 1–8. IEEE, 2015.

[41] Xuehan-Xiong and F. De la Torre. Supervised descent method and its application to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[42] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. *arXiv preprint arXiv:1704.06244*, 2017.

[43] T. Zhang, G.-J. Qi, J. Tang, and J. Wang. Sparse composite quantization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4548–4556, 2015.

[44] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):918–930, May 2016.

[45] X. Zhen, Z. Wang, M. Yu, and S. Li. Supervised descriptor learning for multi-output regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1211–1218, 2015.