Journal of Electronic Imaging

JElectronicImaging.org

# Coupled cascade regression from real and synthesized faces for simultaneous landmark detection and head pose estimation

Chao Gou
Qiang Ji

Chao Gou, Qiang Ji, "Coupled cascade regression from real and synthesized faces for simultaneous landmark detection and head pose estimation," *J. Elecftron. Imaging* **29**(2), 023028 (2020), doi: 10.1117/1.JEI.29.2.023028

# Coupled cascade regression from real and synthesized faces for simultaneous landmark detection and head pose estimation

**Chao Gou[a,*] and Qiang Ji[b]**
[a]Sun Yat-sen University, School of Intelligent Systems Engineering, Guangzhou, China
[b]Rensselaer Polytechnic Institute, Department of Electrical, Computer, and Systems Engineering, Troy, New York, United States

**Abstract** The existing approaches usually perform facial landmark detection and head pose estimation independently and sequentially, ignoring their coupled relations. We introduce a unified framework, named coupled cascade regression (CCR), for simultaneous facial landmark detection and head pose estimation. Based on the cascade regression framework, we propose to learn two separate regressors to update the landmark locations and three-dimensional (3D) face model parameters at each cascade level. To capture the coupled relations of the landmark locations and head pose, we further apply the 3D face projection model to refine the prediction results in each cascade iteration and make them consistent. CCR can leverage both the learning methods and the projection model to simultaneously perform facial landmark detection and pose estimation to enhance the performances of both tasks. We also propose to learn the cascade regressors from the combination of real and synthesized face images to solve the problem of limited variations in head pose for training. Experimental results on Helen, labeled face parts in the wild, 300-W, and Boston University datasets show that our proposed CCR method outperforms other conventional methods both for landmark detection and head pose estimation.  [DOI: 10.1117/1.JEI.29.2.023028]

**Keywords:** facial landmark detection; head pose estimation; coupled cascade regression.

Paper 191101 received Dec. 3, 2019; accepted for publication Apr. 7, 2020; published online Apr. 22, 2020.

## 1 Introduction

Facial landmarks, also known as face fiducial points such as eye corners, mouth corners, and nose tip, play an important role in face recognition,[1–4] facial action unit recognition,[5,6] and three-dimensional (3D) face reconstruction.[7,8] Head pose estimation aims to predict the orientation of the head with respect to the camera coordinate frame. It has been widely used in many scenarios, such as capturing the visual attention of subjects in human–machine intersection, estimating the gaze direction of a driver, and analyzing social event interaction.[9–12]

For facial landmark detection, cascade regression framework has become one of the most effective and efficient frameworks. Typically, as a learning-based approach, it learns regression models at each cascade level to map the local appearance features to target variables (i.e., shape updates). For head pose estimation, the related approaches can be categorized into learning-based methods and model-based methods. The learning-based methods use machine learning techniques to learn the mapping between image observation and head pose, while the model-based methods link them through a computer vision projection model.[13,14] Most of the existing methods either follow the learning-based approach or model-based approach to achieve landmark detection or head pose estimation.[11,14–18] However, since the landmark positions and head pose are related, they should be tracked jointly. In addition, most methods' performance rely on a large scale of labeled data, especially for recent deep learning-based methods for landmark detection and head pose estimation. However, manually labeled training images cannot cover variations in real scenarios. Recently, learning from synthetic data with automatic accurate

*Address all correspondence to Chao Gou, E-mail: gouchao@mail.sysu.edu.cn

1017-9909/2020/$28.00

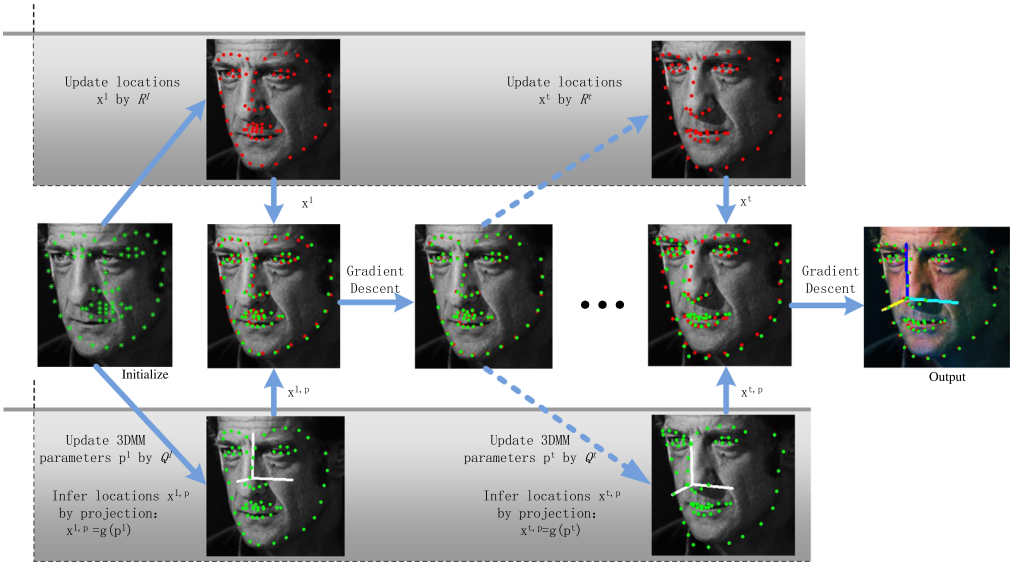



**Fig. 1** The overall architecture of our proposed CCR method for simultaneously estimating the landmark locations and head pose. $R^T$ and $Q^T$ denote two separate learned regressors for updating the landmark locations and 3DMM parameters at iteration $T$, respectively. The head pose can be easily derived from 3DMM. After estimating two different landmark locations $\mathbf{x}^t$ and $\mathbf{x}^{t,p}$ based on regressor $R^T$ and $Q^T$ with projection model $g(\cdot)$, respectively, CCR further updates based on gradient decent to make them consistent with each other (best view in color).

annotations is an effective way to tackle the problem of lack of manual annotations.[13,19–21] Therefore, it only makes sense to combine the cascade learning-based and the model-based method from both real and synthetic data to boost the performance of landmark localization and pose estimation, while simultaneously addressing the inadequate annotation problem.

In this paper, to jointly exploit relationships among facial landmarks and head pose from both real and synthetic images, we propose a unified framework named coupled cascade regression (CCR). It first leverages the benefits from the cascade regression framework to separately estimate the landmark positions and head pose (see Fig. 1). In particular, by leveraging the cascade regression framework, we iteratively learn two regressors to update landmark locations and 3D face model parameters separately. Since two-dimensional (2D) facial landmark and 3D pose parameters are related, which can be derived from the projection model,[13–15] we also link them through the 3D face projection model to further refine the estimations. As a result, CCR can exploit the interactions between landmark points and head pose to perform simultaneous landmark detection and head pose estimation. To overcome the limited head pose variations in training data, we extend the training samples by synthesizing the facial images.

In summary, our main contributions are threefold. (1) The proposed CCR can simultaneously detect the landmark locations and estimate the head pose. Furthermore, CCR can improve the performances of both tasks. (2) CCR combines both learning-based and model-based approaches. (3) CCR learns from both real and synthetic images.

## 2 Related Work

A large number of methods have been proposed to tackle the problem of facial landmark detection. These methods can be classified into three categories: constrained local model (CLM), holistic methods, and cascade regression-based methods. CLM approaches[22,23] estimate landmark locations based on fitting local appearance models and a global shape model. Holistic methods[24] learn models that can capture the global appearance and face shape information. Recently, a cascade regression framework, especially along with deep models,[13,25–30] has achieved impressive performances, where it learns regression models at each cascade level to iteratively map the discriminative local features around landmarks to the ground truth

landmark positions. Dapogny et al.[29] propose to incorporate the landmark-wise attention maps and intermediate supervisions into the deep cascade convolution network for landmark detection. Wan et al.[30] propose to integrate a deep regression module and a deocclusion module into the cascade regression framework to tackle the problem of landmark detection under occlusion.

Head pose estimation approaches in computer vision can be generally categorized into learning-based and model-based methods. The learning-based methods try to map the extracted appearance features [e.g., histogram of oriented gradients and scale-invariant feature transform (SIFT)] to 3D head poses (e.g., pitch, yaw, and roll).[11,20,31,32] Drouard et al.[33] propose to learn a mixture of linear Gaussian regression models, which can simultaneously map high-dimensional features to low-dimensional head pose parameters and face bounding box shifts. Xu et al.[11] propose a deep multitask learning framework for head pose estimation. Head and face regions are jointly fed into the deep model for the tasks of head pose estimation and face verification, respectively. The model-based methods estimate the head pose by linking the 2D observation and 3D face model through the computer vision projection model. These methods typically first perform 2D landmark detection, followed by fitting the 3D face model to estimate the head pose. Sung et al.[34] estimate head pose by fitting related cylindrical models. Some other effective methods[18] decode head pose from 3D face model parameters, which are estimated by minimizing the misalignment error between the ground truth locations and projected locations of the 3D face model on the image plane.

There are a few approaches that simultaneously detect landmarks and predict the head pose. By treating the 3D morphable model (3DMM) and corresponding 3D face model parameters as a representation of the 2D facial shape, some research works[13,14,35,36] proposed to iteratively update the 3D face model parameters for facial landmark detection and pose parameters' estimation. Typically, 3D face model parameters consist of projection matrix parameters and 3D deformable parameters. The head pose can be extracted from the projection matrix parameters. Tulyakov and Sebe[37] propose to estimate the 3D facial shape from a single image based on cascade regression framework using the shape invariant features. They extract head pose from the defined face basis vector, which denotes the rotation angle of the face. Zhu et al.[13] propose to use convolutional neural networks (CNN) as the regressor in the cascaded framework to learn the mapping between the local appearance and 3D face model parameters. They introduce a face profiling method to synthesize the facial images with large poses to tackle the problem of limited training images. Xu and Kakadiaris[38] propose a method called JFA to coarsely estimate the head pose by the global facial CNN features, and further iteratively update the facial landmarks by local CNN feature on the basis of a cascade regression framework. They do not leverage the power of 3D face projection model, which can link the head pose and facial landmark locations. Tran and Liu[14] propose a framework to learn a nonlinear 3DMM from a large set of in-the-wild face images. They introduce a network encoder to estimate the projection, shape, lighting, and albedo parameters. However, their methods need many face images.

Other recent deep-learning-based methods employ multitask learning for landmark detection.[17,27,39–43] Wang et al.[42] propose a recurrent convolutional shape regression method to jointly learn all the shape updates at all cascade levels by using a recurrent network with a gated recurrent unit. Honari et al.[27] propose to learn from the limited number of face images with accurate landmark annotations and other face images with class labels in a semisupervised scheme. Yin et al. propose a unified framework termed JASRNet for simultaneously performing image super-resolution and landmark localization in tiny faces.[17] They introduce a deep shared encoder in the network to capture complementary information for both tasks to boost their performance.[17]

A large amount of training data is crucial for successful supervised learning. With the increasing progress in computer graphics and virtual reality, it is becoming possible to learn from the virtual images that cover the distribution of targets in appearances in the real world.[21,44–46] Learning-by-synthesis is becoming an increasingly active topic for landmark detection and head pose estimation.[13,19,20,47–49] Feng et al.[19] generate synthesized faces by a 3D morphable face model for training a more robust facial landmark detector. Larumbe et al.[48] propose to learn from a synthetic head pose dataset and experimental results validate that learning-by-synthesis can enhance the performance for head pose estimation. Zhu et al.[13] augment the training data with a face profiling method based on 3D image meshing and rotation. Wang et al.[20]

propose to learn a coarse-to-fine deep model from synthetic data for head pose estimation. It coarsely classifies the input image into four categories followed by fine regression of head pose parameters. Yin et al.[49] propose a generative adversarial networks (GAN) based face frontalization method for face data augmentation. The attention mechanism is introduced in the generator and discriminator to learn a richer feature representation for frontal face generation. Another work presented in Ref. 50 also proposes a GAN-based learning scheme to leverage unlabeled data. In addition, they introduce a LaplaceKL loss to optimize the deep model.[50]

Different from most of the aforementioned methods, which address the problem of landmark detection and head pose estimation by learning or modeling from images separately and independently, our proposed CCR combines the learning with a projection model at each cascade level. Joint learning for landmark detection and 3D face model parameters' estimation based on their consistency allows for simultaneous landmark prediction and head pose estimation. In addition, the combination of learning with a projection model can enhance the performance on landmark detection and head pose estimation, which makes CCR achieve preferable results. Motivated by profiling faces for image synthesis[13] and our previous work of learning-by-synthesis,[21,44] we also propose to learn from synthesized facial images for landmark detection and head pose estimation.

## 3 Preliminary

### 3.1 *General Cascade Regression*

Landmark detection aims to estimate the 2D face shape $\mathbf{x} \in \Re^{2\cdot N}$ of $N$ landmarks in an image $\mathbf{I}$. Before we introduce our proposed CCR for simultaneous landmark detection and head pose estimation, we introduce the general cascade regression framework for 2D landmark detection.[16,51] Given an image $\mathbf{I}$, the objective function for face alignment can be formulated as follows:

$$f(\mathbf{x}) = \frac{1}{2}\|\Phi(\mathbf{x}, \mathbf{I}) - \Phi(\mathbf{x}^*, \mathbf{I})\|^2, \tag{1}$$

where $\mathbf{x}$ are the landmark locations, $\mathbf{x}^*$ are the ground truth locations, and $\Phi(\mathbf{x}, \mathbf{I})$ are the local appearance features around the current landmark locations. To solve the optimization problem $\tilde{\mathbf{x}} = \arg\min_{\mathbf{x}} f(\mathbf{x})$, we further take derivation on the Taylor expansion of Eq. (1) and set it to zero, thus we can estimate the landmark updates $\Delta\mathbf{x}$ approximately by a regression model, denoted by $R$. In particular, at iteration $t$ for the cascade regression framework, the landmark updates can be estimated iteratively by

$$\Delta\mathbf{x}^t = R^t[\Phi(\mathbf{x}^{t-1}, \mathbf{I})]. \tag{2}$$

The general cascade regression framework is summarized in Algorithm 1. In general, the idea of cascade regression for face alignment is to learn cascade regressor $R^t$ at iteration $t$, where the regressor can iteratively predict the landmark location updates $\Delta\mathbf{x}^t$ based on local appearance features $\Phi(\mathbf{x}^{t-1}, \mathbf{I})$ extracted around current locations $\mathbf{x}^{t-1}$. Then, the new landmark locations $\mathbf{x}^t$ for the next iteration can be estimated by adding the predicted updates $\Delta\mathbf{x}^t$ to the current landmark locations $\mathbf{x}^{t-1}$. This repeats until it converges or to the maximal iteration $T$.

### 3.2 *Facial Image Synthesis*

Inspired by the key idea of capturing information from images without any manual labor annotation,[21,50] where they learn models from synthetic data or unlabeled real data through an adversarial learning framework, we propose to learn models from both real and synthesized facial images. In this subsection, we give the details for synthetic image generations. Built upon the method presented in Ref. 13, facial images with various head poses are generated.

The key idea is to simulate face images with the help of 3D information. First, different from the conventional facial images synthesis methods[52] that ignores the external face region, the context information beyond the face region is exploited by following the 3D image meshing method.[13,53] Specifically, the depth of the face region and external region is conducted followed

**Algorithm 1** General cascade regression framework for face alignment.

**Input:**

Give the image $\mathbf{I}$. Facial landmark locations $\mathbf{x}^0$ are initialized by mean face.

**Do cascade regression:**

**for** t=1,2,...,T or convergence **do**

Estimate the landmark location updates $\mathbf{x}^t$ given the current landmark locations $\mathbf{x}^{t-1}$,

$$\Delta\mathbf{x}^t = R^t(\Phi(\mathbf{x}^{t-1}, \mathbf{I})),$$

Update the landmark locations,

$$\mathbf{x}^t = \mathbf{x}^{t-1} + \Delta\mathbf{x}^t.$$

**end for**

**Output:**

Landmark locations $\mathbf{x}^T$.

by fitting a 3DMM through the multifeatures framework.[54] It is worth noticing that the 2D ground truth landmarks with respect to the face image offer a solid constraint to ensure the accuracy of fitting. In addition, some anchors beyond face region are marked so that the 2D image can be projected back to a 3D space by the triangulation method (see the step of 3D meshing in Fig. 2). After constructing the depth image, we can rotate it out of plane to synthesize facial images with different head poses (see Fig. 2). Since the rotation may lead to distortion of external face regions, the background anchors need to be adjusted by iteratively optimizing an equation list about the relative positions.[13] Because the 3D meshing method can keep the original facial landmark annotations, we can acquire the ground truth after facial image synthesis. As shown in Fig. 2, we can generate facial images with facial landmark locations, which explicitly enlarge the training data with head pose variations.

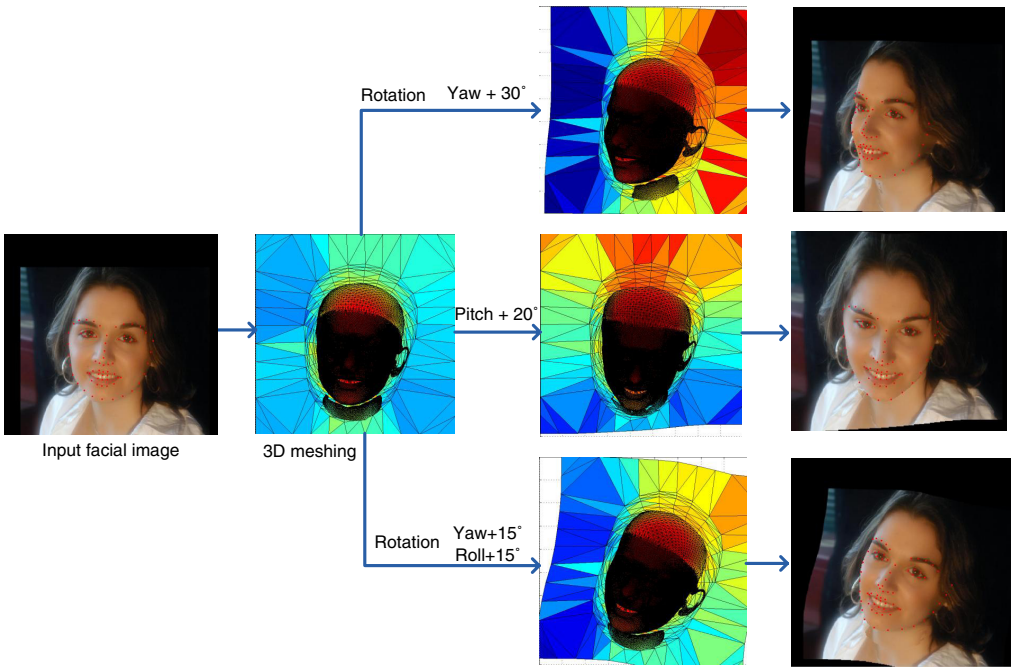



**Fig. 2** Framework of 3D image meshing and rotation for facial image synthesis with accurate facial landmark annotation.

## 4 Coupled Cascade Regression

In this paper, we propose a united framework named CCR for simultaneous landmark detection and head pose estimation. For clarity, all related key variables and model functions are defined, as shown in Table 1.

The overall framework of the proposed CCR method that estimates the target facial shape and head pose is shown in Fig. 1 and Algorithm 2. In particular, we apply a cascaded regression framework to jointly achieve two tasks; one is for facial landmark detection and the other is for 3D face pose parameters' estimation. For facial landmark detection, we use a general cascade regression framework to directly predict the updates of landmark location $\mathbf{x}$ at each iteration. For pose estimation, we propose to estimate the updates of 3D face model parameters $\mathbf{p}$, which encode head pose and deformable information. To capture the coupled relationship between $\mathbf{x}$ and $\mathbf{p}$, we further link the 3D face model parameters $\mathbf{p}$ and 2D landmark locations $\mathbf{x}$ through the projection model $g(\cdot)$ to ensure their consistency.

As shown in Fig. 1 and Algorithm 2, by leveraging the cascade regression framework, we jointly perform two tasks of landmark detection and 3D face model parameter estimation by learning the cascade regression models separately. It begins with an initialization of mean 3D face model parameters $\mathbf{p}^0$, which can map face to mean landmark locations $\mathbf{x}^0$ with the projection model. At each iteration, it updates the landmark locations and 3D face model parameters based on the learned cascade regressors $R^t$ and $Q^t$, respectively. It further applies the projection model through a gradient descent method to ensure consistency between the landmark positions and face pose parameters. As a result, CCR can effectively learn the cascaded regressors by incorporating local appearance and the projection model to improve the performance of both landmark detection and head pose estimation. In the following, we describe our proposed CCR method in detail.

### 4.1 *Update the Landmark Locations*

Various regression functions such as linear regression model, random forest, and neural network can be applied in the general cascade regression framework. One of the most widely used models is the linear regression model,[16] which is effective and efficient. In this work, we also use the linear model in a general cascade regression framework for landmark detection. Hence, the regression model $R^t$ for estimating the updates of landmark locations can be formulated as

**Table 1** The definitions of models and variables in the proposed CCR.

| Variables/models | Definitions |
|---|---|
| $\mathbf{x}$ | Landmark location |
| $\mathbf{x}^*$ | Ground truth of the landmark location |
| $\mathbf{p}$ | 3D face model parameters |
| $R$ | Regression model for estimating landmark location updates |
| $W_R$ | The parameters (weights) of model $R$ |
| $Q$ | Regression model for estimating 3D model parameter updates |
| $W_Q$ | The parameters (weights) of model $Q$ |
| $\mathbf{r}$, $\mathbf{b}$ | The bias parameters |
| $(\cdot)^t$ | The state of variables/models with respect to the $t$'th cascade iteration |
| $g(\cdot)$ | 3D face projection model |
| $\Phi$ | The local appearance features (i.e., SIFT in this work) |

**Algorithm 2** Proposed CCR for simultaneous landmark detection and head pose estimation.

**Input:**

Give the facial image $\mathbf{I}$. Key point locations $\mathbf{x}^0$ are initialized by projection model with mean 3D face model parameters $\mathbf{p}^0$.

**Do cascade regression:**

**for** t=1,...,T **do**

*step 1:* Estimate the landmark location updates $\Delta\mathbf{x}^t$ given the current landmark locations $\mathbf{x}^{t-1}$,

$$\Delta\mathbf{x}^t = R(\Phi(\mathbf{x}^{t-1}, \mathbf{I})),$$

Update the landmark locations,

$$\mathbf{x}^{t*} = \mathbf{x}^{t-1} + \Delta\mathbf{x}^t,$$

*step 2:* Estimate the parameter updates $\Delta\mathbf{p}^t$ given the current landmark locations $\mathbf{x}^{t-1}$,

$$\Delta\mathbf{p}^t = Q(\Phi(\mathbf{x}^{t-1}, \mathbf{I})),$$

Update the parameters,

$$\mathbf{p}^{t*} = \mathbf{p}^{t-1} + \Delta\mathbf{p}^t,$$

*step 3:* Make $\mathbf{p}^t$ and $\mathbf{x}^t$ consistent to be corresponding to each other based on projection model, with initialization of $\mathbf{p}^{t*}$ and $\mathbf{x}^{t*}$,

$$\mathbf{x}^t, \mathbf{p}^t = \arg\min_{\mathbf{x},\mathbf{p}} \varepsilon(\mathbf{x}, \mathbf{p}).$$

**end for**

**Output:**

Acquire the landmark locations $\mathbf{x}^T$ and head pose from $\mathbf{p}^T$.

$$\begin{aligned}\Delta\mathbf{x}^t &= R^t[\Phi(\mathbf{x}^{t-1}, \mathbf{I})] \\ &= W_R^t\Phi(\mathbf{x}^{t-1}, \mathbf{I}) + \mathbf{r}^t,\end{aligned} \tag{3}$$

where $W_R^t$ and $\mathbf{r}^t$ are the weights and bias parameters, respectively, for regressor $R^t$ at iteration $t$.

For the regression model training at cascade level $t$, given $K$ training facial images with ground truth landmark locations $\mathbf{x}^*$, the ground truth updates of landmark location $\Delta\mathbf{x}_i^{t,*}$ in the $i$'th image can be calculated by subtracting the current landmark locations $\mathbf{x}_i^{t-1}$ from the ground truth landmark locations $\mathbf{x}_i^*$. In addition, given the training images with estimated key point locations $\mathbf{x}^{t-1}$, the local appearance features $\Phi(\mathbf{x}_i^{t-1}, I_i)$ of $i$'th image can be extracted. Then, we can learn the corresponding model parameters $W_R^t$ and $\mathbf{r}^t$ by the least-square formation with closed form solution:

$$W_R^{t^*}, \mathbf{r}^{t^*} = \arg\min_{W_R^t, \mathbf{r}^t} \sum_{i=1}^{K} \|\Delta\mathbf{x}_i^{t,*} - W_R^t\Phi(\mathbf{x}_i^{t-1}, I_i) - \mathbf{r}_i^t\|^2. \tag{4}$$

For landmark location prediction at iteration $t$, given previous estimated landmark positions $\mathbf{x}^{t-1}$ and the learned regressor with parameters $W_R^t$ and $\mathbf{r}^t$, we can estimate the updates $\Delta\mathbf{x}^t$ of landmark locations by Eq. (3). Then, we can acquire the new locations by adding the updates $\Delta\mathbf{x}^t$ to previously estimated landmark locations $\mathbf{x}^{t-1}$.

### 4.2 *Update the Head Pose and 3D Morphable Model Parameters*

After updating the landmark locations, we further apply another regressor to update the 3DMM based on the local appearance features at the same cascade iteration.

#### 4.2.1 *3D morphable model*

Based on our previous work,[15,36] we use 3DMM to represent the 3D facial shape. 3DMM is defined as the shape model with a dense mesh and we can further simplify it by the 3D landmark points at the corresponding mesh points. 3DMM can be described with a mean 3D shape and principal component analysis (PCA) space linearly with deformable parameters as

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{B}\mathbf{q}, \tag{5}$$

where $\mathbf{s}$ is the 3D shape of $N$ landmarks in the head coordinate system denoted by $\mathbf{s} = \{x_1, y_1, z_1, \ldots, x_N, y_N, z_N\}$, $\bar{\mathbf{s}}$ is the mean 3D shape, $\mathbf{B}$ are the learned PCA bases, and $\mathbf{q}$ denotes the nonrigid deformable parameters that capture the shape variations.

A 2D face shape is a projection of a 3D face shape $\mathbf{s}$ onto the image plane. Here, we use weak perspective projection and the landmark locations in image plane can be calculated by projection function as

$$\mathbf{x} = g(\mathbf{p}) = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{M}_{2\times 3}(\bar{\mathbf{s}} + \mathbf{B}\mathbf{q}) + \mathbf{t}, \tag{6}$$

where $\lambda_1$ and $\lambda_2$ are scaling factors in row and column directions, respectively, $\mathbf{M}_{2\times 3}$ is the first two rows of rotation matrix $\mathbf{M}$, which is encoded by head pose angle (pitch $\alpha$, yaw $\beta$, and roll $\gamma$), $\mathbf{q}$ denotes the deformable parameters, and $\mathbf{t} = (t_1, t_2)^T$ is the 2D translation vector. We use $\mathbf{p} = \{\lambda_1, \lambda_2, \alpha, \beta, \gamma, t_1, t_2, \mathbf{q}\}$ to represent all the parameters of the model. Hence, we can accurately acquire the 2D landmark locations if we have the 3D face model parameters $\mathbf{p}$ with corresponding 3DMM projection function $g(\cdot)$. On the other hand, we can estimate the 3D face model parameters from 2D landmark locations based on nonlinear optimization. In this paper, the ground truth 3D face model parameters are not provided in the synthesized facial images. Since the synthesized images retain the accurate 2D landmark locations, we generate the related 3D face model parameters $\mathbf{p}$ by our previous work.[15]

#### 4.2.2 *Update the 3D model parameters*

Similar to updating the landmark locations at iteration $t$, we also use a linear regression model to predict the updates of 3D face model parameters by

$$\Delta\mathbf{p}^t = W_Q^t \Phi[g(\mathbf{p}^{t-1}), \mathbf{I}] + \mathbf{b}^t, \tag{7}$$

where $g(\mathbf{p}^{t-1})$ denotes the projected 2D landmark locations $\mathbf{x}^{t-1,p}$ based on current estimated 3D face model parameters $\mathbf{p}^{t-1}$ by Eq. (6), $\Phi[g(\mathbf{p}^{t-1}), \mathbf{I}]$ is the extracted image features at $\mathbf{x}^{t-1,p}$, and $W_Q^t$ and $\mathbf{b}^t$ are the weights and bias parameters, respectively, of cascade regressor $Q^t$ at iteration $t$.

Given $K$ training images with ground truth 3D face model parameters $\mathbf{p}^*$, we can learn the model parameters $W_Q^t$ and $\mathbf{b}^t$ by standard least-square formation with a closed-form solution as

$$W_Q^{t^*}, \mathbf{b}^{t^*} = \arg\min_{W_Q^t, \mathbf{b}^t} \sum_{i=1}^{K} \|\Delta\mathbf{p}_i^{t,*} - W_Q^t \Phi(g(\mathbf{p}_i^{t-1}), I_i) - \mathbf{b}_i^t\|^2. \tag{8}$$

Given the $i$'th training image with current estimated model parameters $\mathbf{p}_i^{t-1}$, we can estimate the current landmark locations $\mathbf{x}_i^{t-1,p}$ from Eq. (6). Hence, the related local appearance features $\Phi(g(\mathbf{p}_i^{t-1}), I_i)$ can be calculated. We can acquire the ground truth updates of 3D face model

parameters by $\Delta\mathbf{p}_i^{t,*} = \mathbf{p}^* - \mathbf{p}_i^{t-1}$. It should be noted that we initialize the landmark location $\mathbf{x}^{0,p}$ based on the mean 3D face model parameters $\overline{\mathbf{p}}$ from the training data by $g(\overline{\mathbf{p}})$.

For the testing at iteration $t$, given the projection model $g(\cdot)$ with current 3DMM parameters $\mathbf{p}^{t-1}$ and learned regressor with parameters $W_Q^t$ and $\mathbf{b}^t$, we calculate the updates $\Delta\mathbf{p}^t$ of 3D face model parameters using the learned regressor by Eq. (7). The new 3D face model parameters can be estimated by $\mathbf{p}^{t*} = \mathbf{p}^{t-1} + \Delta\mathbf{p}^t$.

### 4.3 *Coupling*

We have already described the methods to separately estimate the landmark locations and 3D face model parameters using the cascade regression framework. However, they are performed independently, ignoring the joint relationship among 2D facial landmark detection and 3D head pose estimation. To exploit their dependencies, we further propose to add a third step to make them consistent with each other. We capture the relationship between the first and second steps through the projection model. The whole framework is called CCR, which is summarized in Algorithm 2. At each cascade level $t$, we learn one regressor $R^t$ for updating the landmark locations and learn another regressor $Q^t$ for updating the 3DMM parameters. We further refine $\mathbf{p}^{t*}$ and $\mathbf{x}^{t*}$ through the projection model in Eq. (6) to ensure their consistency. As a result, they are updated to new values of $\mathbf{p}^t$ and $\mathbf{x}^t$ for the next iteration.

Specifically, for the third step, we define the objective function as Euclidean distance of landmark locations from two tasks as

$$\begin{aligned} J(\mathbf{x},\mathbf{p}) &= \arg\min_{\mathbf{x},\mathbf{p}} \varepsilon(\mathbf{x},\mathbf{p}) \\ &= \arg\min_{\mathbf{x},\mathbf{p}} \frac{1}{2}[\mathbf{x} - g(\mathbf{p})]^2, \end{aligned} \tag{9}$$

where both $\mathbf{x}$ and $\mathbf{p}$ are unknowns and $g(\mathbf{p})$ is the projection function, as shown in Eq. (6). For iteration $t+1$, we solve this optimization problem by the gradient descent method with initialization of $\mathbf{p}^{t*}$ and $\mathbf{x}^{t*}$. It is worth noticing that $\mathbf{p}^{t*}$ and $\mathbf{x}^{t*}$ are calculated by the learned regressors $Q^t$ and $R^t$, respectively. We alternatively update 3D face model parameters $\mathbf{p}$ and the location $\mathbf{x}$ as

$$\begin{aligned} \mathbf{p}^{t+1} &= \mathbf{p}^{t*} - \eta \frac{\partial \varepsilon(\mathbf{x}^{t*},\mathbf{p})}{\partial \mathbf{p}}\bigg|_{\mathbf{p}^{t*}} \\ \mathbf{x}^{t+1} &= \mathbf{x}^{t*} - \xi \frac{\partial \varepsilon(\mathbf{x},\mathbf{p}^{t*})}{\partial \mathbf{x}}\bigg|_{\mathbf{x}^{t*}}, \end{aligned} \tag{10}$$

where $\eta$ and $\xi$ are the learning rate parameters. As formulated in Eq. (6), we can easily decode the head pose of "pitch," "yaw," and "roll" from the estimated 3D face model parameters $\mathbf{p}$.

## 5 Experiments

In this section, we describe the implementation details first, followed by some discussions and analyses about the experimental results.

### 5.1 *Implementation Details*

#### 5.1.1 *Datasets*

We conduct experimental comparisons with the state-of-the-art methods on landmark detections and pose estimations on four benchmark datasets, including LFPW,[55] Helen,[56] 300-W,[57] and Boston University (BU).[58] The LFPW dataset contains 811 training images and 224 testing images. The Helen dataset consists of 2000 training images and 330 testing images. 300-W combines images from AFW,[59] LFPW, Helen, and XM2VTS. In particular, we follow the same

protocol in Refs. 60 and 61 so that the whole set of AFW, training images of LFPW, and Helen are used for training, which consists of 3148 faces in total. The IBUG (challenging), testing images of LFPW and Helen (common), are used for testing, which consists of 689 faces in total (full). We follow the same protocol in Ref. 18 on the BU dataset to evaluate the head pose estimation.

#### 5.1.2 *Evaluation metrics*

For landmark detection, we perform evaluations on the 68 facial points. We evaluate the performance by standard mean error, which is the distance between the ground truth landmark locations and estimated locations normalized by the interpupil distance. For head pose estimation, we evaluate the performance by mean absolute error, which is the absolute difference between the ground truth angles and estimated angles in degree.

#### 5.1.3 *Parameters setting*

In Algorithm 2, the number of iterations $T$ for the cascade regression model is set to 6. When learning the PCA bases using the annotations provided by Ref. 37, we retain 95% of the energy and result in 34 dimensions of deformable parameters of $\mathbf{q}$. We generate the related 3D face model parameters for learning by minimizing the misalignment errors for all facial landmarks. All experiments are conducted with nonoptimized MATLAB® codes on a standard PC, which has an Intel i5 3.47 GHz CPU. It takes around 110 ms per-frame on the LFPW dataset.

### 5.2 *Experiments*

#### 5.2.1 *Experimental comparisons*

For fair comparisons of facial landmark detection, we conduct experiments with similar linear cascade regression model-based methods. To demonstrate the effectiveness of our proposed method, we also perform two baselines. We first discard the third step in Algorithm 2 without capturing any joint relationship using the cascade regression framework, where we call this method the CR landmark. In addition, we conduct another experiment where we do cascade regression to update 3D face model parameters only (only the second step in Algorithm 2). We call it CR-3DMP. We finally do the comparisons using the CCR. In addition, to demonstrate the effectiveness of learning from virtual images synthesized from training samples, we call it CCR synthesis. An example of detection results from these three methods is shown in Fig. 3. From the sample image shown in Fig. 3, we can see that the proposed CCR framework performs better over the two baselines that directly update the landmark (CR landmark) or deformable

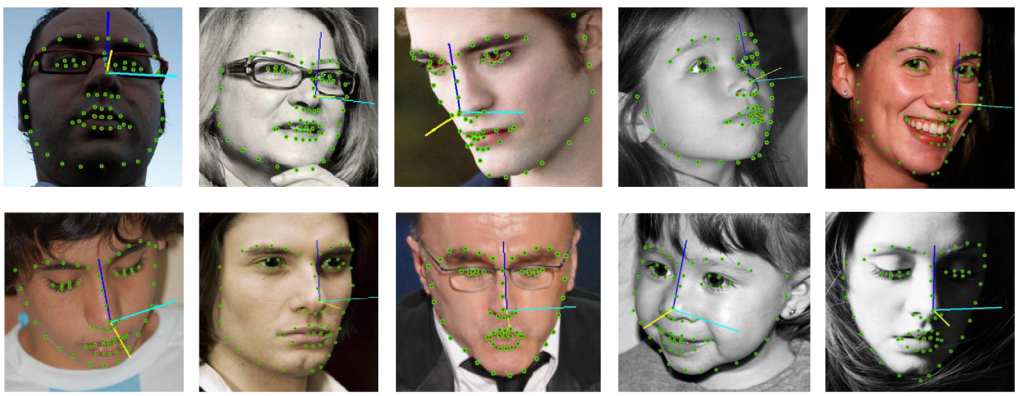


**Fig. 3** Example of detected landmarks by (a) CR landmark, (b) CR-3DMP, (c) CCR, and (d) CCR synthesis. The red points denote the detected points, the white points denote the ground truth, and the green bounding box is the outer boundary of annotated landmarks. The normalized errors are 9.18, 10.80, 6.95, and 5.44 for CR landmark, CR-3DMR, CCR, and CCR synthesis, respectively.

model parameters (CR-3DMP). In addition, the introduced learning from facial image synthesis can improve the performance by effective augmentation of existing training samples. We further report the normalized interpupil errors in Tables 2 and 3. Some image examples are shown in Figs. 4 and 5.

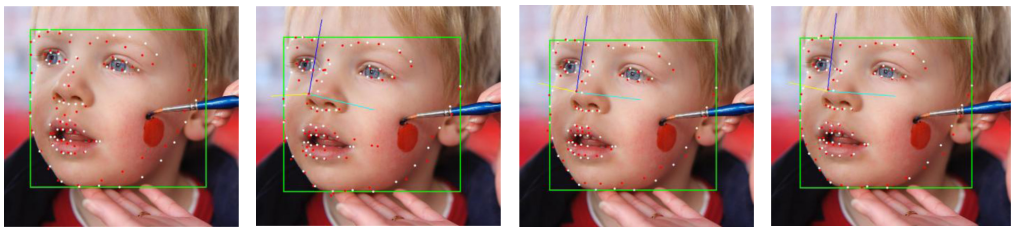
For the experiments on LFPW and Helen, we train the CCR on the corresponding training data provided in LFPW and Helen separately, where LFPW contains 811 training images and

**Table 2** Landmark detection comparison of normalized error (%) on Helen and LFPW datasets, with best result highlighted.

| Method | LFPW dataset | Helen dataset |
|---|---|---|
| RCPR[62] | 6.56 | 5.93 |
| Supervised descent method (SDM)[16] | 5.67 | **5.50** |
| GN-DPM[63] | 5.92 | 5.69 |
| Joint cascade[25] | 5.62 | 5.52 |
| CR landmark (baseline) | 5.75 | 5.88 |
| CR-3DMP (baseline) | 6.15 | 6.41 |
| CCR (proposed) | 5.69 | 5.73 |
| CCR synthesis (proposed) | **5.49** | 5.52 |

**Table 3** Landmark detection comparison of averaged error (%) normalized by interpupil distance on 300-W dataset.

| Method | Common | Challenging | Full |
|---|---|---|---|
| RCPR[62] | 6.18 | 17.26 | 8.35 |
| ESR[51] | **5.28** | 17.00 | 7.58 |
| SDM[16] | 5.57 | 15.40 | 7.50 |
| Joint cascade[25] | 5.54 | 13.75 | 7.15 |
| CR landmark (baseline) | 5.75 | 15.56 | 7.64 |
| CR-3DMP (baseline) | 6.47 | 18.08 | 8.78 |
| CCR (proposed) | 5.55 | 14.01 | 7.24 |
| CCR synthesis (proposed) | 5.47 | **12.63** | **6.90** |



(a) CR-Landmark Error: 9.18 (b) CR-3DMP Error: 10.80 (c) CCR Error: 6.95 (d) CCR-synthesis Error: 5.44

**Fig. 4** Qualitative results of our proposed CCR method for 300-W dataset (best view in color).

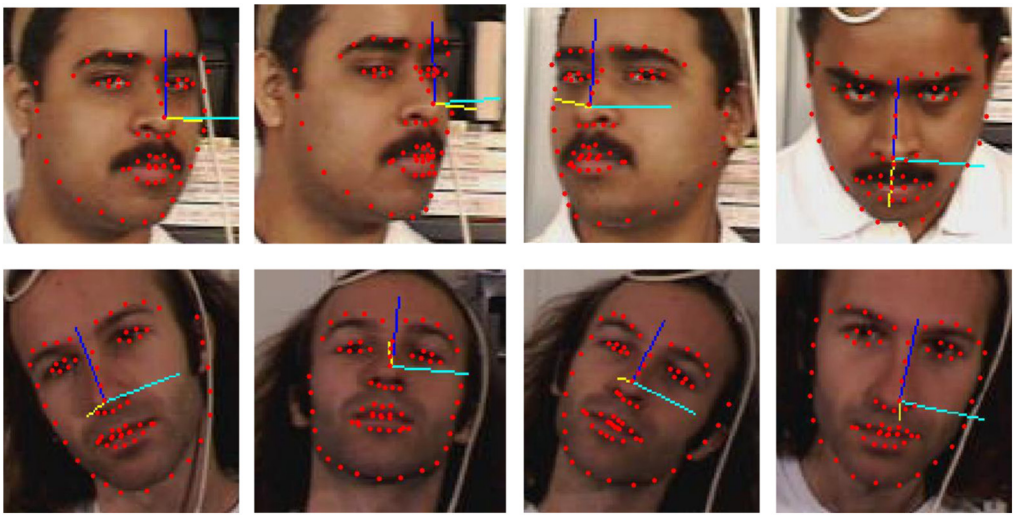


**Fig. 5** Qualitative results of our proposed CCR method for BU dataset (best view in color).

Helen provides 2000 training images, respectively. For the CCR synthesis, the training samples are enlarged, as described in the subsection of implementation details. We compare the proposed method with similar cascade regression framework-based methods. From Table 2, our proposed method can achieve preferable results. SDM[16] performs slightly better than CCR due to limited head poses in the Helen dataset. Our proposed CCR outperforms the baselines and experiments demonstrate the effectiveness of learning with the face projection model. In addition, learning CCR from augmented facial images can further improve the detection performance, where we synthesize the training samples by profiling the facial image.

Results for another large and challenging dataset named 300-W are shown in Table 3. From Table 3, CCR can achieve the best results on the full subset of 300-W with a normalized error of 6.90. For the challenging subset, learning CCR from synthetic data significantly outperforms a similar work[25] by 8.1%. Effective augmentation of facial images with a larger variation of appearance with respect to different head poses further enhances the performance of learning-based landmark detection methods. Upon further investigation, the challenging subset contains many facial images with large head poses and we tackle this problem by introducing the CCR of learning with 3D deformable model. By a combination of cascade regression framework and 3D deformable model, we can leverage the power of learning and model to simultaneously perform the two tasks of landmark detection and head pose estimation.

To demonstrate the effectiveness of CCR for head pose estimation, we conduct comparisons with similar work on the BU dataset. Experimental results are listed in Table 4. We perform detection and tracking on the BU dataset and test on each frame of the videos in BU. Some detection image examples are shown in Fig. 5. It should be noted that the ground truth of 3D face model parameters (head pose and deformable parameters) is generated by the

**Table 4** Head pose estimation comparison of mean absolute error on BU dataset.

| Method | Pitch | Yaw | Roll | Average |
|---|---|---|---|---|
| AAM + Cylindrical[34] | 5.6 | 5.4 | 3.1 | 4.7 |
| SDM + Deformable[18] | 4.3 | 6.2 | 3.2 | 4.6 |
| 3D CLM[64] | 6.0 | **3.9** | 3.7 | 4.5 |
| Joint cascade + deformable[25] | 5.3 | 4.9 | 3.1 | 4.4 |
| CCR (proposed) | 4.8 | 5.1 | 3.3 | 4.4 |
| CCR_synthesis (proposed) | **4.3** | 5.1 | 3.2 | **4.2** |

model-based method. We compare our model for head pose with other model-based approaches. As shown in Table 4, by learning CCR from the combination of real and synthetic images, CCR can achieve the best results with an average mean absolute error of 4.2 in degrees. Different from conventional methods that sequentially perform landmark detection and head pose estimation, CCR achieves two tasks in one step based on shape indexed features. It is worth noticing that we generate the 3D face model parameters by previous work[15] as the ground truth, which can further be improved for head pose estimation with accurate 3D face model parameters.

### 5.2.2 *Further analysis*

As mentioned before and shown in Fig. 3, CCR significantly improves the performance on the landmark detection and head pose estimation by exploiting the inherent dependencies among the landmarks and projection models. We further investigate how the CCR incorporates the local appearance and projection model in a cascade regression framework to enhance the performance on landmark detection. We take one iteration in cascaded regression for example, as shown in Fig. 6. We can see that it converges to a more accurate position (especially for the red eyebrow related landmarks estimated by "step 1," the green cheek points estimated by "step 2") after making them consistent by leveraging the projection model.

Since we leverage the cascade regression framework, we further analyze the convergence performance of CCR. As shown in Fig. 7, CCR improves the performances of facial landmark detection and pose estimation quickly at the first three iterations. Slightly different from a conventional cascade regression framework, where the max iteration is empirically set to 4, it converges to the optimal at the sixth iteration in this work. Hence, the iteration is set to 6. Our proposed framework can easily be generalized for other tasks like eye center and gaze


**Fig. 6** An example of outputs from one iteration in a cascade regression framework. In Algorithm 2: (a) red points are the outputs of "step 1" and green points are the outputs of "step 2" and (b) the output of "step 3."

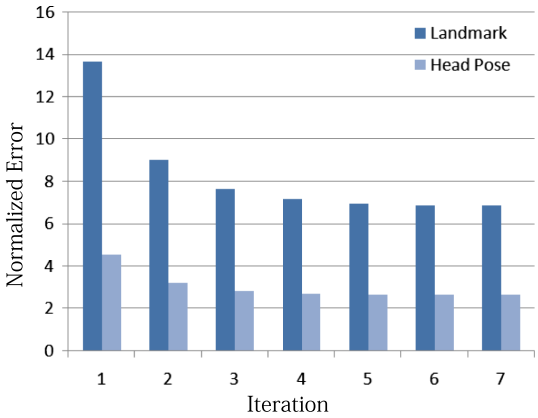



**Fig. 7** Results of CCR at each cascaded iteration on 300-W database. *Y* coordinate denotes the normalized interpupil error (%) for landmark detection, and mean absolute error in degree for head pose estimation. The changes are small after the fourth iteration.

estimation. It is worth noticing that we only compare with a similar cascade regression framework with linear regression models for facial landmark detection. The linear regression model in this framework can also be a nonlinear deep model, which will be part of our future work.

## 6 Conclusion

In this paper, we propose a unified framework of CCR for simultaneous head pose estimation and landmark detection. Different from conventional cascade regression-based methods for landmark detection or learning-based pose estimation, the proposed CCR performs cascade regression to update face model parameters and landmark locations separately, followed by capturing the coupled relation of 3D head pose and 2D landmark locations through the projection model at each cascade level. As a result, CCR incorporates local appearance and 3D face model to achieve simultaneous landmark localization and pose estimation. Thorough experimental results on benchmark datasets also demonstrate that it can further improve both tasks. In addition, effective data augmentation by facial image synthesis for CCR training can improve the performance of detection and head pose estimation.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 61806198 and the Key Research and Development Program 2020 of Guangzhou. Part of the work was performed when the first author was a joint-supervision PhD student at Rensselaer Polytechnic Institute.

## References

1. Y. Su, X. Gao, and X. C. Yin, "Fast alignment for sparse representation based face recognition," *Pattern Recognit.* **68**, 211–221 (2017).
2. H. Li et al., "Cascaded face alignment via intimacy definition feature," *J. Electron. Imaging* **26**(5), 053024 (2017).
3. W. Deng et al., "From one to many: pose-aware metric learning for single-sample face recognition," *Pattern Recognit.* **77**, 426–437 (2018).
4. J. P. Robinson et al., "Face recognition: too bias, or not too bias?" arXiv:2002.06483 (2020).
5. Y. Wu and Q. Ji, "Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3400–3408 (2016).
6. W. Li et al., "EAC-Net: deep nets with enhancing and cropping for facial action unit detection," *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(11), 2583–2596 (2018).
7. J. Roth, Y. Tong, and X. Liu, "Adaptive 3D face reconstruction from unconstrained photo collections," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 4197–4206 (2016).
8. L. Jiang, X.-J. Wu, and J. Kittler, "Pose-invariant three-dimensional face reconstruction," *J. Electron. Imaging* **28**(5), 053003 (2019).
9. A. Narayanan, R. M. Kaimal, and K. Bijlani, "Estimation of driver head yaw angle using a generic geometric model," *IEEE Trans. Intell. Transp. Syst.* **17**(12), 3446–3460 (2016).
10. E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: a survey," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(4), 607–626 (2009).
11. L. Xu, J. Chen, and Y. Gan, "Head pose estimation using deep multitask learning," *J. Electron. Imaging* **25**(1), 013029 (2019).
12. N. Wang et al., "Facial feature point detection: a comprehensive survey," *Neurocomputing* **275**, 50–65 (2018).
13. X. Zhu et al., "Face alignment in full pose range: a 3D total solution," *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(1), 78–92 (2019).
14. L. Tran and X. Liu, "On learning 3D face morphable model from in-the-wild images," *IEEE Trans. Pattern Anal. Mach. Intell.* (2019).

15. C. Gou et al., "Shape augmented regression for 3D face alignment," *Lect. Notes Comput. Sci.* **9914**, 604–615 (2016).
16. X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, IEEE, pp. 532–539 (2013).
17. Y. Yin et al., "Joint super-resolution and alignment of tiny faces," arXiv:1911.08566 (2019).
18. F. Vicente et al., "Driver gaze tracking and eyes off the road detection system," *IEEE Trans. Intell. Transp. Syst.* **16**(4), 2014–2027 (2015).
19. Z. H. Feng et al., "Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting," *IEEE Trans. Image Process.* **24**(11), 3425–3440 (2015).
20. Y. Wang et al., "A deep coarse-to-fine network for head pose estimation from synthetic data," *Pattern Recognit.* **94**, 196–206 (2019).
21. C. Gou et al., "Cascade learning from adversarial synthetic images for accurate pupil detection," *Pattern Recognit.* **88**, 584–594 (2019).
22. J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *IEEE 12th Int. Conf. Comput. Vision*, IEEE, pp. 1034–1041 (2009).
23. C. Lindner et al., "Robust and accurate shape model matching using random forest regression-voting," *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(9), 1862–1874 (2015).
24. S. Lucey et al., "Fourier Lucas-Kanade algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(6), 1383–1396 (2013).
25. Y. Wu, C. Gou, and Q. Ji, "Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 5719–5728 (2017).
26. Y. Wu, S. K. Shah, and I. A. Kakadiaris, "GoDP: globally optimized dual pathway deep network architecture for facial landmark localization in-the-wild," *Image Vision Comput.* **73**, 1–16 (2018).
27. S. Honari et al., "Improving landmark localization with semi-supervised learning," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1546–1555 (2018).
28. Z. Tong et al., "Robust facial landmark localization based on two-stage cascaded pose regression," in *Proc. AAAI Conf. Artif. Intell.*, Vol. 33, pp. 10055–10056 (2019).
29. A. Dapogny, K. Bailly, and M. Cord, "DeCaFA: deep convolutional cascade for face alignment in the wild," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 6893–6901 (2019).
30. J. Wan et al., "Robust face alignment by cascaded regression and de-occlusion," *Neural Networks* **123**, 261–272 (2020).
31. X. Geng and Y. Xia, "Head pose estimation based on multivariate label distribution," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1837–1842 (2014).
32. M. Patacchiola and A. Cangelosi, "Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods," *Pattern Recognit.* **71**, 132–143 (2017).
33. V. Drouard et al., "Robust head-pose estimation based on partially-latent mixture of linear regression," arXiv:1603.09732 (2016).
34. J. Sung, T. Kanade, and D. Kim, "Pose robust face tracking by combining active appearance models and cylinder head models," *Int. J. Comput. Vision* **80**(2), 260–274 (2008).
35. A. Jourabloo and X. Liu, "Large-pose face alignment via CNN-based dense 3D model fitting," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.* (2016).
36. C. Gou et al., "Coupled cascade regression for simultaneous facial landmark detection and head pose estimation," in *IEEE Int. Conf. Image Process.*, pp. 2906–2910 (2017).
37. S. Tulyakov and N. Sebe, "Regressing a 3D face shape from a single image," in *IEEE Int. Conf. Comput. Vision*, IEEE, pp. 3748–3755 (2015).
38. X. Xu and I. A. Kakadiaris, "Joint head pose estimation and face alignment framework using global and local CNN features," in *IEEE Int. Conf. Autom. Face and Gesture Recognit.*, pp. 642–649 (2017).
39. Z. Zhang et al., "Facial landmark detection by deep multi-task learning," *Lect. Notes Comput. Sci.* **8694**, 94–108 (2014).
40. R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(1), 121–135 (2019).

41. K. Zhang et al., "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process Lett.* **23**(10), 1499–1503 (2016).
42. W. Wang, S. Tulyakov, and N. Sebe, "Recurrent convolutional shape regression," *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(11), 2569–2582 (2018).
43. Y. Zhao et al., "Joint face alignment and segmentation via deep multi-task learning," *Multimedia Tools Appl.* **78**(10), 13131–13148 (2019).
44. C. Gou et al., "Learning-by-synthesis for accurate eye detection," in *Int. Conf. Pattern Recognit.*, pp. 3362–3367 (2017).
45. A. Dosovitskiy et al., "Learning to generate chairs, tables and cars with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 692–705 (2017).
46. G. Hu et al., "Frankenstein: learning deep face representations using small data," *IEEE Trans. Image Process.* **27**(1), 293–303 (2018).
47. M. Ariz, A. Villanueva, and R. Cabeza, "A novel 2D/3D database with automatic face annotation for head tracking and pose estimation," *Comput. Vision Image Understanding* **148**, 201–210 (2016).
48. A. Larumbe et al., "Improved strategies for HPE employing learning-by-synthesis approaches," in *IEEE Int. Conf. Comput. Vision Workshop*, pp. 1545–1554 (2017).
49. Y. Yin et al., "Dual-attention GAN for large-pose face frontalization," arXiv:2002.07227 (2020).
50. J. P. Robinson et al., "Laplace landmark localization," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 10103–10112 (2019).
51. X. Cao et al., "Face alignment by explicit shape regression," *Int. J. Comput. Vision* **107**(2), 177–190 (2014).
52. I. Masi et al., "Pose-aware face recognition in the wild," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 4838–4846 (2016).
53. X. Zhu et al., "High-fidelity pose and expression normalization for face recognition in the wild," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 787–796 (2015).
54. S. Romdhani and T. Vetter, "Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior," in *IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognit.*, IEEE, Vol. 2, pp. 986–993 (2005).
55. P. N. Belhumeur et al., "Localizing parts of faces using a consensus of exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(12), 2930–2940 (2013).
56. V. Le et al., "Interactive facial feature localization," *Lect. Notes Comput. Sci.* **7574** 679–692 (2012).
57. C. Sagonas et al., "300 faces in-the-wild challenge: the first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, pp. 397–403 (2013).
58. M. La Cascia, S. Sclaroff, and V. Athitsos, "Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3D models," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(4), 322–336 (2000).
59. X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, IEEE, pp. 2879–2886 (2012).
60. S. Zhu et al., "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 4998–5006 (2015).
61. Z. Zhang et al., "Learning deep representation for face alignment with auxiliary attributes," *IEEE Trans. Pattern Anal. Mach. Intell.* **38**, 918–930 (2016).
62. X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 1513–1520 (2013).
63. G. Tzimiropoulos and M. Pantic, "Gauss-Newton deformable part models for face alignment in-the-wild," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1851–1858 (2014).
64. T. Baltrusaitis, P. Robinson, and L. P. Morency, "3D constrained local model for rigid and non-rigid facial tracking," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2610–2617 (2012).

**Chao Gou** is an assistant professor at Sun Yat-sen University. He received his BS degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2012 and his

PhD from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2017. From September 2015 to January 2017, he was supported by UCAS as a joint-supervision PhD student in Rensselaer Polytechnic Institute, Troy, New York, USA. His research interests include computer vision and machine learning.

**Qiang Ji** is a professor with the Department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute (RPI). He received his PhD in electrical engineering from the University of Washington. He recently served as a director of the Intelligent Systems Laboratory (ISL) at RPI. He is an editor on several related IEEE and international journals, and he has served as a general chair, program chair, technical area chair, and program committee member in numerous international conferences/workshops. His research interests are in computer vision, probabilistic graphical models, information fusion, and their applications in various fields.