International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 29, No. 8 (2015) 1556011 (27 pages)

DOI: 10.1142/S021800141556011X

A Hierarchical Regression Approach for Unconstrained Face Analysis

Yuanyuan Liu*, Jingying Chen†, Cunjie Shan‡,
Zhiming Su§ and Pei Cai¶

National Engineering Research Center for E-Learning
Central China Normal University, Wuhan, P. R. China

Collaborative & Innovative Center
for Educational Technology (CICET)
Wenhua College, Wuhan, P. R. China

*jane19840701@hotmail.com
†chenjy@mail.ccnu.edu.cn
‡shancunjie20@126.com
§happyszm@foxmail.com
¶caipei@mails.ccnu.edu.cn

Received 9 October 2014
Accepted 8 July 2015
Published 23 September 2015

Head pose and facial feature detection are important for face analysis. However, many studies reported good results in constrained environment, the performance could be decreased due to the high variations in facial appearance, poses, illumination, occlusion, expression and make-up. In this paper, we propose a hierarchical regression approach, Dirichlet-tree enhanced random forests (D-RF) for face analysis in unconstrained environment. D-RF introduces Dirichlet-tree probabilistic model into regression RF framework in the hierarchical way to achieve the efficiency and robustness. To eliminate noise influence of unconstrained environment, facial patches extracted from face area are classified as positive or negative facial patches, only positive facial patches are used for face analysis. The proposed hierarchical D-RF works in two iterative procedures. First, coarse head pose is estimated to constrain the facial features detection, then the head pose is updated based on the estimated facial features. Second, the facial feature localization is refined based on the updated head pose. In order to further improve the efficiency and robustness, multiple probabilitic models are learned in leaves of the D-RF, i.e. the patch's classification, the head pose probabilities, the locations of facial points and face deformation models (FDM). Moreover, our algorithm takes a composite weight voting method, where each patch extracted from the image can directly cast a vote for the head pose or each of the facial features. Extensive experiments have been done with different publicly available databases. The experimental results demonstrate that the proposed approach is robust and efficient for head pose and facial feature detection.

Keywords: D-RF; unconstrained face analysis; hierarchical regression; head pose estimation; facial feature detection.

†Corresponding author.

1. Introduction

Human–computer interface is an active research topic in computer vision area.[22,26] Despite recent advances, people still interact with machines through devices like keyboard and mice, which are not part of natural human–computer communication.[12] As people interact by means of many channels, including body posture and facial expression, an important step towards more natural interfaces is the visual analysis of the users movements by the machine.[34,36] Head pose estimation and facial feature detection are important for many applications like natural human–computer interfaces, face recognition, facial expression analysis and visual focus of attention recognition.[1,2,6,8,10,25,29] Most of the existing methods focus on face analysis (e.g. head pose estimation[18,26,31] and facial feature detection[1,6]) in constrained environment, however, face analysis in unconstrained environment remains challenging due to high variations in facial appearance, poses, illuminations, occlusion, expression and make-up.

In recent years, regression Random Forest (RF) is a popular method in computer vision given their capability to handle large training datasets, high generalization power and speed, and easy implementation.[10,13,14,18,40] Some work showed the power of RF in mapping image features to votes in a generalized Hough space[32] or to real-valued functions.[40] Recently, multiclass RF has been proposed in Ref. 18 for real-time head pose recognition from 2D video data and 3D range images.[13,14] Furthermore, Dantone *et al.* proposed a conditional RF to detect facial feature points under various head pose only in the horizontal direction.[10] The accuracy rate reaches 82.3% in natural head poses instead of head pose motion in wide range. Hence, how to detect refined head poses and facial feature points in real time and unconstrained environment remains a problem.

In this work, Dirichlet-tree probabilistic model is introduced into regression RF framework in the hierarchical way to achieve the efficiency and robustness for face analysis. The Dirichlet-tree distribution was proposed by Dennis.[24] It is the distribution over leaf probabilities that result from the prior on branch probabilities. He proved the high accuracy and efficiency of the distribution. Some researchers use a Dirichlet-tree distribution in pose estimation,[22] multi-objects tracking[38] and affective computing.[15] The flowchart of the proposed approach is shown in Fig. 1, which includes two stages for face analysis. First, in order to eliminate the influence of noise

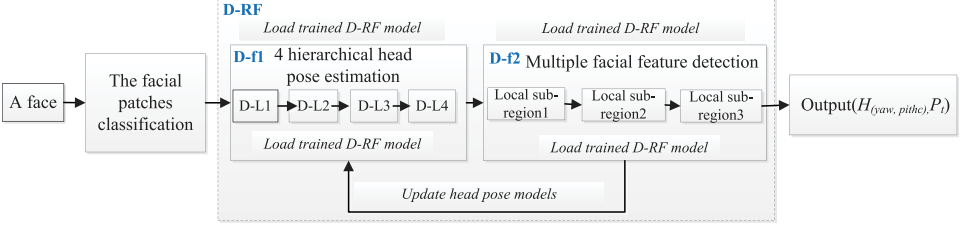


Fig. 1. The flowchart of the proposed approach for face analysis.

in unconstrained environment, facial patches extracted from facial area are classified as positive or negative facial patches using RF and only positive facial patches can be used for face analysis. Then D-RF is proposed to face analysis in the two iterative procedure. As shown in Fig. 1, the trained D-RF model represents the different sub-forest trained by D-RF in the different layer and different iteration. In the first iteration, coarse head poses are initially estimated to constrain facial feature detection in the D-f1 and D-f2 layers of the D-RF, where D-f1 consists of four sub-layers for head pose estimation, i.e. D-L1, D-L2, D-L3 and D-L4. D-L1 and D-L2 are two sub-layers in the horizontal estimation and D-L3 and D-L4 are two sub-layers in the vertical estimation. D-f2 layer is used for multiple facial feature detection within three local sub-regions (i.e. the mouth, nose and eyes sub-regions) under the estimated head pose. In the second iteration, head pose are updated based on the detected facial feature localizations and the refined head pose is used to further improve the accuracy of the facial features using D-RF. The two iterative procedures can obtain high accuracy for unconstrained face analysis rather than cast too much time for running. Hence, two iterations of the D-RF are implemented in this work.

The main contributions of this paper are as follows. We propose a novel approach for face analysis in unconstrained environment based on a hierarchical regression framework. The approach is inspired by our previous conference paper[23] in ICIP2014. Different from our previous work, the hierarchical D-RF is introduced for face analysis in the two iterative procedures. The estimated head pose could provide geometric constraints to the facial features detection, while the detected facial features help to refine the estimated head pose. An achieved multiple-PCA (M-PCA) feature subspace is extracted from positive facial patches to improve efficiency and robustness. Furthermore, multiple probabilitic models are learned in leaves of the D-RF, i.e. the patch's classification, the head pose probabilities, the locations of facial points and face deformation model (FDM) that is defined as the offsets from the centroid of facial patches to tip of the nose under different head poses. Moreover, the composite weighted voting that fuses weighted classification and regression voting is used to vote multiple leaves. Experiments have been carried out to evaluate the performance in terms of coarse and refined head pose estimation, and feature point detection. The results obtained suggest that the approach could estimate the head pose and facial feature locations robustly and efficiently in unconstrained environment.

The rest of this paper is organized as follows: Section 2 discusses relevant work to this study. Section 3 details the hierarchical regression approach for head pose estimation and multiple facial feature detection. In Sec. 4, experiment evaluation and analysis are given. Conclusions and proposals on future work are described in Sec. 5.

2. Related Work

In this section, we highlight three subjects that are the closest to our work, which include RF, head pose estimation and facial feature detection.

RF is a popular method in computer vision given their capability to handle large training datasets, high generalization power and speed, and easy implementation.[10,13,14,18,30,40] It has emerged as a powerful and versatile method successful in real-time human pose estimation, object detection, facial point detection and action recognition.[40] Multi-class RF has been proposed for the real-time determination of head pose from 2D video data.[18] In Ref. 36, a conditional RF has been used for real-time body pose estimation from depth data. A conditional RF also has been proposed to estimate facial features point under various head pose only in the horizontal direction in Ref. 10. It has been shown that the body pose can be estimated more efficiently using regression than using classification forests.[16] Criminisi *et al.*[9] used RF regression to vote for the positions of the sides of bounding boxes around organs in CT images. More information about RF and their application in computer vision can be found in Ref. 36.

Head pose estimation. Head pose estimation is important in many human machine interfaces. Head orientation is related to a person's direction of attention, it can present useful information about what the person is paying attention to. Different methods have been developed for two types of image data, i.e. 2D images or depth data. Methods on depth data can provide high accuracy, however they require special hardware (e.g. expensive depth sensor) and need more computations.[37] In this study, we focus on 2D images. Lots of work have been done on head pose estimation for 2D images, some based on local facial features, while others based on the globe image.[35] Local approaches usually estimate head pose from a set of facial features such as eyes, eyebrows and lips. Pose can be obtained using a different set of 5 points (the inner and outer corners of each eye, and the tip of the nose).[26]

Global approaches use an entire image of face to estimate head pose.[33] The principal advantage of the approaches is that only the face needs to be located. Osadchy *et al.*[28] instead use a convolutional network to learn the mapping for head pose estimation and can achieve real-time performance for the problem. Recently, multiclass RF have been proposed in Ref. 18 for real-time head pose estimation and 3D range images.[13,14,32] Dantone *et al.* proposed conditional RF to estimate head pose under various conditions only in the horizontal direction. The accuracy rate reaches 72.3% with five yaw angle classes.

Facial feature detection. Facial feature detection is often the first step for many applications such as face recognition, facial expression analysis and visual focus of attention recognition.[11] Earlier works can be classified into two categories, depending on whether they use holistic or local features. Holistic methods, e.g. Active Appearance Models,[1,6] use the texture over the whole face region to fit a linear generative model to a test image. Such algorithms suffer from lighting changes, modeling complexity, and a bias towards the average face. Moreover, these methods perform poorly on unseen identities and deal poorly with low resolution images.[1,6] Active Shape Models use a linear Point Distribution Model (PDM) constructed from aligned training shapes, driven to fit a new image thanks to simple models of the

appearance along profiles centered on each landmark.[8] But it is sensitive to head pose rotation on wide range. Dantone *et al.* proposed conditional RFs to detect multiple facial feature points in different head poses.[10] Their algorithm is the most accurate approach up-to-date in the literature, capable of precise localizations even in uncontrolled image conditions, like the ones present in the Labeled Face Parts in the Wild database.[3] Yang and Patras proposed a sieving regression forest voting for facial feature detection in Ref. 39, which achieved the state-of-the-art results on two public challenging datasets with face images in the wild, without resorting to explicit shape models.

In general, most of the existing work focuses on facial analysis which includes head pose estimation and facial feature localization from constrained environment. However, unconstrained face analysis still remains a challenge due to the high variations in facial appearance, poses, illumination, occlusion, expression and make-up.

3. D-RF for Unconstrained Facial Analysis

In Sec. 3.1, we first summarize an overview of the proposed approach for unconstrained facial analysis. In Sec. 3.2, we show how to model the D-RF from the RF framework. In Sec. 3.3, a positive/negative facial patch extraction and classification method is presented. Then, details on coarse head pose estimation and facial feature detection using D-RF are given in Secs. 3.4 and 3.5, respectively. Finally, refined head pose and facial features are obtained using D-RF iteratively in Sec. 3.6.

3.1. *Overview of the approach*

In the work, we propose a hierarchical D-RF framework for unconstrained face analysis in the iterative way. The framework of the D-RF is shown in Fig. 2. The general RF is an ensemble approach in which several tree predictors are combined together to obtain high performance for classification or regression (see Fig. 2(a)). Each tree in the forest is independently generated with random samples selected from the whole data set. The Dirichlet-tree is the distribution over leaf probabilities

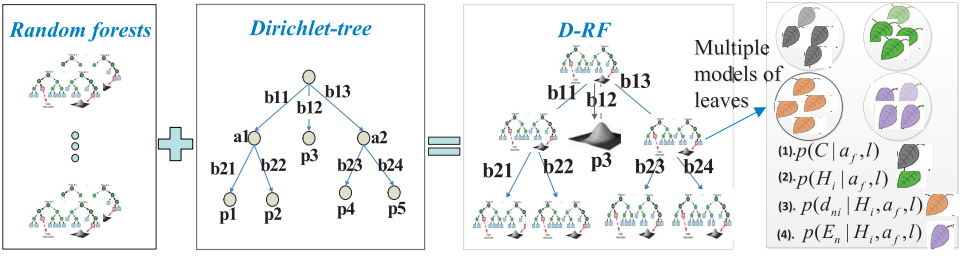


Fig. 2. The proposed D-RF and multiple models in leaves of the D-RF. (a) RFs, (b) the general Dirichel-tree distribution, (c) the D-RF framework and (d) the multiple models in leaves of the D-RF.

$[p_1 \ldots p_i]$ that results from this prior node probabilities $[a_1, a_2, \ldots, a_k]$ on branch probabilities b_{ji},[22] where i is the number of a leaf, k is the number of a prior node, j is the layer of a branch as shown in Fig. 2(b). In order to enhance efficiency and accuracy, Dirichlet-tree distribution is introduced into RF framework as D-RF.[22,23] The node in each sub-layer of D-RF is a regression learning procedure, so the whole D-RF is a hierarchical regression learning based on a tree structure. It is noted that each child node in the sub-layer of the D-RF is related to his parent. Hence, the D-RF only computes the probabilities of the relative trees in the child layer instead of all trees' probabilities in the forest. Therefore, D-RF can provide high accuracy and efficiency. Meanwhile, multiple leaf models of the D-RF for our tasks are shown in Fig. 2(d), which include a patch's classification probability, a head pose probability, a probabilistic regression model for the locations of the base facial points and a FDM model.

To address this face analysis problem in the framework of the proposed approach, we cast it as a joint probability estimation problem and tackle it using the powerful D-RF. Specifically, we can formulate it as follows, i.e.

$$(H_{\text{yaw,pitch}}, P_t) = \arg\max_{H,P} p(H_{\text{yaw,pitch}}, P_t|\Omega, C), \tag{1}$$

where Ω is the corresponding facial feature space, C is the class label of the facial patch, $H_{\text{yaw,pitch}}$, P_t represent head poses and facial feature positions.

Figure 3 gives the overview of the approach. First, in order to eliminate the influence of noise in unconstrained environment, positive/negative patches have been extracted and classified within the face area at the top layer of the hierarchical regression approach (see Fig. 3(a)). Specially, we propose a two iterative procedures for unconstrained face analysis as Figs. 3(b) and 3(c). We first get an initial estimation of the head pose and facial feature localization using a hierarchical D-RF from positive facial patches. The coarse head poses have been estimated at the top layer, which consists of four sub-layers for 25 head poses in the horizontal and

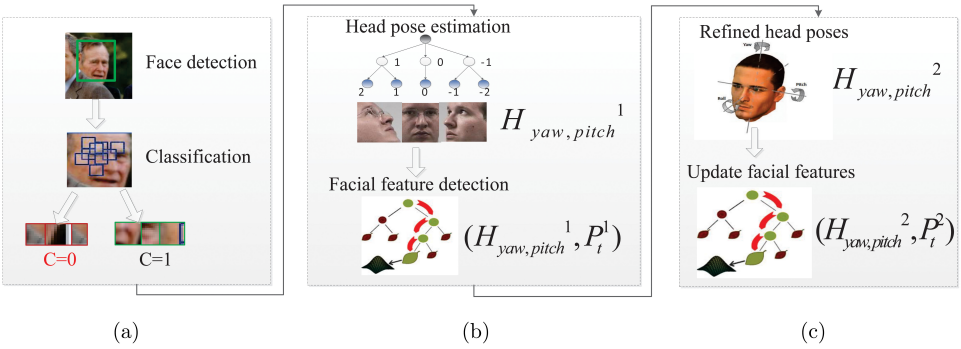


Fig. 3. The iterative procedures of the proposed approach. (a) Facial patch classification, (b) the first iteration-Initialization and (c) the second iteration-Update.

vertical directions. Subsequently, at the bottom layer, under 25 cascaded head poses and FDMs, the posterior probability of facial feature positions can be obtained by D-RF in different local sub-regions, i.e.

$$\begin{aligned} H^1_{\text{yaw,pitch}} &= \arg\max_H p(H_{\text{yaw,pitch}} \,|\, \Omega, C), \\ P^1_t &= \arg\max_{P_t} p(P_t \,|\, H^1_{\text{yaw,pitch}}, \Omega, C). \end{aligned} \tag{2}$$

For the second iteration, as shown in Fig. 3(c), the shape-related geometric features and configuration can be extracted from the previously detected facial feature positions. Refined head poses are updated based on the detected facial feature localizations and the refined head poses are used to further improve the accuracy of the facial features using D-RF. As demonstrated in Ref. 26, the shape related geometric features are more robust to lighting changes and occlusion for poses. Therefore, the refined positions of facial features could be updated and formulated as,

$$P^2_t = \arg\max_{P_t} p(P_t \,|\, H^2_{\text{yaw,pitch}}, P^1_t, \Omega, C). \tag{3}$$

3.2. *D-RF*

Each tree $T = \{T_t\}$ in the D-RF is built and selected randomly from a different subset of the training images. From each facial area, we extract a set of facial patches $\{P_i = \{\Omega_i, C_i, H_i, D_i \,|\, S_j\}\}$. The S_j represents the learned probabilitic model using the prior layer of the D-RF, Ω_i represents texture feature subspace described in Sec. 3.4.1, $C_i = \{0, 1\}$ represent the patch class label, only the patch with $C_i = 1$ can be used for face analysis, H_i represents the annotated head pose parameters, D_i represents geometry features based on coordinates of facial features points.

We define a patch comparison feature as simple binary tests φ, similar to Refs. 10, 13 and 18,

$$|R_1|^{-1} \sum_{k \in R_1} \Omega(k) - |R_2|^{-1} \sum_{k \in R_2} \Omega(k) > \tau, \tag{4}$$

where R_1 and R_2 are two random rectangles within the facial patches, $\Omega(i)$ is the feature space extracted from the P_i, τ is a threshold, and k is the pixel within the rectangles.

The training of a sub-forest in each sub-layer in the D-RF is given below:

(1) Divide the set of patches P into two subsets P_L and P_R for each φ.

$$P_L = \{P \,|\, \varphi < \tau\}, \quad P_R = \{P \,|\, \varphi > \tau\}. \tag{5}$$

(2) Select the splitting candidate φ which maximizes the evaluation function Information Gain (IG).

$$\text{IG} = \arg\max_{\varphi} (H(P \,|\, S_j) - (\omega_L H(P_L \,|\, S_j) + \omega_R H(P_R \,|\, S_j))), \tag{6}$$

where ω_R, ω_L are the ratio between the number of samples in set P_L (arriving to left subset using upper binary tests), set P_R (arriving to right subset using upper binary tests) and set P (total node samples). $H(P|S_j)$ is the defined class uncertainty measure and the entropy of the continuous patch labels under the learned probability model using the prior jth sub-layer of the D-RF.

$$H(P \mid a_j) = -\sum_{i=1}^{N} \frac{\sum_i p(C_i, H_i, D_i \mid S_j, P)}{|P|} \log \left(\frac{\sum_i p(C_i, H_i, D_i \mid S_j, P)}{|P|} \right), \quad (7)$$

where $p(C_i, H_i, D_i \mid S_j, P)$ indicates the probability that the patch P belongs to the head pose class H_i and feature point localization D_i under the prior probability model S_j of the jth sub-layer in the D-RF. D-RF models the Dirichlet-tree probability $p(C_i, H_i, D_i \mid S_j, P)$ in the node and estimates by

$$p(C_i, H_i, D_i \mid S_j, P) = \int \frac{p(C_i, H_i, D_i, S_j|P) \cdot p(S_i|P)}{\sum_{i=1}^{n} p(C_i, H_i, D_i, S_j \mid P) p(S_j \mid P)} d\alpha. \quad (8)$$

(3) Create leaf l when IG is below a predefined threshold or when a maximum depth is reached. Otherwise continue recursively for the two subsets P_L and P_R at the first step. Leaves of the D-RF include four learned models (see Fig. 3): (1) a patch's classification probability, (2) a head pose probabilistity, (3) a probabilistic regression model for the locations of the base facial points and (4) a FDM model.

During testing, multiple privileged probabilities are learned by the hierarchical regression D-RF and then the composite voting model with the estimated state is used. Details on the hierarchical regression approach are given as following for face analysis.

3.3. *Facial patch classification*

In order to locate face area under various poses and conditions, a cascade of boosted classifiers with Haar-like feature[20] has been trained to detect faces with multiple head pose databases. We have achieved the average detection rate of 94.7% in different databases. The detected facial area may include some noise for facial analysis, such as hair, neck and occlusion. In order to eliminate noise, the facial area is segmented into foreground and background areas. The foreground areas include positive and negative facial patches, where the positive facial patches contribute to estimate head pose while the negative facial patches including occlusion or noise may introduce errors for the task. In the work, we segment background areas based on histogram distributions first (see Fig. 4). The process of positive facial patch classification is given in Fig. 5.

Segment background squares: The detected facial area is divided into 6×6 nonoverlapping squares and histogram distributions of the squares are computed as shown in Fig. 4. We analyze the uniformity of histogram distributions of the patches and segment most of the background squares.

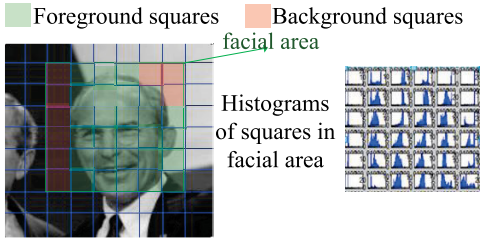


Fig. 4. Foreground and background squares.

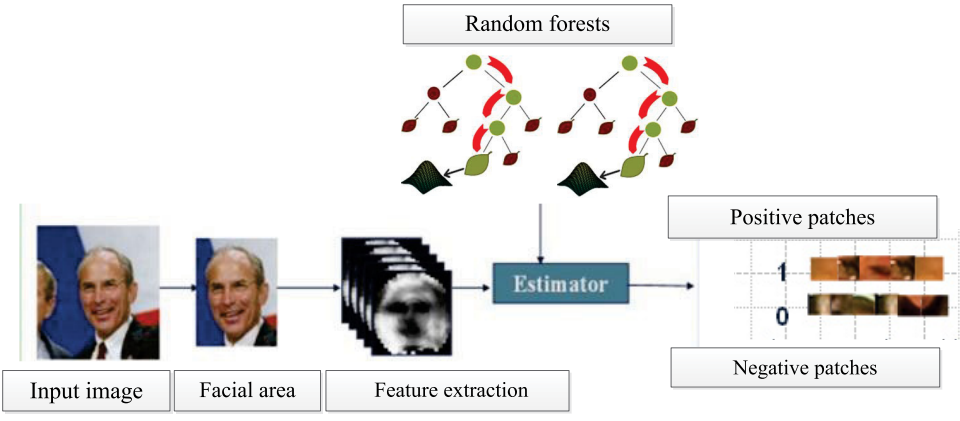


Fig. 5. Positive patches extraction and classification.

Classify positive and negative facial patches: 200 patches are randomly extracted from the rest of areas with background removed, which include positive and negative facial patches. The positive and negative facial patches are classified using RF.[10,13,18] In order to model the random tree, the training set of positive facial patches are labeled as 1 and the negative facial patches are labeled as 0. A tree T grows up by mutiple texture features of the labeled patches. The training and testing using the RF is similar to Refs. 10, 13 and 18. When all test patches arrive at leaves of trees in the forest, we use the probability $p(C \mid l_t(P))$ stored at a leaf to judge whether the test patch belongs to a positive/negative class, where $C = 1$ represents the positive patches while $C = 0$ represents the negative patches. The algorithm diagram is shown in Fig. 5. The facial patch classification probabilities stored in leaves is the privileged information for the next facial analysis. Only the positive facial patch that was labeled $C = 1$ can be used for facial analysis.

3.4. *Head pose estimation*

3.4.1. *Training*

In our head pose estimation setup, a training sample is a combined feature set, multiple texture information built from all the positive facial patches available for

each of the discrete head poses. From each face image, we randomly extract positive facial patches $\{P_i = \{\Omega_i, H_i, D_i \mid S_j, C_i = 1\}\}$. $C_i = 1$ represents the positive facial patch classified from facial foreground area. In the following, we simplify the equations by omitting the decription $C_i = 1$ for reading conveniently. Ω_i is an achieved M-PCA feature sub-space extracted from multiple texture features of positive facial patches. The set of $H_i = \{H_i^1, (H_i^2 \mid H_i^1), (H_i^3 \mid H_i^2, c_i^1), (H_i^4 \mid H_i^3, H_i^2, H_i^1)\}$ is the output space in the Dirichilet-tree distribution, which contains the annotated discrete angles in different sub-layers of the D-RF, where H_i^1 are 3 yaw rotation angles in the first sub-layer of D-f1, $H_i^2 \mid H_i^1$ are 5 yaw angles refined from coarse 3 yaw angles in the second sub-layer, $H_i^3 \mid H_i^2$, H_i^1 are 15 pitch angles under condition of each yaw angle H_i^2 in the third sub-layer, $H_i^4 \mid H_i^3$, H_i^2, H_i^1 are 25 refined angles based on the above annotated angles in the fourth sub-layer of D-f1. D_i represents geometric features based on coordinates of facial features points.

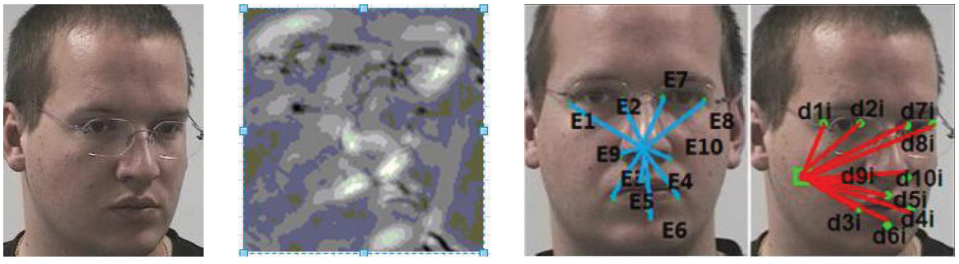
M-PCA texture feature subspace Ω_i: We extract the multiple texture features $x_i^j, j = 1, 2, 3$ in the each facial patch resized of 30×30. x_i^1 contains Gabor feature descriptors at five angles and seven scales with dimension as $35 \times 30 \times 30$, x_i^2 represents Sobel feature descriptors at horizontal and vertical directions with dimension as $2 \times 30 \times 30$ and x_i^3 is the LBPH feature descriptors with dimension as 30×30.

Because of the high dimension in multiple texture features, subspace projection is used to reduce the dimension as well as to extract the most essential information for classification/representation. PCA is a popular subspace representation method. An achieved method M-PCA based on PCA is proposed to extract subspace from texture model. For the clarity of presentation, in the following sections, the data set is denoted as $T_i = \{x_i^j\}, i = 1, 2, \ldots, N_j, j = 1, 2, 3$.

$$\begin{aligned} \sum &= \frac{1}{\sum_{j=1}^{3} \sum_{i=1}^{N_j} N_j} \sum_{j=1}^{3} \sum_{i=1}^{N_j} (x_i^j - \mu)(x_i^j - \mu)^T, \\ \mu &= \frac{1}{\sum_{j=1}^{3} \sum_{i=1}^{N_j} N_j} \sum_{j=1}^{3} \sum_{i=1}^{N_j} x_i^j, \end{aligned} \tag{9}$$

where j is the number of the texture feature channels and N_j is the number of feature dimensions in the jth channel. The principal components are computed by solving the eigenvalue problem: $\sum V = \Lambda V$, where $\wedge = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is the diagonal matrix whose nonzero entries $\lambda_1 > \lambda_2 > \cdots > \lambda_n$ are the corresponding eigenvalues of $\sum$. And V is the matrix whose columns are the corresponding eigenvectors. The reduced PCA subspace is formed by the first P eigenvectors. An achieved M-PCA subspace is projected by the sum of n eigenvectors, which is used to generate the binary testing in the training of the D-RF as Eq. (4) (see Fig. 6(b)).

$$\Omega_i = v_i^T \bullet \sum_{i=1}^{n} (x_i^j - \mu). \tag{10}$$



(a) (b) (c)

Fig. 6. The training data from a combined feature set. (a) Head image, (b) M-PCA from multiple texture models and (c) geometry features.

Geometric features D_i: $D_i = (d_{ni}, E_n \,|\, a_f)$ represents geometric features including Euclidean distances between points and FDM, where d_{ni} represents the N 2D displacement vectors from the centroid of the positive facial patch P_i to each of the facial feature points P_n, and E_n is the FDM that is defined the N vectors of each facial point and the facial center point F (see Fig. 6(c)), where N is the number of facial points. a_f is the sub-forest in different regions of the face in D-RF.

$$d_{ni} = \|P_n - P_i\|_2, \quad E_n = \|G_n - F\|_2, \quad n = 1, 2, \ldots, N. \tag{11}$$

3.4.2. *Testing*

Each positive facial patch is then fed to the trees in the relative sub-layer of the D-RF in D-f1. At each node of a tree, the patches are evaluated according to the stored binary test and passed either to the right or left child until a leaf node is reached. By passing all the positive facial patches down the D-RF for head pose estimation, each positive facial patch P_i ends in a set of leaves L of the different sub-forest of D-RF instead of ending all leaves of the RF. In each leaf l, there are classification probabilities of head pose distributed by a multivariate Gaussian as in Refs. 10 and 18:

$$p(H_i \,|\, l_{S_j}) = N(H_i \,|\, S_j; \overline{H_i \,|\, S_j}, \Sigma_{H_i \,|\, S_j}), \tag{12}$$

where $\overline{H_i \,|\, S_j}$ and $\Sigma_{H_i \,|\, S_j}$ are the mean and covariance matrix of the head pose probabilities of the sub-forest S_j in the jth layer D-RF.

When the patch reaches to the leaves of the sub-forest, the next sub-forest of D-RF should be loaded based on the class decision $C(P)$. The class decision function of the sub-forest is defined as,

$$C(P) = \arg\max_{S_j \in H_i} p(H_i \,|\, S_j, P), \tag{13}$$

where $p(H_i \,|\, S_j, P)$ is the head pose probability of D-RF in condition of sub-forest S_j of the jth sub-layer. The head poses are then obtained by performing adaptive Gaussian mixture model (GMM)[7,22] for voting.

3.4.3. *D-RF for head pose estimation in the first iteration*

In order to obtain initial head pose estimation in the first iteration, the D-RF is trained as described in Sec. 3.2. As shown in Fig. 7, the proposed D-RF consists of four sub-layers for head pose estimation in the D-f1. Since it is difficult to obtain continuous ground truth head pose data from 2D images, we annotate rotation angles as "−1, 0, 1" and "−2, −1, 0, 1, 2" in two layers. "−1, 0, 1" represent yaw rotation angles as "$-90°, 0°, 90°$", and "−2, −1, 0, 1, 2" represent refined yaw rotation angles as "$-90°, -45°, 0°, 45°, 90°$". We store the multivariate adaptive Gaussian distribution in the leaf as define in Eq. (12). The Dirichlet-tree distribution is improved to the D-RF for our task as Fig. 7. Figure 7(a) shows the framework of head pose estimation using D-RF in the horizontal direction, where a is the estimation result in the horizontal direction, and D-L1 and D-L2 are two sub-layers in the horizontal direction in D-f1. Then, five yaw angles can be estimated in the second sub-layer of the D-f1.

After the yaw angles have been classified, pitch angles are estimated under the condition of the classified yaw angles a. Figure 7(b) shows the framework of estimation using D-RF in the vertical direction. And the angle annotations in the vertical direction are similar to horizontal rotation angles. When the patches are sent down through all vertical sub-layers in D-RF, sub-trees are selected from sub-forests in D-L3 and D-L4 sub-layers of the D-RF using Eq. (13). Finally, we can estimate 25 discrete yaw and pitch angles that are stored at leaves of the D-RF, i.e. $\{90°, 90°\}, \{90°, 45°\} \cdots \{0°, 0°\} \cdots \{-45°, -90°\}, \{-90°, -90°\}$.

3.5. *Multiple facial feature detection*

3.5.1. *D-RF for facial features detection in local sub-regions*

After 25 head poses have been estimated, D-RF is also used to detect multiple facial points under the conditions of the estimated head poses in the D-f2. The framework of the D-RF for head pose estimation can be used in principle for predicting

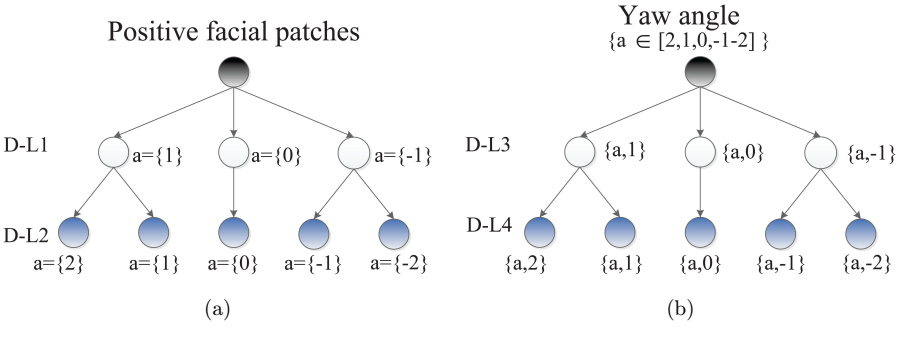


Fig. 7. Head pose estimation by the D-RF in the D-f1 in the first iteration. (a) Head pose estimation in the horizontal direction and (b) head pose estimation in the vertical direction.

Input: positive facial patches from a facial foreground area
Output: the 10 facial feature positions
1. Repeat in the first or second iteration;
2. Loading relative sub-forests of the D-RF based on the head pose parameter in the first or second iteration;
3. Locating the mouth, nose and eyes sub-region using AdaBoost;
4. Selecting patches around each local sub-region;
5. Loading sub-forests in the D-f2 (refer to Figure 1) and testing for facial feature;
6. Voting multiple leaf models to obtain 10 facial feature positions.

Fig. 8. The pseudocode of the D-RF for facial features detection.

continuous parameter of the face, hence the modifications for localizing facial features are straight forward. The set of training facial positive patches is assumed as $\{P_i = \{\Omega_i, D_i \,|\, H_i, C_i = 1\}\}$, as Sec. 3.4.1. $D_i = (d_{ni}, E_n \,|\, a_f)$ represents geometric features including Euclidean distances between points and FDM under the condition of privileged head pose models in different sub-regions of the face. The most discriminative local sub-regions are found using Adaboost with Haar-like features,[20] i.e. the mouth, nose and eyes sub-regions. The algorithm pseudocode is given in Fig. 8. Facial feature detection can be obtained based on estimated head pose by D-RF, and also could help to update refined head pose angles based on their geometry configuration in the second iteration.

Patches from local sub-regions are allowed to predict the locations of local points under estimated head poses. It can avoid a bias towards the average face due to the long distance voting. We reduce the influence of patches from different head poses, face deformation and sub-regions. We measure the confidence pf of a positive patch P for the location of a feature point n by

$$pf \propto \exp\left(\frac{\|\overline{d_{ni}|H_i, a_f}\|^2}{\gamma}\right) \bullet \exp\left(\frac{\|\overline{E_n|H_i, a_f}\|^2}{\gamma}\right). \tag{14}$$

The constant γ is used to control the steepness of this function. A positive patch to vote only is allowed for feature points with a high confidence pf. The probabilistic model of a sub-forest a_f of different sub-regions of D-RF is modeled as

$$p(d_{ni}, E_n \,|\, H_i, a_f, P) = \frac{1}{T_f} \sum_i \sum_{t=1}^{K_f} p(d_{ni}, E_n \,|\, l_{t,a_f,H_i}(P)), \tag{15}$$

where $l_{t,a_f,H_i}(P)$ is the leaf of tree t in a sub-region of D-f2 under head pose H_i. The K_f is the number of trees of a sub-forest. Leaves will be learned if E_n is in the predefined FDM. In a leaf of the sub-forest a_f, $p(d_l(P))$ represents the probability whether the positive facial patch should belong to a feature location

$$p(d_l(P)) = N\left(d_l; \overline{d}_l, \sum_{d_l}\right),$$

$$d_l = d_{ni}, E_n \,|\, l_{t,a_f,H_i}, \tag{16}$$

where $\overline{d_l}$ and $\sum_{d_l}$ denote the mean value and covariance matrix of feature location regression probabilities.

3.5.2. *Composite weighted voting*

We use a composite weighted voting method in a cascaded way. Both classification and regression voting are used to the D-RF. In order to eliminate imbalance of samples, we also store the weight $w_s = P_S/P$ that is defined as the ratio of samples in a subset P_S and full samples P in each single tree of the D-RF. If multiple votes for the feature point P_t is $D_{y_i}(P_t \,|\, a_f, H, S)$ in the patch location y_i, then we set the composite weighted voting model to be given as $V(P_t) \propto K((w_s D_{y_i} - (y_i + \overline{w_s D_{y_i}}))/h_i)$. A Gaussian Kernel K and the bandwidth parameter h_i are given by GMM. In D_{y_i}, $H = \{\text{yaw}, \text{pitch}\}$ represents the classified voting result for head pose, then regression voting can obtain good results by evaluating on sparse positive facial patches, rather than at every positive patch by using all forests. Meanwhile, the competing method is casting GMM voting that is similar to Ref. 23. The 10 facial feature points' locations P_t are obtained by performing mean-shift in $V(P_t)$ for each point t. After facial feature points have been detected, they could help to update head poses in the second iteration.

3.6. *Update head poses and facial feature positions in the second iteration*

After the positions of the facial feature points have been detected, refined head pose angles can be estimated directly from the configuration of these points. Our method for refined head pose estimation uses the extracted feature points (i.e. the inside and outside corners of each eye, the outside corners of the mouth, and the tip of the nose), and the facial symmetry axis is found by connecting a line between the midpoint of the eyes and midpoint of the mouth.[26] Under the assumption that all four eye points are assumed to be coplanar, the yaw angle can be determined from the observable difference in size between the left and right eye due to projective distortion from the known camera parameters. The pitch angle is determined by comparing the distance between the nose tip and the eye-line to an anthropometric model. The update refined head pose could be used to extend facial features using the D-RF in the second iteration (see Fig. 2).

Based on the currently updated head pose $H^2_{\text{yaw,pitch}}$ and the initial facial feature positions P^1_t, the final facial features P^2_t could be improved precisely using updated D-RF in the second iteration. The current sub-forests are updated and loaded based on refined head poses. The testing for facial feature localization is similar to Sec. 3.5. Under refined head poses, we could achieve better performance for the facial feature localization.

4. Experiments

In this section, we thoroughly evaluate the proposed D-RF approach for unconstrained face analysis, i.e. head pose estimation (Sec. 4.1), and facial feature detection (Sec. 4.2) in the two iterative procedures.

The proposed approach have been tested under various experimental conditions, such as Pointing'04 head pose database,[17] LFW database[19] and our lab database (see Fig. 9). The Pointing'04 database consists of 2940 images with different poses and expressions. The LFW database consists of 5749 individual facial images and 13,300 images. The images have been collected in the wild and vary in poses, lighting conditions, resolutions, races, occlusions and make-up. Our lab database includes 3000 images of different students with head poses, expressions and occlusions, and the reference angles have been annotated using the method similar to LFW.[19] When extending the D-RF framework for the purpose of facial feature localization, once again a large dataset of annotated range images of faces is needed. While the LFW is annotated with facial point locations, it has not large variations in range head poses. We additionally annotated 6940 faces taken from Pointing'04 and our lab database with the location of 10 facial feature points shown in Fig. 9. We used Amazon Mechanical Turk, labeling each fiducial point at least three times and taking the mean of the annotations as ground truth.

For evaluation, we divided the databases into a training set and a testing set. The training set consists of 2100 images from Pointing'04 database and 13,000 images

Fig. 9. Examples of images from the datasets, Pointing'04 database (the first row), LFW database (the second row) and our lab database (the third row).

from LWF datasets. The testing set includes the rest of 840 images from Pointing'04 database, 500 images from LFW database and 200 images from our lab database.

4.1. *Head pose estimation*

4.1.1. *Evaluation methodology*

In order to evaluate the proposed approach, estimation accuracy is defined as the ratio of the number of correct estimation samples to the number of testing images. In D-f1, let Y_0, Y_1, Y_2, Y_3, Y_4 be the estimation accuracies of 5 yaw angles and $P_0, P_1, P_2, \ldots, P_n$ be the estimation accuracies of the pitch angles under the correspondent yaw angle. $Q(P_i \,|\, Y_i)$ denotes the final estimation accuracy in leaves of the last sub-layer, which is defined as:

$$Q(P_i \,|\, Y_i) = \frac{\langle P_i, Y_i \rangle \cdot P_i}{\sum_{j=1}^{n} \langle P_j, Y_i \rangle \cdot P_j}. \tag{17}$$

4.1.2. *Experiments in the first iteration*

(1) Training

For the head pose estimation, we trained the trees in three different databases. We fixed some parameters on the basis of empirical observations, e.g. the trees have a maximum depth of 15 and at each node we randomly generate 2000 splitting candidates and 25 thresholds. Other parameters include the number of patches extracted from each image (fixed to 200), the patch size, and the maximum size of the sub-patches defining the areas R_1 and R_2 in Eq. (4).

Figure 10 describes the performance of the algorithm when we varied the size of the facial patches and the number of samples used for training each tree. In Fig. 10(a), the blue, continuous line shows the percentage of estimation accuracy as the patch size, when 300 training images per tree are used. The red, dashed line shows instead the percentage of false estimation rate as the patch size. The plot shows that a minimum size for the patches is critical since small patches cannot capture enough information to reliably predict the head pose. However, there is also a slight performance loss for lager patches. In the case, the trees become more sensitive to occlusions and strong artifacts like holes since the patches cover a large region and overlap more. Having a patch size between 30×30 seems to be a good choice where the patches are discriminative enough to estimate head pose. Figure 10(b) also shows accuracy and mean error rate, this time for 30×30 patches, as a function of the number of training images. It can be noted that the performance increases with more training images in one tree, but it also saturates for training subsets containing more than 300 images.

Each tree grows based on a randomly selected subset of 300 images. Each image in the dataset is manually annotated with one out of the 25 head pose labels. Subforests in different sub-layers of D-RF have been trained independently from the first

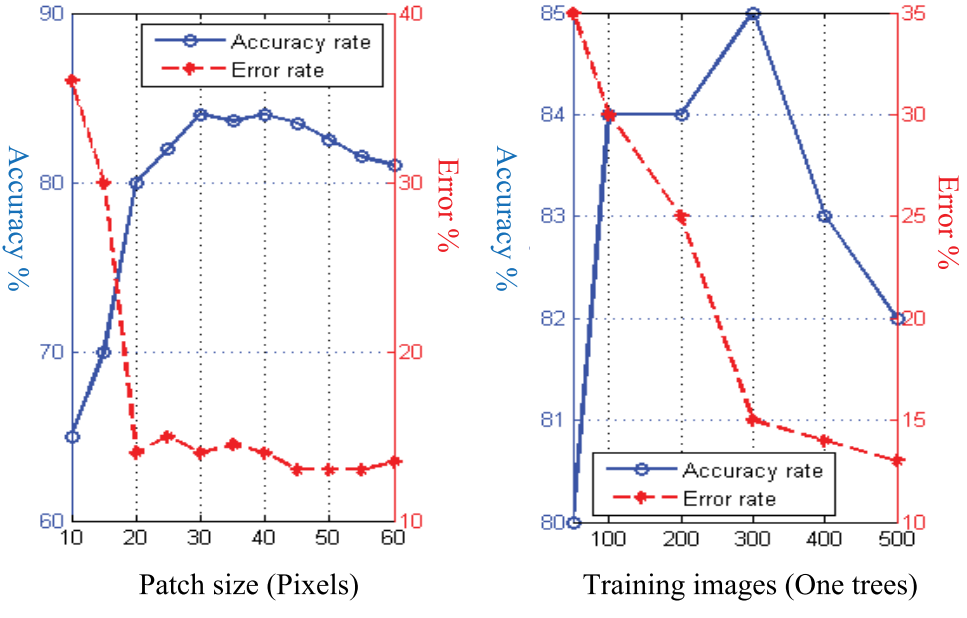


Fig. 10. (Color online) The performance as patch size and training images. (a) Estimation accuracy depending on the patch size (when using 300 training samples), overlaid to the mean error rate. (b) Estimation accuracy and mean error rate depending on the number of training data (for 30×30 patches).

sub-layer to the fourth sub-layer. There are 15 trees in the first sub-layer D-L1, 15 trees in the second sub-layer D-L2 and five trees in each head pose of the prior sub-layer. In the third sub-layer D-L3, we trained 15 trees in total, where three trees in each head pose of the second sub-layer. In the fourth sub-layer D-L4, we trained 30 trees in total, where two trees in each head pose under the prior sub-layers.

(2) Testing

Testing parameters include the training parameters of RF, the Dirichlet-tree layer numbers, and the adaptive GMM parameters. Unless stated otherwise, those parameters in sub-forests were similar to training in all of our experiments. Other parameters are automatically estimated during testing from a validation set generated. Firstly, the faces extracted by our trained Adaboost detector with Haar-features have been normalized to 125×125 pixels. Then, we densely extract positive facial patches from a facial foreground area and sent them down through all sub-layers in the D-f1 in the hierarchical way.

Figure 11 shows accuracies using different sub-layers of the D-RF for 25 head pose classes estimation in the D-f1 of the first iteration. L0 represents the average accuracy of 25 head pose classes using the original RF. While L1 to L4 represent the accuracies of 25 head pose classes using hierarchical 1 to 4 sub-layers of the D-RF.

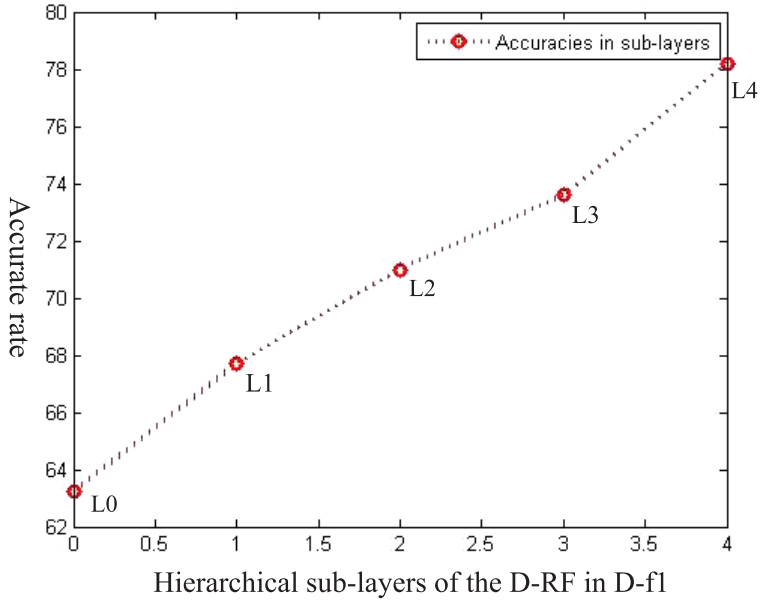


Fig. 11. Accurate comparisons in different sub-layers of the D-RF in D-f1.

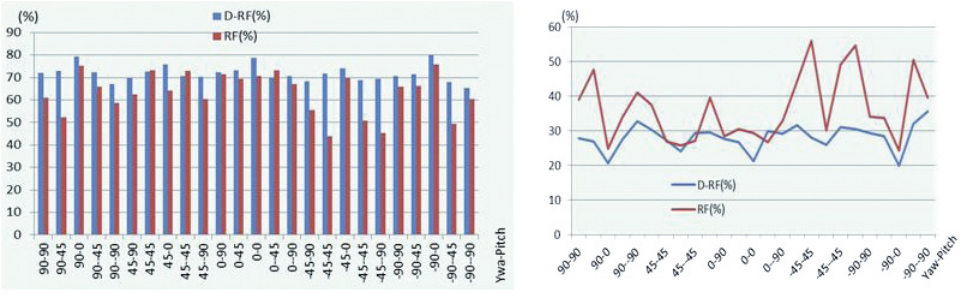
L1 and L2 represent the estimated average accuracies of 25 head pose classes using only one sub-layer (i.e. D-L1) and two sub-layers (i.e. D-L1 and D-L2) in D-RF, respectively. L3 and L4 represent the estimated average accuracies of 25 head pose classes using three sub-layers (i.e. D-L1, D-L2 and D-L3) and four sub-layers (i.e. D-L1, D-L2, D-L3 and D-L4) in D-RF, respectively. As shown in Fig. 11, the final accuracy of original RF reaches to 63.23%, and the proposed approach improves the accuracy with the introduction of the different sub-layers of the Dirichlet-tree. The optimal estimation accuracy is 71.83% using four sub-layers of the D-RF.

(3) Performance comparison

In order to evaluate the contributions of using the D-RF in the D-f1, we compare our method with original RF.[4] Figure 12(a) shows the experiment results, where the blue bars represent estimation accuracies using D-RF and the red bars represent estimation accuracies using RF. Additionally, the mean error of each head pose is shown in Fig. 12(b). The average accuracies of the D-RF and RF are 71.83% and 62.23%, respectively. The D-RF provides higher average accuracy and lower mean error than RF in the horizontal and vertical directions.

4.1.3. *Refined head pose estimation in the second iteration*

To show the generalization capability of our approach, we evaluate the accurate rates of head pose estimation in the first and second iteration through our algorithm on the Pointing'04 database. As shown in Fig. 13, the performance of head pose



(a) (b)

Fig. 12. (Color online) The comparison of the D-RF and RF in the D-f1 of the first iteration. (a) Accuracy and (b) Mean error.

estimation is improved through updated algorithm based on facial features configuration, where the green bar represents the average accuracy in the second iteration by our iterative approach, the orange bar represents the average accuracy in the initial head pose estimation by D-RF in the first iteration. The average accuracy of refined head pose reached 85.9% in the second iteration. Compared to the initial head pose estimation, the performance of refined head pose has higher accuracy.

We have also extensively compared the proposed approach with other state-of-the-art algorithms, i.e. initial D-RF, RF,[4] SVM multi-class.[27] Initial D-RF means the

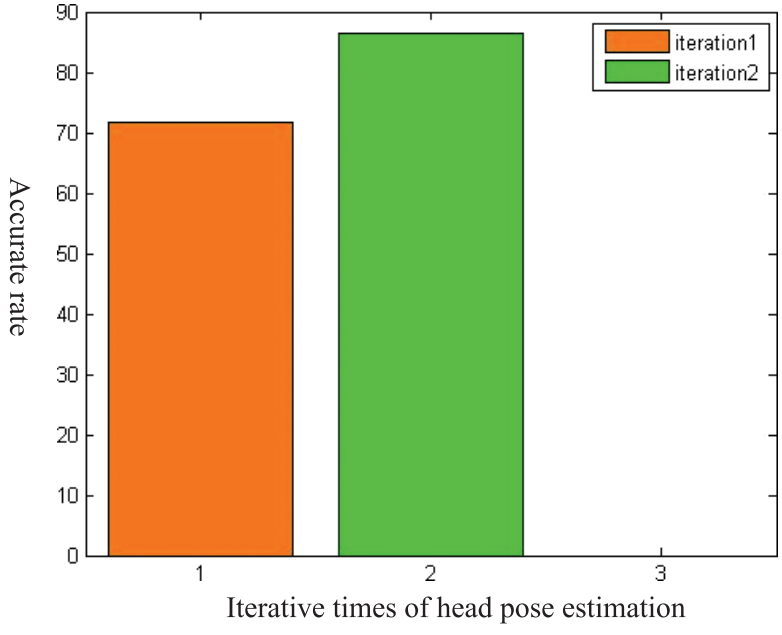


Fig. 13. (Color online) Accuracies of head pose estimation with different iterative times on Pointing'04 database.

Table 1. Comparison of the proposed approach with state-of-the-arts.

Approaches	Accuracy (%)	Mean Error (%)	Time (s)
Our iterative approach	85.9	14.1	0.2115
Initial D-RF	71.83	23.4	0.20375
RF	62.23	36.3	0.99095
SVM multi-class	60.4	38.2	—

D-RF used in the first iteration of our proposed iterative approach. The comparative results are shown in Table 1, including accuracy, mean error and computation time. The experimental result on SVM multi-class is quoted from their papers, whose final results provided the accurate rate of 60.4%. The RF directly estimated 25 head poses in the horizontal and vertical directions simultaneously and provided the accurate rate of 62.23%. Our proposed iterative approach in the case outperformed the other algorithms. The estimated accuracy reached 85.9% and computation time is 0.2115 s. Thanks to facial feature configuration, the results of proposed iterative approach are also better than the results estimated by initial D-RF.

4.2. *Multiple facial feature detection*

4.2.1. *Evaluation methodology*

We measure the localization performance using the inter-ocular distance (IOD) normalized error, define e_n as localization error,

$$e_n = \frac{\|I_n^G - I_n^D\|_2}{I_{\text{IOD}}}, \tag{18}$$

where I_n^G is the ground truth location of point n in the face, I_n^D is the detected location of the point n, and I_{IOD} is the inter-ocular distance, which is defined as the distance between the eye centers. We declare a point correctly detected if the pixel error is below 0.1 IOD.

4.2.2. *Experiments and analysis*

Experimental parameters for the task include the pf (see Eq. (15)) which limits the impact of distant votes is 0.25, the maximize offset distant between facial points and center location of random face patches is 40, the number of mean-shift iterations is 7, the bandwidth of the mean-shift kernel is 10. And the parameters of shape deformation distant of different head poses are automatically estimated during testing a validation set generated from the testing data by randomly extracting positive patches from every testing image. The most important parameter turns out to be pf. When pf is equal to zero, all patches contribute to the mean shift, while only patches in a small neighborhood are taken into account when it is close to one. Figure 14 shows the impact of pf on the average accuracy detection of all facial feature points. If $pf = 0$, it means that the detector tends towards the all patches from a face and

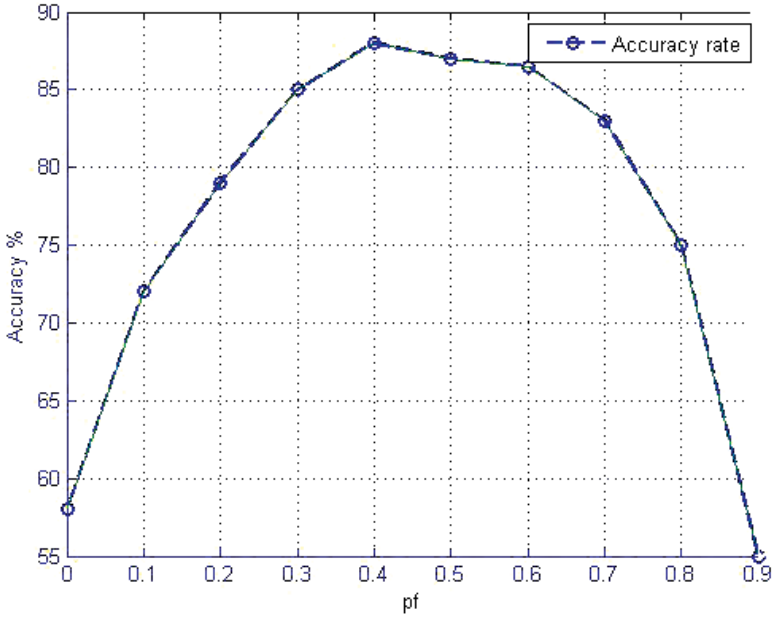


Fig. 14. Averaged accuracies with different values of the parameter pf.

the accuracy is below 60%. When pf moves around 0.2, it means that the detector depends on local patches in the local sub-regions and the performance increases significantly over 70%. When pf comes around 0.4, it means that the detector relies more on the 25 head poses and FDM models in local sub-regions, the accuracy reaches to about 90%. When pf becomes very large (> 0.8), the approach fails due to the small number of patches to vote.

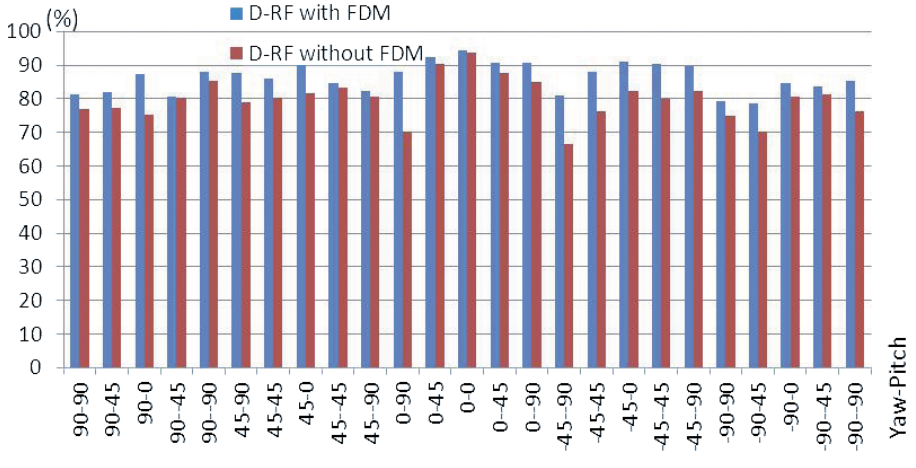


Fig. 15. D-RF with FDM versus without FDM.

Table 2. Accuracy rate of each facial point (%).

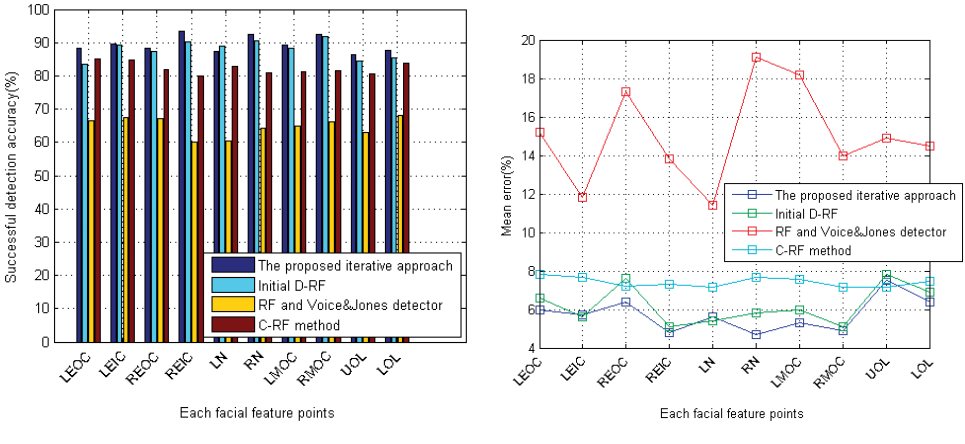
Facial Feature Point	Accuracy	Mean Error
Left eye outside corner (LEOC)	88.3	6
Left eye inner corner (LEIC)	89.7	5.7
Right eye outside corner (REOC)	88.2	6.4
Right eye inner corner (REIC)	93.4	4.8
Left nostril (LN)	87.5	5.6
Right nostril (RN)	92.6	4.7
Left mouth outside corner (LMOC)	89.3	5.3
Right mouth outside corner (RMOC)	92.5	4.9
Upper outside lip (UOL)	86.3	7.5
Lower outside lip (LOL)	87.6	6.4

To evaluate the effectiveness of the predefined FDM, the average detection accuracies of the D-RF with FDM and the D-RF without FDM are given in Fig. 15. The comparison results show that the D-RF with FDM provides higher accuracies than the D-RF without FDM under estimated 25 head poses, particularly, under the head poses of large rotation angles.

Table 2 shows the accuracy of each facial feature point using the proposed approach through two iterative procedures, when $pf = 0.4$. Lower outside lip detection is the most difficult because of facial occlusion and deformation in wide angles of head poses. And the performance is over 89.54% in average using the proposed approach.

4.3. *Comparison with state-of-the-art methods*

In this subsection, we compared the proposed approach with the state-of-the-art methods for final facial feature detection, i.e. D-RF in the first iteration,[23] conditional random forests (C-RF),[10] and RF+Viola&Jones method[14,21] on the



(a) (b)

Fig. 16. (Color online) Comparison to other methods. (a) Detection accuracy of each feature point and (b) mean error of each feature point.

Table 3. Computation time (/second) of the proposed approach, Initial D-RF, C-RF and RF.

Approaches	Positive Patch Classification	Head Pose Estimation	Facial Feature Detection	Total
Iterative approach	0.016914	0.2115	0.705403	0.933817
Initial D-RF	0.016914	0.20375	0.456094	0.676748
C-RF	—	0.57758	0.41337	0.99095
RF	—	0.87659	0.6412	1.51779

Pointing'04 database. The trained D-RF models have been generated from the LFW, Pointing'04 and our lab databases. The comparison results are shown in Fig. 16, where initial D-RF is also the D-RF used in the first iteration of our approach. Figures 16(a) and 16(b) demonstrate the accuracy and mean error of each facial point. Here, the dark blue bars represent the results of the two times interative approach in the paper, the light blue bars represent the results using initial D-RF, the red bars show the results detected by C-RF and the yellow bars show the results of RF+Viola&Jones method. As it can be seen, our approach performs better than the others on the Pointing'04.

The experiments have been conducted on a PC with Intel(R) Core(TM) i5-2400 CPU@ 3.10 GHz. The comparison of computation time is given in Table 3. From the

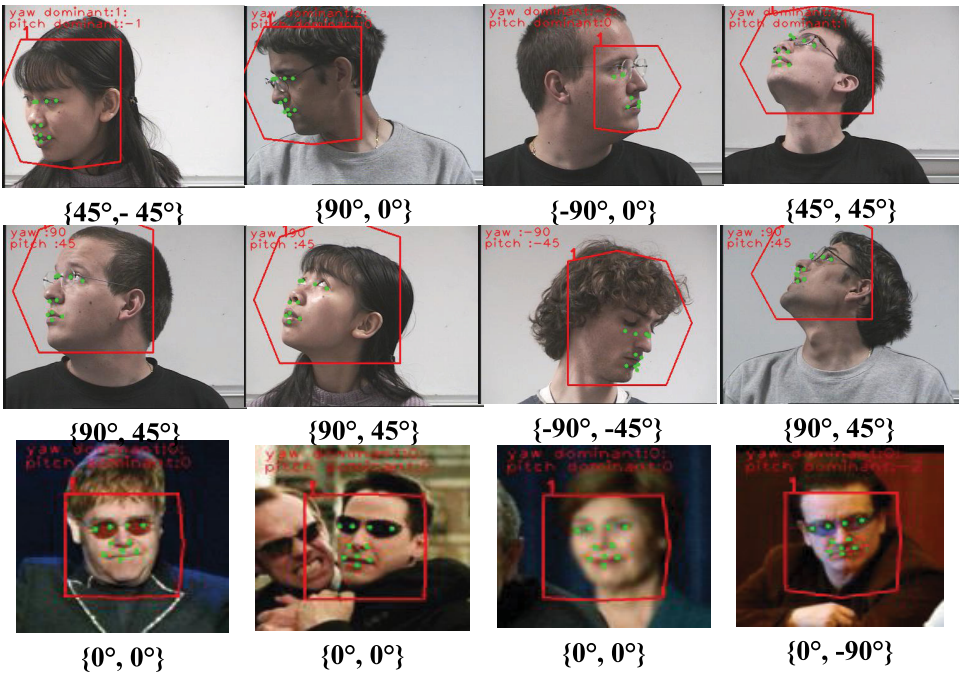


Fig. 17. Examples of the detected facial feature points using the iterative approach.

table, one can see that the D-RF with two iterative procedures is faster than the C-RF and RF but a little bit slower than the initial D-RF. The total computation time for facial analysis is 0.933817 s using our proposed iterative approach, 0.676748 s using initial D-RF,[22] 0.99095 s using C-RF[10] and 1.51779 s using RF.[4] They do not rely on special GPU. Additionally, some examples of detection results are shown in Fig. 17. The proposed approach in two iterative procedures performs well under some wide range head pose variations, occlusion, different illumination and noise.

5. Conclusions

In this paper, we propose a robust and efficient approach for face analysis under unconstrained environment based on a hierarchical regression framework. The proposed D-RF introduces Dirichlet-tree probabilistic model into regression RF framework in the hierarchical way. First, in order to eliminate the influence of noise and background in the facial area, a robust negative/positive facial patch extraction and classification method is proposed. Then, the D-RF works in the two iterative procedures to enhance the accuracy. Coarse head pose is estimated to constrain the facial feature detection, and the head pose is updated based on the detected facial features, iteratively, the facial feature localization is refined based on the updated head pose. Furthermore, in order to improve the efficiency and robustness, multiple probabilistic models are learned in leaves of the D-RF, i.e. the patch's classification, the head pose probabilities, the locations of facial points and a FDM. Moreover, our algorithm takes a composite weight voting method, where each patch extracted from the image can directly cast a vote for the head pose or each of the facial features. Experiment results show that the proposed approach benefits facial analysis in unconstrained environment. The proposed approach outperforms the state-of-the-art approaches on three different databases. In future work, more complex models could be introduced into 2D/3D unconstrained face analysis, such as expression model. This approach could be extended to detect a person's direction of attention in a wide scene, e.g. the students' attention and expression in a classroom scene. Advanced massively parallel computing technologies[41,42] should be incorporated to support real-time applications.

Acknowledgments

This work was supported by the National Key Technology Research and Development Program (No. 2013BAH18F02) and research funds from Ministry of Education and China Mobile (MCM20130601), Research funds from the Humanities and Social Sciences Foundation of the Ministry of Education (No. 14YJAZH005), Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (CCNU13B001), Wuhan Chenguang Project (2013070104010019), Central China Normal University Research Start-up funding (No. 120005030223), the Scientific Research Foundation for the Returned Overseas Chinese Scholars (No. (2013)693),

Natural Science Fund of Hubei Province (2014CFB661), National Key Technology Research and Development Program (No. 2014BAH22F01), and Research Funds of CCNU from the Colleges Basic Research and Operation of MOE (No. CCNU14A05019).

References

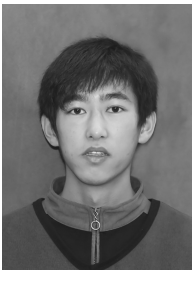
1. S. Baker and I. Matthews, Lucas-Kanade 20 years on: A unifying framework, *Int. J. Comput. Vis.* **56**(1) (2004) 221–255.
2. M. P. Beham and S. M. M. Roomi, A review of face recognition methods, *Int. J. Pattern Recogn. Artif. Intell.* **27**(4) (2013) 254–261.
3. P. Belhumeur, D. Jacobs, D. Kriegman and N. Kumar, Localizing parts of faces using a consensus of exemplars, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(12) (2013) 2930–2940.
4. L. Breiman, Random forests, *Mach. Learn.* **45**(1) (2001) 5–32.
5. J. Chen, D. Chen, X. Li *et al.*, Towards improving social communication skills upon multimodal sensory information, *IEEE Trans. Indu. Informat.* **10**(1) (2014) 323–330.
6. T. Cootes, G. Edwards and C. Taylor, Active appearance models, *IEEE Trans. Pattern Anal. Mach. Intell.* **23** (2001) 681–685.
7. T. F. Cootes, M. C. Ionita, C. Lindner *et al.*, Robust and accurate shape model fitting using random forest regression voting, in *Proc. European Conf. Computer Vision* (Springer, Berlin, Heidelberg, 2012), pp. 278–291.
8. T. Cootes and C. Taylor, Active shape models — 'smart snakes', *BMVC*, BMVC'92. Springer, London (1992), pp. 266–275.
9. A. Criminisi, J. Shotton and E. Konukoglu, Decision forestsfor classification, regression, density estimation, manifold learning and semi-supervised learning, Technical Report TR-2011-114, Microsoft Research (2011).
10. M. Dantone, J. Gall, G. Fanelli *et al.*, Real time facial feature detection using conditional regression forests, *IEEE Conf. on Computer Vision and Pattern Recognition* (CVPR, 2012), pp. 2578–2585.
11. D. F. DeMenthon and L. S. Davis, Model based object pose in 25 lines of code, *Int. J. Comput. Vis.* **15** (1995) 123–141.
12. G. Fanelli, M. Dantone, J. Gall *et al.*, Random forests for real time 3D face analysis, *Int. J. Comput. Vis.* **101**(3) (2013) 437–458.
13. G. Fanelli, J. Gall and L. VanGool, Real time head pose estimation with random regression forests, *IEEE Conf. on Computer Vision and Pattern Recognition* (CVPR, 2011), pp. 617–624.
14. G. Fanelli, T. Weise, J. Gall and L. VanGool, Real time head pose estimation from consumer depth cameras, *Pattern Recognition*, Springer, Berlin, Heidelberg (2011), pp. 101–110.
15. M. Figueiredo and A. K. Jain, Unsupervised learning of finite mixture models, *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(3) (2002) 381–396.
16. R. Girshick, J. Shotton, P. Kohli, A. Criminisi and A. Fitzgibbon, Efficient regression of general-activity humanposes from depth images, *IEEE Int. Conf. Computer Vision* (ICCV, 2011), pp. 415–422.
17. N. Gourier, D. Hall and J. Crowley, Estimating face orientation from robust detection of salient facial features, in *Proc. Pointing 2004, ICPR Int. Workshop on Visual Observation of Deictic Gestures* (2004), pp. 183–191.

18. C. Huang, X. Ding and C. Fang, Head pose estimation based on random forests for multiclass classification, *ICPR* (2010), pp. 934–937.
19. G. Huang, M. Ramesh, T. Berg and E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments, Technical Report, University of Massachusetts, Amherst (2007).
20. M. Jones and P. Viola, Fast multi-view face detection, Technical Report TR2003-096, Mitsubishi Electric Research Laboratories (2003).
21. Y. Li, S. Wang and X. Ding, Person-independent head pose estimation based on random forest regression, *17th IEEE Int. Conf. Image Processing (ICIP)* (IEEE, 2010), pp. 1521–1524.
22. Y. Liu, J. Chen, Y. Liu, Y. Gong and N. Luo, Dirichlet-tree distribution enhanced random forests for head pose estimation, *ICPRAM* (2014), pp. 87–95.
23. Y. Liu, J. Chen and C. Shan, Dirichlet-tree distribution enhanced random forests for facial feature detection, *ICIP* (2014), pp. 234–238.
24. T. Minka, The dirichlet-tree distribution, (1999), Available at http://research.microsoft.com/minka/papers/dirichlet/minkadirtree.pdf.
25. C. Morimoto and M. Mimica, Eye gaze tracking techniques for interactive applications, *Comput. Vis. Image Understand.* **98** (2005) 4–24.
26. E. Murphy-Chutorian and M. Trivedi, Head pose estimation in computer vision: A survey, *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(4) (2009) 607–626.
27. J. Orozco, S. Gong and T. Xiang, Head pose classification in crowded scenes, *BMVC* (2009), pp. 1–3.
28. M. Osadchy, M. L. Miller and Y. LeCun, Synergistic face detection and pose estimation with energy-based models, in *Journal of Machine Learning Research* **8** (2007) 1197–1215.
29. C. Peters, S. Asteriadis and K. Karpouzis, Investigating shared attention with a virtual agent using a gaze-based interface, *J. Multimodal User Interf.* **3**(1–2) (2010) 119–130.
30. N. Quadrianto and Z. Ghahramani, A very simple safe-Bayesian random forest, *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(6) (2014) 1297–1303.
31. R. Ruddarraju, A. Haro and I. Essa, Fast multiple camera head pose tracking, in *Proc. 16th Int. Conf. Vision Interface*, Halifax, Canada, June 2003, pp. 2–10.
32. S. Schulter, C. Leistner, P. M. Roth *et al.*, On-line Hough forests, *BMVC* (2011), pp. 1–11.
33. S. O. Shahdi and S. A. R. Abu-Bakar, Variant pose face recognition using discrete wavelet transform and liniar regression, *Int. J. Pattern Recogn. Artif. Intell.* **26**(6) (2012) 1307–1317.
34. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman and A. Blake, Real-time human pose recognition in parts from single depth images, *Communications of the ACM* **56**(1) (2013) 116–124.
35. C. Stauffer, W. Grimson *et al.*, Adaptive background mixture models for real-time tracking, *IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, Vol. 2 (1999), p. 2246.
36. M. Sun, P. Kohli and J. Shotton, Conditional regression forests for human pose estimation, *IEEE Conf. on Computer Vision and Pattern Recognition*, *CVPR* (2012), pp. 3394–3401.
37. J. G. Wang and E. Sung, EM enhancement of 3D head pose estimated by point at infinity, *Image Vis. Comput.* **25**(12) (2007) 1864–1874.
38. X. Yan and C. Han, Mutiple target tracking by probability hypothesis density based on dirichlet distribution, *J. XiAnJiaoTong Univ.* **45**(2) (2011) 003.
39. H. Yang and I. Patras, Sieving regression forest votes for facial feature detection in the wild, *ICCV* (2013), pp. 1936–1943.

40. D. Zhu and X. Ramanan, Face detection, pose estimation and landmark localization in the wild, in *2012 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (2012), pp. 2879–2886.
41. D. Chen, X. Li, L. Wang, S. U. Khan, J. Wang, K. Zang and C. Cai, Fast and scalable multi-way analysis of neural data, *IEEE Transactions on Computers*, Vol. 64, No. 3 (2015), pp. 707–719.
42. D. Chen, L. Wang, A. Y. Albert, M. Dou, J. Chen, Z. Deng and S. Hariri, Parallel simulation of complex evacuation scenarios with adaptive agent models, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 26, No. 3 (2015), pp. 847–857.

Yuanyuan Liu received her B.E. degree from Nanchang University, Nanchang, China, in 2005, and her M.E. degree from Huazhong University of Science and Technology, Wuhan, China, in 2007. She is currently a doctoral candidate for the Ph.D. at the National Engineering Research Center for E-Learning, Central China Normal University. Her research interests include image processing, computer vision and pattern recognition.

Jingying Chen received her bachelor and master degrees from the Huazhong University of Science and Technology, Wuhan, China, and her Ph.D. from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2001. She was a Postdoctor in INRIA, France, and a Research Fellow with the University of St Andrews and University of Edinburgh, UK. She is currently a Professor with the National Engineering Center for E-Learning, Central China Normal University, China. Her research interests include image processing, computer vision, pattern recognition and human–machine interface.



Cunjie Shan received his B.E. degree from Harbin Normal University, Harbin, China, in 2013. Currently, he is a master student in the National Engineering Research Center for E-Learning, Central China Normal University. His research interests include pattern recognition and image processing.

Zhiming Su received his B.E. degree from Wuhan University of Technology, Wuhan, China, in 2013. Currently, he is a master student in the National Engineering Research Center for E-Learning, Central China Normal University. His research interests include pattern recognition and image processing.

Pei Cai received his Digital Media Technology Bachelor of Engineering degree from Central China Normal University, Wuhan, China, in 2013. Currently, he is a master student in the National Engineering Research Center for E-Learning, Central China Normal University. His research interests include software engineering and graphics image processing.