Neurocomputing 173 (2016) 42–53

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Robust head pose estimation using Dirichlet-tree distribution enhanced random forests ☆

Yuanyuan Liu [a,b,c], Jingying Chen [a,b,*], Zhiming Su [a,b], Zhenzhen Luo [a,b], Nan Luo [a], Leyuan Liu [a,b], Kun Zhang [a,b]

[a] National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, China
[b] Collaborative & Innovative Center for Educational Technology (CICET), China
[c] Wenhua College, Wuhan, China

ARTICLE INFO

Article history:
Received 16 June 2014
Received in revised form
10 March 2015
Accepted 23 March 2015
Available online 4 August 2015

Keywords:
D-RF
HPE
Combined texture
Geometric features
Patch classification
Composite weighted voting

ABSTRACT

Head pose estimation (HPE) is important in human–machine interfaces. However, various illumination, occlusion, low image resolution and wide scene make the estimation task difficult. Hence, a Dirichlet-tree distribution enhanced Random Forests approach (D-RF) is proposed in this paper to estimate head pose efficiently and robustly in unconstrained environment. First, positive/negative facial patch is classified to eliminate influence of noise and occlusion. Then, the D-RF is proposed to estimate the head pose in a coarse-to-fine way using more powerful combined texture and geometric features of the classified positive patches. Furthermore, multiple probabilistic models have been learned in the leaves of the D-RF and a composite weighted voting method is introduced to improve the discrimination capability of the approach. Experiments have been done on three standard databases including two public databases and our lab database with head pose spanning from -90° to 90° in vertical and horizontal directions under various conditions, the average accuracy rate reaches 76.2% with 25 classes. The proposed approach has also been evaluated with the low resolution database collected from an overhead camera in a classroom, the average accuracy rate reaches 80.5% with 15 classes. The encouraging results suggest a strong potential for head pose and attention estimation in unconstrained environment.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Real-time, robust HPE algorithms are very important and an active research topic in computer vision as [1,2]. Knowing human's head poses can provide important cues concerning visual focus of attention and analyzing human's behavior. Also, head pose is crucial to applications like video surveillance, intelligent environments, human machine interfaces and affection recognition [3–6]. Due to its practical signification and challenges, there is a fair amount of work developed fast and reliable algorithms for head pose estimation. However, most of the work has reported good results in constrained environment, the performance could be decreased due to the high variations in unconstrained environment, such as, facial appearance, poses, illumination, occlusion, expression and make-up. Hence, a Dirichlet-tree distribution enhanced Random Forests approach (D-RF) is proposed in this paper to estimate head pose efficiently and robustly in unconstrained environment.

Based on different features, several methods for the problem can be briefly divided into two categories, facial geometric feature and facial texture feature based methods. The methods based on facial geometric features usually require high image resolution for facial feature identification, such as eyes, eyebrows nose or lips [7–9]. These methods can provide accurate estimation results relying on accurate detection of facial feature points and high quality images. Other based on facial texture approaches usually use texture feature from an entire face to estimate head pose [10–13]. It may be good for dealing with low resolution image but not robust to occlusion. In the real life scene, the various illumination, occlusion, low image resolution and wide scene make the head pose estimation difficult. In order to estimate head pose in unconstrained environment, we address the problem based on combined geometric and texture features.

More recently, classification and regression are very popular methods for head pose estimation on low resolution images such as neural networks (NN) [14], support vector machines (SVM) [15,16], nearest prototype matching [7] or random forests [17,10,8,18]. Gourier et al. [14] used an auto-associative network to learn the mapping for

☆Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding author at: National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, China.
E-mail addresses: jane19840701@hotmail.com (Y. Liu), chenjy@mail.ccnu.edu.cn (J. Chen), happyszm@foxmail.com (Z. Su), 13720269596@139.com (Z. Luo), luonanccnu@hotmail.com (N. Luo), lyliu@email.ccnu.edu.cn (L. Liu), zhk@mail.ccnu.edu.cn (K. Zhang).

http://dx.doi.org/10.1016/j.neucom.2015.03.096
0925-2312/© 2015 Elsevier B.V. All rights reserved.

head pose estimation on low resolution images. A simple winner-takes-all process was applied to select the head pose which prototype gives the best match in NN. They achieved a precision of 10.3 degrees in the yaw angle and 15.9° in the pitch angle only on the Pointing'04 database [9]. Orozco et al. [16] trained a multi-class Support Vector Machine for pose classification in crowed scenes. The distance features of each pixel of a head to the mean appearance templates of head images at different poses have been proposed to train a multi-class SVM for head pose classification. The performance on crowd public space and low resolution videos reached 80% accurate rate in 4 head poses classification. In [7], Wu and Trivedi proposed a two stage framework for continuous head pose estimation based on a finer geometrical structure. In the first stage, coarse head poses were classified by nearest prototype matching method, and refined head poses were estimated with a complex geometrical structure in second stage. The total accuracy was 75.4% in yaw and pitch angles. Recently, multi-class random forests become a very popular method in the field owing to their capability to handle large training datasets, their high generalization power and speed, and the relative ease of implementation.

Random forests are a family of ensemble classifiers introduced by Breiman in 2001 [19], which can be used either for multi-class classification [17,10,13], regression [8,11], or even both at the same time [12,18]. Fanelli et al. [18] proposed regression random forests for real-time head pose estimation from depth cameras. They reached 89% accurate rate with head and nose tip successful localization in high quality depth images. Some works [8,20] showed the power of RF in mapping image features to votes in a generalized Hough space [21] or to real-valued functions. Random forests have been combined with the concept of Hough transform for object detection and action recognition. These methods use two objective functions for optimizing the classification and the Hough voting properties of the random forests. Huang et al. [13] proposed Gabor feature based multi-class random forest method for head pose estimation. In order to enhance the discriminative power, they employed LDA technique for node tests. The successful accuracy reached 89% in public high resolution databases. Dantone et al. [12] proposed conditional random forests to estimate head pose under various conditions only in the horizontal direction. They used prior knowledge of some global variable to constrain output. In this case the global variable was the orientation of the head, divided into 5 classes. The accuracy rate reached 72.3% with five head pose classes in the wild database. Hence, head pose estimation in the wild and unconstrained environment is still a challenge and a significant problem.

To improve the accuracy and efficiency in the wild and unconstrained environment, a Dirichlet-tree distribution algorithm is introduced into random forest framework to estimate head pose in this paper. The idea of the paper is to use prior knowledge of some global variable to constrain output based Dirichlet-tree distribution. The Dirichlet-tree distribution was proposed by Minka [22]. It is the distribution over leaf probabilities that result from the prior on branch probabilities. Minka proved the high accuracy and efficiency of the distribution. Some researchers use a Dirichlet-tree distribution in multi-objects tracking [23] facial feature detection [24] and affective computing [25]. In this work, D-RF is proposed to estimate head poses in a coarse to fine way in various and unconstrained environment.

This paper is an extension of a paper presented at conference [10]. The main and different contributions from the conference paper are as follows. First, in order to improve classification, more powerful combined texture and geometric features (i.e., Gabor feature-based PCA, Sobel, LBPH and two geometric features) from positive facial patches are extracted to estimate head pose with D-RF, instead of only the texture features (Gabor feature-based PCA and gray values) used in the ICPRAM paper. Second, in the previous ICRPAM paper, single probabilistic model has been learned in leaves of the D-RF and a GMM method was used to vote the leaves. In this paper, multiple probabilistic models (i.e., head pose angles and two geometric offset vectors) have been learned in leaves of the new D-RF, and a composite weighted voting method that is composited of the classification and regression voting measures is introduced into probabilities voting to improve the discrimination capability of the approach. Third, an additive confidence parameter *pf* in the composite weighted voting method has been used to control the number of positive patches through geometric offset vectors stored at leaves, which can be used to eliminate the influence due to face deformation and wide range head poses. Finally, more detailed experiments have been done on three standard databases and our low resolution database collected from an overhead camera in a classroom, the average accuracy rate reaches 76.2% with 25 classes in standard databases and 80.5% with 15 classes in our collected low resolution database, respectively.

2. D-RF for head pose estimation

The flowchart of the proposed approach is given in Fig. 1. In the first stage, facial patches are extracted and classified to positive/negative patches from detected facial areas, and combined texture and geometric features from positive facial patches have been extracted. In the second stage, a more accurate D-RF approach with combined texture and geometric features is proposed based our previous work to estimate head pose in the horizontal and vertical directions. The proposed D-RF consists of four layers. D-L1 and D-L2 are two layers in the horizontal direction, D-L1 represents coarse classification while D-L2 is refined classification. In D-L2, the yaw angle has been estimated based on the classified result of D-L1. D-L3 and D-L4 are two layers in the vertical direction, D-L3 represents horizontal refined classification and vertical coarse classification, while D-L4 represents final refined classification in two freedom head poses. In each leaf of the D-RF, there are multiple probabilistic models including patch class probability, head poses and two geometric offset vectors. A composite weighted voting method is used to obtain final head pose parameter based multiple probabilistic models in leaves. Finally, the final head pose angles have been obtained in the D-L4 layer of D-RF. Details are given in the following.

2.1. Positive facial patch extraction

A facial area is first detected by Adaboost with Haar-like feature [26], which may include some noise for head pose estimation, 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 patches and negative patches, where the positive patches contribute to estimate head pose while the negative patches including occlusion or noise may introduce errors for the task.

To segment the background, the detected facial area which is normalized as 125∗125 pixels is divided into 6∗6 non-overlapping squares, and histogram distributions of the squares are computed as shown in Fig. 2. We analyze the uniformity of histogram distributions of the patches and segment most of the background patches.

200 patches are randomly extracted from the rest of facial area with background removed, which include positive and negative facial patches. The positive and negative patches are classified using RF [17,19]. In order to model the random tree, positive facial patches are labeled as 1 and the negative facial patches are labeled as 0. A tree T grows up based on Gabor features and histogram of the labeled patches. The training and testing are similar to RF

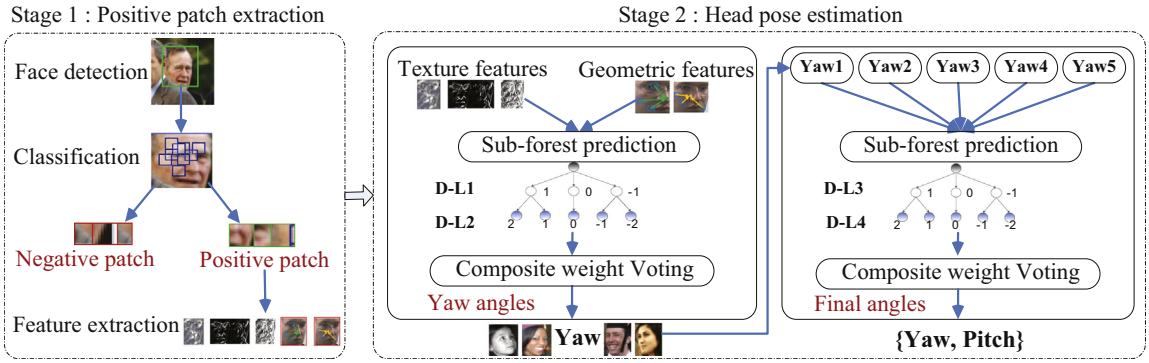


Fig. 1. The flowchart of the proposed approach for robust head pose estimation.

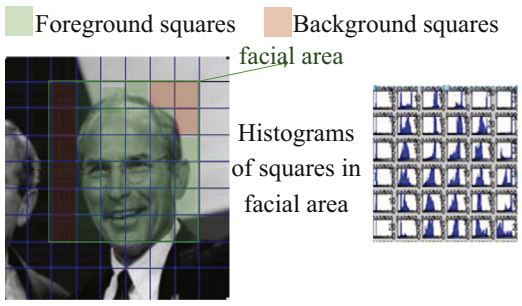


Fig. 2. Foreground and background square segmentation.

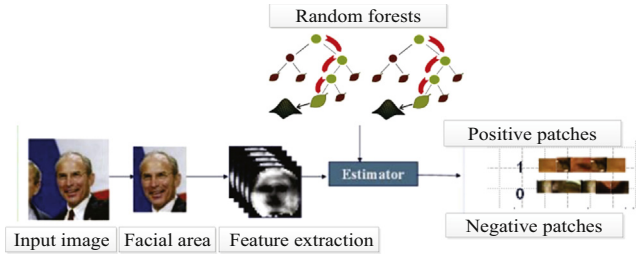


Fig. 3. Facial positive and negative patch classification.

[17,12,18]. When all the test patches arrive at leaves of trees in the forest, we use the probability $p(c=k|l_t(P))$ stored at a leaf to judge whether the test patch belongs to a class k, where $k=1$ represents the positive patch while $k=0$ represents the negative patch. The algorithm diagram is shown in Fig. 3. Only the positive patch is used to estimate head pose using D-RF.

2.2. D-RF for head pose estimation

2.2.1. General idea of the D-RF

The Dirichlet-tree is the distribution over leaf probabilities $[p_1 \ldots p_i]$ that results from this prior node probabilities $[a_1, a_2, a_k]$ 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. 4(a). RF is an ensemble approach in which several tree predictors are combined together to obtain high performance for classification or regression (see Fig. 4(b)). Each tree in the forest is independently generated with random samples selected from the whole data set.

As shown in Fig. 4(c), D-RF arranges random trees of RF as the Dirichlet-tree structure, where each node a_j is a sub-forest of the D-RF. Each tree in a sub-forest has been generated with selected samples in Dirichlet distribution. It is noted that the sub-forest is only related to his prior node. Hence, the D-RF only computes the final probabilities under its prior cascaded sub-forests instead of all trees in the forest. Therefore, D-RF can perform with high accuracy and efficiency. The training and testing of the proposed D-RF is given as below.

2.2.2. Combined texture and geometric features

In this paper, unlike texture features in the ICPRAM [10], more powerful combined texture and geometric features are proposed to estimate head pose from positive facial patches. Texture feature is obtained based on multiple texture descriptions, i.e., Gabor feature-based PCA, LBPH, Sobel descriptor and gray values from positive patches. $F_i = \{f_i^1, f_i^2, f_i^3, f_i^4\}$ represent the extracted texture feature channels. f_i^1 contains Gabor feature-based PCA at five angles and seven scales with dimension as 30∗35, f_i^2 represents Sobel edge descriptor at the horizontal and vertical directions with dimension as 30∗30∗2, f_i^3 is the LBPH descriptor with dimension as 30∗30, and f_i^4 is the raw gray values of the patch (see Fig. 5).

$D_i = \{d_i^1, d_i^2\}$ is the geometric feature set. The components d_i^1, d_i^2 represent two offset vectors from the patch centroid to the tip of the nose and to the face centroid, where d_i^1 denotes the 2 dimension displacement vector from the centroid of the facial patch P_i to the tip of the nose point N, and d_i^2 denotes the displacement vector from the P_i and the facial centroid point F (see Fig. 6)

$$d_i^1 = \|P_i - N\|_2, \quad d_i^2 = \|P_i - F\|_2 \tag{1}$$

2.2.3. Training of the D-RF

Each tree T in the D-RF $T=\{T_t\}$ is built and selected from a different set of the training images. From each image, we extract a set of combined features from positive facial patches $P_i = \{F_i, D_i, C_i | k_i\}$. F_i and D_i are texture and geometric features described as in Section 2.2.2 Only $k_i = 1$ represents the positive facial patch that can be used to estimate head pose. C_i contains the annotated head pose parameter in different layers of the D-RF. In our case, we denote $C_i = \{c_i^1, (c_i^2|c_i^1), (c_i^3|c_i^2, c_i^1), (c_i^4|c_i^3, c_i^2, c_i^1), C_{(x,y)}\}$, where c_i^1 are 3 yaw rotation angles in the first layer of the Dirichlet-tree distribution, $c_i^2|c_i^1$ are 5 yaw angles refined from in the second layer, $c_i^3|c_i^2$, c_i^1 are 15 pitch angles under condition of each yaw angle c_i^2 in the third layer, $c_i^4|c_i^3$, c_i^2, c_i^1 are 25 refined angles based on the above annotated angles at leaves of the Dirichlet-tree in the fourth layer. $C_{(x,y)}$ are the offset vectors from the facial patch centroid to the location of the end of the nose and to the face centroid.

We define a patch comparison feature as our binary tests φ, similar to [12,18]

$$\varphi = |R_1|^{-1} \sum_{j \in R_1} f^n(j) - |R_2|^{-1} \sum_{j \in R_2} f^n(j) \tag{2}$$

where R_1 and R_2 are two random rectangles within the patches, $f^n(j)$ is the texture feature channel $n=\{1, 2\ldots\}$, j is the pixel within the rectangles.

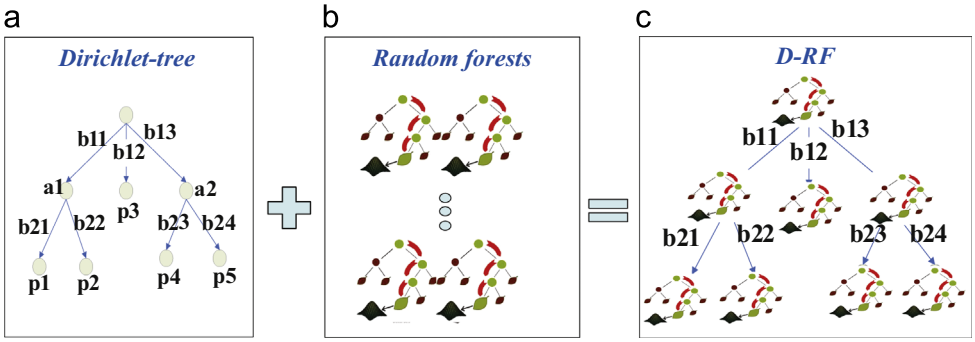


Fig. 4. A general D-RF. (a) A general Dirichlet-tree distribution. (b) A general random forest. (c) The configuration of the D-RF.

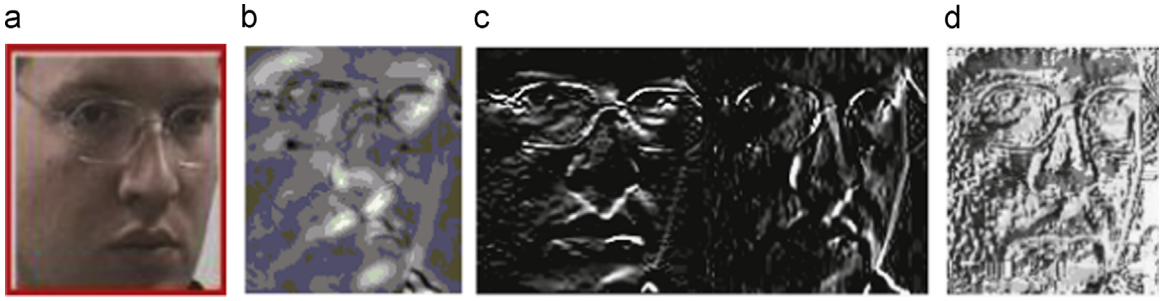


Fig. 5. Multiple texture features. (a) Facial area. (b) Gabor feature-based PCA. (c) Sobel features. (d) LBPH description.

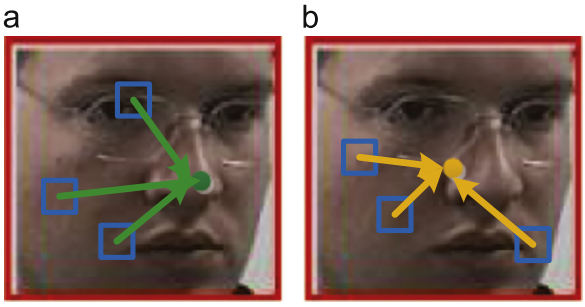


Fig. 6. Geometric features. (a) The offset vector from the patch centroid to the tip of the nose. (b) The offset vector from the patch centroid to the facial centroid.

The training of a sub-forest a_j in the different layers of D-RF is given below:

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

$$P_L = \{P|\varphi < \tau\}, \quad P_R = \{P|\varphi > \tau\} \tag{3}$$

where φ is the patch comparison feature (Eq. (2)) and τ is a predefined threshold.

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

$$IG = \arg\max_{\varphi}(H(P|a_j) - \sum_{S\in\{L,R\}} \frac{|P_S(\varphi)|}{|P|} H(P_S(\varphi)|a_j)) \tag{4}$$

where $H(P|a_j)$ is the entropy of the different set P for annotated patch labels and in this case, all training patches are positive $k_i = 1 \forall i$)

$$H(P|a_j) = -\frac{\sum_i p(C_i|a_j,P_i)}{|P|} \log\left(\frac{\sum_i p(C_i|a_j,P_i)}{|P|}\right) \tag{5}$$

where $p(C_i|a_j,P_i)$ indicates the probability that the patch P_i belongs to the head pose class C_i in the sub-fores a_j of the j-th layer in the D-RF.

3. Create a 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.

A leaf of the D-RF includes three probabilistic models, i.e., (1) a positive/negative patch probability $p(k=1|P_i)$, (2) a head pose parameter probability $p(C_i|a_j,P_i)$, (3) a probability regression model for the location of the tip of the nose, which eliminate the influence of face deformation.

2.2.4. Testing of the D-RF

During testing, k_i and a_j are first estimated, then voting model with the estimated state is used. Details on the testing in each layer are given in the following sections.

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. Each patch P_i ends in a set of leaves L of the relative sub-forests of D-RF instead of all leaves of the D-RF. In each leaf l_{a_j} of the sub-forest a_j, there are multiple probabilities $p(c_i^m, D_i|l_{a_j})$. We simplify the distribution over the leaf models by a multivariate Gaussian as in

$$\begin{aligned} p(c_i^m, D_i|l_{a_j}) &= p(c_i^m|D_i, l_{a_j}) \bullet p(D_i|l_{a_j}) \\ &= N(c_i^m|a_j; \overline{c_i^m|a_j}, \Sigma_{c_i^m|a_j}) \bullet N(D_i|a_j; \overline{D_i|a_j}, \Sigma_{D_i|a_j}) \end{aligned} \tag{6}$$

where $\overline{c_i^m|a_j}$ and $\Sigma_{c_i^m|a_j}$ are the mean and covariance matrix of the head pose probabilities of the sub-forest a_j of the m-th layer in the D-RF, $\overline{D_i|a_j}, \Sigma_{D_i|a_j}$ are the mean and covariance matrix of the offset of the tip of the nose and the facial centroid.

While Eq. (6) only models the probability for a positive patch P_i ending in leaves of a single tree, the probability of the sub-forest a_j is obtained by averaging over the trees in it,

$$P(c_i^m, D_i|a_j, P) = \frac{1}{T_t}\sum_t p(c_i^m, D_i|l_t, a_j(P)) \tag{7}$$

where T_t is the number of trees in the sub-forest a_j, l_t, a_j is the corresponding leaf for patch P in the t-th tree of the sub-forest a_j.

2.2.5. Sub-forest prediction in the D-RF

An adaptive Gaussian mixture model (GMM) is used to predict sub-forests in the next layer, similar to [10]. The D-RF allows to make kinds of predictions in different nodes of each layer at testing.

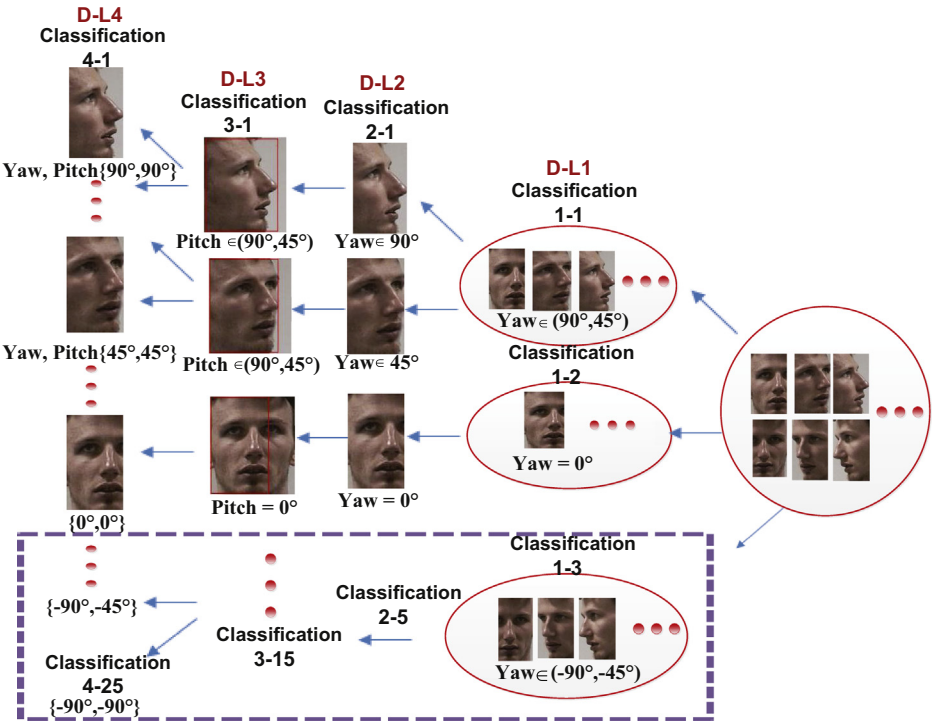


Fig. 7. Head pose estimation in the horizontal and vertical directions. Classification m-n represents the head pose classifiers in the *n*-th node and *m*-th layer of the D-RF.

Fig. 8. Examples of images from three standard databases, Pointing'04 database (the first row), LFW database (the second row), and our lab database from a near camera (the third row).

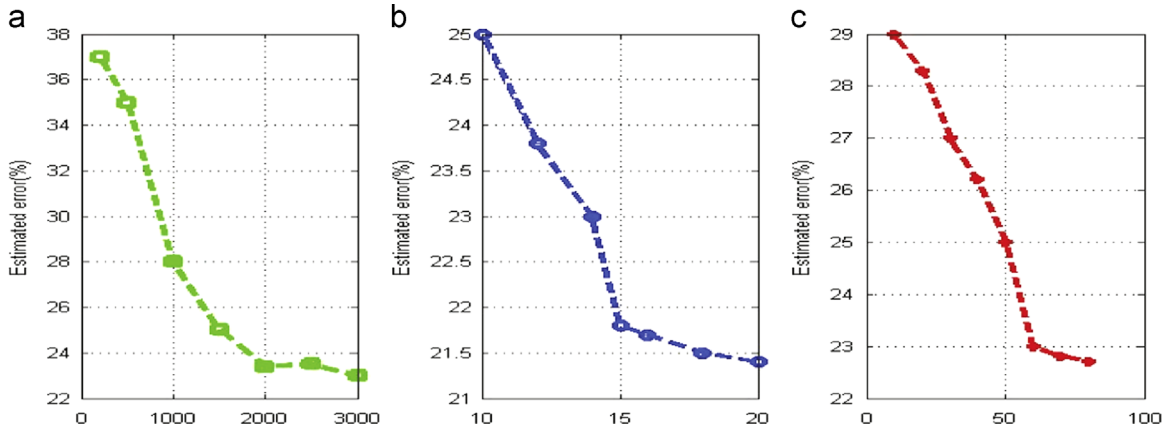


Fig. 9. Mean errors for varied splitting candidates, depth of the trees and the number of trees in the D-RF. (a) The Splitting candidates (Depth=15, Number=60). (b) The Depth of trees (Split=2000, Number=60). (c) The number of trees (Depth=15, Split=2000).

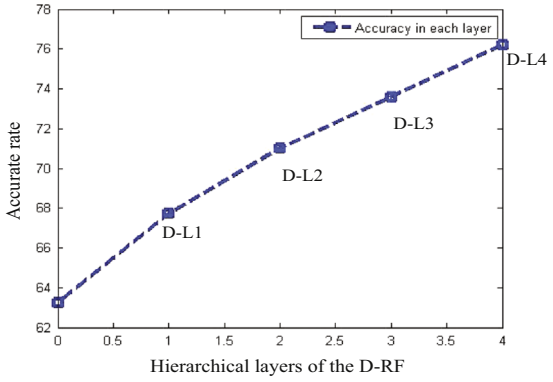


Fig. 10. Accuracy comparison in different layers of the D-RF.

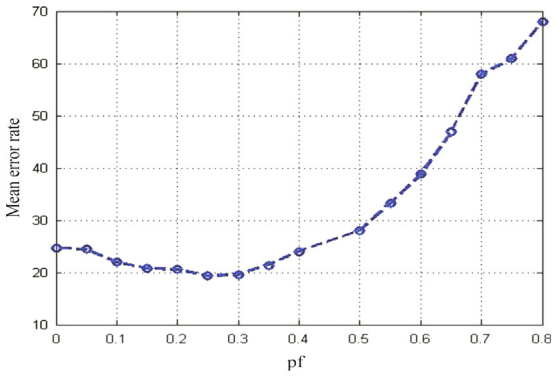


Fig. 11. Mean error rate with different values of the confident parameter *pf*.

The obvious one is regarding the most probable class label given a new data set, i.e. a sub-forest in the new layer.

After the probability of the current sub-forest a_j has been obtained, the new sub-forest a_{j+1} should be selected based on class decision function $C(a_{j+1})$,

$$C(a_{j+1}) = \arg\max_{a_j \in c_i^m} p(c_i^m, D_i|a_j, \mathrm{P}) \tag{8}$$

The GMM can adaptively select trees from the sub-forest a_{j+1} in the new layer node. To this end

$$p(C_i|\alpha_{j+1}, P) = \frac{1}{k_i}\sum_{i=1}^{j+1}\sum_{t=1}^{k_i} p(C_i|l_t, a_{j+1}(P)) \tag{9}$$

where l_t, a_{j+1} is the corresponding leaf for patch P in the predicted sub-forest of the next layer. The discrete value k_i is the number of trees in the predicted sub-forest.

When the patches are passed the sub-forest in the last layer of the D-RF, the head pose angles are then computed by performing a composite weighted voting method.

2.2.6. A composite weighted voting method

Different from a single voting measure in [10], a weighted composition of regression and classification voting measures are used in this paper. In order to eliminate imbalance of samples in a set, we first store the weight $w_s = P_S/P$ that is defined as the ratio of the number of samples P_S in each subset and the total number of samples P in each single tree of the D-RF.

Additionally, geometric offsets stored at leaves can be used to eliminate the influence due to face deformation and large rotation poses. To integrate the votes coming from different patches, we accumulate them based on an additive confidence parameter *pf*,

$$pf \propto \exp\left(-\frac{|d_i^n|}{\gamma}\right) \tag{10}$$

The constant γ is used to control the steepness of this function. A positive patch with a high confidence *pf* is only allowed to vote for head pose estimation.

If votes for the head pose c_i^m is in the patch location y_i, then we set the composite weighted voting model to be given as

$$V(c_i^m) \propto K((w_s V(x,y) - (y_i + \overline{w_s V(x,y)}))/h_j) \tag{11}$$

where $V(x,y) = \sum p(D_i|_t)$, $\overline{w_s V(x,y)}$ is the mean of the geometric offset probabilities in the tree T. A Gaussian Kernel K and the bandwidth parameter h_j are given by Gaussian filter. Finally, regression voting can obtain good results evaluating on sparse patches rather than all patches. The final head pose parameter and nose location are obtained by mean-shift in $V(x,y)$.

2.3. D-RF for head pose estimation in the horizontal and vertical directions

In order to obtain robust head pose estimation in the horizontal and vertical directions in unconstrained environment, D-RF is trained as described in Section 2.2.3. Since it is difficult to obtain continuous ground truth head pose data from 2D images, we annotate rotation angles as three coarse classes, i.e. "−90°,0°,90°" and five refined classes, i.e., "−90°,−45°,0°,45°,90°" in different layers of the D-RF. We store the multivariate adaptive Gaussian distribution in the leaf as defined in Eqs. (7) and (11). Fig. 7 shows the framework of head pose estimation using D-RF in the horizontal and vertical directions, where Yaw and Pitch mean the

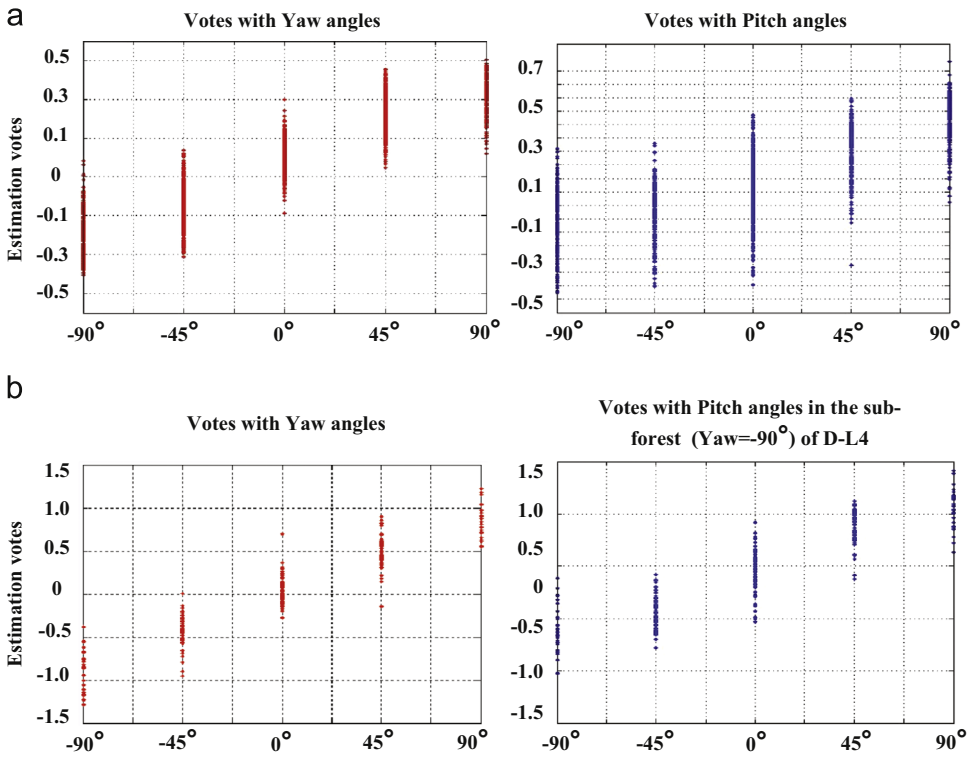


Fig. 12. Voting distributions using RF and D-RF. The horizontal axis represents head pose class distribution, and the vertical axis represents probabilistic values of voting in each head pose class. (a) Voting distribution using RF. (b) Voting distribution using D-RF.

Table 1
The final estimation accuracies (%) using D-RF in 25 head pose classes.

	90°		45°		0°		−45°		−90°	
90°	75.6	61.0	74.1	62.5	77.4	71.6	76.3	55.6	74.9	65.9
45°	78	52.3	78.5	73.1	79.1	69.4	80	43.9	77.5	66.2
0°	80.6	75.2	81.9	64.1	83.7	70.6	81	69.8	80.6	75.7
−45°	78.4	66.0	81.1	78.8	79.2	73.2	77.8	50.7	75	49.4
−90°	76.2	58.8	76.5	60.3	75.7	67	75.6	45.2	74.3	60.4

Table 2
Comparison of the proposed algorithm with state-of-the-arts. The second and third columns show average estimated rates of yaw and pitch accuracies. The fourth column shows the mean error rates. And the last column shows the running time of these algorithms.

Algorithms	Yaw (%)	Pitch (%)	Mean error (%)	Time (s)
Proposed algorithm	86.15	76.2	21.8	0.88995
ICPRAM [10]	83.52	71.83	23.4	0.98995
RF [19]	78.4	62.23	36.3	1.36859
SVM multi-class [16]	80.6	60.4	38.2	–
NN [14]	79.5	56.7	39	–

Table 3
Comparison of different features and methods.

PPC+CGT+CWV+GMM	No PPC	No CGT	No GMM	No CWV
76.2%	67.14%	68.37%	53%	71.3%

estimation results respectively in the horizontal and vertical directions. D-L1 and D-L2 are two layers in the horizontal direction in D-RF, while D-L3 and D-L4 are two layers in the vertical direction in D-RF. Then, five yaw angles can be estimated in the D-L2 layer of the D-RF. After the yaw angles have been estimated, pitch angles are estimated under condition of the estimated yaw angles. When the patches are sent down through all layers in the D-RF, sub-forests will be selected in the cascaded way using Eqs. (9) and (10). Finally, we can estimate 25 pair discrete angles in the D-L4 using composite weight voting as Eq. (11), i.e. $\{90°, 90°\}$, $\{90°, 45°\}\ldots\{0°, 0°\}\ldots\{-45°, -90°\}, \{-90°, -90°\}$.

3. Experiments

In this section, the proposed D-RF for head pose estimation has been thoroughly evaluated with different quality images from three standard databases and our low resolution database collected from an overhead camera in a classroom. Moreover, some result examples of occluded images are given.

3.1. Head pose estimation in three standard databases

In this section, we compare results of some state of the art approaches with our approach using the Pointing'04 database [9], LFW database [27] and our lab database from a near camera (see Fig. 8). The Pointing'04 database consists of 2940 images with different poses and expressions. The LFW database consists of 5749 individual facial images. The images have been collected 'in the wild' and vary in poses, lighting conditions, resolutions, races, occlusions, and make-up. Our lab database has been collected using 20 different persons with different poses, expressions and occlusions, and the reference angles have been annotated using the method similar to LFW [27]. First, we divided the databases into a training set and a testing set. The training set consists of 2100 images from Pointing'04 database and 3000 images from LFW database. The testing set includes the rest of 840 images from Pointing'04 database, 1500 images from LFW database, and 200 images from our lab database.

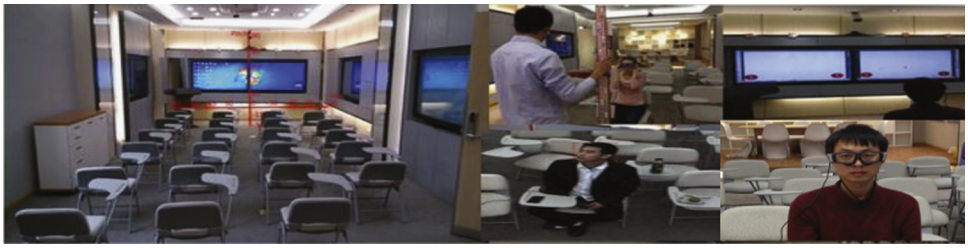


Fig. 13. The classroom database labeling using SMI.

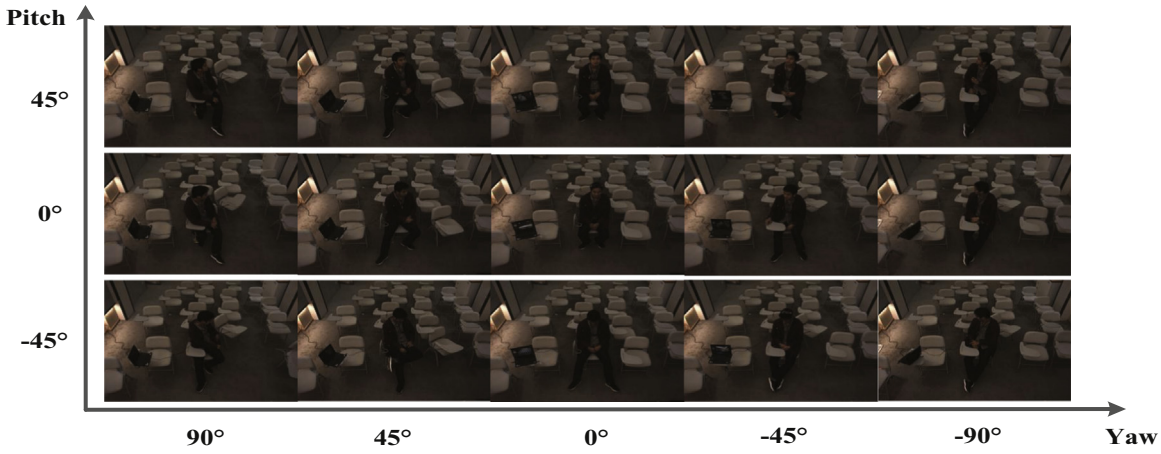


Fig. 14. Sample images from our classroom database.

3.1.1. Training

For training the trees in the Pointing'04 and LFW database, we fixed some parameters as Fig. 9, 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 including the number of patches extracted from each image (fixed to 200) and the patch's size (30∗30) are similar to [18]. Each tree grows based on a selected subset of 186 images in each sub-forest with Dirichlet distribution. 25 sub-forests in different layers of the D-RF have been trained independently. The plots in Fig. 9 show the performance of the algorithm when we varied the depth of each trees in D-RF, the number of trees in D-RF and the times of splitting candidates. One can see that the best performance occurred in the 2000 splitting candidates, 15 depth of the trees and over 60 trees trained for the D-RF.

3.1.2. Testing

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

$$Q(P_i|Y_i) = \frac{\langle P_i, Y_i \rangle \cdot P_i}{\sum_{j=1}^{n} \langle P_j, Y_i \rangle \cdot P_j} \tag{12}$$

Testing critical parameters include the RF parameters, the number of Dirichlet-tree layer L_i, the adaptive GMM parameters and the confidence Pf (0.25). Fig. 10 shows final estimation accuracies with different layers L_i of D-RF for 25 head pose estimation. None represents as using original RF to classify 25 head poses. While D-L1∼D-L4 represent that 1∼4 layers of the D-RF are devoted to classify 25 head poses. The accuracy of original RF reaches to 63.23%, and the proposed approach improves the accuracy with the introduction of the different layers of the Dirichlet-tree. The optimal estimation accuracy is 76.2% by using 4 layers of D-RF. Fig. 11 shows the impact of the value of confident pf on the average accuracy of estimated head poses. pf controlled the number of patches to vote in the D-RF. $pf=0$ means that all positive patches are allowed for voting. $pf \geq 0.5$ means too many patches are lost, which makes the mean error rate decline. In the case, we used $pf=0.25$ to control the number of patches to avoid the influence due to face deformation and large rotation angles.

Table 4
Average accuracies (%) by *pf*.

pf	Yaw	Pitch	Missed	Time (/s)
0	76.3	77.0	25.6	0.9047
0.125	80.2	77.6	23.5	0.8632
0.25	83.8	80.2	18.4	0.7514
0.5	72.0	73.6	29.6	0.6921

In Fig. 12, we plot the voting distribution using RF and D-RF to estimate 25 head poses respectively. The graph's horizontal axis represents head pose class distribution, and the vertical axis represents probabilistic values of voting in each head pose class. In Fig. 12(a), the left image shows the voting values with five yaw angles using RF and the right image shows the voting values with five pitch angles using RF. One can see that voting values in different angle classes of RF overlap about halfway to classify five pitch angles difficulty. Fig. 12(b) shows the estimation probabilistic voting distribution using D-RF. The left image in Fig. 12(b) shows the voting values with five yaw angles using D-RF, and the right image shows the voting values with five pitch angles under the sub-forest (Yaw$=-90^{\circ}$) of D-L4. Note that the D-L4 includes 5 sub-forests to obtain pitch angle votes under the condition of 5 yaw angles. Only the voting distribution in the sub-forest (Yaw$=-90^{\circ}$) is shown in Fig. 12(b) because it is most difficult to classify due to wide range pose variation. And probabilistic values of voting have a discriminative distribution with different pitch

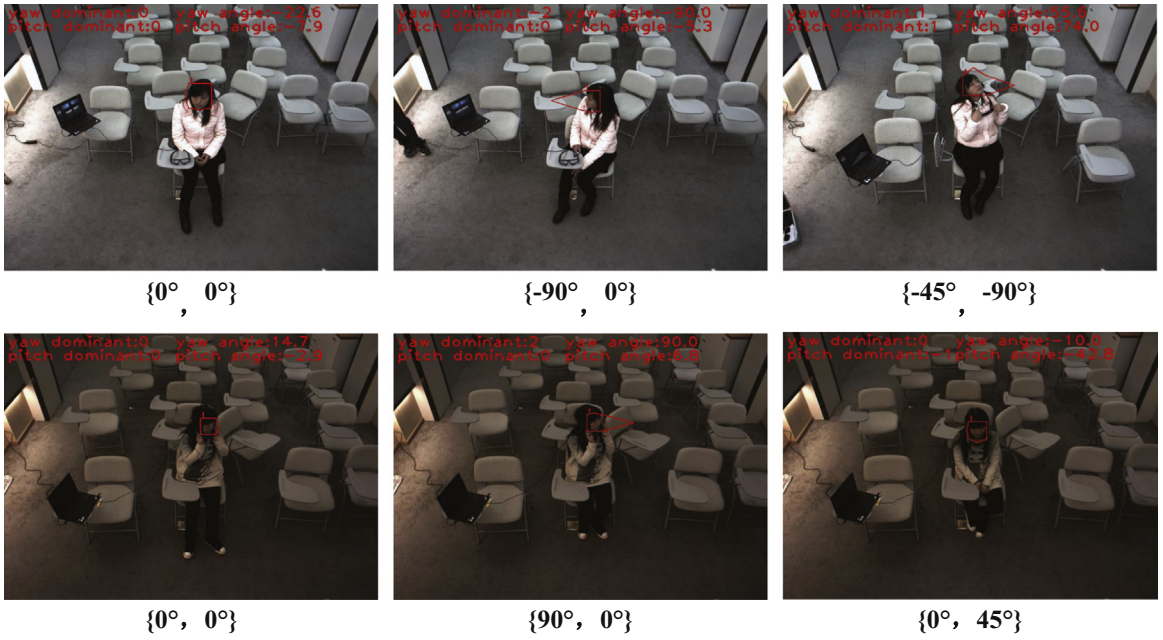


Fig. 15. Examples of successfully estimated images from our classroom database. (For interpretation of the references to color in this figure, the reader is referred to the web version of this paper.)

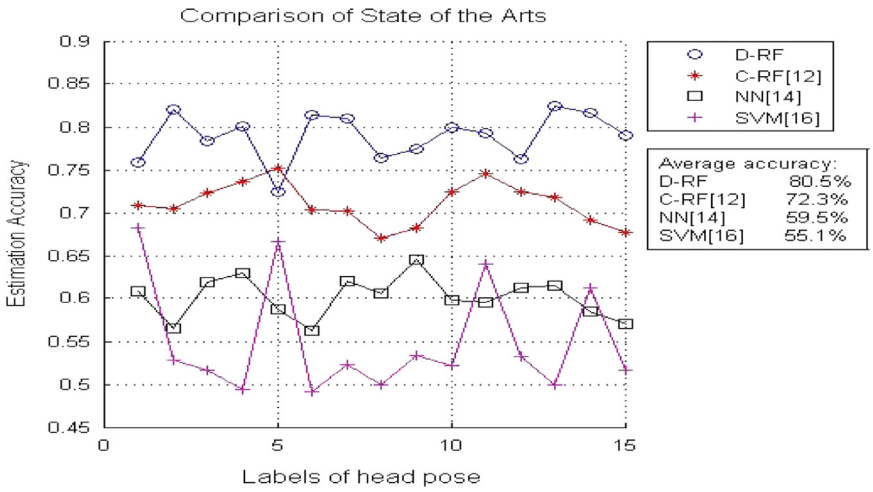


Fig. 16. Accurate rates per pose by different methods, i.e., D-RF, C-RF, NN and SVM.

angles, as shown in Fig. 12(b). One can see that the D-RF have better discrimination capability than RF.

The final estimated results are shown in Table 1, which describes estimation accuracies in 25 class head pose areas. The average accuracy of the D-RF reached 76.2% by the proposed algorithm in the paper.

3.1.3. Comparison with state of the art

We extensively compared our proposed algorithm with other state of the art algorithms, i.e., standard random forests [19], our previous work in ICPRAM conference [10], SVM multi-class [16], neural networks [14]. We declare that the same training and testing sets from three databases are used in the following comparison experiments. The experiments have been conducted on a PC with Intel(R) Core(TM) i5-2400 CPU@ 3.10 GHz.

The comparison of results is shown in Table 2, including successful accurate rate, mean error rate in yaw and pitch angles estimation. The computation time of related algorithms is shown in the last column of Table 2. The experimental methods on SVM multi-class and NN are quoted from their papers, whose final results provided accuracies rate of 60.4% and 56.7% respectively. The RF directly estimated 25 head poses in the horizontal and vertical directions simultaneously and provided a 62.23% accurate rate. Our proposed algorithm in the case outperformed the other algorithms. The estimated accuracy reached 76.2% and computation time is 0.88995 s. Thanks to the combined texture and geometric features and a composite voting method to delete the unwanted patches from face deformation and large rotation angle in unbalanced sample sets, the results of the proposed algorithm are better than the results in ICPRAM. And the computation time is saved because of less patches from a face area.

3.1.4. Results with different combinations of features and methods

Table 3 shows average accuracies obtained with the same databases using different combination features and methods, e.g., positive patch classification (PPC), combined geometric and texture features (GTF), the GMM prediction and composite weight voting (CWV). The first column shows the accuracy using all the

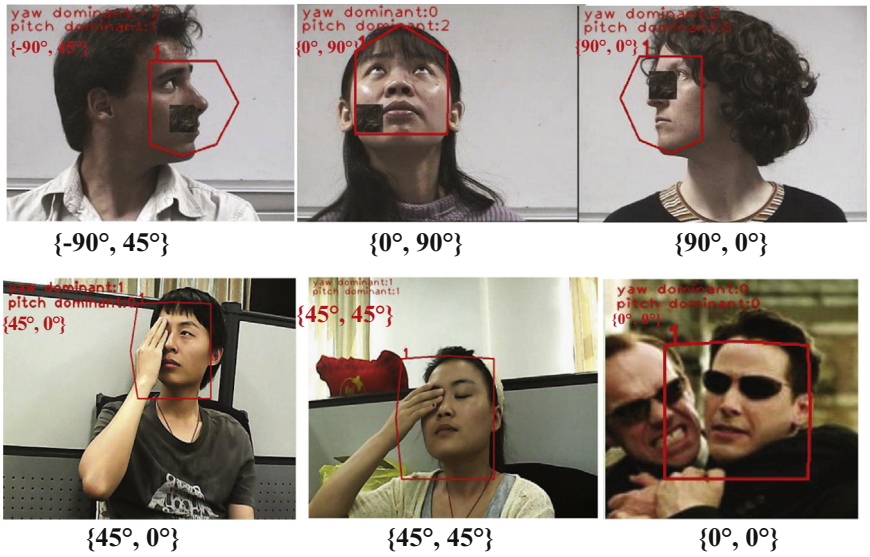


Fig. 17. Example results of the occluded test images. (For interpretation of the references to color in this figure, the reader is referred to the web version of this paper.)

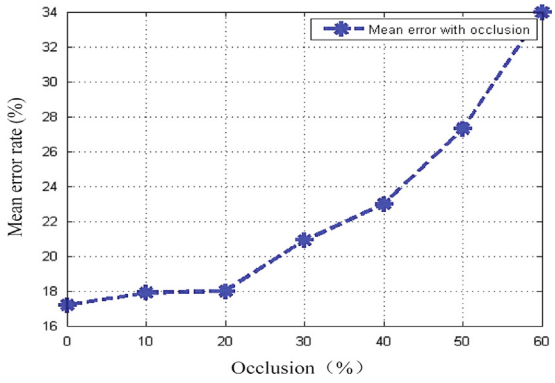


Fig. 18. Mean errors as different size of occluded blocks.

features and methods, the second column shows the accuracy without PPC, the third column shows the accuracy only using texture features rather than GTF, and the fourth column is the accuracy without GMM. The last column shows the results using average voting instead of using CWV. As shown in the table, using all the proposed features and methods performs better than the others.

3.2. Head pose estimation in the classroom database from an overhead camera

To evaluate our approach on the low resolution images, 58 students with 75 different head poses have been collected in a wide classroom from an overhead camera. The classroom database consists of 4350 images with various illuminations, expressions, low resolution and poses. In order to obtain the ground truth of head poses, all head pose parameters including orientation and position have been labeled using SMI Eye Tracking Glasses as in Fig. 13.

The classroom database contains head pose spanning from -90° to 90° in horizontal direction, -45° to 90° in vertical direction. The face size is about 70*80 pixels in the image, as shown in Fig. 14. 3000 images from the database are used for training in 15 class head poses and the rest 1350 are for testing. Each tree is grown by 225 images sampled from a subset. 15 sub-forests have been respectively trained in different layers, where each sub-forest consists of 10 trees.

The training and testing parameter selection is similar to Section 3.1. As shown in Table 4, one can see that accuracies in yaw and pitch orientations are related to the confidence *pf*. $pf=0$ means that all positive patches are allowed for voting. $pf=0.5$ means too many patches are missing, which makes the accuracy decreased. In the case, we used $pf=0.25$ to control the number of patches to avoid the influence due to face deformation and large rotation angles.

Some examples of successful estimation are given in Fig. 15, where the red words in the upper area in these images represent the estimated head pose classes and regression angles, the clearer results are written in the below of images. Our approach can robustly estimate head pose under some wide range pose various, lighting and low resolution.

Besides, some experiments have been compared with C-RF [12], NN [14], and SVM multi-class [16] on this low resolution database for 15 class head pose estimation. The average accuracies are 80.5% using the proposed approach in this paper, 72.3% using C-RF, 55.1% using SVM multi-class, and 59.5% using NN method. The accuracy rate per pose has been shown in Fig. 16. The horizontal axis $x=1,2,..15$ represents each head pose label similar to $\{\theta_{yaw},\theta_{pitch}\}$. The vertical axis represents the estimated accuracy per head pose. One can see that our method is more robust than the other popular methods on the low resolution images.

3.3. Results of the occluded face images

In addition, some experiments have been done in some occluded face images. We randomly add occluded blocks on the images from the databases. Some successful examples result on the occluded test images using the proposed approach are shown in Fig. 17. Where the red words in the upper area in these images represent the estimated head pose, the clearer results are written in the below of images. Fig. 18 shows the mean error, averaged over all head pose angles. The extent of missing data is measured as the percentage of the area covered by the face bounding area. From the plot, the proposed approach is robust to such missing reconstructions. Even when 50% of the face was occluding, we still obtained an average error below 30%.

4. Conclusions

In this paper, we propose a more robust and efficient approach based previous work for head pose estimation in the vertical and horizontal directions in unconstrained environment. First, positive/negative facial patches are extracted and classified to eliminate the influence of noise and occlusion. Then, the D-RF is proposed to estimate the head pose in a coarse-to-fine way. A more powerful combined texture and geometric features are proposed to train and test in the D-RF. Furthermore, multiple probabilistic models have been learned in leaves of the D-RF and a novel composite weighted voting method is introduced to improve the discrimination capability of the approach. The proposed approach has been evaluated with three standard datasets spanning from -90° to 90° in vertical and horizontal directions under various conditions, whose accuracy reaches 76.2% on 25 classes. The proposed approach has also been evaluated with low resolution data collected from an overhead camera in our classroom, the average accuracy rate reaches 80.5% with 15 classes. Experiment results show that the D-RF performs more accurate and efficient than state of the arts, and is more robust to the low quality and occluded images. In future work, we intend to use the head pose and facial feature point to recognize the visual attention of multi-students in a smart classroom in a real time application. In future work, we intend to use the head pose and facial feature point to recognize the visual attention of multi-students in a smart classroom with the support of advanced massively parallel computing technologies [28, 29] in a real time application.

Acknowledgments

This work was supported by the National Key Technology Research and Development Program (no. 2013BAH18F02), Research funds from Ministry of Education and China Mobile (MCM20130601), Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (CCNU13B001), Research funds from the Humanities and Social Sciences Foundation of the Ministry of Education (no. 14YJAZH005), Central China Normal University Research Start-up funding (no. 120005030223), The Scientific Research Foundation for the Returned Overseas Chinese Scholars (no. (2013)693), Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (no. CCNU14A05020, no. CCNU14A05019), National Natural Science Foundation of China (no. 61272206), National Key Technology Research and Development Program (no. 2014BAH22F01).

References

[1] E. Murphy-Chutorian, M. Trivedi, Head pose estimation in computer vision: a survey, Trans. Pattern Anal. Mach. Intell. 31 (4) (2009) 607–626.
[2] J. Chen, D. Chen, X. Li, Towards improving social communication skills upon multimodal sensory information, IEEE Trans. Ind. Inf. 10 (1) (2013) 323–330.
[3] D. Zhu, X. Ramanan, Face detection, pose estimation and landmark localization in the wild, in: Proceedings of IEEE Conference CVPR, 2012.
[4] J. Chen, D. Chen, A feature-based detection and tracking system for gaze and smiling behaviors, Int. J. Comput. Syst. Sci. Eng. (2011) 207214.
[5] O. Sileye, A.J. Ba, Multi-person visual focus of attention from head pose and meeting contextual cues, IEEE Trans. Pattern Anal. Mach. Intell. (2011).
[6] G. Guo, Y. Fu, Head pose estimation: classification or regression?, in: The 19th International Conference on IEEE, 2008.
[7] J. Wu, M.M. Trivedi, A two-stage head pose estimation framework and evaluation, Pattern Recognit. 41 (2008) 1138–1158.
[8] M. Zhang, K. Li, Y. Liu, Head pose estimation from low-resolution image with Hough forest, in: 2010 Chinese Conference on Pattern Recognition (CCPR), IEEE, Chongqing City, China, 2010, pp. 1–5.
[9] N. Gourier, D. Hall, J. Crowley, Estimating face orientation from robust detection of salient facial features, Pointing 2004. in: ICPR International Workshop on Visual Observation of Deictic Gestures, 2004, pp. 183–191.
[10] Y. Liu, J. Chen, Y. Liu, Y. Gong, N. Luo, Dirichlet-tree distribution enhanced random forests for head pose estimation, in: International Conference of Pattern Recognition on Application and Methods (ICPRAM), 2014.
[11] Y. Li, S. Wang, X. Ding, Person-independent head pose estimation based on random forest regression, in: 2010 17th IEEE International Conference on Image Processing (ICIP), IEEE, Hong Kong, China, 2010, pp. 1521–1524.
[12] M. Dantone, J. Gall, G. Fanelli, L. Gool, Real time facial feature detection using conditional regression forests, in: Proceedings of IEEE Conference on CVPR, 2012.
[13] C. Huang, X. Ding, C. Fang, Head pose estimation based on random forests for multiclass classification, in: Conference on ICPR, 2010, 2010, pp. 934–937.
[14] N. Gourier, J. Maisonnasse, D. Hall, Head pose estimation on low resolution images, in: Multimodal Technologies for Perception of Humans, Springer, Berlin, Heidelberg, 2007, pp. 270–280.
[15] Y. Li, S. Gong, H. Liddell, Support vector regression and classification based multi-view face detection and recognition, in: IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 300–305.
[16] J. Orozco, S. Gong, T. Xiang, Head pose classification in crowded scenes, in: BMVC, 2009, pp. 1–3.
[17] G. Fanelli, J. Gall, Real time head pose estimation with random regression forests, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[18] G. Fanelli, T. Weise, J. Gall, L. Gool, Real time head pose estimation from consumer depth cameras, in: Conference on DAGM, 2011.
[19] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[20] S. Schulter, C. Leistner, P.M. Roth, On-line Hough forests, in: Conference on BMVC, 2011, 2011, p. 111.
[21] O. Barinova, V. Lempitsky, P. Kholi, On detection of multiple object instances using hough transforms[J], IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1773–1784.
[22] T. Minka, The Dirichlet-tree distribution, ⟨http://research.microsoft.com/minka/papers/dirichlet/minkadirtree.pdf⟩, 1999.
[23] X. Yan, C. Han, Multiple target tracking by probability hypothesis density based on Dirichlet distribution, J. XiAn JiaoTong Univ. 45 (2), 2011.
[24] Y. Liu, J. Chen, C. Shan, Dirichlet-tree distribution enhanced random forests for facial feature detection, in: ICIP, 2014, pp. 234–238.
[25] M. Figueiredo, A.K. Jain, Unsupervised learning of finite mixture models, IEEE Trans. Pattern Anal. Mach. Intell. 24 (3) (2002) 381–396.
[26] M. Jones, P. Viola, Fast Multi-view Face Detection. Tech. Rep., TR2003-096, Mitsubishi Electric Research Laboratories, 2003.
[27] G. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, Technical Report, University of Massachusetts, Amherst, 2007.
[28] D. Chen, X. Li, L. Wang, S.U. Khan, J. Wang, K. Zeng, C. Cai, Fast and scalable multi-way analysis of neural data, IEEE Trans. Comput. 64 (3) (2015) 707–719.
[29] D. Chen, L. Wang, A.Y. Albert, M. Dou, J. Chen, Z. Deng, S. Hariri, Parallel Simulation of Complex Evacuation Scenarios with Adaptive Agent Models, IEEE Trans. Parallel and Distributed Systems 26 (3) (2015) 847–857.

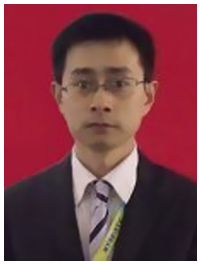
Yuanyuan Liu received B.E.degree from NanChang University, NanChang, China, in 2005, and M.E. degree from Huazhong University of Science and Technology, Wuhan, China, in 2007. She is currently a doctoral candidate for the Ph.D. degree in National Engineering Research Center for E-Learning, Central China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

Jingying Chen received the bachelor's and master's degrees from the Huazhong University of Science and Technology, Wuhan, China, and the Ph.D. degree from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2001. She was a Postdoctor in INRIA, France, and a Research Fellow with University of St. Andrews and University of Edinburgh, U.K. 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.

Zhiming Su received B.E. degree from Wuhan University of Technology, Wuhan, China, in 2013. Now, 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.

Zhenzhen Luo received B.E. degree form Hankou College, Wuhan, China, in 2010, and M.E. degree from Wuhan University of Technology, China, in 2013. Now, she is a doctoral candidate for the Ph.D. degree in National Engineering Research Center for E-Learning, Central China Normal University. Her research interests include pattern recognition and image processing.

Nan Luo received the B.E degree from China University of Geosciences, Wuhan, China in 2007. Now, he is a Master student for the M.E degree from National Engineering Research Center for E-Learning, Central China Normal University. His research interests include pattern recognition and image processing.



Leyuan Liu received the B.S. degree and M.E degree in computer science and engineering from Wuhan Institute of Technology, Wuhan, China, in respectively 2004 and 2007, and the Ph.D. degree in pattern recognition and intelligent system from Huazhong University of Science and Technology, Wuhan, China, in 2012. He is currently an Assistant Professor in the National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, China. His research interests include vision-based human–computer interactive, computer vision and pattern recognition.

Kun Zhang received B.E and Ph.D degrees from Huazhong University of Science and Technology, Wuhan, China, in respectively 2005 and 2010. Now he is an Assistant Professor in the National Engineering Research Center for E-Learning, Central China Normal University. His research interests include optical image processing and image enhancement.