International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 33, No. 7 (2019) 1940006 (30 pages)

DOI: 10.1142/S0218001419400068

Visual Focus of Attention and Spontaneous Smile Recognition Based on Continuous Head Pose Estimation by Cascaded Multi-Task Learning

Yuanyuan Liu*, Xingmei Li†, Fang Fang‡ and Fayong Zhang§

Faculty of Information Engineering
China University of Geosciences
Wuhan, P. R. China
**liuyy@cug.edu.cn*
†*524951054@qq.com*
‡*ffang1014@163.com*
§*zhangfayong@163.com*

Jingying Chen¶ and Zhizhong Zeng‖

National Engineering Research Center for E-Learning
Central China Normal University
Wuhan, P. R. China
¶*chenjy@mail.ccnu.edu.cn*
‖*zzzeng@mail.ccnu.edu.cn*

Received 11 April 2018
Accepted 16 October 2018
Published 27 December 2018

Multi-person Visual focus of attention (M-VFOA) and spontaneous smile (SS) recognition are important for persons' behavior understanding and analysis in class. Recently, promising results have been reported using special hardware in constrained environment. However, M-VFOA and SS remain challenging problems in natural and crowd classroom environment, e.g. various poses, occlusion, expressions, illumination and poor image quality, etc. In this study, a robust and un-invasive M-VFOA and SS recognition system has been developed based on continuous head pose estimation in the natural classroom. A novel cascaded multi-task Hough forest (CM-HF) combined with weighted Hough voting and multi-task learning is proposed for continuous head pose estimation, tip of the nose location and SS recognition, which improves accuracies of recognition and reduces the training time. Then, M-VFOA can be recognized based on estimated head poses, environmental cues and prior states in the natural classroom. Meanwhile, SS is classified using CM-HF with local cascaded mouth-eyes areas normalized by the estimated head poses. The method is rigorously evaluated for continuous head pose estimation, multi-person VFOA recognition, and SS recognition on some public available datasets and real-class video sequences. Experimental results show that our method reduces training time greatly and outperforms the state-of-the-art methods for both performance and robustness with an average

§Corresponding author.

accuracy of 83.5% on head pose estimation, 67.8% on M-VFOA recognition and 97.1% on SS recognition in challenging environments.

Keywords: Head pose estimation; multi-person VFOA recognition; spontaneous smile recognition; cascaded multi-task Hough forest.

1. Introduction

Multi-person Visual focus of attention (M-VFOA) and spontaneous smile (SS) is important for learners' behavior understanding and analysis in class.[9,16] It has great impact on teaching and learning. Hence, it's necessary to recognize M-VFOA and SS in class, which can make learners more engaged in the learning process for productive learning.[27,32] The VFOA of a person can be defined as the person or the object that a person is focusing his visual attention on Ref. 3. Researches on VFOA recognition can be classified into two types, i.e. invasive way based on special sensors and non-invasive way based on visual cues. Invasive systems are generally accurate and reliable depending on expensive sensors (e.g. SMI Eye Tracking Glasses or cameras mounted on a helmet).[33] However, the discomfort and restriction disrupt user's natural behaviors. In most of current work, non-invasive way has been shown that head orientation (head pose and location) can be reasonably utilized as an approximation of people's VFOA.[3,30] In the other word, facial expression is important to reflect a person's emotion on the attention target. Facial expressions recognition aims to classify a given facial image into one of the six commonly used emotion types, which include anger, disgust, fear, smile, sad and surprise proposed by Paul Ekman.[51] Among these six facial expressions, smile is distinct facial configuration. It's very informative in real-life applications. SS recognition can be used to assess a learners' learning interest in technology-enhanced learning (TEL) environment.[23]

In this paper, we focus on an automatic and non-invasive way for M-VFOA and SS recognition based on continuous head pose estimation. The existing techniques for the three tasks achieve satisfactory results in well-designed environments based on special sensors, in which images are usually acquired from a close distance. In a real-world classroom, however, factors, such as far distance, pose variation, occlusion, poor image quality, imbalanced data, etc., make head pose estimation, M-VFOA and SS recognition are much more challenging.[34,38,48] Recently, random forest (RF) has been adopted to estimate head poses, attention and expression from images, and the computational efficiency and readiness of parallel programming make them favorable choices.[8,18,26,29] In the paper, we propose a cascaded multi-task Hough forest (CH-HF) for continuous head pose estimation, facial feature location and SS recognition, which improves accuracies of recognition and makes the training procedure simpler than single task learning via introducing weighted Hough voting and multi-task learning to RFs. Meanwhile, a novel VFOA model is proposed to recognize attention states based on estimated head poses, environmental cues and prior states in the natural classroom.

1.1. *The pipeline of M-VFOA and SS recognition*

Figure 1 gives an overview of the M-VFOA and SS recognition system consisting of three modules, i.e. continuous head pose estimation, M-VFOA recognition and spontaneous expression recognition. Due to occlusions, various poses and low resolution in the natural environment, multi-persons' continuous head pose estimation and location is a very challenging problem in computer vision. In the head pose estimation module, a more discriminative CM-HF method is proposed to estimate continuous head pose and tip of the nose location of each person from the image. Then, M-VFOA recognition module gives the final VFOA targets of persons based on the estimated head poses, visual environment cues and prior state in the natural classroom scene. Concurrently, in the SS recognition module, each person's SS is recognized using CM-HF with local cascaded mouth-eyes areas normalized by the estimated head poses.

Figure 2 gives our system environment and a successful recognized example. In Fig. 2(a), we can see that our system just relies on a wide-angled overhead camera in the ceiling of the classroom without the need for any special hardware (e.g. goggles, head mounted equipment, image processing board or infrared sensitive cameras equipped with infrared LEDs[4,5]). Two recognized examples in different classes are shown in Fig. 2(b). The recognized result set $S_i\{\theta_{\text{yaw}}, \theta_{\text{pitch}}, \text{smile/no}-\text{smile}, T_i\}$

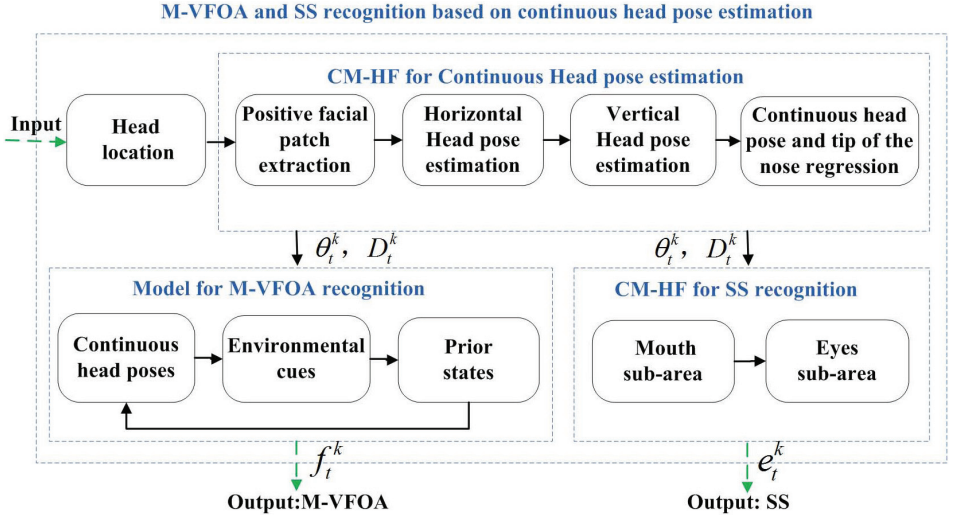
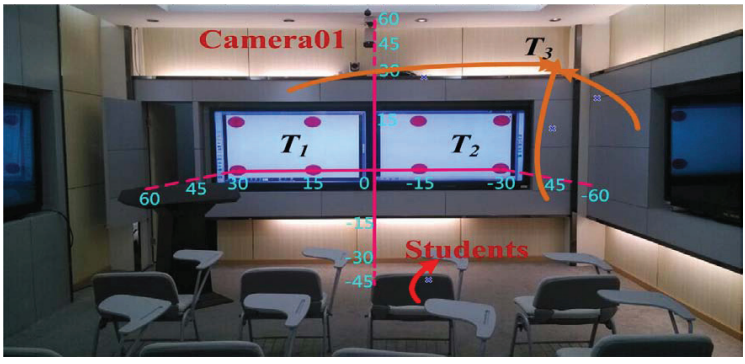
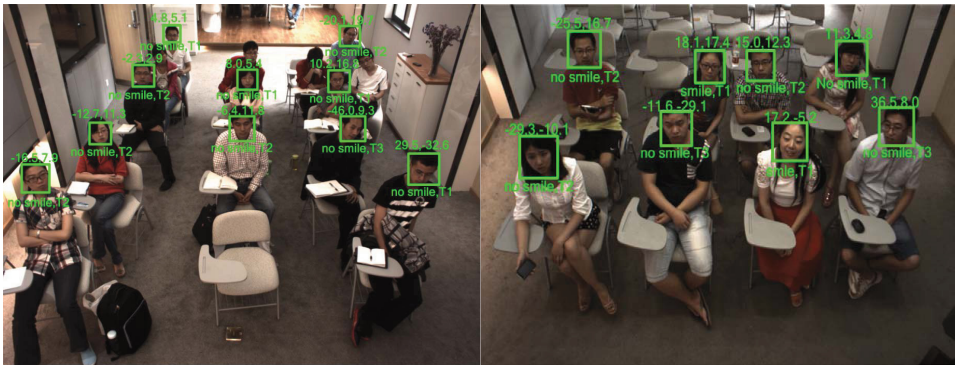


Fig. 1. The overview of the M-VFOA and SS recognition system based on continuous head pose estimation. The pipeline of the system includes three modules, i.e. continuous head pose estimation, M-VFOA and SS recognition. The parameters θ_t^k and D_t^k in head pose estimation module represent the estimated head poses in horizontal and vertical directions and tip of the nose of the person k at time t. The parameters f_t^k, e_t^k, in M-VFOA and SS module, denote the VFOA and SS state of the person k at time t, respectively.



(a)



(b)

Fig. 2. The system environment and recognized example. (a) The system environment (b) The recognition results of head location, poses, M-VFOA and SS in two real classes, where the green rectangles show the located faces, the angles $\{\theta_{\text{yaw}}, \theta_{\text{pitch}}\}$ above the rectangle show the estimated head poses in horizontal and vertical directions, and the $\{\text{smile/no} - \text{smile}, T_i\}$ below the rectangle show smile expression and the VFOA target.

represent each person's head poses, spontaneous expression, and the VFOA target, where $T_i\{T_1, T_2, T_3\}$ represent the VFOA of a person on left slide, right slide and outside position of the white board in the system, respectively.

1.2. *Our contributions and paper organization*

This paper is an extension version of our previous conference paper.[30] The main differences from the conference paper include: (1) This paper proposed a system for M-VFOA and SS recognition, while the conference paper only recognized VFOA; (2) A CM-HF method has been proposed for multi-tasks learning, such as, head pose estimation, tip of the nose regression and SS recognition, instead of single learning for head pose estimation in the previous conference paper; (3) A local cascaded

muti-features based patches from mouth-eyes areas have been presented for reducing the influence of facial appearance. The contributions of the paper are as follows:

(1) We develop a M-VFOA and SS recognition system based on continuous head pose estimation in the natural class scene.
(2) We propose a CM-HF method for head pose angles, tip of the nose regression and SS recognition, which is introduced by the weighted Hough voting and multi-task learning into RFs. The weighted Hough voting is the powerful tool capable of mapping complex input spaces into discrete and, respectively, continuous output spaces. The multi-tasks learning based CM-HF method can be trained for head pose, tip of the nose and SS sub-forests, concurrently, which can reduce training time greatly.
(3) Local cascaded multi-features based patches from mouth-eyes areas have been used for SS recognition, which can reduce the influence of facial appearance noises.

The remainder of the paper is organized as follows. In Sec. 2, we review related work. Section 3 describes the proposed CM-HF for multi-tasks training. Section 4 gives continuous head pose and tip of the nose estimation using CM-HF. Section 5 presents VFOA recognition based on head poses, environmental cues and prior states in the classroom scene. Section 6 details SS recognized by CM-HF. Section 7 discusses the experiments and results, as well as demonstrates uses of the environment. Conclusion and future work are given in Sec. 8.

2. Related Work

In this section, we highlight four key subjects that are the closest to our work, i.e. RF classification and regression, head pose estimation, VFOA recognition and spontaneous expression recognition.

Random Forest. 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.[7,6,10,11,15,41] It has emerged as a powerful and versatile method successful in real-time human pose estimation, object detection, facial point detection and action recognition. In Ref. 41, 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 in Ref. 11. Criminisi *et al.*[10] used RF regression to vote for the positions of the sides of bounding boxes around organs in CT images. Other works showed the power of RF in mapping image features to vote in a generalized Hough space.[16,50] We improved Hough forest (HF) with multi-task learning for continuous head pose estimation and SS expression recognition.

More information about RFs and their application in computer vision can be found in Ref. 47.

Head pose estimation. Head pose estimation is usually achieved using template matching, subspace embedding, and tracking methods.[28,32] Recently, classification and regression methods, such as neural networks (NN),[20] support vector machines (SVM),[34] nearest prototype matching,[46] and RF,[22,29,50] have gained popularity for head pose estimation in natural environment. Gourier *et al.*[20] applied an auto-associative network to learn the mapping for head pose estimation on low-resolution images. The method achieved a precision of 10.3° degrees in the yaw angle and 15.9° in the pitch angle on the Pointing'04 database. Orozco *et al.*[34] trained a multi-class SVM for head pose classification in crowded scenes. The performance on videos acquired in crowded public spaces with low resolution reached 80% accuracy rate in a four-pose classification. Wu *et al.*[46] proposed a two-stage framework for head pose estimation based on a geometrical structure. Multi-class and regression RFs become a popular method for head pose estimation on low resolution images owing to their robustness. Liu *et al.*[29] extended RF by introducing a Dirichlet-tree distribution. Zhang *et al.*[50] developed a head pose estimation based on HFs on low-resolution images.

Recently, deep Convolutional Neural Networks (DNNs) is applied for head pose estimation,[1,35,44] which achieved improved performance in cases such as multi-view occlusion and low image resolution. In Ref. 44, Venturelli *et al.* tackled the problem of head pose estimation through a CNN on the public Biwi Kinect Head Pose dataset. In Ref. 35, Patacchiola and Angelo proposed an approach based on DNNs supplemented with the most recent techniques adopted from the deep learning community. Moreover, they investigated the use of dropout and adaptive gradient methods giving a contribution to their ongoing validation. The improvement, however, heavily relies on the large number of training sets and high performance computing power.

Head Pose estimation is inevitably affected by environmental noises, hence, how to address head pose estimation in some challenging environment and limited training sets for improved performance is still an open problem.

VFOA recognition. VFOA is an important non-verbal communication cue with functions such as establishing relationships, regulating the course of interaction, retrieving information and exercising social control. Researches on VFOA recognition can be classified into two types, i.e. invasive recognition based on special sensors and non-invasive recognition based on visual cues. In this paper, we focus on non-invasive recognition methods. Non-invasive recognition includes the methods based on eye gaze direction or head orientation. Gaze estimation requires high resolution close-up views, which are generally not available in practice.[13] In real applications, it has been shown that head orientation can be reasonably utilized as an approximation of people's VFOA.[3,40,45] In Ref. 45, Michael Voit. *et al.* deduced the VFOA from coarse head pose and the accurate rate reached 57% in all frames. Authors in Ref. 3 presented investigations about VFOA recognition in general meeting scenario comprising head poses, a projection screen and a table as visual targets.

The averaged accuracy reached 55.1% with four users spontaneously. In Ref. 40, Smith *et al.* proposed the GMM and HMM methods for modeling VFOA in the wild from head poses. One can see that the non-invasive VFOA recognition is still a challenging task in natural and unconstrained large-scale scene. In this paper, we recognize multi-persons' VFOA based on the estimated head poses, visual environment cues and prior state in the natural classroom. Our previous work on VFOA recognition can be found in Ref. 30.

Spontaneous expression recognition. Among six facial expressions (smile, angry, surprise, neutral, fear, disgust), smile is a natural spontaneous facial expression and very informative in a class learning process.[2,36] Lots of works have existed and obtained excellent results on facial expression recognition in some special environment.[12,37,43,49] However, only a small part of work address some challenging issues in natural and unconstrained environment.[27] For spontaneous expression recognition, a general recognition framework appeared in most of previous works can be divided into two major steps, i.e. feature extraction and classifier construction. Ito *et al.*[23] detected facial features, i.e. eyes, nose and mouth, and generated feature vectors, the lip lengths, lip angles and mean intensities of the cheek areas, firstly. Then, they recognized the smile expression using a linear discrimination function based on obtained feature vectors. However, their facial feature detection method is sensitive to various illuminations. In Ref. 14, Fanelli *et al.* presented a HF for facial expression recognition from image sequences, which achieved a recognition rate of 76% on MMI spontaneous expression database. Linear Discriminant Analysis (LDA) and SVM are used as classifiers with Local Binary Pattern (LBP) features for smile detection on GENKI-4K dataset by An *et al.*,[2] which achieved 85% of the average accuracy. It's noted that the good results usually have happened in constrained environments or depending on multi-cameras in special scenes. How to address spontaneous expression recognition problem in crowd and unconstrained environment for improved performance is still an open problem.

3. The CM-HF Training for Multi-Tasks

A CM-HF method has been proposed for multi-tasks learning, such as, head pose estimation, tip of the nose regression and SS recognition, instead of single-task learning for head pose. We train the CM-HF based on multi-task learning for head pose, tip of the nose and SS, concurrently. The trained CM-HF includes head pose sub-forest, tip of the nose sub-forest, and SS sub-forest in different hierarchical layers of the whole forest (see Fig. 3). In order to train multi-task sub-forests in different layers of the CM-HF, the training images have been divided into hierarchical training sub-sets, including head pose training sub-set, tip of the nose training sub-set and SS sub-set. For each facial area, we normalize it as the size of 125*125, and randomly extract 200 facial patches $P = \{p_i\}$ from the whole face, where $p_i = \{(F_i; k_i, H_i^m, D_i, E_i)\}$. The patch appearance F_i is defined as multiple texture feature channels $F = \{F_i^1, F_i^2, F_i^3\}$. F_i^1 contains the gray values of the raw facial patch with

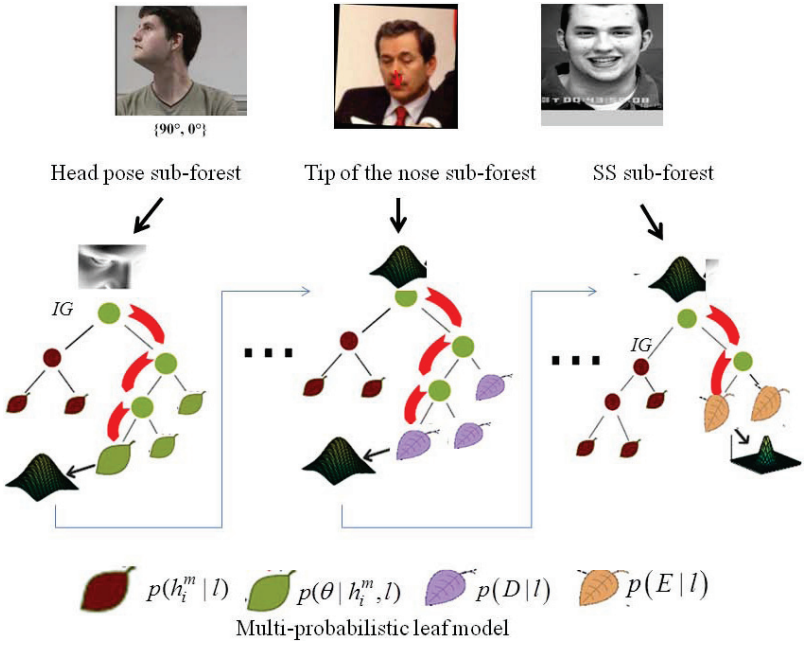


Fig. 3. The cascaded multi-task Hough forest training.

dimension as 31*31. F_i^2 represents the Gabor feature-based PCA (Principal component analysis) of positive facial patches with dimensions as 35*12, where Gabor filter bank with eight different rotations and five different phase shifts. F_i^3 is the histogram distributions of the patches. The $k_i = \{1, 0\}$ are labeled as facial patch classes, where $k_i = 1$ represent positive facial patches that are real facial patches and $k_i = 0$ represent negative facial patches that include occlusion or background noise from the detected facial areas. The channel $H_i^m = \{h_i^m, \theta_{\text{yaw,pitch}}\}$ is the contextual head pose label, which contains the head pose class and angles. The annotations of h_i^m in different layers of CM-HF are referred to Ref. 29, $\theta_{\text{yaw,pitch}}$ represents the head pose angles in the horizontal and vertical directions. $D_i = p_i - N$ is the offset vector from a patch centroid p_i to the tip of the nose N. E_i is labeled as SS expression class, where $E_i = 1$ represents smile expression and $E_i = 0$ represents non-smile expression. It is noted that the facial patches from mouth and eyes areas can be trained for SS sub-forest layer. The training procedure for multi-tasks is given in the following.

First, we define a patch comparison feature φ as our binary tests:

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

where R_1 and R_2 are two random rectangles within the patches, $F^n(j)$ is the texture feature channel ($n \in 1, 2, 3$), j is a pixel point from two rectangles.

To construct a tree in each sub-forest, a node divides a set of training patches P into two subsets P_L and P_R for each comparison feature φ. Thus, we have

$$P_L = \{P|\varphi < \tau\}, \quad P_R = \{P|\varphi > \tau\}, \tag{2}$$

where τ is a predefined threshold.

Then, selecting a classification splitting candidate φ that best splits the feature set P into two subsets P_L and P_R to maximize the evaluation function Information Gain (IG)

$$\arg\max_{\varphi} (H(P|a_j) - \sum_{S\in\{L,R\}} \frac{|P_S|}{|P|}(H(P|H_i^m) + H(P|D_i) + H(P|E_i)), \tag{3}$$

where $H(P|H_i^m)$,$H(D_i)$, $H(P|E_i)$ are the entropies of the different set P for annotated head pose parameters, offset vector of tip of the nose and SS expression.

Different from the previous work,[29] our labels are modeled as a continuous distribution including angles, geometric offset of tip of the nose and SS expression. So, we improved integrated entropy as

$$\begin{aligned}
H(P|H_i^m) &= -\frac{\sum_i p(H_i^m|a_j, k=1, P)}{|P|} \log\left(\frac{\sum_i p(H_i^m|a_j, k=1, P)}{|P|}\right),\\
H(P|D_i) &= -\frac{\sum_i p(D_i|a_q, k=1, P)}{|P|} \log\left(\frac{\sum_i p(D_i|a_j, k=1, P)}{|P|}\right),\\
H(P|E_i) &= -\frac{\sum_i p(E_i|a_r, k=1, P)}{|P|} \log\left(\frac{\sum_i p(E_i|a_j, k=1, P)}{|P|}\right),
\end{aligned} \tag{4}$$

where $p(H_i^m|a_j, k=1, P)$ indicates the probability that the facial positive patch belongs to the head pose parameter H_i^m in the sub-forest a_j of the jth layer, $p(D_i|a_q, k=1, P)$ is the probability that the positive patch belongs to the tip of the nose in the sub-forest a_q, and $H(P|E_i)$ is the probability of SS expression in the forest a_r. The parameters j, q, r are the hierarchical layers within the CM-HF.

The training continues until the tree reaches the maximum depth or the number of samples in a node falls below a threshold. The construction process creates a leaf l when IG is below a predefined threshold or when the maximum depth is reached. A leaf node stores the multi-probabilities for multi-tasks as the facial patch class $p(k_i|l)$, head pose classification $p(h_i^m|l)$, the distant offset of nose tip $p(D|l)$, continuous head pose angles $p(\theta|h_i^m, l)$, and SS expression $p(E|l)$.

4. Continuous Head Pose Estimation

In this section, the CM-HF and weighted Hough voting is used to estimate continuous head pose and tip of the nose in a hierarchical way. Different from our previous work Dirichlet-tree distribution enhanced random forests (D-RF),[29] in this paper, the weighted Hough voting and multi-task learning are introduced into cascaded RF

framework as CM-HF for head pose angles, tip of the nose regression and SS recognition. The weighted Hough voting is the powerful tool capable of mapping complex input spaces into discrete and, respectively, continuous output spaces. The CM-HF estimates continuous head pose and tip of the nose from cascaded classification to regression via multi-task forest. We describe the testing procedure of CM-HF sub-forests for head pose and tip of the nose in details.

4.1. *Face location*

In a natural and crowd class scene, multi-persons' face detection and tracking is a challenging task due to wide angles and low resolution. In this case, a cascade of boosted classifiers with Haar-like feature[24] has been trained to detect faces with several facial databases collected from Internet and real-life videos. In order to decrease false detection, detected face areas were averaged in the initial 30 frames of a sequence. Then, robust facial detection is performed within sub-windows, which are extensions of the average face areas. After face detection, a mean shift method with skin color and motion information is used to track face areas.[47] In the system, it is able to detect tracking failures using constraints derived from the distance between persons' positions. Once the tracking failure has been detected, re-initialization and face detection is needed.

4.2. *Continuous head pose angles and tip of the nose regression using CM-HF*

Figure 4 shows that continuous head poses including angles and tip of the nose are estimated by the CM-HF in the horizontal and vertical directions. It includes five cascaded sub-layers, where L1 and L2 sub-layers are estimation in the horizontal direction, L3 and L4 sub-layers are estimation in the vertical direction, and L5 sub-layer is rotating angles and tip of the nose regression. By passing all the patches down all the trees in the CM-HF, the patches end in a set of leaves of different trees within the relative sub-layer.

In L1 to L4 sub-layers, we simplify the distribution over the discrete head pose class by an adaptive multi-variance Gaussian Mixture Model (GMM)[29]:

$$p(h^m|l_m) = N(h^m; \bar{h}^m, \Sigma_{l_m}), \tag{5}$$

where $\bar{h}^m$ and Σ_{l_m} are the mean and covariance matrix of the contextual head pose class. The l_m represents the leaf probability in the mth sub-layer.

In L5 sub-layer of the CM-HF, a weighted Hough voting method is proposed for head pose angles and tip of the nose regression, simultaneously. In order to eliminate imbalance of samples, we also store the weight $w_s = P_S/P$ that is defined as the ratio of the samples in each subset P_S and the total samples P in each single tree of the CM-HF.

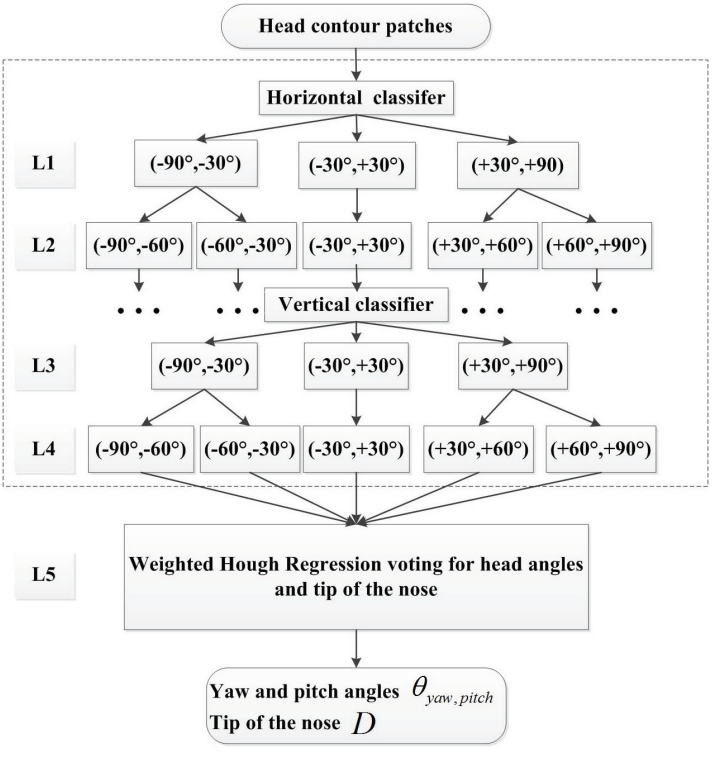


Fig. 4. The flow chart of continuous head pose estimation and tip of the nose location, which includes five cascaded sub-layers. L1 and L2 sub-layers are horizontal classification, and L3 and L4 sub-layers are vertical classification. In the L5 sub-layer, the head pose angles and tip of the nose can be regressed under the contextual head pose class. For clarity, only the trees of L3 sub-layer under the horizontal class -30° to 30° are shown.

To integrate multi-probabilistic votes by different patches P, we accumulate them in an additive way into a 3D Hough image $V(p_i, D, \theta)$, where we sum up the votes in the sub-forest $\{T_t\}_{t=1}^{N}$ for each pixel-location p_j,

$$V(p_j) = \sum_{P} p(D, \theta | P; \{T_t\}_{t=1}^{N}). \tag{6}$$

The estimation procedure simply computes the Hough image V. Then, it returns the set of its maximum patch locations as the estimation hypotheses. The $V(p_j)$ values serve as the confidence measures for each hypothesis vote. In a Hough image, an alternative way to find the maximum would be to use the mean-shift procedure as it is done in other Hough voting-based frameworks.[16] To vote the continuous head pose and tip of the nose in the patch location p_j, we set the weighted Hough voting model as

$$V(D, \theta) \propto K((w_S V(p_j) - (p_j + \overline{w_S V(p_j)}))/h_j), \tag{7}$$

where a Gaussian Kernel K and the bandwidth parameter h_j are given by Gaussian filter, $\theta = \theta_{\text{yaw,pitch}}$ represents the Hough voting result for head pose angles in a Hough image, and D is the position of the patch that belongs to tip of the nose. Then, a regression voting method provides good results by evaluating sparse positive patches in the fifth sub-layer, rather than all patches in the sub-forest.

5. M-VFOA Recognition

In this section, we develop an M-VFOA recognition system based on continuous head pose estimation in the natural class scene. This system can simultaneously recognize the M-VFOA targets in the classroom by a monocular overhead camera. A novel VFOA model is proposed to recognize and track attention based on head poses, prior state and visual environmental cues as shown in Fig. 5. $\theta_t^k = \{\theta_{\text{yaw,pitch}}\}$ represent estimated head poses of the person k in horizontal and vertical directions at time t. c_t represents the environmental cues currently. f_t^k and f_{t-1}^k denote the VFOA states of person k at time t and prior time $t-1$. Environmental cues c_t include the physical placement of targets and the participant's 3D positions in the classroom. Prior state comprises the prior recognized attention state and some prior 3D cues in the classroom.

5.1. *Environmental cues*

Due to unknown 3D cues, M-VFOA recognition from head poses captured by a monocular camera is difficult. In order to solve this problem, we introduce an approximated method to obtain the environmental cues $c_t = \{T_i, B_t^k\}$ based on some prior state. $T_i = \{T_1, T_2\}$ are the physical placement of attention targets in the white board and could be measured previously (see Fig. 1). $B_t^k = \{B_x, B_y, B_z\}$ is the 3D position of a person estimated from a monocular camera at time t in class. According to averaging 100 persons' sitting height (the height of tip of the nose) in the classroom, we assumed a 2D reference plane with its height $B_y = 120\,\text{cm}$ as the prior state of 3D cues, firstly. Then, 35 different points in this 2D reference plane were labeled according to their image coordinates.

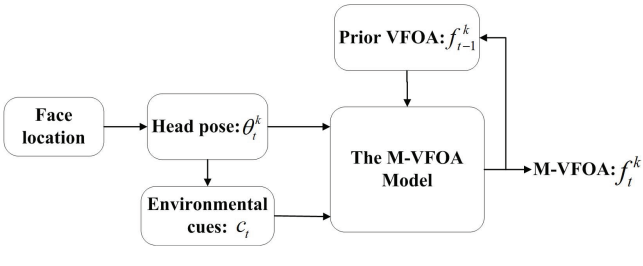


Fig. 5. The VFOA model for M-VFOA (f_t^k) recognition based on head poses (θ_t^k), prior state (f_{t-1}^k) and visual environmental cues (c_t).

Homography matrix H between the 2D reference plane and the image was computed based on these labeled points by the affine transformation $h(\bullet)$, firstly. In the recognizing step, the reverse procedure could be performed. Each person's position $B_t^k = \{B_x, B_y = 120, B_z\}$ can be obtained based on located tip of the nose D_t^k and the computed Homography matrix H by the affine transformation $h(\bullet)$.

$$B_t^k = h(D_t^k, H), \text{ with } B_y = 120. \tag{8}$$

5.2. *M-VFOA model*

In order to recognize M-VFOA in the natural classroom, the attention point $T(x, y)$ of a person is computed under the estimated head pose $\theta_{\text{yaw,pitch}}$ and his position B_t^k in the scene. The geometric model between $T(x, y)$, $\theta_{\text{yaw,pitch}}$, and B_t^k is defined as follows,

$$T(x, y) = \left\{ \frac{B_z}{\cot(\theta_{\text{yaw}})} + B_x, B_y + \frac{\tan(\theta_{\text{pitch}}) \times B_z}{\cos(\theta_{\text{yaw}})} \right\}. \tag{9}$$

This model is shown in Fig. 6.

If the attention point $T(x, y)$ belongs to the white board $T_i, i = 1, 2$, it means that VFOA of a person is within the white board T_i. Then VFOA (f_t^k) of a person k could be recognized as:

$$f_t^k = \sum_t \sum_{k=1, i \neq k}^{K} \delta(T_t^k - T_i), \quad k = 1, \ldots, K; i = 1, 2, \tag{10}$$

where $\delta(T_t^k - T_i) = 1$ represents that a person focuses the targets T_1 or T_2. While $\delta(T_t^k - T_i) = 0$ represents that a person does not focus the white board and VFOA directly outputs T_3.

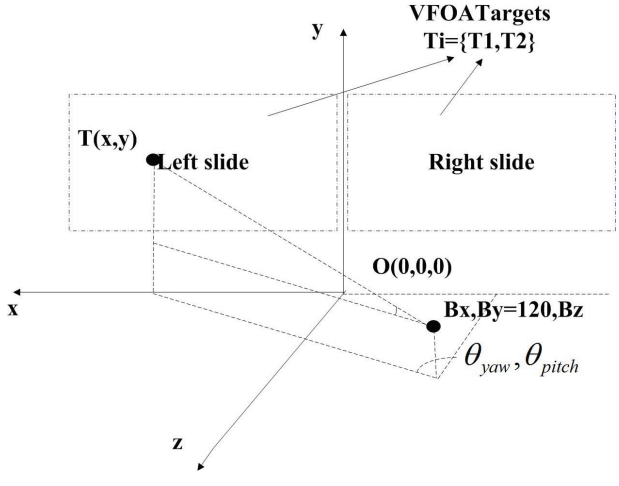


Fig. 6. The M-VFOA Geometric model between the attention point $T(x, y)$, a person's position B_t^k and the head pose $\theta_{\text{yaw,pitch}}$.

5.3. *M-VFOA tracking with the prior state*

To recognize M-VFOA in a video sequence from a classroom, we rely on GMM to track the jointed model displayed in Fig. 5. Estimating the visual focus can be posed in a probabilistic framework as finding the VFOA state maximizing the a posteriori probability,

$$f_t^k = \underset{f_t^k \in T_i}{\arg\max}\ p(f_t^k|\theta_t^k) \text{ with } p(f_t^k|\theta_t^k) \propto p(\theta_t^k|f_t^k)p(f_t^k). \tag{11}$$

For each person's VFOA, it can be modeled as a Gaussian distribution of head pose $N(\theta_t^k|f_t^k;\mu_i,\Sigma_i)$ with the mean μ_i and the covariance matrix Σ_i. Thus, in the modeling, the head pose distribution is represented as a GMM, with the mixture index (i) denoting the focus target,

$$p(\theta_t^k|f_t^k) = \sum_{i=1}^{t-1} \pi_i N(\theta_t^k|f_t^k;\mu_i,\Sigma_i), \tag{12}$$

where π_i represents the parameter set of linear relation where head pose relate to the VFOA direction. t is the number of frames in the sequence. The property is used for the method recognizing and tracking M-VFOA from head pose that we consider as our baseline model.

6. SS Recognition

In this module, the CM-HF method is used for SS recognition in challenging environments. Local cascaded multi-features based patches from mouth-eyes areas have been used for SS recognition, which can reduce the influence of facial appearance noises. The CM-HF training for SS recognition is shown in Sec. 3, which is multi-task learning. For the SS sub-forest learning, CM-HF was carried out using 5700 smiling images and 14500 non-smiling images, which were obtained from different sources, e.g. public expression databases, Internet and captured video sequences in a natural classroom scene. For local cascaded feature extraction, the local mouth and eyes areas have been detected by a cascade of Adaboost tree classifiers with Haar-like features[24] and are normalized based on the estimated head pose $\theta_{\text{yaw,pitch}}$. Figure 7

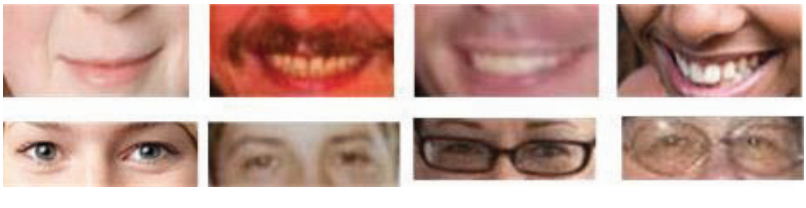


Fig. 7. Examples of mouth and eyes areas normalized by the estimated head poses from smile images.

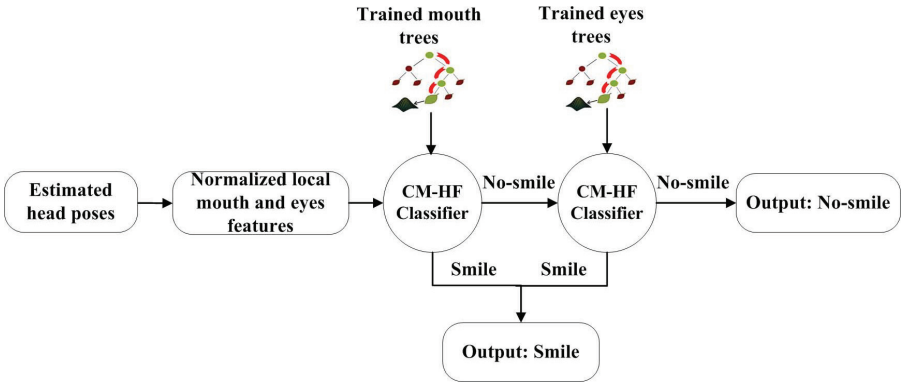


Fig. 8. The pipeline of SS recognition by CM-HF.

shows some examples of normalized local mouth and eyes areas extracted from smiling images. The SS sub-forest consist of mouth trees and eyes trees. The former is trained with mouth areas based patch features, while the latter is trained with eyes areas based patch features.

The testing framework for SS recognition is shown in Fig. 8. In order to eliminate the influence of head poses, we normalize the local mouth and eyes areas, by estimated head poses. After pose normalization, the CM-HF is used to recognize each person's SS expression state in a cascaded way. Mouth trees are firstly used to classify smile/no smile in the mouth area. When the classified result is "smile", the testing procedure is ending and outputs smiling state. When the classified result is "no smile", eyes trees compensate to classify the expression state in the eyes areas in a cascaded way. The probability of SS in the CM-HF can be achieved as

$$\begin{aligned} p(E_i = 1|\theta_{\text{yaw,pitch}}) &= p(E_i = 1|\theta_{\text{yaw,pitch}}, f_{\text{mouth}}) \\ &\quad + p(E_i = 0|\theta_{\text{yaw,pitch}}, f_{\text{mouth}}) p(E_i = 1|\theta_{\text{yaw,pitch}}, f_{\text{eyes}}); \\ p(E_i = 0|\theta_{\text{yaw,pitch}}) &= p(E_i = 0|\theta_{\text{yaw,pitch}}, f_{\text{mouth}}) p(E_i = 0|\theta_{\text{yaw,pitch}}, f_{\text{eyes}}), \end{aligned} \tag{13}$$

where $p(E_i = 1|\theta_{\text{yaw,pitch}})$ represents the probability of the SS expression stored in leaves of SS sub-forest, $p(E_i = 0|\theta_{\text{yaw,pitch}})$ represents the probability of no-smile expression, $\theta_{\text{yaw,pitch}}$ is the estimated head poses in horizontal and vertical directions, f_{mouth} and f_{eyes} represents the mouth and eyes trees, respectively.

The weighted Hough voting method is used to vote the leaves' probabilities in the sub-forest. We can obtain the final expression by averaging the leaves of trees:

$$p(E|\theta_{\text{yaw,pitch}}) = \frac{1}{T} \sum_{T} \sum_{i} p_i(E|\theta_{\text{yaw,pitch}}, V), \tag{14}$$

where T is the number of trees within the sub-forest and V is the Hough space of the image patches.

7. Experiments

In this section, we thoroughly evaluated the proposed approach for head pose estimation, M-VFOA and SS recognition on some public available datasets and real videos in challenging environments and the natural classroom scene.

7.1. *Datasets and settings*

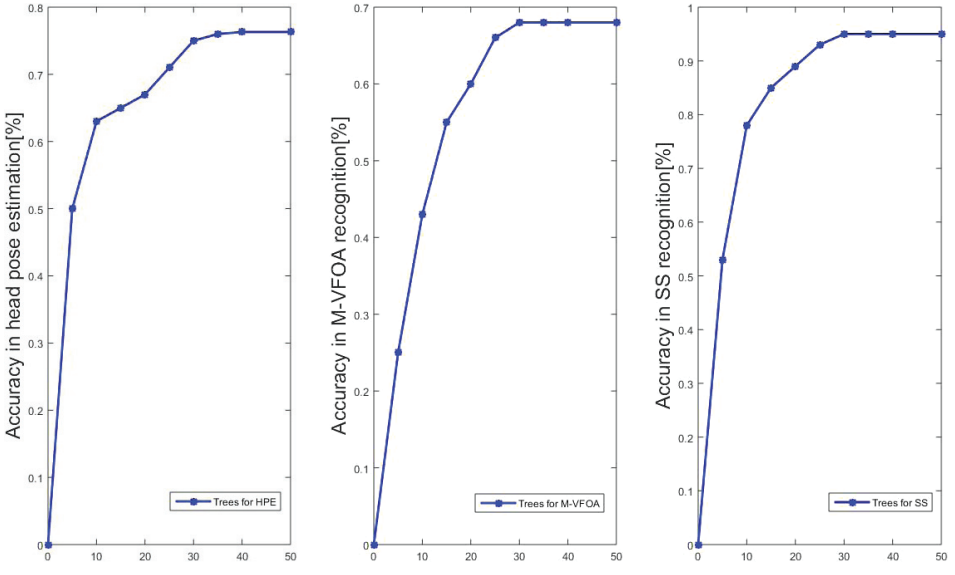
To evaluate our method on M-VFOA recognition and head pose estimation, four head pose datasets were used: Pointing'04 dataset,[19] LFW dataset,[21] IDIAP head pose dataset,[39] and CCNU head pose dataset in the wide classroom.[29] The Pointing'04 head pose dataset is a benchmark of 2790 monocular face images of 15 people with variations of pan and tilt angles from -90° to $+90^\circ$. For every person, two series of 93 images (93 different poses) are available. The LFW dataset consists of 5749 individual facial images. The images were collected in the wild, and varied in poses, lighting conditions, resolutions, races, occlusions, and make-up. The IDIAP dataset was collected along with a head pose ground truth and each participant's discrete VFOA ground truth. The dataset is comprised of eight meetings involving four people each, recorded in a meeting room, ranged from 7 to 14 minutes. The CCNU dataset was collected including an annotated set of 58 people with 75 different head poses from an overhead camera in the wide scene. It contains head poses spanning from -90° to 90° in horizontal direction, and -60° to 90° in vertical direction. Our method for VFOA recognition was trained with 2000 images from Pointing'04, 5000 images from LFW dataset, and 4000 images from CCNU dataset. In evaluation, 500 images from Pointing'04 dataset, 1000 images from LFW dataset, 1500 images from CCNU dataset, three meeting videos from IDIAP dataset, and real life videos were used.

To evaluate our method on SS recognition, three public facial expression datasets were used: Cohn-Kanade (CK+) dataset,[31] GENKI-4K dataset,[25] CCNU dataset, and real class videos. The CK+ database is a widely used benchmark for evaluating expression recognition techniques, which contains 593 image sequences across 128 subjects, which contains images from neutral face to peak expression face. GENKI dataset was collected from Internet and captured video sequences in real life, which was labeled as SS and no-smile expressions. For testing, we randomly sample 500 images from CK dataset, 500 images from GENKI dataset, and six class video sequences captured from an overhead camera in the natural and crowd classroom scene.

The proposed method has been implemented using C++, OpenCV library, Boost library under MS windows environment with Intel(R) Core(TM) i5-2400 CPU@ 3.10 GHz, RAM 8 GB. And an overhead camera is used to capture images and videos on the ceiling of the crowd classroom scene.

7.2. *Training*

For multi-task training, we give the accuracies with regard to the number of training trees as Fig. 9, for head pose estimation, M-VFOA recognition, and SS recognition.



(a) The number of the trees (b) The number of the trees (c) The number of the trees

Fig. 9. Accuracies with regard to the number of trees in the CM-HF, (a) For head pose estimation, (b) for M-VFOA recognition, (c) for SS recognition.

Other parameters including the number of patches extracted from each image (fixed to 200), the splitting candidates in a tree (2000), depth in a tree (15), and the patch's size (30*30) are similar to Ref. 29. Each tree grows based on a selected subset of 300 images. The plots in Fig. 9 show the performance of the method when we varied the number of trees in CM-HF. One can see that the best performance occurred in over 40 trees trained for the CM-HF. Compared to over the number of trees 100 in RF[6] training and 60 in D-RF training for each single task,[29] our multi-task training costs fewer trees, which can reduce training time greatly.

7.3. *Continuous head pose estimation*

7.3.1. *Face location*

In the location step, we have achieved the average detection rate of 94.7% in LFW dataset, CCNU classroom dataset, and class videos. The average tracking time is 0.007 s per frame. The proposed method is able to detect tracking failures using constraints derived from the distance between persons' positions. Once the tracking failure has been detected, re-initialization and face detection is needed.

7.3.2. *Head pose estimation and tip of the nose location on Pointing'04, LFW, and CCNU datasets*

Table 1 lists the accuracies with respect to the yaw and pitch rotations of our CM-HF method as well as the average errors in terms of degrees. A 4-fold cross-validation was

Table 1. Accuracies and average errors (degrees) of CM-HF method for head pose estimation and tip of nose location on different datasets.

Datasets	θ_{yaw} (%)	θ_{pitch} (%)	$\theta_{\text{yaw,pitch}}$ (%)	Ave. Error	STD.	Tip of the Nose
Pointing'04	90.8	94.2	83.5	6.5°	3.4°	0.14 pixels
LFW	85.4	91.4	73.8	12°	6.1°	0.12 pixels
CCNU	84.2	90.5	74.0	13.5°	5.9°	0.6 pixels

conducted. Among the three datasets, our method achieved the greatest performance with Pointing'04. The accuracy in yaw and pitch angles were in the range of 80–90%, respectively. Note that both LFW and CCNU datasets consist of great variation of poses, lighting, occlusions, etc. For these two more challenging datasets, the accuracy was also above 80% for yaw rotation and above 70% for yaw and pitch rotation. The average error reached 6.5° for the Pointing'04 and those of the other two were close to 10°. In all cases, the standard deviation of the average error was fairly low. The average location error on tip of the nose reached 0.4 pixels. The examples of continuous head pose estimation and tip of the nose location are shown in Fig. 10, where the first row is estimated results on LFW dataset, the second row is on Pointing'04 dataset, and the last row is on CCNU dataset. One can see that our proposed CM-HF

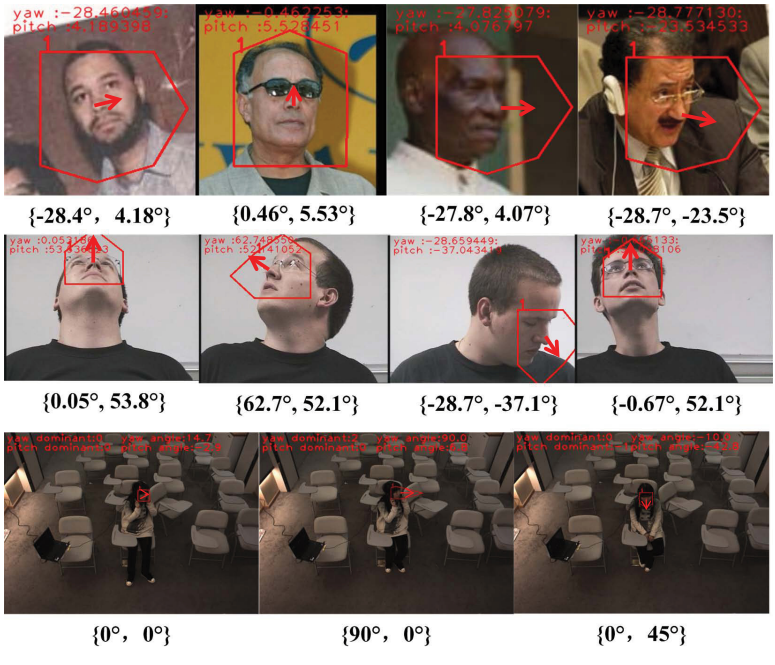


Fig. 10. The successful examples of continuous head pose estimation and tip of the nose location on LFW, Pointing'04, and CCNU datasets. The estimated head pose angles are shown in the images, and more clearer results can be seen below the images. The red arrows are the results of tip of the noses (color online).

method can estimate continuous head pose angles and locate tip of the nose in some challenging environments including occlusions, expressions, poses, low resolution, and make-ups, etc.

7.3.3. *Compare to the state-of-the-art methods for continuous head pose estimation and tip of the nose location*

In comparison with the state-of-the-art head pose estimation methods, we conducted experiments using the Dirichilet-tree enhanced random forest (D-RF),[29] Approximate view manifolds (AVM),[42] HF,[17] Tree-structured parts model (TSPM) of Ref. 53, Multi-class RF,[22] multi-class SVM,[34] NN[20] and CNN.[1] We employed a 4-fold cross-validation and used the same training and testing datasets. Table 2 lists the average accuracies and errors across three head pose datasets, including Pointing'04, LFW and CCNU head pose datasets. D-RF[29] and HF[17] yielded comparable results with an accuracy of approximately 70% in yaw and pitch rotations. AVM,[42] TSPM of Ref. 53, Multi-class SVM,[34] Multi-class RF,[22] and NN[20] produced similar accuracy in the range of 60%. The CNN[1] in this experiment contains three convolution layers followed by three max-pooling layers and two fully connected layers and obtains the accuracy of 63.29%. CM-HM exhibited the highest accuracy of 76.3% and the accuracy of the estimation of the yaw component reached 86.5%. The weighted Hough voting method removes the unwanted patches from face deformation and large rotation angle in unbalanced sample sets, which ensures improved accuracy in our proposed method. The improvement with respect to the second best is about 9%. We get the same observation from the average estimation error. The average error of CM-HM method was 9.5°. In addition, the standard deviation of CM-HM indicates that CM-HM achieved the greatest consistency with a small STD. It is evident that our CM-HM improved the head pose estimation with great robustness. Meanwhile, the average location errors of tip of the nose have been compared to relative methods, i.e. D-RF, TSPM, HF, and RF. The best performance is 0.4 pixels by using the CM-HF.

Table 2. Accuracy and average error of head pose and tip of the nose estimation using different methods.

Methods	θ_{yaw} (%)	θ_{pitch} (%)	$\theta_{\text{yaw,pitch}}$ (%)	Ave. Error	STD.	Tip of the Nose
AVM[42]	80.56	74.75	58.33	17.2°	6.7°	—
D-RF[29]	83.52	86.0	71.83	13.5°	5.5°	1.3 pixels
HF[17]	82.3	85.6	70.5	13.7°	5.2°	0.5 pixels
TSPM of[53]	81.9	72.4	57.8	15.3°	10.2°	1.2 pixels
Multi-class SVM[34]	80.6	74.9	60.4	20.2°	5.7°	—
Multi-class RF[22]	78.4	79.3	62.23	26.3°	8.4°	1.5 pixels
CNN[1]	81.4	73.94	63.29	19.5°	6.8°	—
NN[20]	79.5	71.3	56.7	29°	7.5°	—
Our CM-HF	86.5	88.2	76.3	9.5°	5.0°	0.4 pixels

7.4. *M-VFOA recognition*

Evaluation was conducted using our CCNU classroom dataset, the IDIAP dataset, and real-class videos in the natural and crowd classroom. The recognition rate R_f in the sequence is defined as

$$R_f = \frac{\sum_t N_{\text{matched}}}{\sum_t N_{T_i}}, \tag{15}$$

where $\sum_t N_{\text{matched}}$ represents the number of correct recognized targets in the recognized sequence t that match the same target in the T_i, and N_{Ti} is the number of target events in the ground truth T_i. In order to obtain the ground truth of VFOA, all attention parameters including orientation and position have been labeled using SMI Eye Tracking Glasses.

Some VFOA recognition results are provided in Fig. 11. The experiments have been performed on six videos in real classes. Each recording includes eight persons in ten minutes long. The average accuracy reaches 67.8% using the proposed approach. The individual accuracy on T_1, T_2, T_3 are 62.28%, 60.81% and 80.2% correspondingly. The accuracies on T_1 and T_2 are inferior to T_3 due to their smaller scales in the classroom.

Figure 12 presents the plot of the VFOA recognition rate with different B_y values of environmental cues in the classroom. It shows that the $B_y = 120$ cm could obtain the better performance. Additionally, we report the results obtained when setting the prior VFOA states. Figure 13 shows the M-VFOA recognition results with prior VFOA states and without prior states in real-class videos. The blue bars represent the recognition rate in different target with prior states, and the red bars represent

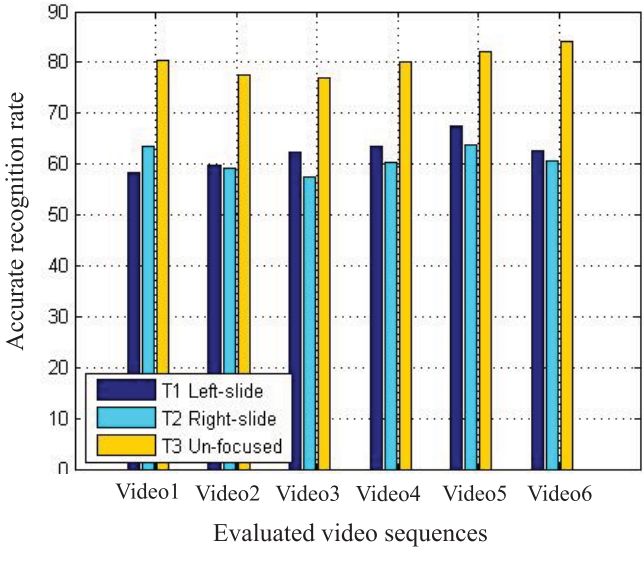


Fig. 11. M-VFOA recognition rate (%) in different target event.

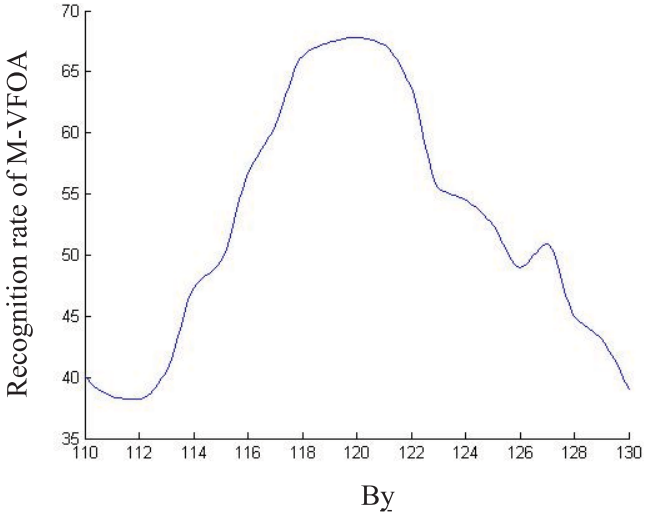


Fig. 12. M-VFOA recognition rate with environmental cues B_y.

the recognition rate in different target without prior states. One can see that the recognition results with prior states take more robust performance.

Table 3 provides the VFOA recognition results obtained using different head pose estimation methods on both real class videos and IDIAP meeting videos, including the CM-HF proposed in this paper, D-RF,[29] HF,[17] RF.[22] One can see that the best recognition rate can be obtained using our proposed CM-HF method.

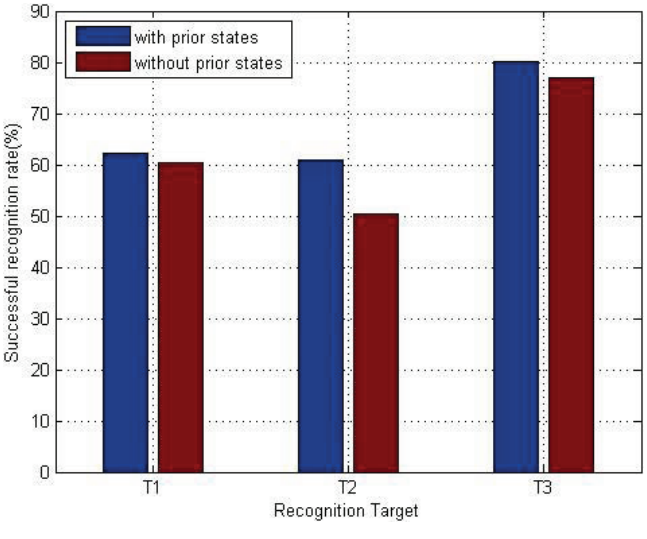


Fig. 13. M-VFOA recognition results with prior states vervus without prior states. The recognition results with prior states are more robust than the without prior states ones.

Table 3. VFOA recognition rates with different head pose estimation methods on two datasets.

	Class Videos		IDIAP Dataset	
Methods	Recog. Rate (%)	Ave. Error (%)	Recog. Rate (%)	Ave. Error (%)
D-RF[29]	63.5%	38.4%	68.6%	35.0%
HF[17]	60.2%	39.8%	68.5%	35.2%
RF[22]	50.2%	46.5%	55.0%	44.7%
Our CM-HF	67.8%	32.3%	71.4%	30.3%

Additionally, we compared our proposed M-VFOA model with the model proposed by Ba and Odobez[3] and the system by Smith and Ba.[40] Due to different application scenes, the compared results are using different datasets, models and methods. The compared results are show in Table 4. Smith and Ba[40] estimated the wandering visual focus of attention for multiple people on an outdoor advertisement poster. The accuracies on two recognized results (focused on the advertisement poster and unfocused on the advertisement poster) are, respectively, 70.56% and 75.3%. Our results seem quite far from that reported by Smith and Ba. Several factors may explain the difference. First, outdoor with multi-persons were studied and only a recognized advertisement poster was in front of them. The two recognition rates were focused on the poster and unfocused on the poster. Second, people were recorded from a camera placed directly in front of them. Hence, it was easier to recognize than our scene. Besides, our results have outperformed the recognized results by Ba and Odobez.[3] Ba and Odobez[3] recognized four person' VFOA in the natural meeting from head poses, where their recognition rates were 48.3% on the target S. people (speaking people), 38.2% on the target Table and 77.5% on the target Slide. The recognized four persons in the natural meeting room were sitting at equally spaced around table. The head poses in horizontal and vertical directions were estimated by a dynamic Bayesian network approach. In our work, we recognized eight persons in a natural classroom simultaneously for three recognized targets, i.e. the right slide (T_1), the left slide (T_2) and outside of the double slide (T_3). The persons in the natural and crowd classroom were recorded from an overhead camera in the ceiling. The more powerful CM-HF method has been used to estimate continuous head pose instead of the dynamic Bayesian network approach. The averaged recognition rate is 64.5% on the CCNU datasets.

7.5. *SS recognition*

SS recognition has been tested from 500 images from CK+ expression dataset, 500 images from GENKI dataset, and 500 images from CCNU classroom dataset. A 4-fold cross-validation was conducted. Table 5 lists the accuracies with respect to smile and non-smile expression of our CM-HF method on different datasets. One can see that the best performance is on the CK+ expression dataset, where the average

Table 4. Compared to another VFOA model and system.

Approaches	Recognition Rate (%)		
Our M-VFOA model	T1	T2	T3
	50.81	58.64	84.14
Ba and Odobez[3]	S. People	Table	Slide
	48.3	38.2	77.5
Smith and Ba[40]	Focused Adv.		Unfocused Adv.
	70.56		75.3

accuracy is 97.8%. Owing to more noise and spontaneous expression on CCNU dataset, however, where the average accuracy still achieves to 91.7% recognized by the CM-HF method. The each STD is about 2.0%, which indicates the proposed method of robustness on different datasets. The examples of SS recognition are shown in Fig. 14, where the first row is recognized results on GENKI dataset, the second row is on CK+ dataset, and the last row is on CCNU dataset. One can see that our proposed CM-HF method can recognize SS expression in some challenging environments including occlusions, expressions, poses, low resolution, make-ups, etc.

7.5.1. *Compare to the state-of-the-art methods for SS recognition*

Different methods (SVM, Boosting different,[36] IMRF,[52] ELM[2]) have been compared to our method on CK+ and GENKI datasets. We employed a 4-fold cross-validation

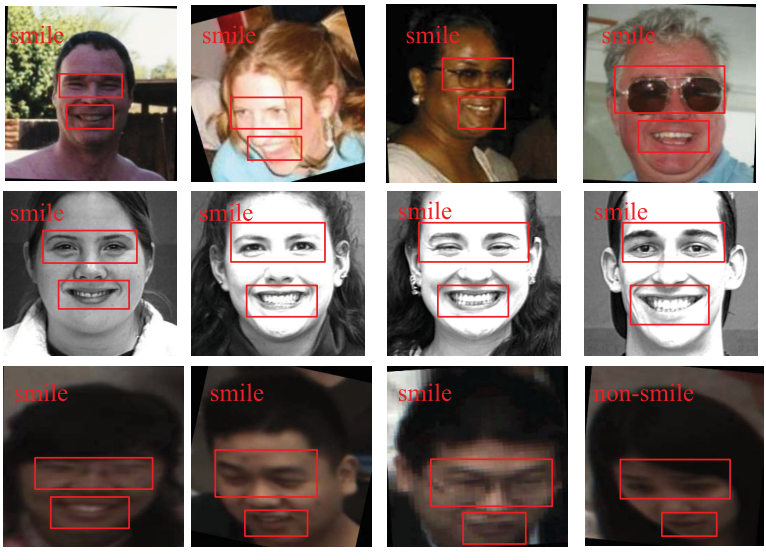


Fig. 14. The successful examples of SS recognition on GENKI, CK+, and CCNU datasets. The red rectangles are the eyes and mouth areas. The proposed CM-HF method can recognize spontaneous smile expression in some challenging environments including occlusions, illuminations, poses, low resolution, and make-ups, etc (color online) .

Table 5. Accuracy (%), average errors (%) and STD. (%) with respect to smile and non-smile expression of the CM-RF method on three challenging datasets.

Datasets	Smile	Non-Smile	Ave. Acc.	Ave. Error	STD.
CK+ dataset	97.8	96.9	97.1	3.0	1.2
GENKI dataset	95.3	93.8	94.0	5.7	2.6
CCNU dataset	92.1	91.5	91.7	7.2	2.7

Table 6. Comparison of different methods on the CK+ and GENKE datasets.

Different Methods	Smile (%)	Non-Smile (%)	Ave. Acc. (%)	STD. (%)
SVM	84.4	83.6	84.1	3.0
Boosting different[36]	88.3	89.9	89.7	2.6
IMRF[52]	85.7	85.2	85.5	2.7
ELM[2]	90.2	87.6	88.5	2.4
Our CM-HF	95.6	94.5	95.3	1.8

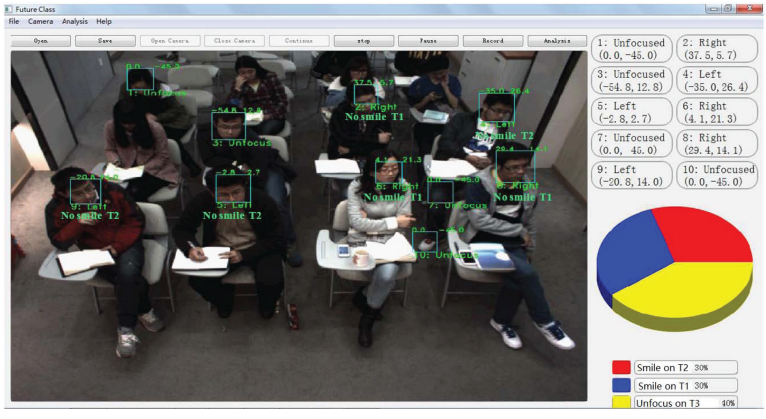
and used the same training and testing datasets. Table 6 shows the comparison of different methods on the two datasets. In ELM,[2] LDA and SVM are used as classifiers with LBP features. The average accuracy reached to 88.5%. In Boosting different,[36] pixel intensity difference (PID) is extracted as feature, while AdaBoost is used as classifier, whose average accuracy is 89.7%. IMRF[52] proposed iterative multi-output RFs, which can obtain 85.5% of the average accuracy. In this paper, the cascaded CM-HF based on mouth and eyes sub-forests achieved average 95.3% recognition rate and obtained the best performance. Additionally, the smallest STD (1.8%) indicates that our CM-HF improved the SS recognition with great robustness. Table 7 gives the average accuracy with respect to features based on whole face area, local mouth area, local eyes area and cascaded local mouth-eyes areas using our proposed method. From the table, one can see that our method with features based on cascaded local mouth-eyes areas can obtain the best performance, whose average accuracy reaches to 95.3%, and STD is 1.8%.

7.6. *System analysis*

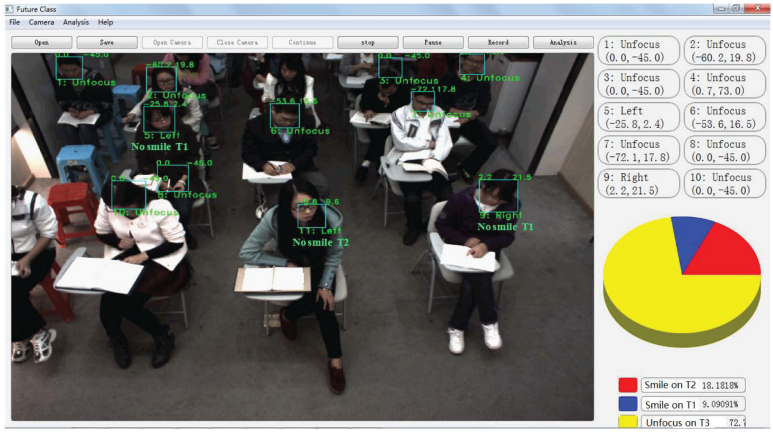
The system is developed based on continuous head pose estimation for M-VFOA and SS recognition. Figure 15 illustrates two recognized frames in different real-class

Table 7. Comparison of patches based a whole face area, mouth area, eyes area and cascaded mouth-eyes areas.

Different Feature Areas	Smile	Non-Smile	Ave. Acc.	STD
Full Face	80.4	83.5	82.9	3.3
Mouth	86.7	88.3	88.5	2.6
Eyes	83.6	68.1	75.7	3.5
Cascaded mouth-eyes areas	94.6	93.5	94.3	2.1



(a)



(b)

Fig. 15. Examples of M-VFOA with SS recognition in the system, where Left, Right, Unfocused represent the VFOA on the left slide T_2, right slide T_1, and outside of the white broad, respectively. (a) The frame in a class video, (b) The frame in another class video.

videos. The green rectangles are the tracked faces. The estimated head pose angles are above the rectangles. Meanwhile, the recognized M-VFOA target and SS expression are under the rectangles. The computation time per frame is 0.8794 s with 10 persons in the classroom. The pie chart in the lower right corner of the system interface gives the M-VFOA and SS distributions of persons in class, where the green pie indicates that persons' spontaneous expression states are "smile" on the VFOA target T_1 (right slide), the red pie indicates that person' spontaneous expression states are "smile" on the attention target T_2 (left slide), and the yellow pie indicates that the persons' VFOA target is T_3 (outside of the slides). It can help a teacher understands student's learning interest and provides appropriate teaching to make

learning efficient. In Fig. 15, the multi-persons' face detection and pose estimation in the crowd classroom scene remain some errors due to non-visible face areas and noises. In future, we will research on more efficient and real-time multi-persons' face detection and pose estimation algorithms in challenging environments.

8. Conclusion

In this paper, we present a M-VFOA and SS recognition system based on continuous head pose estimation in the natural and crowd classroom scene. The multi-persons' VFOA and SS can be automatically obtained by the system without using any special hardware or manual intervention. The system is developed for recognizing M-VFOA and spontaneous expression from continuous head pose estimation with only an over-head camera in the ceiling. A CM-HF combined with weighted Hough voting and multi-task learning is proposed to train for continuous head pose estimation, tip of the nose location and SS recognition, concurrently, which improves accuracies of recognition and reduces training time. The proposed solution estimates continuous head poses by the CM-HF method with extracted multiple features, firstly. Then, the M-VFOA is recognized based on head poses, environmental cues and prior states in the natural classroom. Meanwhile, SS expression is classified by the CM-HF method with local cascaded mouth-eyes areas normalized based on the estimated head poses.

Experimental results show that our method can automatically recognize M-VFOA and SS expression based on continuous head pose estimation and outperform the compared methods for both performance and robustness at reasonable costs in class. On several public and collected datasets and videos, our method performs with an average accuracy of 83.5% on head pose estimation in the horizontal and vertical directions, 67.8% on M-VFOA recognition and 97.1% on SS recognition. The computation time per frame is 0.8794 s with about 10 persons.

In future work, some research can be investigated to improve the performance of our system and model. First, CM-HF with deep CNN features based multi-task learning approach could lead to better performance for non-visible face location and head pose estimation. Second, more other contextual activity cues could be introduced to the VFOA model, such as, the position tracking of the teacher, the speaking person as the moving target and so on, which may bring more robust performance. Finally, more complicated spontaneous expression recognition will be built for learning behavior analysis in class.

Acknowledgments

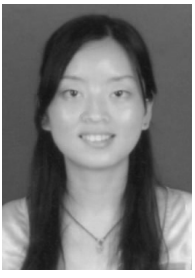
This work was supported by the National Natural Science Foundation of China (Nos. 61602429 and 61877026), Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (CCNU16A02020).

References

1. K. Alex, I. Sutskever and G. E. Hinton, ImageNet classification with deep convolutional neural networks, *Adv. Neural Inf. Process. Syst.* **25** (2012) 1097–1105.
2. L. An, S. Yang and B. Bir, Efficient smile detection by extreme learning machine, *Neurocomputing* (2015) 354–363.
3. S. Ba and J. Odobez, Multiperson visual focus of attention from head pose and meeting contextual cues, *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(1) (2011) 101–116.
4. J. Bailenson, A. C. Beall and J. Loomis, Transformed social interaction decoupling representation from behavior and form in collaborative virtual environments, *PRESENCE Teleoperators and Virtual Environments* **13**(4) (2004) 428–441.
5. J. Odobez and S. Ba, A cognition and unsupervised map adaptation approach to the recognition of focus of attention from head pose, *Proc. IEEE Conf. Multimedia and Expo*, (Beijing, China, 2007), pp. 183–191.
6. L. Breiman, Random forests, *Machine Learning* **45**(1) (2001) 5–32.
7. A. Ito, X. Wang, M. Suzuki and S. Makino, Smile and laughter recognition using speech processing and face recognition from conversation video, *Proc. IEEE Int. Conf. Cyberworlds* (Singapore, 2005), pp. 437–444.
8. J. Orozco, S. Gong and T. Xiang, Head pose classification in crowded scenes, *Proc. British Mach. Vis. Conf.* (London, UK, 2009), pp. 1–3.
9. K. Smith, S. Ba and J. Odobez, Multi-layer temporal graphical model for head pose estimation in real-world videos, *Proc. IEEE Conf. Image Process* (Paris, 2014), pp. 3392–3396.
10. A. Criminisi, J. Shotton and E. Konukoglu, Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning, Rep. TR-2011-114, Microsoft Research Cambridge **5**(6) (2011) 12.
11. X. Chu, W. Ouyang, H. Li and X. Wang, Structured feature learning for pose estimation, *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (Las Vegas, USA, 2016), pp. 4715–4723.
12. B. Glocker, O. Pauly and E. Konukoglu, Joint classification-regression forests for spatially structured multi-object segmentation, *Proc. Europe Conf. Computer Vision* (Florence, Italy, 2012), pp. 870–881.
13. A. Doshi and M. M. Trivedi, Attention estimation by simultaneous observation of viewer and view, *IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition Workshops* (2010), pp. 21–27.
14. G. Fanelli, A. Yao and P. L. Noel, Hough forest-based facial expression recognition from video sequences, *Trends and Topics in Computer Vision* (2012), pp. 195–206.
15. G. Fanelli, M. Dantone and J. Gall, Random forests for real time 3d face analysis, *Int. J. Comput. Vis.* **101**(3) (2013) 437–458.
16. S. R. Bulo and P. Kontschieder, Neural decision forests for semantic image labelling, *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (Columbus, USA, 2014), pp. 81–88.
17. M. Garca-Montero, C. Redondo-Cabrera, R. Lpez-Sastre and T. Tuytelaars, Fast head pose estimation for human computer interaction, *Pattern Recogn. Image Anal.* (2015), pp. 101–110.
18. M. Dantone, J. Gall, G. Fanelli and L. VanGool, Real time facial feature detection using conditional regression forests, *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (Providence, USA, 2012), pp. 2578–2585.
19. M. Sun and P. Kohli, Conditional regression forests for human pose estimation, *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (Providence, USA, 2012), pp. 3394–3401.

20. N. Gourier, J. Maisonnasse and D. Hall, Head pose estimation on low resolution images, *Multi-modal Technologies for Perception of Humans* (2007), pp. 270–280.
21. 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), pp. 7–49.
22. J. Gall and V. Lempitsky, Class-specific hough forests for object detection, *Proc. Decision Forests for Computer Vision and Medical Image Analysis* (Miami, USA, 2013), pp. 143–157.
23. M. Zhang, K. Li and Y. Liu, Head pose estimation from low-resolution image with hough forest, *Proc. IEEE Conf. Chinese Conf. Pattern Recognition* (Chongqing, China, 2010), pp. 1–5.
24. C. Yang, R. Duraiswami and L. Davis, Efficient mean-shift tracking via a new similarity measure, *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (San Diego, USA, 2005), pp. 176–183.
25. N. Gourier, J. Maisonnasse and D. Hall, Head pose estimation on low resolution images, *Proc. Multi-Modal Technologies for Perception of Humans* (Southhampton, UK, 2007), pp. 270–280.
26. S. Kaymak and I. Patras, Multimodal random forest based tensor regression, *IET Comput. Vis.* **8**(6) (2014) 650–657.
27. C. Huang, X. Ding and C. Fang, Head pose estimation based on random forests for multiclass classification, *Proc. IEEE Conf. Pattern Recognition* (Istanbul, Turkey, 2010), pp. 934–937.
28. A. Doshi and M. M. Trivedi, Attention estimation by simultaneous observation of viewer and view, *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (San Francisco, USA, 2010), pp. 21–27.
29. M. Voit and R. Stiefelhagen, Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios, *Proc. 10th Int. Conf. Multi-Modal Interfaces* (Chania, Crete, Greece, 2008), pp. 173–180.
30. Y. Liu, L. Liu, J. Chen, Z. Su and K. Zhang, Multi-person visual focus of attention from head pose on a natural classroom, *IEEE Conf. Pattern Recogn. Appl. Methods* (2016) 205–213.
31. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion specified expression, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (2010), pp. 94–101.
32. A. Dapogny and K. Bailly, Pairwise conditional random forests for facial expression recognition, *Proc. IEEE Int. Conf. Computer Vision* (Santiago, Chile, 2015), pp. 3783–3791.
33. J. Odobez and S. Ba, A cognition and unsupervised map adaptation approach to the recognition of focus of attention from head pose, *IEEE Int. Conf. Multimed. Expo* (2007) 183–191.
34. J. Orozco, S. Gong and T. Xiang, Head pose classification in crowded scenes, *British Machine Vision Conference* (2009), pp. 1–3.
35. G. Fanelli, A. Yao and P. L. Noel, Hough forest-based facial expression recognition from video sequences, *Proc. Trends and Topics in Computer Vision* (Heraklion, Crete, Greece, 2012), pp. 195–206.
36. C. Shan, Smile detection by boosting pixel differences, *IEEE Trans. Image Process.* (2012) 431–436.
37. N. Gourier, D. Hall and J. Crowley, Estimating face orientation from robust detection of salient facial features in pointing, *Proc. Int. Conf. Pattern Recognition Workshop on Visual Observation of Deictic Gestures* (Cambridge, UK, 2004), pp. 1379–1382.

38. G. Huang, M. Ramesh, T. Berg and E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments, *Proc. Technical Report* (University of Massachusetts, Amherst, USA, 2007), pp. 7–49.
39. www.idiap.ch/headposedatabase.
40. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, The extended cohn-kanade dataset(ck+): A complete dataset for action unit and emotion specified expression, *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops* (San Francisco, USA, 2010), pp. 94–101.
41. S. E. Kahou, P. Froumenty and C. Pal, Facial expression analysis based on high dimensional binary features, *Proc. European Conf. Computer Vision* (Zurich, Switzerland, 2014), pp. 135–147.
42. Y. Liu, L. Liu, J. Chen, Z. Su and K. Zhang, Multi-person visual focus of attention from head pose on a natural classroom, *Proc. IEEE Conf. Pattern Recognition Appl. Methods* (Rome, Italy, 2016), pp. 205–213.
43. K. Sundararajan and D. Woodard, Head pose estimation in the wild using approximate view manifolds, *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops* (Boston, USA, 2015), pp. 50–58.
44. M. Garca-Montero, C. Redondo-Cabrera, R. Lpez-Sastre and T. Tuytelaars, Fast head pose estimation for human computer interaction, *Proc. Pattern Recog. Image Analysis* (Spain, 2015), pp. 101–110.
45. X. Zhu and D. Ramanan, Face detection, pose estimation and landmark localization in the wild, *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (Providence, USA, 2012), pp. 2879–2886.
46. J. Wu and M. Trivedi, A two-stage head pose estimation framework and evaluation, *Pattern Recogn.* **41** (2008) 1138–1158.
47. X. Zhao, T. Kim and W. Luo, United face analysis by iterative multi-output random forests, *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (Columbus, USA, 2014), pp. 1765–1772.
48. X. Yuan and A. Mohamed, A multi-class boosting method for learning from imbalanced data, *Int. J. Granular Comput., Rough Sets Intell. Syst.* **4** (2015) 13–29.
49. J. Zeng, W. S. Chu, F. De la Torre, J. F. Cohn and Z. Xiong, Confidence preserving machine for facial action unit detection, in *Proc. IEEE Int. Conf. Computer Vision* (2015), pp. 3622–3630.
50. M. Zhang, K. Li and Y. Liu, Head pose estimation from low-resolution image with hough forest, *IEEE Conf. Chin. Conf. Pattern Recogn.* (2010) 1–5.
51. T. Zhang, W. Zheng, Z. Cui, Y. Zhong, J. Yan and K. Yan, A deep neural network driven feature learning method for multi-view facial expression recognition, *IEEE Trans. Multimed.* (2016).
52. X. Zhao, T. Kim and W. Luo, Unified face analysis by iterative multi-output random forests, *IEEE Conf. Comput. Vis. Pattern Recogn.* (2014) 1–8.
53. X. Zhu and D. Ramanan, Face detection,pose estimation and landmark localization in the wild, *IEEE Conf. Comput. Vis. Pattern Recogn.* (2012) 2879–2886.



Yuanyuan Liu received B.E. degree from Nan-Chang University, Nan-Chang, China, in 2005, M.E. degree from Huazhong University of Science and Technology, Wuhan, China, in 2007, and Ph.D. degree from Central China Normal University. She is currently a lecturer in China University of Geosciences. Her research interests include image processing, computer vision and pattern recognition.

Xingmei Li received B.E. and M.E. degree from China University of Geosciences, Wuhan, China, in 2000 and 2003, and Ph.D. degree from Huazhong University of Science and Technology, Wuhan, China, in 2011. She is currently an associate professor in China University of Geosciences (Wuhan). Her research interests include image processing and artificial intelligence.

Fang Fang received a B.S. degree in Computer Science and Technology in 1998 and a Ph.D. degree in Management Science and Engineering in 2012 from the China University of Geosciences, Wuhan, China. She is currently an Associate Professor at the Department of Information Engineering at the China University of Geosciences. Her research interests include spatial data mining, machine learning, and GIS application.

Fayong Zhang received B.E. degree from China University of Geosciences, Wuhan, China, in 1996, M.E. degree from China University of Geosciences, Wuhan, China, in 2001, and Ph.D. degree from China University of Geosciences, Wuhan, China, in 2009. He is currently an associate professor in faculty of information engineering, China University of Geosciences. His research interests include GIS, Machine learning, underground pipeline and geological information.

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 Post-doctor 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, educational technology and human-machine interface.

Zhizhong Zeng received B.E. degree from Wuhan University of Science & Technology, Wuhan, Chin, Mater and Ph.D. degree from Huazhong University of Science and Technology, Wuhan, China. He is currently a lecturer in Central China Normal University. His research interests include machine learning, data mining, and computing intelligence.