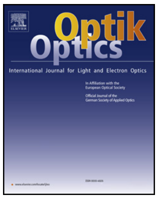
Optik - International Journal for Light and Electron Optics 182 (2019) 647–657

Contents lists available at ScienceDirect

Optik

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



Original research article

Real-time pose invariant spontaneous smile detection using conditional random regression forests

Leyuan Liu[a,b], Wenting Gui[a], Li Zhang[a], Jingying Chen[a,b,*]

[a] National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China
[b] National Engineering Laboratory for Technology of Big Data Applications in Education, Central China Normal University, Wuhan 430079, China

ARTICLE INFO

Keywords:
Smile detection
Head pose
Conditional random regression forests
Real-time

ABSTRACT

Detecting spontaneous smile in unconstrained environment is a challenging problem mainly due to the large intra-class variations caused by head poses. This paper presents a real-time smile detection method based on conditional random regression forests. Since the relation between image patches and smile intensity is modelled conditional to head pose, the proposed smile detection method is not sensitive to head poses. To achieve high smile detection performance, techniques including regression forest, multiple-label dataset augmentation and non-informative patch removement are employed. Experimental results show that the proposed method achieves competitive performance to state-of-the-art deep neural network based methods on two challenging real-world datasets, although using hand-crafted features. A dynamical forest ensemble scheme is also presented to make a trade-off between smile detection performance and processing speed. In contrast to deep neural networks, the proposed method can run in real-time on general hardware without GPU.

1. Introduction

Smile is the most common facial expression of human beings, and it often indicates the emotion and intention of a person. It is therefore unsurprising that smile detection has many applications such as smile shutter for digital camera [1], affect-sensitive e-learning [2], student's interest analysis [3], viewer experience understanding [4], etc. Although many smile detection methods [5–8] have reported promising results on face images with nearly frontal head pose, having a frontal head pose is not an actual assumption for most real-world applications. Moreover, many existing smile detection methods only deal with acted facial expressions captured under highly controlled environment, while people in the real-world applications demonstrate their facial expressions in a spontaneous way. Recently, deep neural networks have achieved impressive progress in spontaneous facial expression recognition under unconstrained environment [9–12]. However, most of these deep neural networks cannot run in real-time on general hardware without GPU. As a consequence, there is an ever-growing need for real-time pose invariant spontaneous smile detection methods [13].

Detecting smile faces with various poses from images is a challenging problem, because faces with various poses are inherently multimodally distributed in feature spaces [14–16]. Such multimodality usually leads to a large intra-class variation, which imposes a great challenge to build an effective model for smile detection. Three categories of methods, i.e., the pose-transformation method [17,18], the pose-specific method [11,12] and the pose-invariant method [19,9,10] have been proposed to handle the multi-modal problem. In the pose-transformation method, the features extracted from the input face images are first transformed to the

[*] Corresponding author at: National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China.
E-mail addresses: lyliu@mail.ccnu.edu.cn (L. Liu), chenjy@mail.ccnu.edu.cn (J. Chen).

https://doi.org/10.1016/j.ijleo.2019.01.020
Received 12 November 2018; Received in revised form 26 December 2018; Accepted 13 January 2019

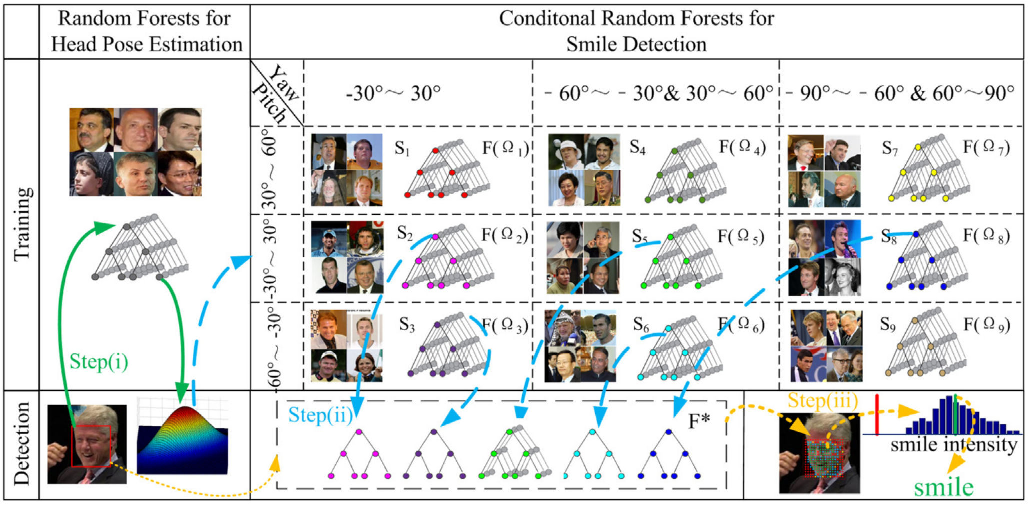


Fig. 1. The framework of the proposed Conditional Random Regression Forests based smile detection method.

corresponding features of the faces seen in a frontal pose, and then the model trained by frontal faces can be used to detect smile faces in images with various head poses. However, transferring features extracted from an image with non-frontal head pose to that would have been extracted if the head pose was frontal in principle is an ill-posed problem [14]. In the pose-specific method, the training dataset is split into several subsets according to head poses, and a pose-specific classifier is trained on each subset. However, the head pose of a face image is usually ambiguous (that is, does not belong to any of the subset exactly). To the authors' best knowledge, there is no work that has explored how to combine multiple pose-specific classifiers for improving smile detection accuracy. The pose-invariant method tries to learn generic image representations that are insensitive to variations across modalities. To learn such pose-invariant image representations, deep neural networks are usually employed. As well-known, extremely large training datasets are required by deep neural networks. Unfortunately, large-scale labeled training datasets are not yet available for smile detection. Detecting smile faces with various poses using a relatively small dataset remains a significant research challenge.

In this paper, a real-time pose invariant smile detection method is proposed. To achieve high smile detection accuracy in real-world applications while keeping real-time running speed on general hardware, we employ Conditional Random Regression Forests (CRRFs) [20] rather than deep neural networks. The framework of the proposed method is illustrated in Fig. 1. In the training phase, the training dataset is split into 9 sub-datasets according to yaw and pitch angles of head pose, and 9 CRRFs-based pose-specific classifiers are trained independently on these sub-datasets. In the detection phase: (i) The probability of the head pose is estimated from the test face image by a random forest. (ii) A certain number of regression trees are randomly selected from the corresponding CRRF according to the probability of head pose; Meanwhile, a dynamic random regression forest is constructed by these selected trees. (iii) Image patches densely sampled from the test image are fed to the dynamic random regression forest to cast votes for estimating smile intensity, and then the informative votes are selected and averaged to make the final smile/non-smile prediction.

There are mainly four contributions in this paper. First, the relation between image patches and smile intensity is modelled conditional to head pose by a group of random regression forests, hence the intra-class variation can be significantly reduced by head pose division and forest regression. Moreover, multiple pose-specific classifiers (i.e., regression trees) are dynamically selected and combined for improving smile detection accuracy. Second, a multi-label strategy for augmenting training dataset is proposed, so that high accuracy pose-specific classifiers can be trained on a small dataset. Third, non-informative image patches are picked out and discarded in the detection phase, thus the noise to the prediction can be removed. Fourth, to guarantee real-time processing speed, a ensemble scheme which selects only a fraction of trees from the trained CRRFs for constructing the dynamic random regression forest is presented. The proposed method has been extensively evaluated on two challenging real-world datasets, and experimental results have proved that the proposed method can detect smile face with various poses effectively and efficiently.

The rest of the paper is arranged as follows: the related work is introduced in Section 2, the details of the proposed smile detection method are described in Section 3, the experiments are presented in Section 4, and the conclusions are given in Section 5.

2. Related work

2.1. Smile detection

In the last decade, lots of researchers have employed various machine learning tools for smile detection. Shan [5] proposed an AdaBoost based smile detection method, in which the gray intensity differences between pixels in the face images are used as features. Liu et al. [6] combined AdaBoost and SVM classifiers for smile detection, and achieved improved performance on the GENKI4K database. An et al. [7] employed Extreme Learning Machine (ELM) for smile detection, and achieved an accuracy of 88.2% on the GENKI4K database. Luo et al. [3] designed a smile detection method based on random forests, and the experimental results have shown that the method works well on “in-the-wild” face images.

Most of existing methods considered smile detection as a binary decision problem, while several works [21] treated smile detection as a regression problem. Girard et al. [21] employed intensity-trained multi-class and regression models for smile detection, and experimental results have shown that multi-class and regression models outperformed binary-trained classifiers on smile detection.

To address the problem of lack of a large-scale training dataset, many researchers adopted handcrafted features for smile detection. Zhang et al. [11] used the 2D SIFT features extracted from facial landmark points as the input data for the deep neural network. Levi et al. [22] converted the LBP codes of original facial images into a 3D metric space and then used the 3D codes as inputs for convolutional neural networks. Lopes et al. [23] applied some preprocessing techniques in order to extract only expression specific features from a face image and explored the presentation order of the samples while training a convolutional neural network. Recently, Chen et al. [24] constructed a deep convolutional network called Smile-CNN to perform feature learning and smile detection simultaneously. Although experimental results demonstrate that the Smile-CNN model can effectively deal with “small data”, this work was only evaluated on nearly frontal face images.

Many researchers have observed that only a few facial regions are informative for smile detection. Du et al. [25] reported that smile involves only three action units, i.e., AU12 (Lip Corner Puller), AU25 (Lips Part) and AU6 (Cheek Raiser). Cui et al. [26,27] believed that the mouth shape can effectively reflect a persons smile state, and extracted a snappy set of features from a few of facial landmarks around the mouth for smile detection. Luo et al. [3] combined two classifiers which respectively extract features from the eyes region and mouth region for improving smile detection accuracy.

2.2. Conditional random forest

Conditional random forest [28] is derived from random forest [29]. While a random forest is trained on the entire training set, a conditional random forest consists of a group of random forests that are trained on sub-datasets split according to a latent condition (e.g., head pose in this work). It has proved that conditional random forest is a powerful and versatile tool for solving cross-modal problems in high-level tasks of computer vision. Sun et al. [28] employed a conditional random forest for real-time human pose estimation from depth images. Experimental results have shown that the incorporation of latent conditions such as human height and torso orientation can improve the performance of human pose estimation. Dantone et al. [20] proposed a conditional regression forest for locating facial feature points from “in-the-wild” 2D images. They used head pose as a latent condition and demonstrated that conditional regression forests outperform regression forests for facial feature points detection. Tang et al. [30] designed a structured latent regression forests for estimating 3D articulated hand posture. Experiments have shown that the latent regression forests out-performs state-of-the-art methods in both accuracy and efficiency. Recently, Liu et al. [12] proposed a conditional convolutional neural network enhanced random forest (CoNERF) for facial expression recognition, which achieved an average accuracy of 94.09% on the multi-view BU-3DEF dataset. Conditional random forest is not only powerful to tackle with cross-modal problems in computer vision, but also inherits the advantages from random forest including the capability to handle over-fitting problem, high generalization power and efficiency.

3. The proposed method

From the view of machine learning, the essential task of smile detection is learning the probability $p(\theta|P)$ that maps a given face image (P) to its corresponding smile state (θ). Many smile detection methods [7,8] learn this probability directly from the entire training set. However, face images captured in many real-world applications are usually with various head poses, as people tend to move their heads freely. The head pose variety causes large intra-class variations which make it difficult to learn the probability $p(\theta|P)$ directly. A convenient idea is to build a serial of pose-specific smile models, so that the intra-class variations faced by each pose-specific smile model will be significantly reduced. To this end, conditional random forest [20] is employed in this work to learn the pose-specific smile model $p(\theta|\omega, P)$, and then to estimate the probability $p(\theta|P)$ by

$$p(\theta|P) = \int p(\theta|\omega, P)p(\omega|P)d\omega \tag{1}$$

where ω corresponds to the head pose that can be estimated by a head pose estimation algorithm [31]. For the convenience of implementation, the head pose space is discretized into N disjoint subspaces $\{\Omega_n, n = 1, \ldots, N\}$, then Eq. (1) becomes

$$p(\theta|P) \approx \sum_{n=1}^{N} \left(p(\theta|\Omega_n, P) \int_{\omega\in\Omega_n} p(\omega|P)d\omega \right) \tag{2}$$

In order to learn the pose-specific smile model $p(\theta|\Omega_n, P)$, the training dataset is also divided into N sub-datasets $\{S_n, n = 1, \ldots, N\}$ according to head poses, and then N groups of conditional random forests $\{\mathcal{F}(\Omega_n), n = 1, \ldots, N\}$ can be trained on these subsets respectively.

3.1. Training conditional random forests for smile detection

Each tree $F_t(\Omega_n)$ in the conditional random forest $\mathcal{F}(\Omega_n)$ is constructed from a set of patches $\{P_j, j = 1, \cdots, M\}$, which are randomly sampled from the training images in the sub-dataset S_n. A patch is denoted as $P_j = (\{I_j^k\}_{k=1}^K, \theta_j)$, where $\{I_j^k\}_{k=1}^K$ are a set of visual features extracted from the sampled image, and θ_j is a label that indicates the smile state of the sampled face image. Most existing smile detection methods use a binary label to denote smile state, i.e., $\theta = 1$ for smile and $\theta = 0$ for non-smile. However, the appearances of smile faces with different intensities are various. Hence, a continuous label that indicates smile intensity is used to denote smile state in this work. As a consequence, regression trees rather than decision trees are trained.

The trees in each conditional random forest $\mathcal{F}(\Omega_n)$ are built independently in the following procedure:

(1) Build a pool of binary decision candidates $\{h_i(P, \tau_i)\}$, where $h(P, \tau)$ is a weak classifier and τ is a threshold. In this work, the patch comparison function [32] is adopted as the weak classifier. The patch size and position in the patch comparison function as well as the threshold τ are generated randomly.

(2) Recursively select the best binary decision h^* which maximizes the information gain (ψ), and use it to divide the set of patches $\mathcal{P}_p$ at the current node into two subsets $\mathcal{P}_l$ and $\mathcal{P}_r$. The information gain is defined as:

$$\psi = \mathcal{H}(\mathcal{P}_p) - \frac{|\mathcal{P}_l|}{|\mathcal{P}_p|}\mathcal{H}(\mathcal{P}_l) - \frac{|\mathcal{P}_r|}{|\mathcal{P}_p|}\mathcal{H}(\mathcal{P}_r) \tag{3}$$

where $\mathcal{H}(\mathcal{P})$ corresponds to the uncertainty measure of a set of patches $\mathcal{P}$ .The uncertainty measure plays a key role in the training procedure, and it has a profound impact on the discriminative power of the tree. In contrast to the covariance-based uncertainty measure used in [32] and the classification objective uncertainty measure used in [20], we propose a clustering-based uncertainty measure:

$$\mathcal{H}(\mathcal{P}) = -\sum_{c=1}^{C} \frac{\sum_{P_j \in \mathcal{P}} p(\theta_c|P_j)}{|\mathcal{P}|} \log(\frac{\sum_{P_j \in \mathcal{P}} p(\theta_c|P_j)}{|\mathcal{P}|}) \tag{4}$$

where C is the number of clusters, θ_c is the center of the cth cluster. The K-Means + + algorithm [33] is employed for clustering. $p(\theta_c|P_j)$ indicates the probability that the patch P_j is cropped from a face whose smile intensity is θ_c, and it is defined as:

$$p(\theta_c|P_j) \propto \exp\left(-\frac{|\theta_c - \theta_j|}{\lambda}\right) \tag{5}$$

where λ is a control factor. This clustering-based certainty measure is somehow similar to a classification objective uncertainty measure but avoids a hard label assignment of the patches, it therefore can result in a good regression of smile intensities. In our experiments, it performed better than both the covariance-based and the classification objective uncertainty measure.

(3) Create leaf nodes and construct the pose-specific smile model, when information gain is smaller than a threshold. Patches that reach the leaf node $l_{t,n}$ in the tree $F_t(\Omega_n)$ are denoted as $\mathcal{P}_{t,n}$. The pose-specific smile model on the leaf node $l_{t,n}$, i.e. the probability $p(\theta|\Omega_n, l_{t,n})$, is modelled by a Gaussian over $\mathcal{P}_{t,n}$ and stored at the leaf node $l_{t,n}$:

$$p(\theta|\Omega_n, l_{t,n}) = \mathcal{N}(\theta; \bar{\theta}, \sigma^2) \tag{6}$$

3.2. Smile detection using conditional random forests

Given a test face image, a set of patches are densely sampled from it. Each patch P_j is then fed to each trained tree $F_t(\Omega_n)$ in each conditional random forest $\mathcal{F}(\Omega_n)$, and finally ends in a set of leaf nodes $\{l_{t,n}(P_j)\}$. The probabilities $p(\theta|\Omega_n, l_{t,n}(P_j))$ stored at the leaf nodes $l_{t,n}(P_j)$ that the patch P_j reached are assembled to cast a vote for predicting the smile state of the given face image:

$$p(\theta|\Omega_n, P_j) = \frac{1}{T_n}\sum_{t=1}^{T_n} p(\theta|\Omega_n, l_{t,n}(P_j)) \tag{7}$$

where T_n is the number of trees in the conditional random forest $\mathcal{F}(\Omega_n)$. Considering the head pose, we substitute Eq. (7) into Equation(2):

$$p(\theta|P_j) = \sum_n \left(\frac{1}{T_n}\sum_{t=1}^{T_n} p(\theta|\Omega_n, l_{t,n}(P_j))\right)\left(\int_{\omega\in\Omega_n} p(\omega|P)d\omega\right) \tag{8}$$

Then, the smile intensity predicted by a single image patch P_j is

$$\hat{\theta}_{P_j} = \arg\max_{\theta} p(\theta|P_j) \tag{9}$$

$\hat{\theta}_{P_j}$ is called as a vote from the patch P_j.

Not all the densely sampled patches contribute to smile prediction. According to our observation, three kinds of patches are not informative and may even add noise to the prediction: (1) Ambiguous patches that end in leaf nodes with a high uncertainty. (2) Singular patches that cast quite different smile intensities from those casted by most of the other patches. (3) Isolated patches that have few neighboring “unambiguous” and “nonsingular” patches. In order to reduce the influence of votes from these kinds of patches, we first discard the ambiguous patches that end in leaf nodes with a uncertainty greater than an empiric threshold $\tau_{\mathcal{H}}$, filter

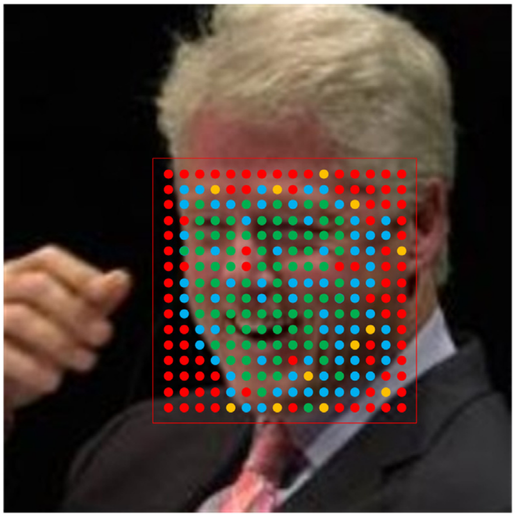


Fig. 2. An example test image: the red, blue and yellow dots represent the centers of the discarded ambiguous, singular and isolated patches respectively, and the green dots are the centers of the patches reserved to cast votes for smile prediction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

out the singular patches by the MeanShift algorithm [34], and then remove the isolated patches using a scanning window. An example test image is shown in Fig. 2, where the red, blue and yellow dots represent the centers of the discarded ambiguous, singular and isolated patches respectively, and the green dots are the centers of the patches reserved to cast votes for smile prediction. Finally, the smile intensity of a test image is predicted by averaging all the votes casted by the reserved patches $\mathcal{P}_r$:

$$\hat{\theta} = \frac{1}{|\mathcal{P}_r|} \sum_{P_j \in \mathcal{P}_r} \hat{\theta}_{P_j} \tag{10}$$

Algorithm 1 summarizes the procedures of smile prediction using conditional random forests.

Algorithm 1. Smile detection using conditional random forests (V1).

```
Input:
    The trained random forest for head pose estimation ($\mathcal{F}_h$);
    The set of trained conditional random forests for pose-specific smile prediction ($\{\mathcal{F}(\Omega_n), n = 1, ..., N\}$);
    The test face image ($I$).
Output:
    The predicted smile intensity of the test image ($\hat{\theta}$).
____________________________________________
Densely sample a set of patches $\mathcal{P} = \{P_j, j = 1, ..., M\}$ from $I$;
Estimate head pose distribution $p(\omega|P)$ by $\mathcal{F}_h$;
for $j = 1, \ldots, M$ do
    for $n = 1, \ldots, N$ do
        Estimate $p(\theta|\Omega_n, P_j)$ by $\mathcal{F}(\Omega_n)$ using Eq. (7);
    end for
    Compute $p(\theta|P_j)$ using Eq. (8);
    Compute the vote $\hat{\theta}_{P_j}$ using Eq. (9);
end for
Remove non-informative patches, denote the reserved patches as $\mathcal{P}_r$;
Assemble all the votes from $\mathcal{P}_r$ and compute $\hat{\theta}$ using Eq. (10);
return $\hat{\theta}$.
```

3.3. An ensemble scheme for speed-up

Algorithm 1 described in Section 3.2 is computation consuming, as each image patch needs to be fed to all the $T_1 + T_2 + \cdots + T_N$ trees in all the N groups of conditional random forests. To achieve real-time processing speed, a ensemble scheme is proposed. Inspired by [20], we totally select $T(T \ll T_1 + T_2 + \cdots + T_N)$ trees from the N groups of conditional random forests. The number of trees selected from the conditional random forest $\mathcal{F}(\Omega_n)$ is assigned according to the estimated head pose distribution of the test image:

$$\tilde{T}_n = \left\lfloor T \int_{\omega \in \Omega_n} p(\omega|P) d\omega + 0.5 \right\rfloor \tag{11}$$

As $\int_{\omega \in \Omega_n} p(\omega|P) d\omega \approx \tilde{T}_n / T$ in this ensemble scheme, Eq. (8) can be transformed to

$$p(\theta|P_j) \approx \frac{1}{T} \sum_{n=1}^{N} \sum_{t=1}^{\tilde{T}_n} p(\theta|\Omega_n, l_{t,n}(P_j)) \tag{12}$$

This means that $p(\theta|P_j)$ can be estimated only using the T selected trees. Using this ensemble scheme, the computational load will theoretically fall to about $T/(T_1 + T_2 + \cdots + T_N)$ of that consumed by Algorithm 1. However, the accuracy of smile prediction may decrease when using this ensemble scheme, since less trees are used.

To minimize the accuracy decrease caused by the ensemble scheme, we empirically evaluate the average regression error ($\varepsilon_{t,n}$) of each tree $F_t(\Omega_n)$ in each group of conditional random forest $\mathcal{F}(\Omega_n)$, and then set a selected probability ($\rho_{t,n}$) to each tree according to its average regression error:

$$\rho_{t,n} \propto \frac{1}{\varepsilon_{t,n}} \tag{13}$$

$$\varepsilon_{t,n} = \frac{1}{M'} \sum_{i=1}^{M'} |\hat{\theta}_{t,n,i} - \theta_i| \tag{14}$$

where M is the number of face images used for empirically evaluation, $\hat{\theta}_{t,n,i}$ is the smile intensity predicted by the tree $F_t(\Omega_n)$, and $\boldsymbol{\theta}_i$ is the labeled smile intensity. Compared to the ensemble scheme which selects trees from conditional random forests randomly (or with a uniform probability) [20], selecting trees with the above regression-error-related selected probability results in a smaller accuracy decrease and performs more stably.

Algorithm 2 summarizes the procedures of smile prediction using conditional random forests with the proposed ensemble scheme.

Algorithm 2. Smile detection using conditional random forests (V2)

```
Input:
    The trained random forest for head pose estimation ($\mathcal{F}_h$);
    The set of trained conditional random forests for pose-specific smile prediction ($\{\mathcal{F}(\Omega_n), n = 1, \ldots, N\}$);
    The test face image ($I$).
Output:
    The predicted smile intensity of the test image ($\hat{\theta}$).
Densely sample a set of patches $\mathcal{P} = \{P_j, j = 1, \ldots, M\}$ from $I$;
Estimate head pose distribution $p(\omega|P)$ by $\mathcal{F}_h$;
Compute $\tilde{T}_n$ using Eq. (11);
Select $\tilde{T}_n$ trees with a probability $\rho_{t,n}$ from $\mathcal{F}(\Omega_n)$, and construct a dynamic random forest $\mathcal{F}^*$ by these selected trees.
for $j = 1, \ldots, M$ do
    Compute $p(\theta|P_j)$ by $\mathcal{F}^*$ using Eq. (12);
    Compute the vote $\hat{\theta}_{P_j}$ using Eq. (9);
end for
Remove non-informative patches, denote the reserved patches as $\mathcal{P}_r$;
Assemble all the votes from $\mathcal{P}_r$ and compute $\hat{\theta}$ using Eq. (10);
return $\hat{\theta}$.
```

3.4. Training data augmentation

As large-scale labeled dataset is currently not available for smile detection, the number of training images in each sub-dataset is usually small. With a relatively small set of training images, the over-fitting problem may occur [35] and may result in a low smile detection accuracy. Moreover, the "ground truth" labels that indicates the smile intensities of the training images are usually neither objective nor accurate, since the labels are obtained in a rather subjective way. Unfortunately, most regression methods including the conditional random forest used in this work are sensitive to the accuracy of ground truth labels. To address these problems, the multi-label strategy [36] is employed to augment the training data. For each face in the training images, a set of additional"soft" labels $\{\theta_i, i = 1, \ldots, m\}$ are sampled following a Gaussian distribution centered at the "ground truth" label θ_0:

$$p(\theta_i|\theta_0) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\theta - \theta_0)^2}{2\sigma^2}\right), \quad i = 1, \ldots, m \tag{15}$$

Scaling (0.90–1.10) and rotation ($-5°$ to $+5°$) transformations are randomly performed on each of the original image in the training sub-datasets m times, and then a "soft" label is assigned to each transformed image. Accordingly, the amount of images in each training sub-dataset can be extended m-folds. This training data augmentation technique can not only increase the number of images in the training set, but also can alleviate the problem caused by inaccurate smile intensity labels.

4. Experiments

4.1. Datasets

In order to evaluate the performance of the proposed smile detection method on realistic data, the Labeled Faces in the Wild (LFW) dataset [37] and CCNU-Class datasets were used in our experiments. The LFW is a publicly available dataset which consists of 13,232 face images with various head poses and spontaneous facial expressions. The CCNU-Class dataset is collected by ourselves, and it contains 500 images captured in real classes over a span of time. As each image contains 8–12 students, there are totally 5557 faces in the CCNU-Class dataset. We cropped all face regions from the images in these two datasets, and assigned each face with a label $\theta_0 (\theta_0 \in [0, 3])$ by averaging the smile intensities labeled by five experts.

To assess cross-dataset performance of the proposed method, the conditional random regression forests were trained only on the LFW dataset, while tested on both the LFW and CCNU-Class datasets. When building the training dataset, 10,000 face images were selected from the LFW dataset randomly, and each selected face image was augmented 3 times by the data augmentation scheme presented in Section 3.4. This means that the training dataset totally contains 40,000 samples. The rest of 3232 face images in LFW and all the 5557 face images in CCNU-Class were used as two testing datasets.

4.2. Experimental settings

In our experiments, head poses were discretized into 5 yaw angles (i.e. profile-left ($-90°$ to $-60°$), left ($-60°$ to $-30°$), front ($-30°$ to 30°), right (30–60°)), profile-right (60–90°)) and 3 pitch angles (i.e. down ($-60°$ to $-30°$), equal ($-30°$ to 30°), up (30–60°)). The Dirichlet-tree distribution enhanced random forest (D-RF) algorithm [31] was employed for head pose estimation. The accuracies of head pose estimation on the LFW and the CCNU-Class datasets are respectively 82.72% and 83.23%.

Since a head pose is a combination of a yaw angle and a pitch angle, there are 15 (5 × 3) discrete head poses in our case. This means that the training dataset should be divided into as more as 15 sub-datasets for training the pose-specific smile models (i.e., the conditional random regression forests). Fortunately, the profile-left/left faces and profile-right/right faces of the same subject are nearly horizontal symmetry. To keep more training samples in each sub-dataset, we first converted the profile-left/left faces in the training dataset to profile-right/right ones by a horizontal mirror transformation, and then divided the training dataset into only 9 (3 × 3) sub-datasets. Correspondingly, 9 conditional random regression forests were trained independently on these 9 training sub-datasets.

For training the conditional random regression forests, all the cropped face images in the training sub-datasets were resized to 125 × 125 pixels. Each tree was constructed based on 2000 face images randomly selected from the corresponding sub-dataset. After extracting visual features including normalized gray intensity, Sobel gradients, LBP [38],and HOG [39], 256 patches whose size is 20 × 20 were densely sampled from each selected face image. The parameters with respect to random forests were set empirically, e.g., the number of trees in each conditional random regression forest was set to 20, the maximum depth of each tree was set to 15, the size of binary decision candidate pool was set to 3000, the number of clusters in the proposed uncertainty measure was set to 3.

During detecting, faces in testing images were first located using the fast face detection algorithm presented in [40], then the smile intensity of each located face was predicted following the procedures described in Algorithm 2. In order to compare the proposed method with other methods that use binary smile labels, we also output a binary smile prediction for each testing face image. That is, if the predicted smile intensity $\hat{\theta}$ is larger than a threshold τ_θ, the binary output is 1 (smile), otherwise the binary output is 0 (non-smile). In our experiments, the threshold τ_θ was fixed at 0.75. Detection accuracy were used in our experiments as the quantitative evaluation criterion.

The number of trees (T) selected for constructing the dynamic random forest $\mathcal{F}^*$ makes a trade-off between detection accuracy and running speed. Fig. 3 illustrates the average detection accuracy and the running speed on the LFW and CCNU-Class testing datasets when conducted on a computer with a Core i7 4.2GHz CPU. It is obvious that a higher number of trees improves the detection accuracy but decreases the running speed. When the number of trees is larger than 20, the running speed is lower than 15 frames per second (FPS) but the detection accuracy is not improved much. Hence, to keep a real-time running speed, the number of

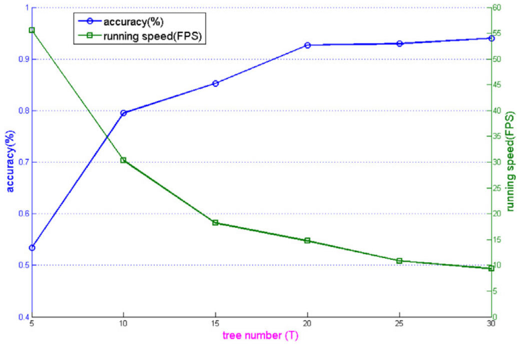


Fig. 3. Trade-off between detection accuracy and running speed.

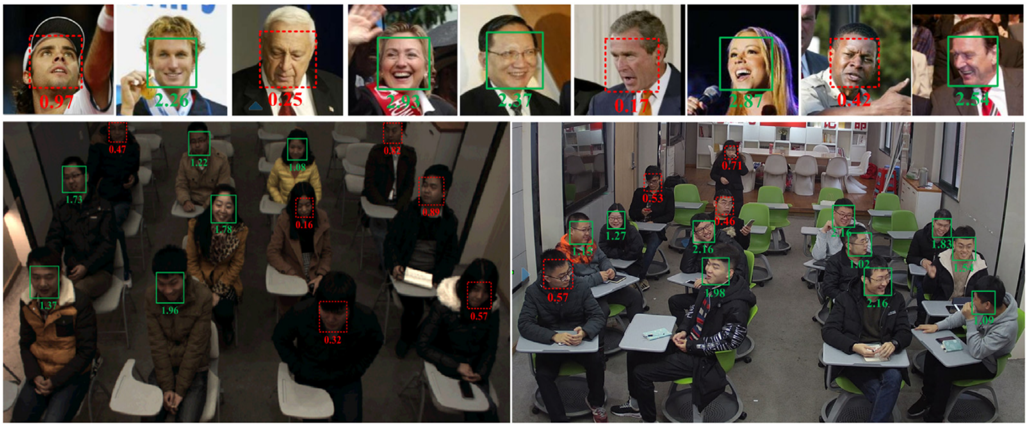


Fig. 4. Examples of smile detection results on the LFW (the first row) and CCNU-Class (the second row) datasets. The solid green boxes indicate smile faces, and dashed red boxes indicate non-smile faces. The number below each box indicates the estimated smile intensity.

trees (T) selected for constructing the dynamic random forest $\mathcal{F}^*$ was fixed to 20 in our experiments.

4.3. Experimental results

4.3.1. Qualitative results

Fig. 4 shows some examples of smile detection results on the LFW and CCNU-Class datasets. It demonstrates that the proposed smile detection method performs well on both the two datasets and does not suffer from various head poses.

4.3.2. Analysis of features

To investigate the influence of image features on the smile detection accuracy, experiments with different deep features and classical features were conducted. We fine-tuned the VGG-face model [41] for smile detection using the augmented LFW training set, and then extracted deep features from the FC1, FC2 and FC3 layers of the fine-tuned VGG-face networks. Deep features as well as classical features including gray intensity, Sobel gradients, HOG [39], and LBP [38] were used to train the conditional random regression forests. Table 1 lists the detection accuracy achieved by the proposed method with seven single features and two combination features on the LFW and CCNU-Class testing datasets. It is clear that (1) the combination of features perform better than single features; and (2) single deep features perform much better than single classical features; but (3) the combination of classical features achieves competitive accuracy with combine deep features. However, extracting deep features takes more than 400 ms on a Core i7 4.2GHz CPU (it takes about 16 ms on a GeForce GTX 1080Ti GPU), while extracting the combination classical features only takes about 9 ms on the same CPU. Therefore, we adopted combination classical features for smile detection in our experiments.

4.3.3. Regression forests V.S. classification forests

To validate that regression forests trained on face images with continuous labels benefit to smile detection, we conducted an experiment for comparing the performances between regression forests and classification forests. Besides the regression forests, a group of conditional classification forests [3] were also trained using the augmented LFW training dataset with binary labels. The results listed in Table 2 indicate that the accuracy achieved by conditional random regression forests is about 4% higher than that achieved by conditional random classification forests.

Table 1
Accuracy (%) of the proposed method using different image features.

Feature	LFW	CCNU-Class
FC1 from fine-tuned VGG-face	94.35	92.06
FC2 from fine-tuned VGG-face	94.60	93.34
FC3 from fine-tuned VGG-face	94.80	93.77
FC1 + FC2 + FC3 from fine-tuned VGG-face	**95.80**	**93.94**
Gray intensity	69.45	66.42
Sobel gradients	66.95	62.53
HOG	81.30	82.46
LBP	83.25	81.43
Gray + Sobel + HOG + LBP	94.05	92.17

Bold values signifies the best results.

Table 2
Accuracy (%) of regression forests and classification forests.

Method	LFW	CCNU-Class
Regression forests	**94.05**	**92.17**
Classification forests	90.10	87.62

4.3.4. *Analysis of uncertainty measure*

The uncertainty measure has a profound impact on the performance of random forests. Hence, we trained three groups of conditional random forests with different uncertainty measures (i.e., the covariance-based uncertainty measure [32], the classification objective uncertainty measure [20] and the proposed clustering-based uncertainty measure), and investigated performances of these three groups of conditional random forests. As shown in Table 3, the conditional random forests trained with the proposed clustering-based uncertainty measure outperforms the conditional random forests trained with the other two uncertainty measures.

4.3.5. *Comparisons of ensemble schemes*

We compared five ensemble schemes of forests, i.e., the full soft ensemble scheme described in Algorithm 1, the speed-up soft ensemble scheme described in Algorithm 2, a hard ensemble scheme that select all the trees in the conditional forests of the highest head pose probability, a blind ensemble scheme that randomly selects trees without taking head pose into account, and a full ensemble scheme that selects all trees in all the 9 conditional random forests. As listed in Table 4, the soft and speed-up soft ensemble scheme perform much better than the other ensemble schemes, and the speed-up soft ensemble scheme runs about 9× faster than the full soft ensemble scheme.

4.3.6. *Evaluation of data augmentation strategy*

To evaluate the effect of the data augmentation strategy described in Section 3.4, we trained two groups of conditional random forests with or without augmented data. The results listed in Table 5 show that using the data augmentation strategy improves the performance of the conditional random forest by 5.65%.

4.3.7. *Comparisons of different methods*

The proposed method was compared with deep learning based methods as well as the classical methods. Table 6 lists the accuracy achieved by the proposed method, the VGG16 [41], VGG-SVM [42], DeepNDF [43], c-CNN [14], random forests (RF) [3], support vector machine (SVM) and AdaBoost. The VGG16 [41], VGG-SVM [42], DeepNDF [43], c-CNN [14] were fine-tuned on the augmented LFW training dataset. The proposed method, RF [3], SVM and AdaBoost were trained using the same image features (i.e., Gray + Sobel + HOG + LBP). The fine-tuned c-CNN [14] achieves the best performance among these eight methods, because c-CNN learns expression representations and the pose-specific adaptive routes that reveal the distribution of underlying modalities. Although using hand-crafted image features, the proposed method achieves competitive results with the c-CNN. The other methods that don't deal with head poses don't achieve satisfying results on both the two testing datasets. Emphasize again, the proposed method can run in real-time on generic CPUs, while the deep learning based methods cannot.

5. Conclusions

This paper has presented a real-time pose invariant spontaneous smile detection method based on conditional random regression forests. Since the relation between image patches and smile intensity is modelled conditional to head pose, the proposed smile detection method is not sensitive to various head poses. Several techniques including regression forests, multiple-label dataset augmentation and non-informative patch removement have been employed to improve the performance of smile detection. Although using hand-crafted features, the proposed method achieves competitive performance to state-of-the-art deep neural network based methods on two challenging real-world datasets. The ensemble scheme, which dynamically selects a fraction of trees from the trained CRRFs for estimating smile intensity, makes a good trade-off between smile detection performance and processing speed. Hence, the proposed method can run in real-time on general hardware without GPU, which is a significant advantage in contrast to deep neural networks. In our future work, we intend to combine a lite neural network with conditional random regression forests to learn image representation and detect smile face in an end-to-end way.

Table 3
Accuracy (%) of forests trained with different uncertainty measures.

Method	LFW	CCNU-Class
Clustering-based	**94.05**	**92.17**
Covariance-based	92.70	91.37
Classification objective	90.10	87.62

Bold values signifies the best results.

Table 4
Accuracy (%) and running speed (FPS) of different ensemble schemes.

Ensemble schemes	LFW	CCNU-Class	Running speed
Full soft	**94.75**	**92.81**	1.61
Speed-up soft	94.05	92.17	13.70
Hard	88.65	85.52	13.27
Blind	80.55	79.93	13.59
Full	81.80	81.52	**1.52**

Table 5
Accuracy (%) of forests trained with and without augmented data.

Forests	LFW	CCNU-Class
With augmented data	**94.05**	**92.17**
Without augmented data	91.75	86.52

Table 6
Accuracy (%) of different methods.

Methods	LFW	CCNU-Class
The proposed	94.05	92.17
VGG16 [41]	89.65	88.77
VGG-SVM [42]	90.20	90.13
DeepNDF [43]	91.35	91.82
c-CNN [14]	**95.85**	**94.36**
RF [3]	81.75	73.52
SVM	83.25	72.38
AdaBoost	71.95	66.27

Bold values signifies the best results.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61702208), the National Key R&D Program of China (2017YFB1401300, 2017YFB1401303), the Natural Science Foundation of Hubei Province (2017CFB504, 2018CFB404), and the Research Funds of CCNU from the Colleges Basic Research and Operation of MOE (CCNU17QN0003, CCNU17ZDJC04).

References

[1] J. Whitehill, G. Littlewort, I. Fasel, Toward practical smile detection, IEEE Trans. Pattern Anal. Mach. Intell. 31 (11) (2009) 2106–2111, https://doi.org/10.1109/TPAMI.2009.42.
[2] J. Chen, N. Luo, Y. Liu, L. Liu, K. Zhang, J. Kolodziej, A hybrid intelligence-aided approach to affect-sensitive e-learning, Computing 98 (1-2) (2016) 215–233, https://doi.org/10.1007/s00607-014-0430-9.
[3] Z. Luo, L. Liu, J. Chen, Y. Liu, Z. Su, Spontaneous smile recognition for interest detection, Chin. Conf. Pattern Recogn. (2016) 119–130, https://doi.org/10.1007/978-981-10-3002-4_10.
[4] T. Senechal, J. Turcot, R. Kaliouby, Smile or smirk? automatic detection of spontaneous asymmetric smiles to understand viewer experience, IEEE Int. Conf. Workshops Autom. Face Gesture Recogn. (2013) 1–8, https://doi.org/10.1109/FG.2013.6553776.
[5] C. Shan, Smile detection by boosting pixel differences, IEEE Trans. Image Process. 21 (1) (2012) 431–436, https://doi.org/10.1109/TIP.2011.2161587.
[6] H. Liu, Y. Gao, P. Wu, Smile detection in unconstrained scenarios using self-similarity of gradients features, IEEE Int. Conf. Image Process. (2014) 1455–1459, https://doi.org/10.1109/ICIP.2014.7025291.
[7] L. An, S. Yang, B. Bir, Efficient smile detection by extreme learning machine, Neurocomputing 149 (1) (2015) 354–363, https://doi.org/10.1016/j.neucom.2014.04.072.
[8] Y. Gao, H. Liu, P. Wu, C. Wang, A new descriptor of gradients self-similarity for smile detection in unconstrained scenarios, Neurocomputing 174 (2) (2016) 1077–1086, https://doi.org/10.1016/j.neucom.2015.10.022.
[9] H. Ding, S. Zhou, R. Chellappa, Facenet2expnet: regularizing a deep face recognition net for expression recognition, IEEE Conf. Autom. Face Gesture Recogn. (2017) 118–126, https://doi.org/10.1109/FG.2017.23.
[10] A. Mollahosseini, D. Chan, M. Mohammad, Going deeper in facial expression recognition using deep neural networks, EEE Winter Conf. Appl. Comput. Vis. (2016) 118–126, https://doi.org/10.1109/WACV.2016.7477450.
[11] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, K. Yan, A deep neural network-driven feature learning method for multi-view facial expression recognition, IEEE Trans. Multimedia 18 (12) (2016) 2528–2536, https://doi.org/10.1109/TMM.2016.2598092.
[12] Y. Liu, X. Yuan, X. Gong, Z. Xie, F. Fang, Z. Luo, Conditional convolution neural network enhanced random forest for facial expression recognition, Pattern Recogn. 84 (7) (2018) 251–261, https://doi.org/10.1016/j.patcog.2018.07.016.
[13] C. Corneanu, M. Simón, J. Cohn, Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: Hhistory, trends, and affect-related applications, IEEE Trans. Pattern Anal. Mach. Intell. 38 (8) (2016) 1548–1568, https://doi.org/10.1109/TPAMI.2016.2515606.
[14] C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan, T. Kim, Conditional convolutional neural network for modality-aware face recognition, IEEE Int. Conf. Comput. Vis. (2015) 3667–3675, https://doi.org/10.1109/ICCV.2015.418.
[15] J. Ma, J. Zhao, Y. Ma, J. Tian, Non-rigid visible and infrared face registration via regularized Gaussian fields criterion, Pattern Recogn. 48 (3) (2015) 772–784.

[16] J. Ma, J. Zhao, H. Guo, J. Jiang, H. Zhou, Y. Gao, Locality preserving matching, Int. J. Comput. Vis. (2018), https://doi.org/10.1007/s11263-018-1117-z.

[17] W. Zheng, Multi-view facial expression recognition based on group sparse reduced-rank regression, IEEE Trans. Affect. Comput. 5 (1) (2014) 71–85, https://doi.org/10.1109/TAFFC.2014.2304712.

[18] M. Jampour, V. Lepetit, T. Mauthner, H. Bischof, Pose-specific non-linear mappings in feature space towards multiview facial expression recognition, Image Vis. Comput. 58 (2) (2016) 38–46, https://doi.org/10.1016/j.imavis.2016.05.002.

[19] J. Ma, J. Zhao, J. Tian, A. Yuille, Z. Tu, Robust point matching via vector field consensus, IEEE Trans. Image Process. 23 (4) (2014) 1706–1721.

[20] M. Dantone, J. Galln, G. Fanelli, Real-time facial feature detection using conditional regression forests, IEEE Conf. Comput. Vis. Pattern Recogn. (2012) 2578–2585, https://doi.org/10.1109/CVPR.2012.6247976.

[21] J. Girard, J. Cohn, F.D. la Torre, Estimating smile intensity: a better way, Pattern Recogn. Lett. 66 (11) (2015) 13–21, https://doi.org/10.1016/j.patrec.2014.10.004.

[22] G. Levi, T. Hassner, Emotion recognition in the wild via convolutional neural networks and mapped binary patterns, ACM Conf. Int. Conf. Multimodal Interaction (2015) 503–510.

[23] A. Lopes, E.A. ad, A. Souza, T. Oliveira-Santos, Facial expression recognition with convolutional neural networks: coping with few data and the training sample order, Pattern Recogn. 61 (2017) 610–628.

[24] J. Chen, Q. Ou, Z. Chi, H. Fu, Smile detection in the wild with deep convolutional neural networks, Mach. Vis. Appl. 28 (1-2) (2017) 173–183, https://doi.org/10.1007/s00138-016-0817-z.

[25] S. Du, Y. Tao, A.M. Martinez, Compound facial expressions of emotion, PNAS (2014) 1454–1462, https://doi.org/10.1073/pnas.1322355111.

[26] D. Cui, G. Huang, T. Liu, Smile detection using pair-wise distance vector and extreme learning machine, Int. Joint Conf. Neural Netw. (2016) 2298–2305, https://doi.org/10.1109/IJCNN.2016.7727484.

[27] D. Cui, G. Huang, T. Liu, Elm based smile detection using distance vector, Pattern Recogn. 79 (7) (2018) 356–369, https://doi.org/10.1016/j.patcog.2018.02.019.

[28] M. Sun, P. Kohli, J. Shotton, Conditional regression forests for human pose estimation, IEEE Conf. Comput. Vis. Pattern Recogn. (2012) 3394–3401.

[29] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[30] D. Tang, H.J. Chang, A. Tejani, T.K. Kim, Latent regression forest: structured estimation of 3D articulated hand posture, IEEE Conf. Comput. Vis. Pattern Recogn. (2014) 3786–3793.

[31] Y. Liu, J. Chen, Z. Su, Z. Luo, N. Luo, L. Liu, K. Zhang, Robust head pose estimation using Dirichlet-tree distribution enhanced random forests, Neurocomputing 173 (1) (2016) 42–53, https://doi.org/10.1016/j.neucom.2015.03.096.

[32] G. Fanelli, J. Gall, L.V. Gool, Real time head pose estimation with random regression forests, IEEE Conf. Comput. Vis. Pattern Recogn. (2011) 617–624, https://doi.org/10.1109/CVPR.2011.5995458.

[33] D. Arthur, S. Vassilvitskii, K-means+ +: the advantages of careful seeding, Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (2007) 1027–1035.

[34] D. Arthur, S. Vassilvitskii, Mean shift, mode seeking, and clustering, IEEE Trans. Pattern Anal. Mach. Intell. 17 (8) (1995) 790–799, https://doi.org/10.1109/34.400568.

[35] Y. Gao, J. Ma, A.L. Yuille, Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples, IEEE Trans. Image Process. 26 (5) (2017) 2545–2560.

[36] X. Geng, Y. Xia, Head pose estimation based on multivariate label distribution, IEEE Conf. Comput. Vis. Pattern Recogn. (2014) 1837–1842, https://doi.org/10.1109/CVPR.2014.237.

[37] G. Huang, M. Ramesh, T. Berg, et al., Labeled faces in the wild: a database for studying face recognition in unconstrained environments, Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition (2008) 1–14.

[38] M. Pietikäinen, Local binary patterns, Scholarpedia 5 (3) (2010) 9775, https://doi.org/10.4249/scholarpedia.9775.

[39] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, IEEE Conf. Comput. Vis. Pattern Recogn. (2005) 886–893, https://doi.org/10.1109/CVPR.2005.177.

[40] L. Liu, J. Chen, C. Gao, N. Sang, A low-cost real-time face tracking system for ITSS and SDASS, Softw.: Pract. Exp. 47 (8) (2017) 1111–1126, https://doi.org/10.1002/spe.2455.

[41] O. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, Proc. Br. Mach. Vis. Conf. (2015) 1–12, https://doi.org/10.5244/C.29.41.

[42] J. Chen, R. Xu, L. Liu, Deep peak-neutral difference feature for facial expression recognition, Multimedia Tools Appl. 77 (22) (2018) 29871–29887, https://doi.org/10.1007/s11042-018-5909-5.

[43] P. Kontschieder, M. Fiterau, A. Criminisi, R. Bulo, Deep neural decision forests, Proc. Int. Conf. Comput. Vis. (2015) 1467–1475, https://doi.org/10.1109/ICCV.2015.172.