3D Model-Based Continuous Emotion Recognition

Hui Chen[1], Jiangdong Li[2], Fengjun Zhang[3], Yang Li[1], Hongan Wang[2,3]
Beijing Key Lab of Human-computer Interaction, Institute of Software, Chinese Academy of Sciences[1]
University of Chinese Academy of Sciences[2]
State Key Lab of Computer Science, Institute of Software, Chinese Academy of Sciences[3]
Beijing, China, 100190

Abstract

We propose a real-time 3D model-based method that continuously recognizes dimensional emotions from facial expressions in natural communications. In our method, 3D facial models are restored from 2D images, which provide crucial clues for the enhancement of robustness to overcome large changes including out-of-plane head rotations, fast head motions and partial facial occlusions. To accurately recognize the emotion, a novel random forest-based algorithm which simultaneously integrates two regressions for 3D facial tracking and continuous emotion estimation is constructed. Moreover, via the reconstructed 3D facial model, temporal information and user-independent emotion presentations are also taken into account through our image fusion process. The experimental results show that our algorithm can achieve state-of-the-art result with higher Pearson's correlation coefficient of continuous emotion recognition in real time.

1. Introduction

Continuous emotion analysis refers to acquire and process long unsegmented naturalistic inputs and to predicate affective values represented in dimensional space [15]. It has been recognized that computers which can understand emotions in natural interactions have the ability to make smarter decisions and provide better interactive experiences [17, 27]. By classifying emotions as different categories, some human-centered systems like [12, 20] have been designed to react differently for different user emotion categories, which provide better interactive experiences for users and show the importance and necessity of emotion estimation in human-computer interactions. In natural communications, people talk and think continuously, and the human emotions are also revealed naturally. Thus emotions in natural communications should be estimated as real values of different affective dimensions for higher quality of human-computer interactions.

Visual signals have been proved to be the most effective and important cues for emotion recognition [1, 18, 22]. Presented in a spontaneous way, emotions in natural communications tend to change more slowly than acted ones, leading to more subtle sequential expressions. The significant presentations in natural exchanges are usually not fully-expressed, resulting in fuzzy difference between different emotion states. Additionally, people express their emotions in variable ways which introduces larger confused information of similar emotions and brings greater challenge about how to link a certain user's expression with more common presentations. Besides, emotions captured from natural interactions are always with large changes, such as more freely head rotations, fast head motions, partial facial occlusions, etc. These characteristics increase the complexity of continuous emotions and make it hard to estimate emotions in natural communications accurately and robustly.

To meet these challenges, we propose a real-time 3D model-based method that recognizes human emotions in dimensional space under natural communications. Our approach introduces 3D facial model into continuous emotion recognition, which brings higher robustness to handle changes of large head rotations, fast head motions and partial facial occlusions. User-specific temporal features and user-independent emotion presentations are also constructed to describe emotions more precisely. The emotions are estimated by a novel random forest-based framework, in which the 3D facial tracking and continuous emotion estimation are taken simultaneously in a regression way.

2. Related work

Human emotions are usually represented in two ways: categorical and dimensional. According to the Facial Action Coding System (FACS) proposed by P. Ekman [9], emotions can be categorized as six classes: happiness, sadness, anger, surprise, fear and disgust. Naturalistic human emotions are complex with fuzzy boundaries in expressions, thus discrete categories may not reflect the subtle emotion transitions and the diversity emotions. Therefore, many

978-1-4673-6964-0/15/$31.00 ©2015 IEEE

works use dimensional representations to interpret human emotions in different affective dimensions. The PAD emotion space [33] is a typical one, which describes continuous emotions in three dimensions of Pleasure, Arousal and Dominance. Fontaine *et al*. [13] described continuous emotions in four dimensions as Arousal, Valence, Power and Expectancy. Dimensional representations can analysis emotions on several continuous scales and describe emotion transitions better, which accordingly is more suitable to represent emotions in natural human-computer interactions.

Most existing emotion recognition algorithms use 2D features extracted from images to predict emotions, which can be subdivided by using appearance features and geometric features [11]. For instance, Wu *et al*. [36] used intensity after Gabor Motion Energy Filters to classify emotions. Kapoor *et al*. [18] took the pixel difference of mouth region to estimate emotions. Such algorithms based on appearance features and achieved good results when the facial pose are consistent. Some works used the 2D facial geometric features. Valstar and Pantic [34] used the geometry feature of 20 2D facial points to predict emotions. Kobayashi and Hara [19] used facial geometric model to recognize emotions. There are also some works like [1, 32] estimating emotions using 2D hybrid features of appearance and geometry such as Active Appearance Model (AAM). 2D features can be directly extracted, but as Sandbach *et al*. [28, 29] pointed out in their surveys, they are not stable enough for the large changes in communications and a consistent facial pose is necessary when 2D features are used, which showed that 2D feature-based algorithms are not adequately robust to recognize continuous emotions.

3D features have also been integrated in many algorithms. Compared with approaches using 2D image features, 3D feature-based approaches are more robust and powerful for emotion recognition. Works using 3D features can be categorized as shape-based and depth-based. Shape-based algorithms use parameters of 3D curve shapes, the positions of 3D landmarks or the changes of 3D landmarks to classify emotions. For instance, Huang *et al*. [31, 39] used Bézier volume to describe facial expressions and took the changes of manifold parameters as symbols of the changes of emotions. Their experimental results showed that the Bézier volume based approaches worked well on classifying spontaneous emotions. Some other works took facial depth features to recognize emotions. Fanelli *et al*. [10] used depth information to classify emotions into discrete categories. The existing 3D feature-based algorithms are adequately robust, but they are rarely used for continuous emotions recognition. In this paper, we present an effective regressive approach that use 3D facial information to estimate continuous emotion in dimensional space.

Continuous emotion presentations are sequential actions. Fused emotion presentations have been designed in order to include dynamic temporal information, eliminate user-dependent information and conquer the large changes of communication environment. Yang and Bhanu [37, 38] showed their image fusion method which used SIFT-flow algorithm combining images from one video clip into one image. SIFT-flow [23] is a robust algorithm for 2D images alignment and also works well in face registration, but it is comparatively time-consuming. With the help of the 3D model, we propose a real-time image fusion method to represent continuous emotions and user-independent emotions respectively.

A lot of methods have been designed to recognize continuous emotions [16]. Some typical schemes are: Support Vector Regression (SVR) [30], Relevance Vector Machines (RVM) [25], Conditional Random Fields (CRF) [2] and so on. As a popular method, random forest [4] has been widely used in both classification and regression tasks. Random forest consists of several classification and regression trees (CART). It can deal with large amount of training samples without over-fitting [8] and has the characteristics of robustness, high-efficiency and powerful ability of regression. Due to the structure of binary trees, it can achieve result with little time cost. Fanelli *et al*. [10] proposed a random forest based framework to estimate the posture of a head by regression from data captured by depth camera. The output showed that random forest can handle facial regression problems in high quality. In our work, we further exploit the regression ability of random forest, wherein 3D facial tracking and continuous emotion estimation could take effect jointly.

3. Proposed method

We propose a random forest-based algorithm that can recognize emotions in dimensional space under natural communications. Different from existing algorithms like [1, 2, 32], our approach uses 3D facial model reconstructed from 2D image, which maintains the positional relationship of facial landmarks and provides more robust clues to overcome changing environments. The continuous emotion presentation and user-independent emotion presentation are also taken into account via 3D head model-based image fusion to describe emotions more precisely.

The framework of our work is shown in Figure 1. During training period, the 3D facial model of input images are firstly restored. Then continuous emotion presentation (CEP) and the user-independent emotion presentation (UIEP) are constructed by image fusion. The 3D facial shapes, CEP images together with their emotion values constitute an augmented training set with which the random forest is constructed. In emotion estimation period, two regressions are taken in the random forest simultaneously: one is for tracking the 3D facial expression, the other is for recognizing the current emotion. The CEP image of current time

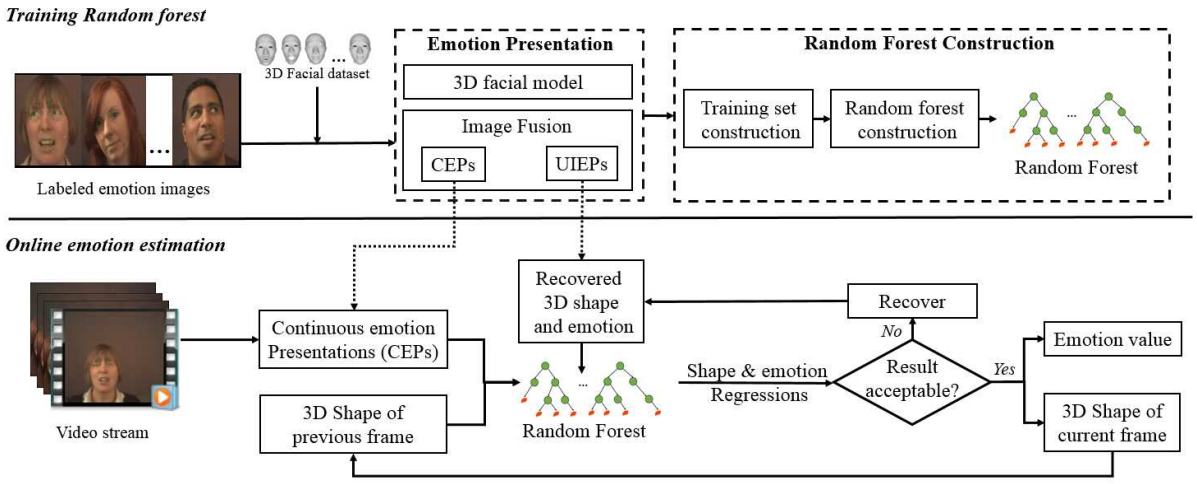

Figure 1. Framework of our 3D model-based continuous emotion recognizing and tracking approach.

step and 3D facial shape of previous time step are taken as the inputs, then the affective value and 3D facial shape of current time step are calculated as outputs. When there are no acceptable outputs of random forest, the recovery operation is taken with the help of UIEP images to achieve recovered 3D facial shape and emotion.

3.1. Data preparing

Continuous emotion dataset are always presented as video clips, which contain too many images and a large part of these images are very similar in emotion value and appearance. In order to reduce the data redundancy and improve the representativeness of training data, the reduced training images are firstly picked from all the frames. During the image picking step, we make sure that the selected images have evenly distributed affective values, cover the entire emotion range and retain the different head postures in natural communicates. For every affective dimension, relatively small training images, around 160, are firstly picked in our method. Then, facial landmarks of every selected image are automatically detected via the algorithm proposed by Baltrusaitis *et al*. [3]. Considering the fact that emotion information are mostly showed by mouth, eyes, and eyebrows, only 42 inner landmarks are chosen in our method, including 8 eyebrows landmarks, 12 eye corner landmarks, 4 nasal landmarks and 18 lip landmarks. Figure 2 shows the labelled landmarks of some selected images.

3.2. Restore 3D facial model

In our method, 3D facial shapes of every labelled images are restored with the help of FaceWarehouse [6], which is a 3D facial dataset containing 3D facial models of 150 subjects from various ethnic backgrounds and every subject has

Figure 2. Facial landmarks of some selected images.

47 FACS blendshapes with 11K vertices. It can be described as a third order tensor:

$$F = C_r \times w_{id}^T \times w_{exp}^T \tag{1}$$

where C_r is a 3D facial blendshape with 11K vertices, and w_{id}^T, w_{exp}^T are the column vectors of identity weights and expression weights in the tensor respectively.

According to works proposed by Cao Chen *et al*. [5], constructing 3D model from 2D image can be separated into two steps: the first step is to calculate the optimal w_{id}^T. With the optimal blendshapes of w_{id}^T, the 3D facial shape of every picked image are constructed in the second step. These two steps both work in an iterative way.

Different from the work of Cao Chen *et al*. which focus on specific user, we want to represent the input images from different persons in a uniform way, we consider that all the input images should be constructed by blendshapes of the same w_{id}^T in FaceWarehouse. So when calculating the optimum w_{id}^T, an energy formula is defined considering this constraint as:

$$E_{id} = \sum_{i=1}^{N} \sum_{b=1}^{42} \left\| P(M^i(C_r \times w_{id}^T \times w_{exp,i}^T)^b - u_i^b) \right\|^2 \tag{2}$$

where N is the number of the picked images; P means the projection matrix; M^i means the extrinsic parameter matrix of camera which can be computed via EPnP algorithm [21]; $w_{exp,i}^T$ stands for the most similar expression for the i^{th} image; u_i^b is the b^{th} landmark on image. The identity w_{id}^T which has the least energy E_{id} are considered the optimal identity and the 47 blendshapes of the optimal w_{id}^T are taken as the fundamental blendshapes for 3D facial model reconstruction.

Once the fundamental blendshapes are acquired, the 3D facial model of every image can be restored via the linear interpolation of fundamental blendshapes as [21] did and the 3D facial models of the picked 2D images can be all reconstructed.

3.3. Image fusion

With the help of the 3D emotion presentations, an image fusion method is implemented. Figure 3 shows the pipeline of our image fusion method. First of all, we label the landmarks of input images using algorithm [3] and reconstruct the 3D facial model. Then the 3D facial shape is transformed to the orthogonal position of space coordinate system and projected to the 2D facial coordinate system as the following formula:

$$u_i^{OP_b} = P(M_{R|t} * V^b) \tag{3}$$

where P means the projection matrix of camera, $M_{R|t}$ represents the transform matrix for a 3D shape from its original position to the orthogonal position of current space coordinate system. V^b means the b^{th} landmark on the 3D facial shape.

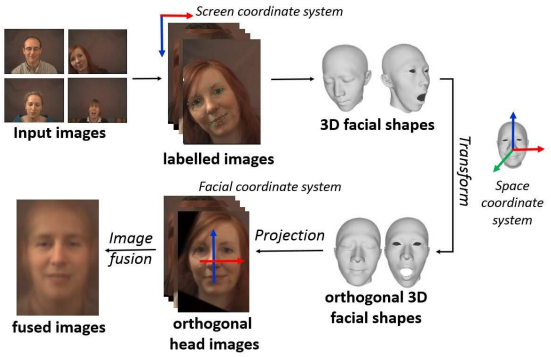

Figure 3. Image fusion pipeline.

With the original landmarks and the projected landmarks $u_i^{OP_b}$, the homographic transform matrix from the original screen space to the facial coordinate space is acquired. The facial part of original image is unified into the 2D facial coordinate system. After transforming all the facial parts of original images to the unified facial coordinate system, these images are superposed and result in one fusion presentation.

For different goals, the image fusion method is used to generate user-specific continuous emotion presentation and user-independent emotion presentations in our work. Continuous emotion presentation (CEP) merges several continuous adjacent frames from a video clip, which is used to contain the dynamic feature and temporal context of emotions. User-independent emotion presentation (UIEP) fuses different images selected from different videos with the same emotion value into one image presentation, which is used to retain the prominent features of same emotion state and eliminate the differences among different persons.

3.4. Training

Training set construction. Random forest is made up of several classification and regression trees (CARTs). As Section 3.1 stated, a relatively small number of emotion samples have been picked, which is not enough to guarantee the robustness and precision of CART. So we firstly expand the emotion samples in order to make them large enough for training.

Suppose $\{CEP_i, M_i, S_i, A_i\}$ is the emotion sample of the i^{th} training image, where CEP_i is the fused continuous emotion presentation of the i^{th} image; S_i is the reconstructed 3D emotion shape; A_i is the labelled affective value; and M_i is the identity matrix. We firstly translate 3D emotion shape S_i along three coordinate axes respectively and get $M-1$ additional 3D emotion shapes, which expands the number of training samples to $N \times M$ as $\{CEP_{ij}, M_{ij}, S_{ij}, A_i\}$, where M_{ij} is the transformation matrix that maps S_{ij} back to S_i. The corresponding homography matrix M_{HOMO} of M_{ij} is then computed out. With M_{HOMO}, CEP_i can be transformed to CEP_{ij}, which is used as the continuous emotion presentation of S_{ij}.

Then several most similar emotion samples of each transformed emotion sample $\{CEP_{ij}, M_{ij}, S_{ij}, A_i\}$ were found. Suppose $\{CEP_{ij}, M_{ij}^l, S_l, A_l\}$ represents another emotion sample, the differences between two emotion samples is evaluated as follows:

$$E_l = \sum_{b=1}^{42} \left\| S_{ij}^b - S_l^b \right\|^2 + w_a \left\| A_i - A_l \right\| \tag{4}$$

$$S_l = M_{ij}^l S_{ij} \tag{5}$$

where superscript b means the b^{th} landmark on the 3D facial shape S_{ij} and S_l; A_i and A_l are the affective values of corresponding shape respectively; M_{ij}^l is the transform matrix between 3D shapes; and w_a is an empirical weight to balance the influences of shape diversity and emotion diversity, here is 350. The most-like emotion samples can be found through minimizing above energy E_l. Then the emotion samples can be extended to $\{CEP_{ij}^l, M_{ij}^l, S_{ij}^l, A_l\}$. Finally, transform S_{ij}^l along three

coordinate axes respectively and randomly pick K shapes from its transformed shapes, we will get the augmented emotion shapes $\{CEP_{ij}^{lk}, M_{ij}^{lk}, S_{ij}^{lk}, A_l\}$. After augmentation, the number of training emotion samples is extended from N to $N \times M \times L \times K$. Here, we set $N = 160, M = 9, L = 3$ and $K = 7$.

With the augmented emotion samples, training patches are then constructed in order to train the random forest. As for emotion sample $\{CEP_{ij}^{lk}, M_{ij}^{lk}, S_{ij}^{lk}, A_l\}$, several training patches reflecting the displacement of 3D emotion shape, difference in affective value, and appearance of image are generated. The displacements of each facial landmarks on emotion shape S_{ij}^{lk} from original shape S_i are recorded as $Dis_s(S_{ij}^{lk}, S_i)$. The difference between affective values is presented as $Dis_a(A_l, A_i)$. To represent appearance of 2D image, we randomly choose Q points from facial area in CEP_{ij}^{lk} and concatenate the intensity values as an intensity vector $Int(CEP_{ij}^{lk})$, where Q is fixed to 400 in our test. Thus, a patch vector is set up as $P = \{Int(CEP_{ij}^{lk}), Dis_s(S_{ij}^{lk}, S_i), Dis_a(A_l, A_i)\}$. Figure 4 indicates the example of generating training patches for one emotion sample. We randomly pick Z intensity vectors in each CEP and get Z patches $\{P_z \mid 1 \leq z \leq Z\}$ in every emotion sample, where Z is set to 100. Finally, a training set including $N \times M \times L \times K \times Z$ training patches are constructed.

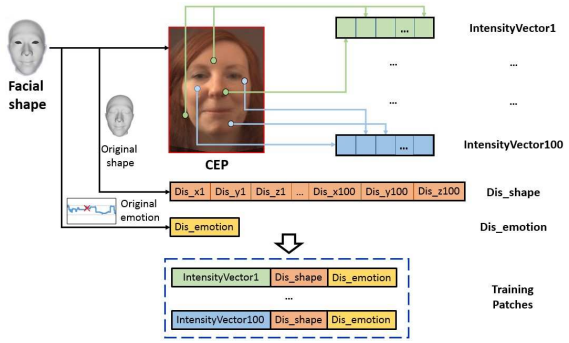

Figure 4. Training patches construction.

Random forest construction. With the generated patches, random forest with several CARTs is constructed. When training every CART, only 70 percent of the patches are used to avoid over-fitting.

In every non-leaf node, a binary test is conducted to split training patches, which is defined as follows:

$$|F_1|^{-1} \sum_{q1 \in F_1} Int_{q1} - |F_2|^{-1} \sum_{q2 \in F_2} Int_{q2} > \tau \quad (6)$$

where F_1 and F_2 means two fragments from current training patch, Int represents the intensity vector and τ is a random threshold. In our test, the length of F_1 and F_2 is set to 60 and the range of binary test threshold is from $[-30, 30]$.

For each non-leaf node, we generate 2000 binary tests $\{t^x\}$ by randomly choosing the parameters of F_1, F_2 and τ. The quality of every binary test is evaluated by regression uncertainty U_R, which consists two parts: the shape regression uncertainty U_{R_s} and the affect regression uncertainty U_{R_a}. These two regression uncertainties are defined as:

$$U_{R_s}(P \mid t^x) = H(P)_s - w_L H(P_L)_s - w_R H(P_R)_s \quad (7)$$

$$U_{R_a}(P \mid t^x) = H(P)_a - w_L H(P_L)_a - w_R H(P_R)_a \quad (8)$$

where $H(P)$ means the differential entropy of patch set and w_L, w_R are the ratio of patches sent into left and right child node respectively. It is assumed that the distribution of training set is normal distribution. So the regression uncertainties can be computed in following formulas:

$$U_{R_s}(P \mid t^x) = \log(|\Sigma^s|) - \sum_{i=\{L,R\}} w_i \log(|\Sigma^s|) \quad (9)$$

$$U_{R_a}(P \mid t^x) = \log(|\Sigma^a|) - \sum_{i=\{L,R\}} w_i \log(|\Sigma^a|) \quad (10)$$

where Σ^s and Σ^a are the covariance matrices of the displacements of shape landmarks and affects. Then total uncertainty U_R can be presented as:

$$U_R(P \mid t^x) = U_{R_s}(P \mid t^x) + \lambda U_{R_a}(P \mid t^x) \quad (11)$$

where λ is an empirical weight which equals to 1. By maximizing U_R, we can minimize the determinants of these covariance matrices and find the best binary test of current node, which is described as t^{opt}.

Once t^{opt} is found, we save the parameters of the optimal binary test as a part of random regression forest and split the training patches of current node into its left child and right child. We take a node as a leaf if it reaches the deepest level L_{max} or the number of patches it contains is less than the minimum threshold P_{min}. Here L_{max} is set to 15 and P_{min} is 20. A leaf node stops splitting and saves the information about patches it holds including the mean and covariance of shape displacements $\{Ave_s, |\Sigma^s|\}$ together with the average and covariance of affect displacements $\{Ave_a, |\Sigma^a|\}$.

3.5. Online emotion estimation

Preparation works. Before online emotion estimation, some preparation works should be done. During emotion recognition, the situation with the lost will appear. In that case, the 3D facial model and affective value need to be restored. So we prepare the shape recovery set and emotion recovery set in advance. To prepare shape recovery set, some frames are picked at a fixed time interval from input video. The 3D facial shapes of these picked images are reconstructed as a shape recovery set R_{shape}. For emotion recover set, images with similar affective values are generated

and stored into several groups. User-independent emotion presentations (UIEPs) of these groups are then calculated. Facial landmarks of every UIEP are automatically detected and the LBP features of each landmark region (a set of 10×10 points around the landmark) are saved as LBP emotion presentation. The LBP emotion presentations and their corresponding emotion values of all UIEPs are collected as emotion recovery set $R_{emotion}$.

Another preparation work is to generate the 3D emotion shape and the emotion value of first frame. We use the method proposed in Section 3.1 to restore the 3D shape of first frame. When computing the emotion value of the first frame, we calculate its LBP emotion presentation and find out its emotion value through comparing the similarity of the LBP emotion presentation of first frame and the LBP emotion presentation of UIEPs in $R_{emotion}$.

Emotion estimation. Taking CEP_t at the current time step, the 3D emotion shape and the affective value $\{S_{t-1}, A_{t-1}\}$ at the previous time step as input, the 3D emotion shape and affective value $\{S_t, A_t\}$ of current time step t can be estimated in a regression way.

Given the input emotion shape S_{t-1} and affective value A_{t-1} at the previous time step, several most-like 3D emotion shapes with their accordingly affective labels $\{S^w, A^w\}$ in the training dataset are picked out. Then the affine matrix M_w from S_{t-1} to S^w and the corresponding homography matrix of M_w is then generated, through which CEP_t is transformed to CEP_t^w as the continuous emotion presentation of S^w.

From S^w, we randomly choose 400 points from facial area and generate a patch set $P^w = \{Int(CEP_t^w), Dis_s(S^w, S_{t-1}), Dis_a(A_w, A_{t-1})\}$. Each test patch will be put into random forest and leaf node of every CART is achieved with the covariances of shape displacement $|\Sigma^s|$ together with the covariance of emotion displacement $|\Sigma^a|$. Thresholds $\theta_s = 10$ and $\theta_a = -1.5$ are set for picking acceptable leaves. If $\log(|\Sigma^s|)$ is more than θ_s or $\log(|\Sigma^a|)Ashraf09$ is more than θ_a, we discard the leaf. Finally a set of acceptable leaves is generated. Then the regression value of shape and affect are calculated through averaging shape displacements and emotion displacements separately. Add them to S^w and A^w separately, the new shape S^{w*} and affective value A^{w*} are achieved.

With all the similar 3D shapes chosen above, we can finally get a set of estimated value of 3D shape and emotion $\{S^{w*}, A^{w*}\}$. The median of these results $\{S^{w'}, A^{w'}\}$ is picked out as the final estimation of current 3D shape and emotion. Transform $S^{w'}$ using the inverse matrix M_w^{-1}, the 3D shape S_t will be achieved. The emotion value of current time step A_t is equal to $A^{w'}$.

As the variation trend of continuous emotions is usually placid [11], we calculate the mean of current emotion value with its previous 500 emotion values and take the result as the final emotion value of current time step.

Algorithm 1 Emotion estimation

$\{S^w, A^w\} \leftarrow$ a set of most-like 3D shapes and emotion values from training set $\{CEP_{ij}^{lk}, M_{ij}^{lk}, S_{ij}^{lk}, A_l\}$
for each simple in $\{S^w, A^w\}$ **do**
  $M_w \leftarrow$ affine transformation matrix from S_{t-1} to S^w
  $P^w \leftarrow \{Int(CEP_t^w), Dis_s(S^w, S_{t-1}), Dis_a(A^w, A_{t-1})\}$
  for $n = 1$ to N **do**
    $Leaf_n \leftarrow$ leaf node of the n^{th} CART reached by P^w
    if $Leaf_n \rightarrow |\Sigma^s| > \theta_s$ or $Leaf_n \rightarrow |\Sigma^a| > \theta_a$ **then**
      discard $Leaf_n$
    end if
  end for
  $\{Leaf_n\} \leftarrow$ the acceptable leaves
  $Reg_s \leftarrow (\Sigma_{n=1}^{N} ave_s^n)/N$
  $Reg_a \leftarrow (\Sigma_{n=1}^{N} ave_a^n)/N$
  $S^{w*} \leftarrow S^w + Reg_s$
  $A^{w*} \leftarrow A^w + Reg_a$
end for
$\{S^{w*}, A^{w*}\} \leftarrow$ alternative results from random forests
$\{S^{w'}, A^{w'}\} \leftarrow$ median result of $\{S^{w*}, A^{w*}\}$
$S_t \leftarrow M_{w'}^{-1} S^{w'}$
$A_t \leftarrow A^{w'}$
$return(S_t, A_t)$

Recovery. There are two situations that we need to recover the 3D facial shape and emotion value. One is the situation that no acceptable leaf is achieved. In this case, recoveries of both shape and emotion should be taken.

During shape recovery, we find out the 3D shape which is nearest to the current time step from shape recover set R_{shape} and take it as the new input shape. With the new input shape and the current CEP_t, the LBP emotion presentation LBP_t of CEP_t is calculated. Taking the affective value of last frame E_{t-1} as constraint, the energy E_a is defined as:

$$E_a = \|LBP_t - LBP_i\|^2 + \beta \|A_{t-1} - A_i\|^2 \qquad (12)$$

where LBP_i is one of the LBP emotion presentation of emotion recovery dataset $R_{emotion}$; and β is an empirical weight, which is set to 45. Through minimizing energy E_a, several UIEPs that most similar to current frame presentation are found. The mean of their affective values are then taken as the recovered affective value.

Another situation is that a large variation of affective value is detected between adjacent frames. As continuous emotions change subtly, if the difference between the adjacent emotion values are larger than an empirical threshold θ_{diffA}, we suppose that the estimated affective value is wrong and take the emotion recovery as above. In our test, θ_{diffA} is set to 0.2. The pseudo-code of online emotion estimation is showed in Algorithm 1.

4. Experiment

To evaluate the feasibility of the proposed method, we developed a prototype system and evaluated our method from three aspects: 1) the precision of 3D facial tracking; 2) the correlation coefficient of the emotion recognition; and 3) the computational performance of our method.

Our continuous emotion recognizing and tracking system is implemented on a PC with dual Intel Xeon CPUs (3.2GHz) and 4GB RAM.

4.1. Dataset

The Audio/Visual Emotion Challenge (AVEC 2012) Database [30] is a public continuous emotion dataset, which recorded audio-video sequences with the SEMAINE corpus [24]. In the database, emotions of every frame are annotated by humans in dimensions like Arousal, Valence and so on. The length of every video is about 3 to 5 minutes, each image in the video has the resolution of 780*580 and the frame rate is 50 fps.

We test our method using AVEC 2012 and evaluate the ability of emotion recognition by the Pearson's correlation coefficient. Since arousal and valence are more frequently used in emotion representation, we test our method on the these two dimensions and compare our result with the baseline of AVEC 2012 and several reported best results which also test on the same dataset.

4.2. Experiment results

The continuous emotion is estimated largely based on the positions of facial landmarks, we firstly compare the facial tracking precision of our algorithm with several typical works. In Table 1, we measure the RMSE (in pixels) for the landmarks on images of different algorithms compared with the ground truth positions. The results of facial tracking methods including multilinear models [35], 2D regression [7] and the state-of-the-art result of landmark tracking using 3D regression method [5] are referenced. From the table we can find our algorithm is more robust and precise to track the landmarks than the 2D tracking method and can achieve similar levels to the best result of 3D tracking, which means that our algorithm is precise enough in facial tracking for emotion estimation.

Figure 5 shows some results of our method in 3D facial tracking. The red dots represent the ground-truth of landmarks and the green ones are the tracking result of our method. From the outputs we can find that our 3D head model based method can retain good performance under the changing interaction environment like out-of-plane head rotations, fast head motions and partial facial occlusions.

Table 2 shows emotion estimation results of our method compared with several typical algorithms. Line 1 shows the result of the baseline of AVEC 2012 competition [30] which used SVR as the regressor; line 2 [26] shows the result of Multiscale Dynamic Cues method; the third line is the result of Continuous Facial Expression Representation (CFER) [32] and the fourth line is the result of Continuous Conditional Random Fields (CCRF) [3]. From the comparison we can see that our algorithm outperforms the other four methods in continuous emotion estimations.

In Figure 6 two subplots of our emotion estimation of a video compared with the ground-truth are showed. The red solid line is the result of our method and the blue dotted

Figure 5. Results of 3D facial tracking. Red dots: the ground-truth; Green dots: the tracking result.

RMSE	<3 pixels	<4.5 pixels	<6 pixels
Multilinear Model [35]	20.8%	24.2%	41.7%
2D Regression [7]	50.8%	64.2%	72.5%
3D Regression [5]	73.3%	80.8%	100%
Our Method	70%	83.93%	94.91%

Table 1. Percentages of frames with RMSE in facial tracking.

Correlation coefficient	Arousal	Valence	Mean
SVR [30]	0.151	0.207	0.179
Multiscale Dynamic Cues [26]	0.509	0.314	0.4165
CFER [32]	0.30	0.41	0.355
CCRF [2]	0.341	0.343	0.342
Our Method	0.564	0.454	0.509

Table 2. Pearson's correlation coefficient of typical emotion regression methods tested on AVEC 2012 dataset.

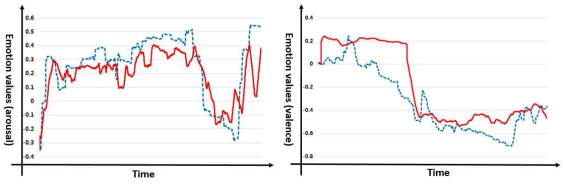

Figure 6. Comparison of our emotion estimation result and the ground-truth. Left: Arousal dimension, Right: Valence dimension.

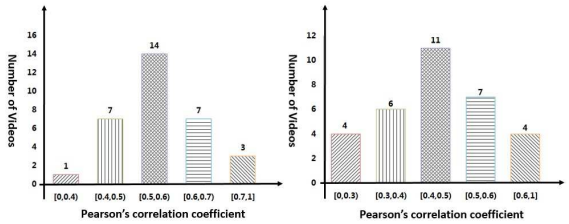

Figure 7. Distribution of correlation coefficients of arousal dimension (left) and valence dimension (right).

line is the ground-truth. The left one is the result in arousal dimension and the right one is in valence dimension. From the figure we can see that our algorithm can recognize the emotion value precisely.

Histograms in Figure 7 show the number of videos with different correlation coefficients of arousal dimension and valence dimension respectively. From these two histograms, we can see that the result of our system is accord with normal distribution approximately, which means the algorithm we proposed is robust and stable. Under most circumstances, our algorithm can achieve the correlation coefficient that between 0.5 and 0.6 in arousal dimension and 0.4 to 0.5 in valence dimension, which equals to the mean results in Table 2.

Time performance of our system depends mainly on the number of CARTs. Table 3 shows the efficiency of our system with different number of CARTs and the Pearson's correlation coefficient under these circumstances. It can be seen that with the increasing of CART numbers, the efficiency of system is falling down and the correlation of emotion estimation is growing better. When the number of CART is larger than 6, the estimation accuracy is almost stable to around 0.51. Since working in real time is necessary for an interaction system, we use 6 CARTs in practice as a trade-off between efficiency and accuracy. Then our method can handle emotion recognition at the speed of around 20 fps, which is usable for most interaction systems.

Number of CARTs	Frames handled in one second	Pearson's correlation coefficient
3	25	0.413
5	23	0.472
6	21	0.511
8	18	0.512
10	17	0.507
12	15	0.514

Table 3. Time performance using different CARTs.

5. Conclusions

In this paper, we propose a 3D model based continuous emotion recognition approach. We introduce 3D facial expression model into our work and restore the 3D facial model from 2D images. With the reconstructed 3D facial shape, an image fusion method is proposed to generate a user's continuous emotion expressions (CEP) and user-independent emotion expressions (UIEP). With the 3D facial models and fused images, a random forest which integrates two regressions for both 3D landmarks tracking and emotion estimating simultaneously is constructed.

Our algorithm has been tested on the video part of AVEC 2012 dataset. The experimental results showed that our real-time approach is powerful to achieve preferable result in continuous emotion estimation. Furthermore, the high efficiency of our system make it possible to provide intelligent responds timely in human-computer interactions.

Although our algorithm is based on 3D facial model, only 2D images or 2D video stream are used as inputs, which gets rid of the bondage of equipments. Due to the usability of our system, it is promising to deploy our system on mobile devices such as smartphones, tablet personal computers and so on.

The image features extracted in our work are intensity of images, which are the simplest features of an image. In future work, we will try to use more robust features such as ones extracted from restored 3D facial model, which may lead to a better performance.

In this paper, we just test our method on the AVEC 2012 dataset. It is because that the AVEC 2012 is typical, popular and widely accepted by researchers of continuous emotion recognition. There are still some other great datasets that can be used to test our method. In the feature, we will test our algorithm on more datasets.

Furthermore, researchers have pointed out that emotion dimensions are correlated with each other [14]. In the future work, we will focus on exploiting and modelling the relationships between different emotion dimensions for a better emotion analysis.

Acknowledgements

The authors gratefully acknowledge SSPNET for providing the AVEC 2012 dataset and Cao Chen, et al. for providing the FaceWarehouse dataset. We also thanks the source code about facial landmarks labeling provided by T. Baltrusaitis et al. This research was supported by by the National Natural Science Foundation of China: 61135003, 61232013, 61173059 and 973 Project: 2013CB329305.

References

[1] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon. The painful facecpain

expression recognition using active appearance models. *Image and Vision Computing*, 27(12):1788–1796, 2009. 0, 1

[2] T. Baltrusaitis, N. Banda, and P. Robinson. Dimensional affect recognition using continuous conditional random fields. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–8, 2013. 1, 6

[3] T. Baltrusaitis, P. Robinson, and L. P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *IEEE International Conference on Computer Vision Workshops*, pages 354–361, 2013. 2, 3, 6

[4] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 1

[5] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics*, 32(4):41, 2013. 2, 6

[6] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: a 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. 2

[7] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014. 6

[8] A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2–3):81–227, 2012. 1

[9] P. Ekman. An argument for basic emotions. cognition and emotion. *IEEE Transactions on Cybernetics*, 6(3–4):169–200, 1992. 0

[10] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. V. Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101(3):437–458, 2013. 1

[11] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259–275, 2003. 1, 5

[12] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3):143–166, 2003. 0

[13] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007. 1

[14] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, 1(1):68–99, 2010. 7

[15] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136, 2013. 0

[16] H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: a survey. In *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops*, pages 827–834, 2011. 1

[17] E. Hudlicka. To feel or not to feel: The role of affect in humanccomputer interaction. *International Journal of Human-Computer Studies*, 59(1):1–32, 2003. 0

[18] A. Kapoor, W. Burleson, and R. W. Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8):724–736, 2007. 0, 1

[19] H. Kobayashi and F. Hara. Facial interaction between animated 3d face robot and human beings. In *Proceedings of the International Conference on Systems, Man and Cybernetics*, pages 3732–3737, 1997. 1

[20] B. Kort and R. Reilly. Analytical models of emotions, learning and relationships: towards an affect-sensitive cognitive machine. In *Conference on Virtual Worlds and Simulation*, 2002. 0

[21] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate O(n) solution to the PNP problem. *International Journal of Computer Vision*, 81(2):155–166, 2009. 3

[22] G. C. Littlewort, M. S. Bartlett, and K. Lee. Faces of pain: automated measurement of spontaneous all facial expressions of genuine and posed pain. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, pages 15–21, 2007. 0

[23] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2011. 1

[24] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1):5–17, 2012. 6

[25] M. A. Nicolaou, H. Gunes, and M. Pantic. Output-associative rvm regression for dimensional and continuous emotion prediction. *Image and Vision Computing*, 30(3):186–196, 2012. 1

[26] J. Nicolle, V. Rapp, K. Bailly, and et al. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pages 501C–508, 2012. 6

[27] C. Peter and R. Beale. Affect and emotion in human-computer interaction: From theory to applications. *Lecture Notes in Computer Science*, 4868, 2008. 0

[28] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert. Recognition of 3d facial expression dynamics. *Image and Vision Computing*, 30(10):762 – 773, 2012. 1

[29] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683 – 697, 2012. 1

[30] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pages 449–456, 2012. 1, 6

[31] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang. Authentic facial expression analysis. *Image and Vision Computing*, 25(12):1856–1863, 2007. 1

[32] C. Soladié, H. Salam, C. Pelachaud, N. Stoiber, and R. Séguier. A multimodal fuzzy inference system using a

continuous facial expression representation for emotion detection. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pages 493–500, 2012. 1, 6

[33] P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394, 1994. 1

[34] M. F. Valstar and M. Pantic. Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. *HumanCComputer Interaction*, pages 118–127, 2007. 1

[35] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. *ACM Transactions on Graphics*, 24(3):426–433, 2005. 6

[36] T. Wu, M. S. Bartlett, and J. R. Movellan. Facial expression recognition using gabor motion energy filters. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 42–47, 2010. 1

[37] S. Yang and B. Bhanu. Facial expression recognition using emotion avatar image. In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, pages 866–871, 2011. 1

[38] S. Yang and B. Bhanu. Understanding discrete facial expressions in video using an emotion avatar image. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(4):980–992, 2012. 1

[39] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang. Audiocvisual affective expression recognition through multistream fused hmm. *IEEE Transactions on Multimedia*, 10(4):570–577, 2008. 1