The Visual Computer
https://doi.org/10.1007/s00371-018-1482-1

ORIGINAL ARTICLE

Emotion information visualization through learning of 3D morphable face model

Hai Jin[1] · Xun Wang[1] · Yuanfeng Lian[1] · Jing Hua[1]

Abstract
Analysis and visualization of human facial expressions and its applications are useful but challenging. This paper presents a novel approach to analyze the facial expressions from images through learning of a 3D morphable face model and a quantitative information visualization scheme for exploring this type of visual data. More specifically, a 3D face database with various facial expressions is employed to build a nonnegative matrix factorization (NMF) part-based morphable 3D face model. From an input image, a 3D face with expression can be reconstructed iteratively by using the NMF morphable 3D face model as a priori knowledge, from which basis parameters and a displacement map are extracted as features for facial emotion analysis and visualization. Based upon the features, two support vector regressions are trained to determine the fuzzy valence–arousal (VA) values to quantify the emotions. The continuously changing emotion status can be intuitively analyzed by visualizing the VA values in VA space. Our emotion analysis and visualization system, based on 3D NMF morphable face model, detect expressions robustly from various head poses, face sizes and lighting conditions and is fully automatic to compute the VA values from images or a sequence of video with various facial expressions. To evaluate our novel method, we test our system on publicly available databases and evaluate the emotion analysis and visualization results. We also apply our method to quantifying emotion changes during motivational interviews. These experiments and applications demonstrate the effectiveness and accuracy of our method.

Keywords 3D morphable face model · Facial expression analysis · Emotion visualization

1 Introduction

Understanding emotion status has always been an interesting yet challenging research topic in the past decades [39]. More recently, with the development of more human-centered services, such as targeted advertisement, trending analysis and self-emotion tracking, automatic emotion detection has become increasingly important. Various types of data sources, e.g., images of facial expressions, speech, electroencephalogram (EEG), electrocardiogram (ECG), can be utilized to analyze the emotions [25]. However, speech, EEG and ECG do not always present in daily routine occasions. Thus, using only images or videos for emotion detection is the most feasible and reliable way in many cases. Additionally, it has been proven that the most effective and natural means for classifying emotions are based upon the facial expressions [2,19]. Therefore, there is growing interest in methods of extracting the underlying information of facial images to analyze emotional states [46]. Information visualization of this type of visual data is also becoming more important lately.

Many research efforts have been made to explore the facial information through 2D facial images, including face recognition [43], age detection [20] and facial expression recognition [3]. Facial feature extraction from 2D images has been intensively studied and is proven effective. 2D-based methods such as active appearance model (AAM) [34], active shape model (ASM) [2] and constrained local model (CLM) [16] are successfully applied to face tracking and face recognition applications. More recently, convolutional neural network (CNN) became more popular in the field of computer vision and was actively applied to object detection and recognition tasks [30]. However, 2D-based methods suffer from their fundamental challenges: illumination and orientation variance may result in very different images for

✉ Jing Hua
jinghua@wayne.edu

[1] Wayne State University, Detroit, MI, USA

Published online: 20 February 2018

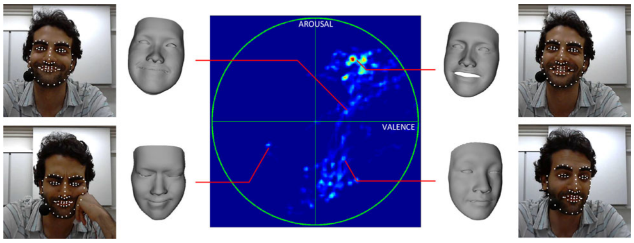

Fig. 1 Interactive 3D emotion query and visualization in valence–arousal (VA) space. The images are the input frames from videos with extracted feature landmarks (as shown in white points). The corresponding 3D faces are reconstructed 3D models

the same individual, which make the classification inaccurate and unstable.

To solve the problems existing in 2D-based methods, a potential solution is to extract features from 3D space directly, including 3D mesh and point set surfaces [23,24]. Directly using 3D scanners, Cohen et al. [13] proposed a 3D data-based facial expression recognition for human–computer interactions (HCI). Blanz et al. [7] presented a 3D face recognition approach based on principal component analysis (PCA) decompositions of a 3D face database. Recently, with the development of 3D capturing devices such as RGBD cameras, acquiring depth information of the face has become much easier than ever. Newcombe et al. [36] presented a 3D object reconstruction method which can be used for face modeling. And Chen et al. [12] showed a facial performance capturing method using RGBD camera. These methods heavily rely on the quality of the 3D data, and the accuracy of analysis also greatly depends on the registration outcome of the testing face to the reference face. In addition, 3D data acquisition is still inconvenient as compared to 2D images or videos.

To combine the advantages from both 2D and 3D approaches, we utilize our 3D face fitting method to construct a 3D face from an input image, which generates a dense correspondence to a reference 3D face. Using a fitted 3D face will not only provide a well registered 3D mesh surface, but also can decompose it into an uniform basis space to obtain normalized features. The generated 3D face can be represented by the weighted sum of the basis functions, and the weight vectors can be used as one of the features to classify the expressions, which can be further translated to emotional status. Our system allows the users to analyze emotions continuously by quantifying and visualizing the detected emotion in valence–arousal space. Figure 1 illustrates our interactive 3D emotion query and visualization system.

1.1 Related work

Facial expression is the most direct reflection of emotion and shares common meanings across different races and cultures. According to Ekman and Friesen's (1975) study of human facial expressions, there are so-called universal facial expressions which represent those common emotions of people: happiness, anger, sadness, fear, surprise and disgust. This study justified the emotion recognition through facial expressions. Ekman and Friensen proposed a facial action coding system (FACS) to describe facial expressions in a standard measurement, which is widely used in image-based emotion classification methods [17].

Extracting 2D features, such as displacement of feature points and intensity change from images, for emotions estimation is the most popular method. For example, Kwang-Eun Ko et al. [28] used Active Shape Model (ASM) to extract facial geometry features to classify emotions. Kobayashi et al. [29] and Valstar et al. [45] used the 20 feature points on the face for emotion classification. Liao et al. [33] presented a method for enhancing the geometry of 3D face meshes based on these 2D feature points. Some work also used intensities around the feature points to enhance the features for predicting emotions. For example, Kapoor et al. [27] used pixel intensity difference to classify the emotions of human subjects. However, this method is heavily dependent on the image quality. There are also some hybrid methods that combine the geometry features and the pixel intensity features to estimate the emotions. Developed from ASM, active appearance model (AAM) [14] fits the facial image with not only geometric feature points, but also the pixel intensities. Lucey et al. [34] showed the capability of AAM-based emotion detection method using Cohn–Kanade dataset. Although 2D features are easy to extract from the images directly, they are unstable under the change in illumination or face pose as shown in Sandbach et al. [38]. Therefore, 3D features like curvature, volume and displacement are used in many 3D-based approaches. These 3D features are more stable and robust than 2D features since they are pose and illumination invariant in nature. Huang et al. [40] extracted the Bézier volume change as the features of the emotions, and Fanelli et al. [18] used the depth information of pixels to classify emotions.

More recently, 3D face modeling from images has made a significant progress, which provides new ideas for emotion analysis [4–6]. Lei et al. [32] presented a face shape recovery method using a single reference 3D face model. Prior knowledge of face is required for these methods to recover the missing depth information from a 2D image. Learning through a face database is another effective approach for tackling this problem. Along this direction, statistical face models based on principal component analysis (PCA) were proposed and constructed [6]. Similar face fitting methods, such as piecewise PCA sub-models [42], were proposed later. Since these approaches based on 2D images achieved plausible 3D face reconstruction results, Blanz et al. applied the 3D face reconstruction method to facial recognition problem [7]. However, these methods are still limited to neutral expression and produce poor reconstruction results on faces with expressions.

Therefore, many studies have been devoted to represent 3D faces with arbitrary expressions. Based on a collection of images or a clip of videos of a person, Suwajanakorn et al. [41] presented novel 3D face reconstruction technique. A base shape was learned using a template 3D face for the subject, which was then used to fit the images of the same subject with various expressions. They used a shape-from-shading method to fine-tune the details of the shape. The high computational cost is the main drawback, which makes this method not realistic for emotion tracking and analysis. Another popular approach for generating 3D face with an expression at high speed and low cost is blendshape [26], which has been proposed for fast and robust 3D facial performance animation in industry-level applications. Cao et al. [9] proposed a system to animate an avatar face using a single camera and blendshape. Their work focused on tracking a user's facial expressions with a 2D camera and then synthesizing the corresponding expression geometry in an avatar [8,11]. But reconstructing 3D models from images for emotion analysis is rarely studied and its capability in capturing discriminative features is under-explored.

On the other hand, emotion is rarely visualized in an intuitive way. Visualizing emotion information and status is another interesting topic in emotion analysis. The most general measurement of emotion is in valence–arousal space (VA space) [37]. Generally, valence value indicates the pleasantness in emotions from negative to positive, while arousal value evaluates the intensity of the emotions: from calm, peaceful to alert and exciting. The universal expressions can be translated into this two-dimensional plotting system, which is clear and intuitive to users perception.

In this paper, we present a novel robust approach that measures and visualizes the emotion status continuously in VA space. We use a 3D facial expression database to build a 3D part-based morphable face model which can be used to reconstruct 3D faces from input facial images. Then, we decompose the reconstructed 3D face to obtain its coefficient vector as well as displacement map for emotion quantification. Finally, we demonstrate the continuous emotion change by visualizing the emotion measurement in VA space.

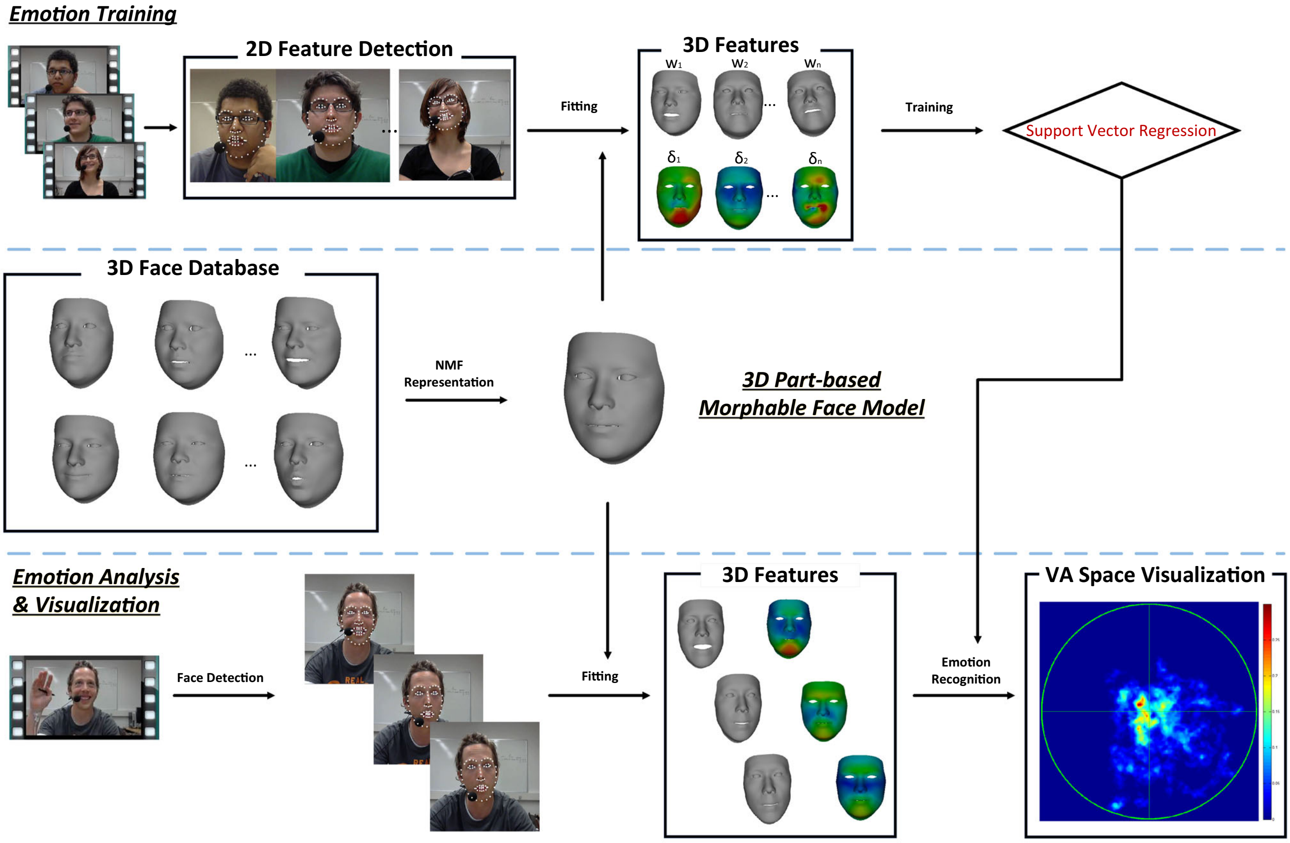

Fig. 2 Pipeline of our emotion analysis and visualization framework with learning of 3D morphable face model. Based on a 3D face database, a 3D part-based morphable face model is built. Using the 3D morphable face model as prior, the 3D face is reconstructed for each input video frame and the coefficient vectors and the displacement maps are obtained as features. The features are used to train a support vector regression (SVR) in the emotion training phase. Based on the trained SVR and the 3D morphable model, the emotion VA values are estimated and visualized in the online emotion analysis phase

Figure 2 illustrates the entire process, which is fully automated without users' interventions.

The contributions of this paper can be summarized as follows:

- We propose an emotion analysis method based on a novel part-based 3D morphable model, which is robust to variations of face poses, illuminations and sizes.
- Our emotion analysis method based on 3D morphable face model can extract more sensitive and reliable features from reconstructed 3D face to classify different emotions. It presents a fully automated 3D face reconstruction technique for 3D facial expression decomposition and feature extraction.
- We provide a robust VA value computation and visualization method to measure the emotion changes continuously in VA space. Our system enables users to monitor and analyze emotions robustly and intuitively from a single camera.

The rest of the paper is organized as follows: in Sect. 2, we first introduce the 3D face database which is used to create our NMF part-based 3D face model. Then, we describe the process to build our part-based 3D morphable face representation based on nonnegative matrix factorization of a 3D face database. In Sect. 2.3, we demonstrate how the part-based 3D face model can be used as a morphable model to reconstruct a high-quality face with an expression from a single image. In Sect. 4, we present the training process of expression classification with support vector regression (SVR, and show the details of testing new images for emotion analysis using our method. In Sect. 5, we explain our visualization method for the continuous emotion measurement. Finally, we demonstrate our experimental results and an example application in Sect. 6.

2 Facial expression reconstruction through morphable 3D face model

In this section, we present our 3D morphable face model for reconstructing 3D facial expression models which will be used for feature extraction and emotion analysis.

2.1 NMF part-based 3D face model

In our 3D expression analysis method, a 3D face model is first constructed from an input image. To accurately fit 3D face model to an input facial image, a 3D face database with various expressions is needed to generate a 3D morphable face model for reconstruction. In this paper, we adopt FaceWarehouse [10] as the 3D face database to build the 3D morphable face model. FaceWarehouse is a 3D facial dataset containing 3D facial models of 150 subjects from various ethnic backgrounds and every subject has 47 FACS blendshapes with 11,000 vertices. The original data is provided in a tensor form: $V = C_r \times w_{id}^T \times w_{\exp}^T$, where C_r is a bilinear face model, w_{id}^T is the column vector of identity weights and $w_{\exp}^T$ is the expression weights.

As Cao et al. presented in [10], a 3D face model is iteratively reconstructed from a 2D input image in two steps: First, they use a neutral expression image to optimize the identity weight w_{id}^T; secondly, they optimize the expression weight $w_{\exp}^T$ for the images with any new expressions. Although the bilinear face model method works well for their person-specific performance animation, the generated 3D face is constrained to the predefined blendshapes, which is not preferred in emotion analysis. Also, 3D models in FaceWarehouse contain the entire heads including ear and neck, which has redundant information for our expression analysis purpose. Therefore, we only retain the face part. To represent a 3D face model, we decompose the dataset using the part-based nonnegative matrix factorization (NMF) in order to model subtle shape displacements. This method provides rich local details for emotion analysis.

Figure 3 shows a subset of 3D faces that we use. We reorganize the database as a matrix $S = 3N \times M$, where N is the number of vertices in a 3D face and M is the number of face examples in the database. Using matrix S, we reconstruct a 3D morphable face model as follows:

$$S = HW, \tag{1}$$

where H is the basis matrix and W is the associated weights. Each face in the dataset can be restored by its weighted sum of basis function H. The model is used for reconstruction of any

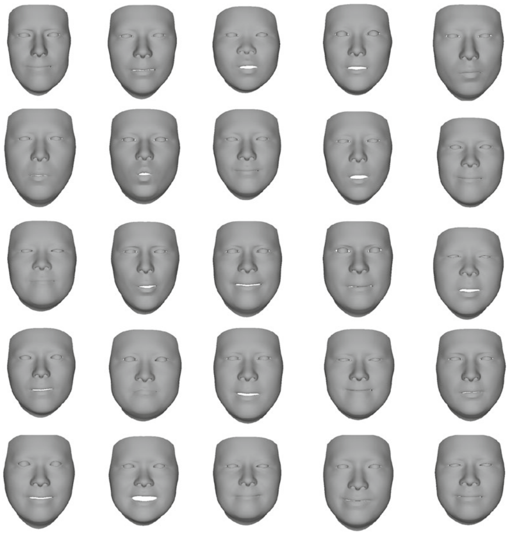

Fig. 3 Randomly selected 25 sample 3D faces in our database

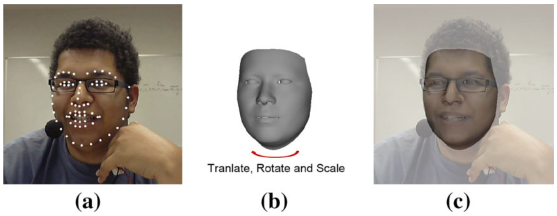

Fig. 4 Feature point detection and initial alignment. **a** Detected features points and the estimated pose according to the input image. **b** Affine transformed template face based on the estimated parameters according to (**a**). **c** Aligned template face to the input face

arbitrary face from the input image iteratively, and we will explain the optimization details for the model in Sect. 2.3.

2.2 Face tracking and alignment

The initial alignment is essential for both the training and testing stages of emotion recognition since the face pose parameters must be estimated before the reconstruction process to improve the fitting quality and reduce the optimization iterations. We use the constrained local model (CLM) [16] to find the 2D landmarks P_n (n is the number of points), with which the initial template pose parameters Ω_i, translation (T), rotation (R) and scale (S) are estimated automatically. Figure 4 illustrates the detected landmarks P_n on the input facial image and the posing direction of the face.

Many studies have shown how to estimate head poses from single facial images [15,35], and we use a similar method as in [15] for our work. The relation between the landmark layout and the head pose is learned using the ground-truth data scanned via $Kinect$ available in FaceWarehouse. A set of training data $F_i = \{P_{ni}, \Omega_i\}$ is obtained for each key frame I_i, where P_{ni} is the facial landmarks detected by CLM and Ω_i is the affine transformation parameters including translation, rotation and scale. Since the translation T and scale S can be directly obtained by the face detection algorithm, we only focus on learning the relation between the landmark layout and the rotation R. We subdivide the learning into 3 parts because the rotation can be decomposed into three angles θ_x, θ_y and θ_z around x, y and z axis, respectively. Thus, we train three independent regressions with respect to θ_x, θ_y and θ_z.

The layout of the facial landmarks is represented as

$$P = \bar{P} + Q\beta, \tag{2}$$

where $\bar{P}$ is the average layout of facial landmarks among all the faces in the training database, Q is the PCA basis matrix, and β is the coefficient which controls shape variation of the feature points. The prediction model can be established via the following regression model:

$$\beta = \mathbf{a} + \mathbf{b}\cos(\theta) + \mathbf{c}\sin(\theta), \tag{3}$$

where $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are the parameters to be trained from the training set F. Equation 3 can be solved by $(\cos(\theta), \sin(\theta))^T = R^{-1}(\beta - \mathbf{a})$, where R^{-1} is pseudo inverse of $(\mathbf{b}|\mathbf{c})$, i.e., $R^{-1}(\mathbf{b}|\mathbf{c}) = I_2$. Hence, we first compute the coefficient vector β using Eq. 2 from the detected facial landmark layout P; then, we compute θ based on the trained predictive model, i.e., Eq. 3.

To obtain the corresponding vertices V_n on the template 3D face, we detect the 2D pixel landmarks on the rendered 3D face image using CLM. Since we know the projection correspondence between those landmark pixels and the 3D vertices, the 2D landmarks can be traced back to 3D face model. Therefore, the one-to-one correspondence can be found for the 3D vertex coordinate vector V_n and the 2D facial landmarks P_n. Once we aligned the 3D template face to the input 2D image as an initialization, we can start to carry out the reconstruction process.

2.3 3D face reconstruction based on part-based 3D morphable face model

In this section, we explain our part-based 3D morphable face model reconstruction method in detail.

Building part-based morphable model Different from the widely used decomposition method such as PCA and vector quantization (VQ) on 3D database, we present a method with nonnegative matrix factorization (NMF) in this paper. PCA learns the face data globally and decomposes it into 'eigenfaces', whereas NMF decomposes it in localized shapes. Utilizing this advantage, we construct a 3D morphable face representation by parts based on NMF to improve the reconstruction quality of the 3D face. It is possible to represent any face in the form of a linear combination of the basis parts, and since the part basis supports better local control, it produces more accurate and robust fitting result on the target facial image.

To build a 3D morphable face model from pre-scanned database, a dense registration across all the scans is essential and the quality of the dense correspondence will affect the factorization result significantly [21]. Since FaceWarehouse is a 3D face database with dense correspondence, we skip this step in our system. And because we only need the faces for our expression analysis task, we only use the frontal subset vertices from the entire head data. Based on m face samples in the database, we construct a $3n \times m$ matrix S, where n is the number of vertices in a 3D face. Each column represents the geometry of one face sample with 3D coordinates in vector form: $s_i = \{x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_n, y_n, z_n\}^T \in \mathbf{R}^{3n}$. Then, the data matrix S is decomposed by nonnegative factorization as $S \approx HW$ and the basis matrix H and the weight W are obtained. The decomposition can be represented as follows:

$$S_{ij} \approx (HW)_{ij} = \sum_{k=1}^{r} H_{ik} W_{kj}, \tag{4}$$

where r is the rank of factorized basis. Using the basis matrix and the weight $\mathbf{w}_n$, we can restore the corresponding 3D data sample by

$$s_n = H\mathbf{w}_n, \tag{5}$$

where $\mathbf{w}_n = (w_1, w_2, \ldots, w_i)^T$ is the nth column in the weight matrix W. To generate a new 3D face, we can manipulate the weight vector and compute the linear combination of the basis face.

We solve the following optimization problem to find a nonnegative factorization,

$$(H, W) = \underset{H \geq 0, W \geq 0}{\operatorname{argmin}} \|S - HW\|^2. \tag{6}$$

According to the theorem in [31], the Euclidean distance $\|S - HW\|$ does not increase under the following update rules:

$$W_{kj} \leftarrow W_{kj} \frac{(H^T S)_{kj}}{(H^T HW)_{ik}}, \qquad H_{ik} \leftarrow H_{ik} \frac{(SW^T)_{ik}}{(HWW^T)_{ik}}. \tag{7}$$

We initialize H and W with random dense matrices and use a simple additive update rule for weight W as follows,

$$W_{kj} \leftarrow W_{kj} + \xi_{kj}[(H^T S)_{kj} - (H^T HW)_{kj}], \tag{8}$$

where

$$\xi_{kj} = \frac{W_{kj}}{(H^T HW)_{kj}}. \tag{9}$$

Based on the NMF decomposition of face database, any new face S_{new} can be parameterized based on the basis matrix H and represented by a weight vector $\mathbf{w}_{\text{new}}$, which means that the contribution of each basis face to the entire 3D face model can be controlled by varying the weight values. Therefore, we name $s = H\mathbf{w}$ as a part-based 3D morphable face representation which carries the prior knowledge of faces for nonrigid fitting.

In this work, we use $m = 150$ scanned face data with $n = 6000$ vertices in each sample for training the part-based 3D morphable face model. To meet the NMF requirements for nonnegative elements in the data matrix, we transform the data into cylindrical coordinate system to make all the elements in the matrix positive. As shown by the empirical result, choosing $k = 100$ in the NMF decomposition process can generate good approximation of the original database. Figure 3 shows some samples of our 3D face database used for training in this paper. Note that, all the faces used in training the part-based 3D morphable face model are not employed for testing the performance of our system.

Iterative 3D face reconstruction We fit the 3D morphable face model to an input 2D image by iteratively optimizing the weight vector $\mathbf{w}$ based on the decomposed basis matrix H. Using the previously estimated initial parameters T, R, S (see Sect. 2.2), we first align the template 3D face model to the input 2D image by translation, rotation and scale. To find the best fitting to the 2D image, we minimize the following least square energy,

$$\mathbf{w} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{k=1}^{m} \|P_n^k - \mathbb{P}^k(\mathbb{F}(RH\mathbf{w} + T))\|^2, \tag{10}$$

where H is the basis of the part-based 3D face representation, R is a rotation and scaling matrix, T is a translation matrix, m is the number of images used for fitting, $\mathbb{F}$ is landmark extraction operation (in Sect. 2.2), $\mathbb{P}$ is the projection operation, and k is the index of a perspective view. In general, our fitting algorithm supports 3D reconstruction from multiple views, and using more views can improve the reconstruction quality and also increase the computational cost. For our video-based emotion estimation task, only one view is available for each frame, therefore, $m = 1$.

Based on gradient descent method, we compute the partial derivative of the energy term in Eq. 10 with respect to the weight vector $\mathbf{w}$ as follows,

$$\begin{aligned} \triangledown E(\mathbf{w}) &= \frac{\partial E}{\partial \mathbf{w}} \\ &= -2 \sum_{k=1}^{m} (P_n^k - \mathbb{P}^k(\mathbb{F}(RH\mathbf{w}+T))) \frac{\partial \mathbb{P}^k(\mathbb{F}(RH\mathbf{w}+T))}{\partial \mathbf{w}}, \end{aligned} \tag{11}$$

where

$$\frac{\partial \mathbb{P}(\mathbb{F}(RH\mathbf{w} + T))}{\partial \mathbf{w}} = \mathbf{c} \tag{12}$$

is a constant vector $\mathbf{c}$ for each image, which means the update step is only determined by the distance between projected feature vertices in the current step and the target landmarks in the input image. Therefore, in each iteration i, the weight vector $\mathbf{w}$ is updated by

$$\mathbf{w} \leftarrow \mathbf{w} - \triangledown E(\mathbf{w}_i). \tag{13}$$

Algorithm 1 shows the iterative updating process of the weight vector $\mathbf{w}$. This process takes around 10 iterations to converge. Figure 5 shows the iterative 3D face reconstruction results for 3 different frames. Once the 3D face is reconstructed from the input image along with the weight vector

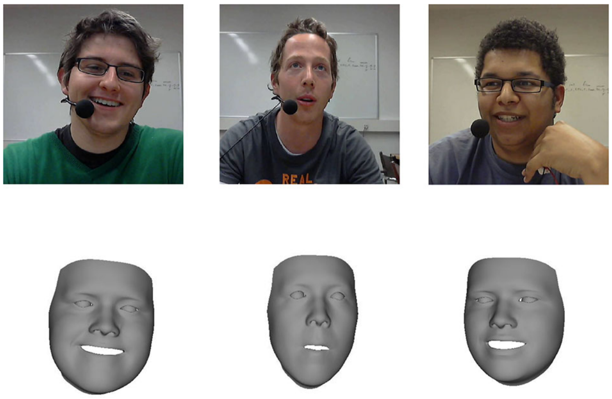

Fig. 5 Iterative 3D face reconstruction results for 3 significant frames

$\mathbf{w}$, we can use it as a part of the feature vector $\boldsymbol{f}$ for expression training in Sect. 4.

Algorithm 1 Iterative Reconstruction

```
procedure ITERATIVE 3D FACE RECONSTRUCTION
    $\mathbf{w}, T, R, \mathbb{P}, \mathbb{F}, threshold \leftarrow initialization$
    Compute the gradient $\triangledown E(\mathbf{w}_i)$ using Eq. 11
    while $\triangledown E(\mathbf{w}_i) > threshold$ do
        $\mathbf{w} \leftarrow \mathbf{w} - \triangledown E(\mathbf{w}_i)$
        Re-compute $\triangledown E(\mathbf{w}_i)$ using Eq. 11
    end while
    $S = H\mathbf{w}$
end procedure
```

3 Feature extraction from 3D NFM face model

After we reconstruct the 3D face model from the input monocular image, we are ready to extract the 3D features from it. As our reconstructed 3D face model S_n is represented by $H\mathbf{w}_n$, where the corresponding weight vector $\mathbf{w}_n$ carries the essential information of the shape, it can be used as a part of the feature vector. Since we use part-based decomposition method to model the 3D face, the weight vector $\mathbf{w}_n$ contains localized feature coding information. Another advantage of using the weight vector is that the 3D model decomposition shares the same basis H; thus, all fitted models are naturally normalized, which makes the features robust for classification. In this paper, we take the first 200 dimensions of the basis to represent the 3D face, which means the dimension of weight vector $\mathbf{w}_n$ is 200. Along with the weight vector, we also combine the displacement of vertices to the feature vector. Note that, as we use the NMF part-based model to keep reconstructing all the frames, the reconstructed model has point-to-point correspondence. For each subject, we compute the spatial displacement as $\boldsymbol{\delta}_n = V_n - V_0$, where V_n is the reconstructed shape and V_0 is the neutral expression of the same subject which is manually selected as reference. We down sample $\boldsymbol{\delta}_n$ to 300 dimensions to reduce the feature dimension in our experiment. We combine the weight vector $\mathbf{w}_n$ and the displacement vector $\boldsymbol{\delta}_n$ as the final feature vector $\boldsymbol{f}_n$.

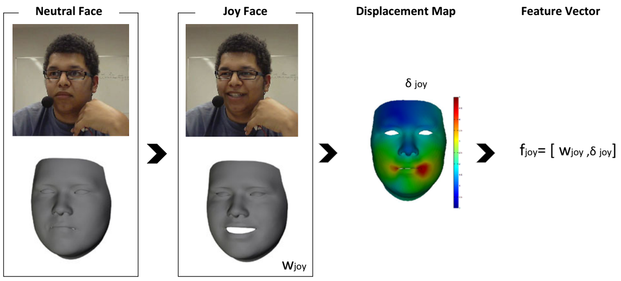

Fig. 6 Illustration of feature vector construction for a joy face: the spatial displacement $\boldsymbol{\delta}_{\text{joy}}$ is computed between the joy face and neutral face, so the feature vector $\boldsymbol{f}_{\text{joy}}$ is composed by shape coefficient $\mathbf{w}_{\text{joy}}$ and displacement map $\boldsymbol{\delta}_{\text{joy}}$

Figure 6 shows an example for constructing the feature vector $\boldsymbol{f}_{\text{joy}}$ for the joy expression of a subject, which is composed by the weight vector $\mathbf{w}_{\text{joy}}$ and the displacement map $\boldsymbol{\delta}_{\text{joy}}$. For the visualization purpose, we only show the absolute distance in the color map for the displacement map. Note that, in our algorithm, the displacement is a three-dimensional vector containing x, y, z components and stored in a vector form. We use the constructed feature vectors of each video frame to train the support vector regression as described in the next section.

4 Emotion analysis using SVR

We adopt a standard support vector regression as proposed by Valstar et al. [44] to establish our 3D NMF morphable model-based emotion analysis method. Our emotion analysis method includes a training step for valence value regression and arousal value regression, and then a runtime quantification step for estimating emotion VA values.

Training data preparation Our training algorithm uses the feature vectors constructed from the generated 3D shapes, which is reconstructed from the video frames of the public database. For each video frame I, we obtain a feature vector $\boldsymbol{f}_i = [\mathbf{w}_i, \boldsymbol{\delta}_i]$ along the manually labeled valence–arousal values V_i and A_i to form a training tuple $(\boldsymbol{f}_i, V_i, A_i)$. We use these training tuples to train two $SVRs$ for valence and arousal, respectively, namely SVR_V, SVR_A.

To improve the training robustness, we generate some randomness (±5%) to the 3D face reconstruction process to obtain more training datasets: (1) Add a random rotation ΔR to the initial face alignment. (2) Add a random translation ΔT to the initial face alignment. (3) Add a random scaling ΔS to the initial face alignment. Since the 3D reconstruction is sensitive to the initial alignment, by adding these random noise we can obtain more training datasets. In this work, we prepare approximately 100 frames for each dataset by sampling the video at 1fps. Then, we generate one random variation for each case at each video frame, so we have four reconstruction results for each frame and approximately 400 3D models for each dataset. We use 10 datasets with 4000 3D models in each one for SVR training. Feature vectors are extracted from the 3D models using our feature extraction method, as described in Sect. 3, and the training matrix $M_{\text{training}} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n\}^T$ is constructed.

Emotion quantification The online testing process takes the video frames as the input to our algorithm. We first employ the OpenCV implementation of face detection method with local binary pattern (LBP) [1] to detect the region of interest (ROI) for the human face. Within the ROI, the 68 facial landmarks are detected by CLM and the head pose is estimated using the method presented in Sect. 2.2. Then, we reconstruct the 3D face using our NMF part-based morphable 3D face model, from which we extract the feature vector $\boldsymbol{f} = [\mathbf{w}, \boldsymbol{\delta}]$. Feeding the feature vector $\boldsymbol{f}$ to SVR_V and SVR_A, we can obtain the estimated VA values. Finally, the VA values are visualized in the VA space for user analysis.

5 Interactive emotion visualization

Many emotion detection methods quantify emotions by giving probability scores for the six fundamental expressions. Recent psychological studies show that quadrants of emotion wheel [22] is more accurate and intuitive than using the six fundamental expressions. Figure 7 shows an example of the emotion wheel, which represents the VA space. All the common emotions including the fundamental ones can be plotted in the VA space with certain translation rules [25]. As shown in Fig. 7, the first quadrant represents the positive emotions, such as joy and surprise, while the third quadrant represents the negative emotions like sadness and disgust.

Our visualization is based on the VA space plotting to reveal the high-level information of the emotion status of the testing subjects from images or videos. We design two types of VA space plotting methods: emotion distribution plotting (EDP) and emotion trajectory plotting (ETP). EDP emphasizes the emotion distributions for the subject during the testing session, while ETP emphasizes the individual VA value. Figure 9 shows the EDP, and Fig. 10 shows the ETP for 4 testing subjects. The density of each coordinate, d_{xy}, on the EDP is computed as follows,

$$d_{xy} = \frac{\Omega_r(p_{xy})}{N}, \tag{14}$$

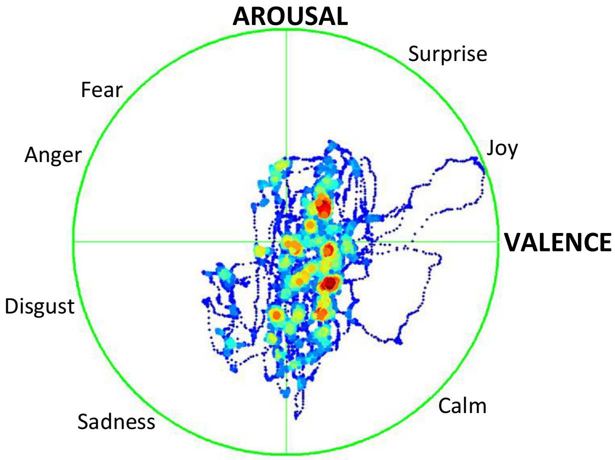

Fig. 7 Illustration of the VA space

where $\Omega_r(p_{xy})$ is the number of VA data points within the distance r from coordinate p_{xy} and N is the total number of VA data points. In this paper, we sample the VA space from -0.5 to 0.5 with a step size of 0.01 and interpolate the intermediate coordinates. Instead of computing the density at every coordinate in the VA space, ETP only computes the density at each data point. We compute the density at data point p_i by

$$d_i = \frac{\Omega_r(p_i)}{\Omega_r(p)_{\max}}, \tag{15}$$

where $\Omega_r(p_i)$ is the total number of data points within the vicinity of p_i with radius r and $\Omega_r(p)_{\max}$ is the largest number of data points across all such p_i's vicinities with the same radius. The density is converted to color map for final visualization.

To facilitate interactive visual information visualization of emotion data, our EDP and ETP visualization platform supports interactive user query. Details related to a VA value, including the original video frames I, reconstructed 3D faces S, weights w and the corresponding displacement maps $\boldsymbol{\delta}$, can be provided to user by clicking the data point in the plotting. Our system also supports data point comparison which allows users to compare two data points from the plotting by selecting the source and target points. Our system can also provide frames, 3D faces, weights and the displacement map between the two shapes.

6 Experiments

We have conducted extensive experiments using public datasets to evaluate the accuracy and effectiveness of our method for emotion analysis and visualization. We also demonstrate some exemplar applications using our system.

Our experiments have been performed on a regular PC with 3.0 GHz 8 Core CPU, 8 GB memory and GeForce 980 graphics card. The computation time of our method is mainly spent on the 3D face reconstruction process. The reconstruction time is approximately 3 s for each frame. We have implemented SVR for regression training and used the trained coefficients in our C++ implemented system. It takes about 5 s to generate the EDP and 3 for ETP with 5000 data points. The query for a data point costs about 3 s since we re-compute the 3D face model to reduce the memory cost.

6.1 Evaluation of the emotion analysis and visualization method

The audio/visual emotion challenge database (AVEC) [39] is a multimodal dataset for continuous emotion detection using audio and video sequences. In this paper, we have only used the videos in the latest AVEC 2015 dataset for the experiments. Each set of data in the database includes a 5 min 1280 × 720 resolution video with the frame rate of 24 fps, and the valence and arousal values are manually annotated for each frame. There are 9 datasets for training, 9 datasets for test and 9 datasets for development use. We have used the AVEC 2015 data for training and evaluated our emotion recognition algorithm on the test data as well as on our in-house recorded data. We have randomly selected 10 datasets including 5 training dataset and 5 development dataset for training purpose and tested on 4 subjects.

To evaluate our emotion analysis method, we first compare it with the expression recognition method using 2D features only. For a fair comparison, both methods use the same SVR. Figure 8 shows the comparison of our 3D model-based method to bilinear face model method and 2D landmark-based method proposed in [39]. To compare with the ground truth, we compute the mean square error of the estimated VA values as follows,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}((V_i - \hat{V}_i)^2 + (A_i - \hat{A}_i)^2), \tag{16}$$

where V_i, A_i are the estimated VA values and $\hat{V}_i$, $\hat{A}_i$ are the ground-truth VA values. Our 3D face model-based method achieves lower MSE than 2D landmark-based method [39].

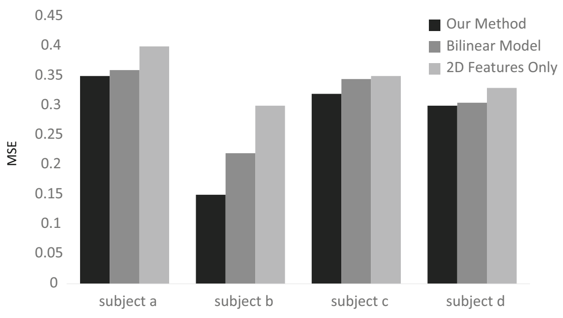

Fig. 8 Mean square errors of detected VA values compared with the ground truth: 1. our 3D feature-based method (NMF), 2. bilinear face model method and 3. the 2D landmark-based method [39]

6.2 VA space analysis and visualization

Based on the estimated VA values, we use our EDP and ETP for visualization. Figure 9 shows the EDP for 4 subjects from the development datasets in AVEC. The colormap shows the data density in certain location in VA space: red indicates high density, while blue indicates low density. The upper row shows the ground-truth VA values plotted in EDP, and the lower row is the estimated VA values using our method. As shown in the result, our method can achieve very similar distribution heatmap to the ground truth. From the EDP, users can understand the global trending of the subjects' emotions. For instance, the emotion data points of subject 1 are mostly located near the center and the region of 'calm,' which means the user is neutral in most of the time during the session. For subject 4, the data is scattered in the positive region, which shows the subject is joyful throughout the session.

For the same datasets, we show the EDP in Fig. 10, where the upper row is the ground-truth trajectory and the lower row is the emotion trajectory estimated using our method. The ETP focuses more on the trajectory of emotion changing so that users can understand the details of the emotion change. For example, for subject 1, there is a clear emotion path from neutral to joy, and then back to neutral, whereas for subject 3, there is an emotion change from neutral to sadness. In order to understand these details in emotion status change, we utilize the emotion status query system which allows users to visualize the underlying information on the data points. By querying the specific data points in the trajectory, the users may understand when the change happens and the subject's condition. For subject 3, we execute two queries for illustration. We first execute one query to check the status of an upper right data point as shown in Fig. 11. The query returns the detected facial feature points, reconstructed 3D face, weight vector and the displacement map for a joy emotion. Since the data point is far from the neutral state, the displacement map shows high intensity around mouth and eyes. We are also provided with a frame ID, with which we could locate the video and see what is the actual situation causing this emotion. Query 2 (Fig. 12) illustrates the emotion comparison of two data points by showing both frames side by side. We select two data points (in red and orange, respectively) to see the emotion differences. The red data point shows a near neutral emotion, while the orange data point shows a negative emotion. The orange frame has a very different head pose

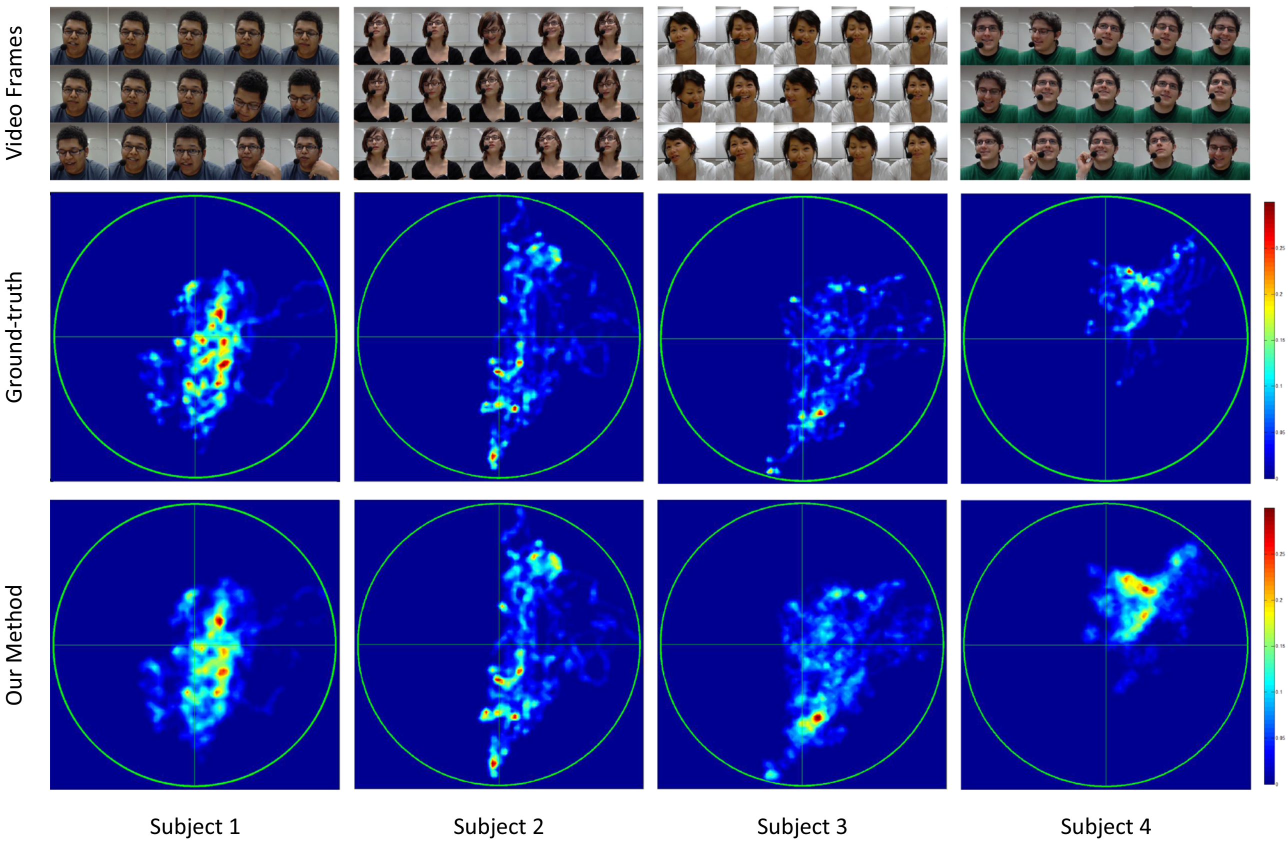

Fig. 9 Emotion distribution plot (EDP) of VA values for 4 subjects. The top row shows the video frames for subjects 1–4 (from left to right). The middle row shows the EDP of the ground-truth VA values, and the bottom row shows the EDP of the estimated VA values using our method

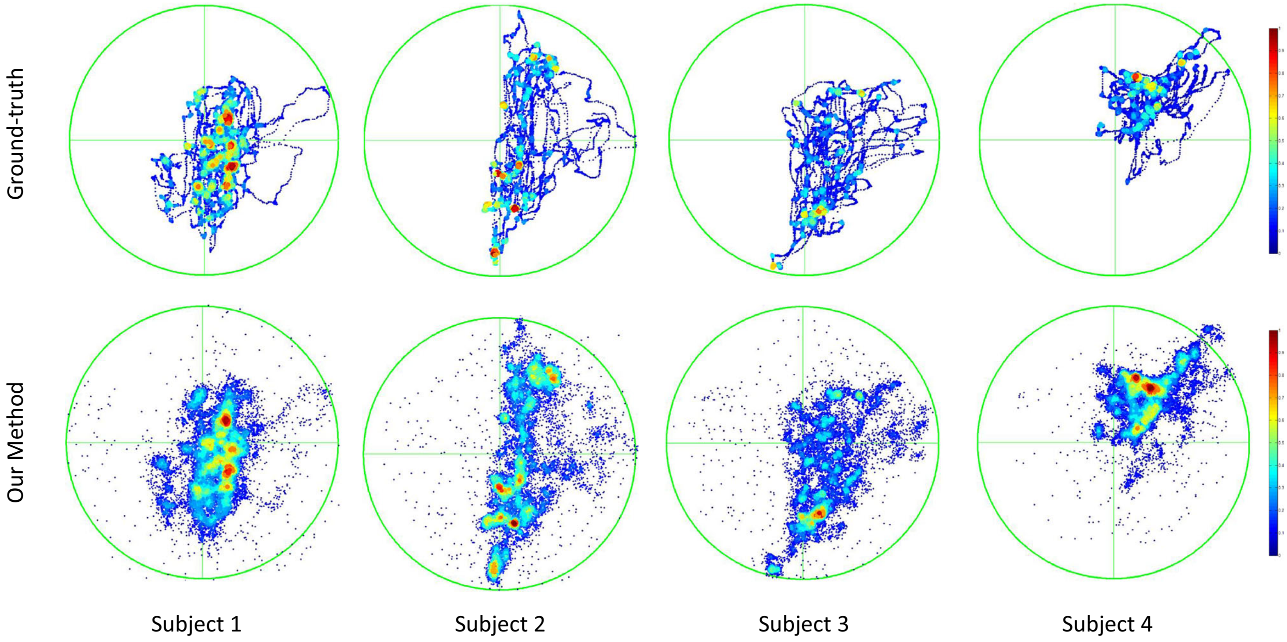

Fig. 10 Emotion trajectory plot (ETP) of VA values for 4 subjects. The top row shows the ETP of the ground-truth VA values for subjects 1–4 (from left to right), and the bottom row shows the ETP of the estimated VA values using our method

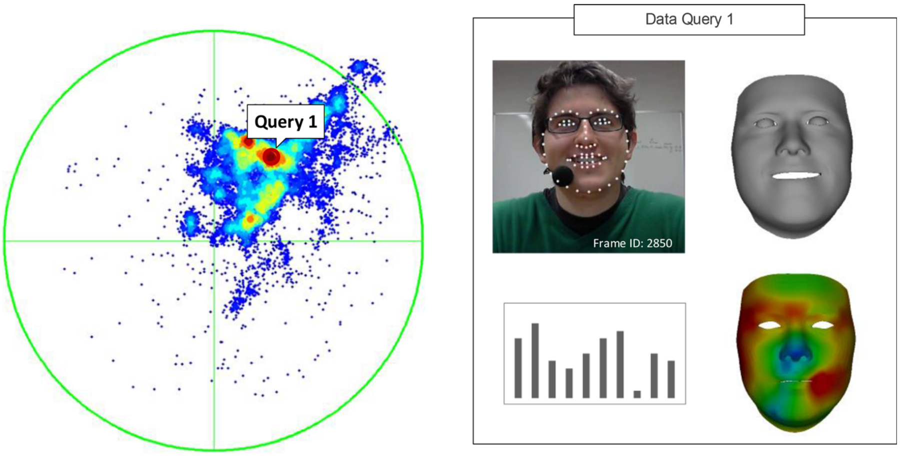

Fig. 11 Querying and visualizing a data point in VA space. The query shows the original frame with detected face landmarks, reconstructed 3D shape, weight vector and the displacement map of the data point. The displacement map shows a high intensity for a smiling face

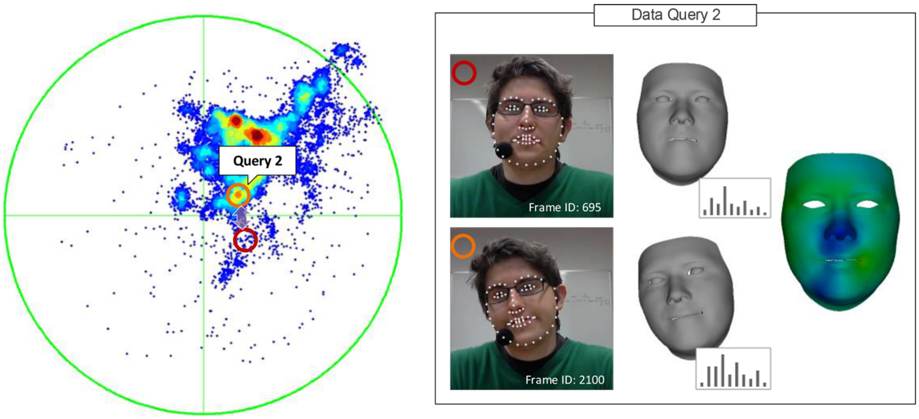

Fig. 12 Selection of two data points. The red data point shows a near neutral emotion, and the orange data point shows a positive emotion. The displacement map shows a small intensity as the two data points are close to each other despite the rotation of the head in orange frame

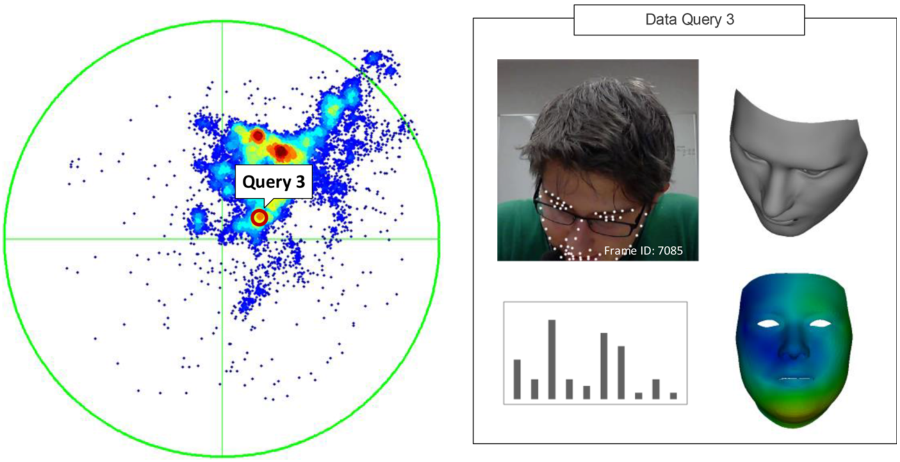

Fig. 13 A query example on a large head rotation data point. Our method provides correct emotion estimation with the extreme head rotation

compared to the red frame, which is difficult to compare the two images directly. Using our 3D morphable face model, we reconstruct two 3D faces by deforming the template face, which gives us a dense correspondence among the 3D faces. Using the dense correspondence of the two 3D face, we can compute the shape difference between two frames. As the two frames are very close in VA space, the displacement map shows small values. Figure 13 shows a query on a large rotation of the head. Our method can handle different poses effectively since our method is invariant to the head poses. Also, based on the query information, it is possible to correct the misclassified data manually by relabeling or re-extracting the 3D features to improve the accuracy.

6.3 Application to motivational interview

We have also applied our method to quantifying the effectiveness of motivational interviews. We have used the videos

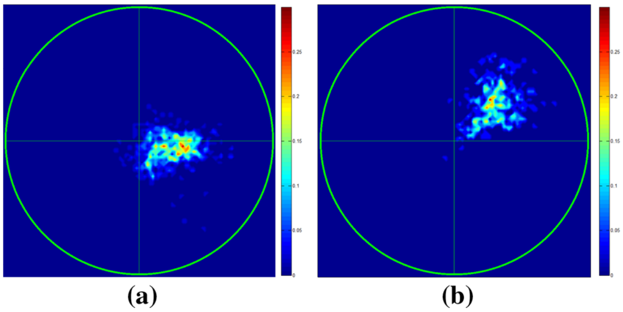

Fig. 14 EDP visualization of a subject's emotions during a motivational interview. **a** Beginning phase; **b** final phase

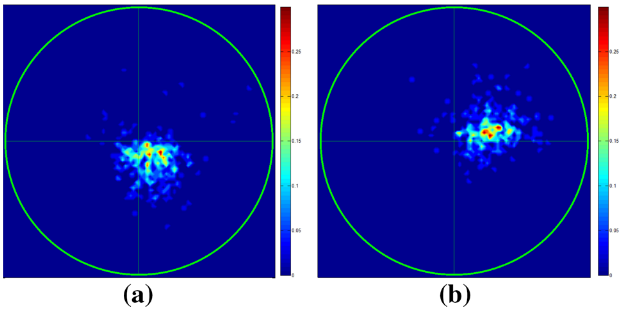

Fig. 15 EDP visualization of another subject's emotions during a motivational interview. **a** Beginning phase; **b** final phase

of patients who were interviewed by professional counselors. The beginning phase of the interview is compared with the final phase in order to quantitatively measure and visualize the effectiveness of this interview process. Figure 14 shows that the emotion of a subject mostly concentrates on calm and neutral status at the beginning phase (Fig. 14a) while exhibiting joyful status at the final phase (Fig. 14b), which indicates an effective therapy. The outcome can be quantitatively computed as the average VA value improvement. Figure 15 shows another case, where the emotion of the subject mostly stays in calm and neutral at the beginning (Fig. 15a) and is improved a little bit at the end (Fig. 15b). The outcome is not as good as the case in Fig. 14. Our method provides a viable tool for the doctors to quantitatively evaluate the patients' emotion status. After evaluating our tool on 20 cases and comparing the results with the professional scoring records, the software tool has been adopted by the Department of Psychiatry in the medical school of our university for quantitative assessment of emotion status.

7 Conclusion

In this paper, we have presented a 3D morphable face model-based approach for emotion analysis and information visualization in VA space. We have built a NMF part-based morphable 3D face model for reconstructing. Based on the input image, a 3D face with expression is reconstructed iteratively using the morphable 3D face model, from which basis parameters and a displacement map are extracted as features for emotion analysis. We have trained two support vector regressions for the fuzzy valence and arousal values, respectively, using the composed feature vectors. The states of the continuous emotions can be effectively visualized by plotting them in the VA space. Our method is fully automatic to compute the VA values from images or a sequence of video with various expressions. And our visualization system also provides the expression details such as the image frames and generated 3D faces interactively by interacting with the VA plot. The experiment results have shown that our method has achieved a remarkable emotion estimation accuracy, and our visualization method can provide a clear understanding of continuous emotion data.

We use a standard SVR as our emotion regression model in this paper, and there is a potential improvement by applying other regressions such as fuzzy neural network. The current 3D face reconstruction step in our system normally takes about 3 s, which does not allow our system to process real-time streaming data on the fly. In our future work, we will parallel our 3D model reconstruction algorithm so that we can run our system in real time for emotion monitoring.

Acknowledgements We would like to thank the reviewers for their valuable suggestions which helped to improve this paper. This work is supported in part by the following grants: LZ16F020002 and NSF CNS-1647200.

References

1. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12), 2037–2041 (2006)
2. Ashraf, A.B., Lucey, S., Cohn, J.F., Chen, T., Ambadar, Z., Prkachin, K.M., Solomon, P.E.: The painful face-pain expression recognition using active appearance models. Image Vis. Comput. **27**(12), 1788–1796 (2009)
3. Bartlett, M.S., Littlewort, G., Fasel, I., Movellan, J.R.: Real time face detection and facial expression recognition: development and applications to human computer interaction. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop **5**, 53–53 (2003)
4. Beeler, T., Bickel, B., Beardsley, P., Sumner, B., Gross, M.: High-quality single-shot capture of facial geometry. ACM Trans. Graph. **29**(4), 40 (2010)
5. Beeler, T., Hahn, F., Bradley, D., Bickel, B., Beardsley, P., Gotsman, C., Sumner, R.W., Gross, M.: High-quality passive facial performance capture using anchor frames. ACM Trans. Graph. **30**(4), 75 (2011)
6. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: SIGGRAPH, pp. 187–194 (1999)
7. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1063–1074 (2003)

8. Cao, C., Hou, Q., Zhou, K.: Displaced dynamic expression regression for real-time facial tracking and animation. ACM Trans. Graph. **33**(4), 43 (2014)
9. Cao, C., Weng, Y., Lin, S., Zhou, K.: 3D shape regression for real-time facial animation. ACM Trans. Graph. **32**(4), 41 (2013)
10. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: a 3D facial expression database for visual computing. IEEE Trans. Vis. Comput. Graph. **20**(3), 413–425 (2014)
11. Cao, C., Wu, H., Weng, Y., Shao, T., Zhou, K.: Real-time facial animation with image-based dynamic avatars. ACM Trans. Graph. **35**(4), 126 (2016)
12. Chen, Y.L., Wu, H.T., Shi, F., Tong, X., Chai, J.: Accurate and robust 3D facial capture using a single RGBD camera. In: IEEE International Conference on Computer Vision, pp. 3615–3622 (2013)
13. Cohen, I., Garg, A., Huang, T.S., et al.: Emotion recognition from facial expressions using multilevel HMM. In: Neural Information Processing Systems, vol. 2 (2000)
14. Cootes, T.F., Edwards, G.J., Taylor, C.J., et al.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001)
15. Cootes, T.F., Wheeler, G.V., Walker, K.N., Taylor, C.J.: View-based active appearance models. Image Vis. Comput. **20**(9), 657–664 (2002)
16. Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. In: British Machine Vision Conference, vol. 1, p. 3 (2006)
17. Ekman, P.: An argument for basic emotions. Cognit. Emot. **6**(3–4), 169–200 (1992)
18. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Van Gool, L.: Random forests for real time 3D face analysis. Int. J. Comput. Vis. **101**(3), 437–458 (2013)
19. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recognit. **36**(1), 259–275 (2003)
20. Geng, X., Zhou, Z.H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. IEEE Trans. Pattern Anal. Mach. Intell. **29**(12), 2234–2240 (2007)
21. Granger, S., Pennec, X.: Multi-scale EM-ICP: a fast and robust approach for surface registration. In: International Conference on Computer Vision, pp. 418–432 (2002)
22. Gunes, H., Pantic, M.: Automatic, dimensional and continuous emotion recognition. Int. J. Synth. Emotions. **1**(1), 68–99 (2010)
23. Guo, X., Hua, J., Qin, H.: Scalar-function-driven editing on point set surfaces. IEEE Comput. Graph. Appl. **24**(4), 43–52 (2004)
24. Guo, X., Hua, J., Qin, H.: Touch-based haptics for interactive editing on point set surfaces. IEEE Comput. Graph. Appl. **24**(6), 31–39 (2004)
25. Ioannou, S.V., Raouzaiou, A.T., Tzouvaras, V.A., Mailis, T.P., Karpouzis, K.C., Kollias, S.D.: Emotion recognition through facial expression analysis based on a neurofuzzy network. Neural Netw. **18**(4), 423–435 (2005)
26. Joshi, P., Tien, W.C., Desbrun, M., Pighin, F.: Learning controls for blend shape based realistic facial animation. In: SIGGRAPH, p. 8 (2005)
27. Kapoor, A., Burleson, W., Picard, R.W.: Automatic prediction of frustration. Int. J. Human Comput. Stud. **65**(8), 724–736 (2007)
28. Ko, K.E., Sim, K.B.: Development of the facial feature extraction and emotion recognition method based on ASM and bayesian network. In: IEEE International Conference on Fuzzy Systems, pp. 2063–2066 (2009)
29. Kobayashi, H., Hara, F.: Facial interaction between animated 3D face robot and human beings. IEEE International Conference on Computational Cybernetics and Simulation **4**, 3732–3737 (1997)
30. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
31. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems, pp. 556–562 (2001)
32. Lei, Z., Bai, Q., He, R., Li, S.: Face shape recovery from a single image using cca mapping between tensor spaces. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7 (2008)
33. Liao, Q., Jin, X., Zeng, W.: Enhancing the symmetry and proportion of 3D face geometry. IEEE Trans. Vis. Comput. Graph. **18**(10), 1704–1716 (2012)
34. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 94–101 (2010)
35. Morency, L.P., Rahimi, A., Darrell, T.: Adaptive view-based appearance models. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. I-803. IEEE (2003)
36. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: reconstruction and tracking of non-rigid scenes in real-time. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 343–352 (2015)
37. Nicolaou, M.A., Gunes, H., Pantic, M.: Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Trans. Affect. Comput. **2**(2), 92–105 (2011)
38. Sandbach, G., Zafeiriou, S., Pantic, M., Rueckert, D.: Recognition of 3D facial expression dynamics. Image Vis. Comput. **30**(10), 762–773 (2012)
39. Schuller, B., Valster, M., Eyben, F., Cowie, R., Pantic, M.: Avec 2012: the continuous audio/visual emotion challenge. In: ACM International Conference on Multimodal Interaction, pp. 449–456 (2012)
40. Sebe, N., Lew, M.S., Sun, Y., Cohen, I., Gevers, T., Huang, T.S.: Authentic facial expression analysis. Image Vis. Comput. **25**(12), 1856–1863 (2007)
41. Suwajanakorn, S., Kemelmacher-Shlizerman, I., Seitz, S.M.: Total moving face reconstruction. In: European Conference on Computer Vision, pp. 796–812 (2014)
42. Tena, J.R., De la Torre, F., Matthews, I.: Interactive region-based linear 3D face models. ACM. Trans. Graph. **30**(4), 76 (2011)
43. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–591 (1991)
44. Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., Pantic, M.: Avec 2014: 3D dimensional affect and depression recognition challenge. In: International Workshop on Audio/Visual Emotion Challenge, pp. 3–10 (2014)
45. Valstar, M.F., Pantic, M.: Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In: International Workshop on Human-Computer Interaction, pp. 118–127 (2007)
46. Wang, H., Ahuja, N.: Facial expression decomposition. In: IEEE International Conference on Computer Vision, pp. 958–965 (2003)

Hai Jin is a Ph.D. candidate at Computer Science Department in Wayne State University. He is currently a member of the Graphics and Imaging Laboratory. He received his M.S. degree from Nagoya University, Japan, in 2012 and B.S. degree from Shanghai Jiao Tong University, China, in 2008. His current research interests include geometric computing, computer graphics and scientific visualization.

Xun Wang is currently a visiting professor at Wayne State University. He is also a professor at the School of Computer Science and Information Engineering, Zhejiang Gongshang University, China. He received his BSc in mechanics, MSc and Ph.D. degrees in computer science, all from Zhejiang University, Hangzhou, China, in 1990, 1999 and 2006, respectively. His current research interests include mobile graphics computing, image/video processing, pattern recognition and intelligent information processing. In recent years, He has published over 80 papers in high-quality journals and conferences.

Yuanfeng Lian is a visiting scholar at Department of Computer Science at Wayne State University. He is also an Associate Professor at China University of Petroleum. He received his Ph.D. degree from Beihang University, China, in 2012 and M.S. degree from Changchun University of Technology, China, in 2003. His current research interests include computer graphics and image processing.

Jing Hua is a Professor of Computer Science and the founding director of Computer Graphics and Imaging Lab (GIL) and Vision Lab (VIS) at Computer Science at Wayne State University (WSU). He received his Ph.D. degree (2004) in Computer Science from the State University of New York at Stony Brook. His research interests include computer graphics, visualization, image analysis and informatics, computer vision, etc. He has authored over 100 papers in the above research fields. He received the Gaheon Award for the Best Paper of International Journal of CAD/CAM in 2009, the Best Paper Award at ACM Solid Modeling 2004 and the Best Demo Awards at GENI Engineering Conference 21 (2014) and 23 (2015), respectively. His research is funded by the National Science Foundation, National Institutes of Health, Michigan Technology Tri-Corridor, Michigan Economic Development Corporation and Ford Motor Company.