Automatic 2.5-D Facial Landmarking and Emotion Annotation for Social Interaction Assistance

Xi Zhao, *Member, IEEE*, Jianhua Zou, *Member, IEEE*, Huibin Li, *Student Member, IEEE*, Emmanuel Dellandréa, *Member, IEEE*, Ioannis A. Kakadiaris, *Senior Member, IEEE*, and Liming Chen, *Senior Member, IEEE*

***Abstract*—People with low vision, Alzheimer's disease, and autism spectrum disorder experience difficulties in perceiving or interpreting facial expression of emotion in their social lives. Though automatic facial expression recognition (FER) methods on 2-D videos have been extensively investigated, their performance was constrained by challenges in head pose and lighting conditions. The shape information in 3-D facial data can reduce or even overcome these challenges. However, high expenses of 3-D cameras prevent their widespread use. Fortunately, 2.5-D facial data from emerging portable RGB-D cameras provide a good balance for this dilemma. In this paper, we propose an automatic emotion annotation solution on 2.5-D facial data collected from RGB-D cameras. The solution consists of a facial landmarking method and a FER method. Specifically, we propose building a deformable partial face model and fit the model to a 2.5-D face for localizing facial landmarks automatically. In FER, a novel action unit (AU) space-based FER method has been proposed. Facial features are extracted using landmarks and further represented as coordinates in the AU space, which are classified into facial expressions. Evaluated on three publicly accessible facial databases, namely EURECOM, FRGC, and Bosphorus databases, the proposed facial landmarking and expression recognition methods have achieved satisfactory results. Possible real-world applications using our algorithms have also been discussed.**

***Index Terms*—Assistive technology (AT), automatic emotion annotation, facial expression recognition (FER), facial landmark localization, portable RGB-D cameras.**

I. Introduction

MANY people experience difficulty in perceiving or interpreting facial expressions of emotion due to their congenital or acquired diseases. People with autism spectrum disorder (ASD) have dysfunction in recognizing emotional and social information from faces [1]. People with low vision (LV) have reduced vision acuity and may not perceive expressions displaying on the faces around them. People with Alzheimer's disease (AD) may have an impaired interpersonal communication function in recognizing others' facial expressions [2]. They have needs in assistance on emotion annotation to improve the quality of their social lives. People with LV have ranked "knowing the facial expressions of the person standing in front of you" as the second place in their needs during social interaction [3]. Assistive technology (AT) in recognizing facial expressions of emotion can help people with ASD to make connections between facial expressions and the emotions.

Annotating emotions, or affective states, is often conducted as facial expression recognition (FER) in the literature. FER has been tested on facial data in both 2-D and 3-D [4]–[8]. FER methods on 2-D facial images and videos are in general computational efficient, and thus can be easily adopted into applications. However, their performance suffers from the classic difficulties in face analysis (i.e., head pose, lighting condition, and face scale). Recently, these difficulties are eased or sometimes even overcome in 3-D FER methods owing to facial shape information captured by the 3-D scanners. 3-D FER methods are inherently more robust to variations in lighting conditions and head poses [9]. Their recognition accuracies are better because of the extra shape information compared to 2-D methods. However, 3-D scanners are still expensive and difficult to take by people. Moreover, richer information in 3-D facial scans also increases the computational cost and time, especially when processing the 3-D facial videos. These factors limit the widespread use of 3-D FER methods.

As advancement in imaging technology, several types of RGB-D cameras have been released as commercial products in the market. It captures 2-D videos as well as aligned depth image sequences, which record the distance of the surfaces of scene objects from the camera's viewpoint. As the most popular RGB-D camera, the Kinect developed by Microsoft has been widely used in the entertainment industry and academic research. The size of RGB-D cameras has been reduced in Creative Senz3-D [10]. To meet the demanding need for wearable sensors, the compact RGB-D camera module has been announced recently [11], which can be integrated into mobile devices or head-mounted wearable devices

Manuscript received January 18, 2015; accepted July 12, 2015. Date of publication August 26, 2015; date of current version August 16, 2016. This work was supported in part by the National Nature Science Foundation of China under Grant 61303121 and Grant 11401464, in part by the Microsoft Research Asia Collaborative Research Award, and in part by the French Research Agency, Agence Nationale de Recherche through the Jemime Project under Grant ANR-13-CORD-004-02. This paper was recommended by Associate Editor S. Zafeiriou.

X. Zhao is with the School of Management, Xi'an Jiaotong University, Xi'an 710048, China (e-mail: zhaoxi@ieee.org).

J. Zou is with the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710048, China (e-mail: jhzou@sei.xjtu.edu.cn).

H. Li is with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710048, China (e-mail: huibinli@mail.xjtu.edu.cn).

E. Dellandréa and L. Chen are with Ecole Centrale de Lyon, Lyon 69130, France (e-mail: emmanuel.dellandrea@ec-lyon.fr; liming.chen@ec-lyon.fr).

I. A. Kakadiaris is with the University of Houston, Houston, TX 77004 USA (e-mail: ikakadia@central.uh.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2015.2461131

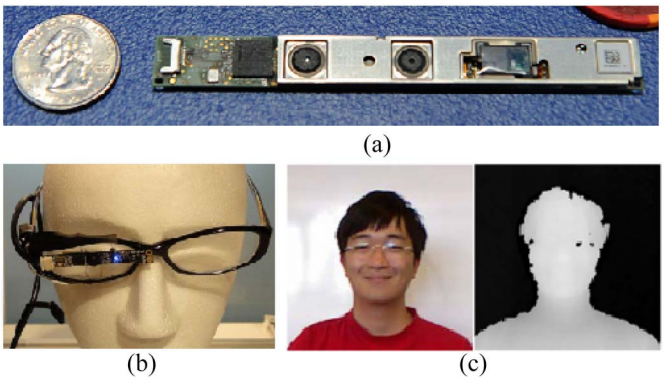


Fig. 1. Depiction of a wearable RGB-D camera. (a) Portable RGB-D module announced by Intel in CES 2014. (b) Concept demonstration of visual data capture from the first-angle view using the RGB-D camera [12]. (c) 2.5-D facial data collected by the camera [13] (left: the texture image and right: the depth image).

[Fig. 1(a) and (b)]. 2.5-D facial data are captured as video streams from these cameras [depicted as in Fig. 1(c)]. These data have the advantages in overcoming the challenges in head pose and illumination while these can also be pervasively captured using affordable RGB-D cameras. It is anticipated that the portable/wearable RGB-D camera-based emotion annotation techniques will become practically available in the near future.

To develop the portable/wearable emotion AT, automatically annotating facial landmarks and expressions are among those crucial tasks. A facial landmark is referred as the fiducial points defined by anthropometry that have consistent reproducibility even in adverse conditions such as facial expression or occlusion [14]. Landmarking accuracy has a direct impact on subsequent tasks like FER. FER deals with the classification of facial feature deformation during expression occurrence into abstract emotion classes that are purely based on visual information [4]. In this paper, we propose an automatic emotion annotation solution for 2.5-D facial data which can be adopted in the aforementioned portable/wearable emotion ATs. The method comprises two components, respectively, automatic facial landmark localization and FER. These two components process the facial data sequentially. During online stage, the output landmarks from the first component are used to extract facial features, which are used in recognizing facial expressions. More specifically, a deformable partial facial model (DPFM) is built being able to deform its shape and appearance. The landmarks on a new face are localized by searching the best fitting of DPFM on it. A 3-D landmarking method has been proposed in [14] using the similar principle. The DPFM-based method differs from it in two aspects. First, instead of building deformable face models in the 3-D coordinate system, we build DPFM in the 2-D image coordinate system. Procrustes analysis and extra transformation parameters are considered during its fitting process. Second, local region resampling process has been excluded from the training and fitting processes as used in [14], which simplify the computation of the landmarking method. To achieve FER, a multidimensional action unit (AU) space is proposed. Hyperplanes are learnt to segment the AU space into different emotion regions using the AU coordinates computed from facial features. During online stage, AU coordinates are estimated by AU detectors. The region, where this coordinate vector located implicates the expression displaying on it.

The major contributions of this paper include the following.

1) A 2.5-D facial landmarking method has been proposed for RGB-D cameras which builds a deformable partial facial model and fits the model to new faces for landmark localization.
2) An AU space has been proposed, where faces displaying a single or a combination of AUs can be represented as multidimensional AU coordinate vectors.
3) An AU coordinate computation method and a novel manner to interpret facial affective states using AU space have been proposed.
4) Systematical evaluation on the proposed methods using three publicly available databases and discussions on future emotion ATs using the proposed emotion annotation solution.

The rest of this paper is organized as follows. Section II discusses the related work. Section III presents the 2.5-D facial landmarking method and Section IV describes the use of the AU space for FER. The experimental results are presented in Section V. Discussions on future applications have been described in Section VI. We draw the conclusion in Section VII.

II. Related Work

Adopting AT to improve social interaction has been investigated in several studies. Peterson *et al.* [15] emphasized the needs from the people with AD for FER during their social lives. Krishna *et al.* [16] proposed the concept of a wearable system to improve the social interaction quality for visually impaired people by recognizing facial expressions using a head-mounted 2-D camera. They further improved the system to provide rehabilitative feedback for reducing stereotypic body mannerisms in [17]. Liu *et al.* [18] proposed the computer-based affect-sensitive ASD intervention tools by designing therapist-like affective models of the children with ASD. Wainer and Ingersoll [19] reviewed the studies using interactive computer programs for delivering direct intervention on development of social and communication skills to the autistic.

While most of these studies adopted 2-D cameras as data capturing devices, RGB-D cameras have garnered increased attention in recent years [20]. The cameras have been widely used in many computer vision problems [21], such as scene reconstruction [22], gesture recognition [23], face analysis [24]–[26], and object recognition [27]. Meanwhile, RGB-D databases have been collected on faces [13], [28], [29], objects [30], and human activities [31]. Limited by the camera's portability and power requirement, most of the studies addressed the visual content collected from a fix point of view or the third person point of view. As Intel announced the availability of the coin-sized RGB-D camera module, highly portable or wearable devices with integrated RGB-D cameras are expected to hit the market in the near future. They are able to collect the egocentric visual contents. These devices

will significantly facilitate the creation of robust assistive technology (AT) on automatic emotion annotation during the social interaction.

The key problems in developing such AT include automatic facial landmarking [14], [37] and FER [9] on the facial data with shape information. These facial data could be categorized into shaped 2-D faces (2-D facial images with depth images from the RGB-D cameras) or textured 3-D faces (3-D facial shape scans with 2-D facial texture images captured from 3-D face scanner). Shaped 2-D faces can also be computed from textured 3-D facial data by keeping the geometry information along the z-direction while neglecting geometry information along the X- and y-direction. Between these two types of facial data, most studies in the literature conducted facial analysis on textured 3-D faces thanks to the well established publicly accessible 3-D facial databases.

In the literature, there exists abundant work on facial feature detection both on 2-D intensity face images [72]–[74], [78] or 2.5-D or 3-D shape-based face data [14], [37], [75]–[77], [79]. While fast and effective methods for 2-D face alignment (see [72]–[74], [78]), have emerged recently, it is also well known, as highlighted in [14], [37], and [75]–[79], that pure 2-D texture related techniques can severely suffer from illumination variations. With commodity depth cameras readily available, a promising research axis is to rely on 3-D depth face data, which capture the geometric shape of faces and thereby enable face landmarking solutions with an improved robustness to lighting changes. Our proposed face landmarking techniques based on RGB-D images is under such a perspective. Facial landmarking on shape-based face data can be local feature oriented or data-driven based through learning. Local feature oriented approaches are effective for shape prominent face landmarks, e.g., nose tip, as they try to embed *a priori* knowledge on facial landmarks by computing local 3-D shape related response, including (e.g., spin image [32], effective energy [33], Gabor filtering [34], generalized Hough Transform [35], local gradients [36], HK curvature [37], shape index [38], and curvedness index [39]). Data-driven-based approaches generally attempt to fit a deformable model in optimizing an objective function with a very first work in [14], followed by others (see [75]–[77], [79]). Besides, as texture and shape carry complementary information, several works (see [14], [75], [76], [79]), also show that their fusion in the fitting process can result in a further improved landmark localization accuracy.

In the literature of FER, two main streams exist including judgement-based approaches and sign-based approaches. Methods in the first group directly classify specific facial patterns into the facial expressions. Most studies on 3-D FER can be categorized into judgement-based approaches which recognize the six universal expressions [40]–[45]. The sign-based approaches abstracted and coded facial muscle activities by facial AUs in facial action coding system (FACS) [46] and then interpreted emotions as combinations of detected AUs using emotional FACS rules [47]. Studies on sign-based FER were mostly conducted on 2-D facial images or videos [5], [48]–[51]. A few works have investigated AU detection, including [52], [53], and [69]. These studies achieved satisfactory results on recognizing over 20 AUs. They either extracted local binary pattern (LBP)-based features from whole faces or built holistic deformable models. However, they rarely addressed the problem of further recognizing facial expressions using their detected AU combinations. In this paper, we opt to explore an automated sign-based approaches using both global and local information and interpret facial expressions using AU combinations in a novel manner. Features extracted using the landmarks are used to describe facial deformation. Instead of just detecting AU occurrence, we further interpret facial emotions as AU combinations in a novel AU space-based manner. The flexibility of our method allows recognizing various facial expressions to satisfy the needs for real world application, rather than just recognizing the six universal ones or single occurred AUs as in the literatures.

III. Facial Landmark Localization

Facial landmarks localized on 2.5-D facial data, such as eye corners, mouth corners, nose tip, etc., can be utilized to extract facial features for automatic face recognition or FER. In the following section, we first build a deformable partial face model to learn local appearance variations around each landmark from both facial texture images and depth images, as well as variations on global landmark shapes. Then we localize landmarks on a 2.5-D face by fitting this model to it. During fitting, the similarity between the model and the face is optimized by driving the model parameters to deform the DPSM.

A. Feature Extraction

The facial depth images can be corrupted by the camera noises, creating holes, and roughness on facial surfaces. Before extracting local feature for landmark representation, we first apply a 2-D median filter on depth images to smooth the face surface. Then, the holes are located by morphological reconstruction [54] and filled by cubic interpolation.

Nineteen landmarks distributed on the major facial components constitute the landmark set used in this paper, as depicted in Fig. 2(a). The local regions around these landmarks sufficiently covered 19 regions of interests on the eyebrows, the eyes, the nose, and the mouth, where nonrigid facial deformation normally occurs in facial expressions. Their coordinates on 2-D texture images are manually annotated for the training data. Since pixel-level correspondence between texture images and depth images can be easily obtained from RGB-D cameras, the landmark coordinates on depth images can be automatically obtained. Thus, corresponded local patches around landmarks can be extracted from both facial texture and depth images, as depicted in Fig. 2(b) and (c).

A landmark can be featured by its local texture and depth patches. However, variations on face scale change the content in these patches. To compensate for it, an estimate of the face region, face location, and scale is computed on texture images using the PittPatt face detector [55]. The faces can then be normalized to a constant scale in both depth and texture images. Additionally, illumination variations of facial

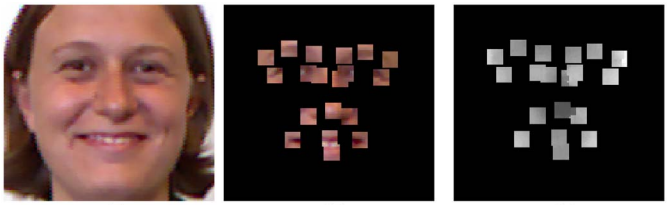


Fig. 2. Depiction of 19 landmarks distributed on the major facial components. (a) Manually labeled landmarks on a facial texture image. (b) Nineteen local texture patches centered on these landmarks. (c) Nineteen local depth patches centered on these landmarks. All texture and depth patches have the same size.

texture are standardized using an image-specific low-pass filtering [56].

After scale and illumination normalization, we extract the local patches around these 19 landmarks from both texture and depth images. The intensity and depth values on the local patches are concatenated into two vectors G and Z as in (1), respectively, where m are the total pixel number across all patches

$$G = (g_1, g_2, \ldots, g_m)^T, \quad Z = (z_1, z_2, \ldots, z_m)^T. \tag{1}$$

B. Deformable Partial Face Model

Provided a training database and its manual landmark sets, we first concatenate the landmark coordinates on each texture image into a vector X, as in (2), where N is the number of landmarks

$$X = (x_1, y_1, x_2, y_2, \ldots, x_N, y_N)^T. \tag{2}$$

Then, all X vectors in the image coordinate system are mapped into S in the shape coordinate system using Procrustes analysis [57] so that 2-D global variations are removed including in-plane rotation and translation.

To build the DPFM, principal component analysis (PCA) is applied to the training features S, G, and Z, where 95% major components are preserved in PCA

$$S = \bar{S} + P_s b_s \tag{3}$$

$$G = \bar{g} + P_g b_g, Z = \bar{z} + P_z b_z \tag{4}$$

where $\bar{S}$, $\bar{g}$, and $\bar{z}$ are the mean shape, mean normalized intensity, and mean depth value separately; P_s, P_g, and P_z are the basis of the shape, texture, and depth variation spaces, respectively; b_s, b_g, and b_z are sets of control parameters of shape, intensity, and depth values.

All the b_i parameters from PCA, where $i \in (s, z, g)$ are independent. It is supposed that all these b_i parameters follow Gaussian distributions with zero mean and standard deviation σ_i [58] (estimated from training features). Figs. 3–5 depict the first two modes at their left and right ending variation $(-3\sigma_j, 3\sigma_j)$, where $j \in (b_{s1,2}, b_{z1,2}, b_{g1,2})$, respectively, for shape variations, texture variations, and depth variations in local patches.

C. Estimation of Texture and Depth Instances

Given the parameter vector b_s, we can generate a new shape instance S by (3). Four parameters are further required to recover the shape S in the image coordinate system, which can be viewed as a reverse transformation of the Procrustes analysis as in (5). These parameters include a pair of translation parameters (C_x, C_y), a scale parameter s and an in-plane rotation parameter ρ

$$X = s \cdot \left(R(\rho) \cdot (\bar{S} + P_s b_s) + C\right) \tag{5}$$

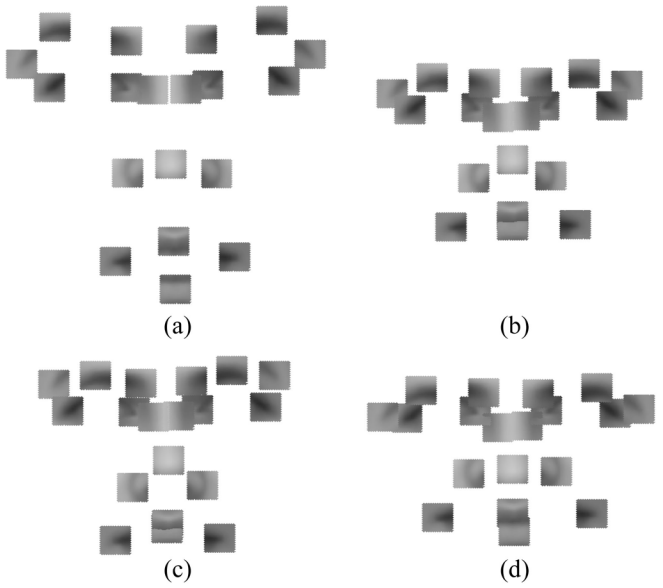


Fig. 3. Depiction of the learnt modes of shape variations. We fix all texture, depth parameters, and change the first two shape parameters b_{s1} and b_{s2}, respectively. Shape instance generated using (a) $b_{s1} = -3\sigma_{s1}$, (b) $b_{s1} = 3\sigma_{s1}$, (c) $b_{s2} = -3\sigma_{s2}$, and (d) $b_{s2} = 3\sigma_{s2}$. The first mode mainly interprets the shape variation in the vertical direction, while the second mode interprets the shape variation in the horizontal direction.

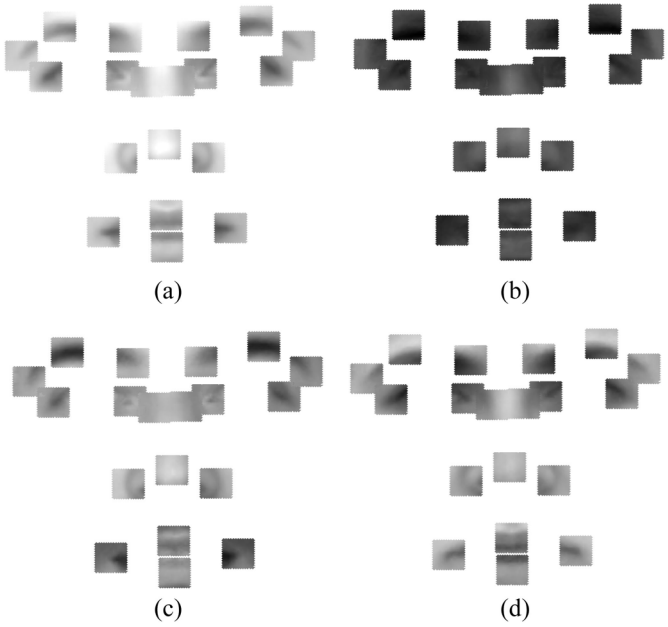


Fig. 4. Depiction of the learnt modes of texture variations. We fix all shape, depth parameters, and change the first two texture parameters b_{g1} and b_{g2}, respectively. Texture instance generated using (a) $b_{g1} = -3\sigma_{g1}$, (b) $b_{g1} = 3\sigma_{g1}$, (c) $b_{g2} = -3\sigma_{g2}$, and (d) $b_{g2} = 3\sigma_{g2}$. The first mode mainly interprets the variation on skin color difference, while the second mode mainly interprets the variation on texture around ocular and mouth regions.

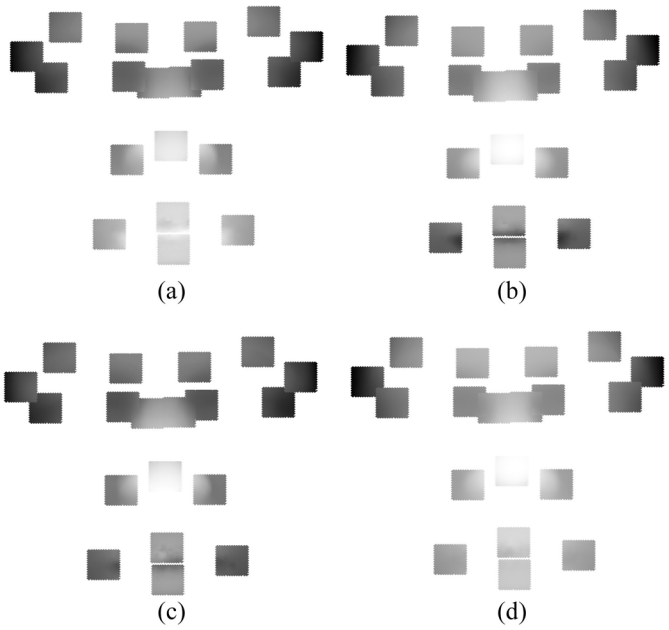


Fig. 5. Depiction of the learnt modes of depth variations. We fix all shape, texture parameters, and change the first two depth parameters b_{z1} and b_{z2}, respectively. Depth instance generated using (a) $b_{z1} = -3\sigma_{z1}$, (b) $b_{z1} = 3\sigma_{z1}$, (c) $b_{z2} = -3\sigma_{z2}$, and (d) $b_{z2} = 3\sigma_{z2}$. Both modes interpret the depth variation around most of the regions. However, the first mode contains more depth variation in nasal area and the second mode contains more variation in the mouth area.

where X is the created shape instance and $R(\rho)$ is the rotation matrix. The shape transformation parameters and shape parameters (b_s) are concatenated into a vector $S_h = (b_s^T | C^T | s | \rho)^T$.

Given a shape instance X and a 2.5-D face, we obtain its vectors G and Z (1) through the same feature extraction process as described in Section III-A. They are further used to estimate b_g and b_z

$$b_g = P_g^T(G - \bar{g}), b_z = P_z^T(Z - \bar{z}). \tag{6}$$

To ensure plausibility of the estimated instance, we limit b_i into the depth $\pm 3\sigma_i$. Any b_i ($i \in (g, z)$) exceeding its corresponding boundary is replaced by its closest boundary. Then, we can generate texture and depth instances $\hat{G}$ and $\hat{Z}$ by (4) using the constrained b_g and b_z.

D. Model Fitting

To locate landmarks on a new 2.5-D face, the learnt DPSM is fit by deforming it to find the maximum similarity on the face. This fitting problem can be considered as an optimization problem with an objective function driving the variable S_h. By initializing S_h, we generate an initial shape instance, and create texture and depth instances $\hat{G}$ and $\hat{Z}$ as described in Section III-C. We then compute normalized cross correlation [59] as texture and depth similarity $F_g(G, \hat{G})$ and $F_z(Z, \hat{Z})$, respectively, and use their weighted sum as the objective function

$$f(S_h) = \alpha F_g + \beta F_z \tag{7}$$

where α and β are a pair of weight parameters. In order to ensure the shape instance is plausible, we also limit b_i, where $i \in b_s$ within the boundary of $\pm 3\sigma_i$. All trespassing b_i is replaced by the its closest boundary. The fitting algorithm is depicted in Algorithm 1. The optimization in step six is processed by the Nelder–Mead [60] simplex algorithm.

Algorithm 1 DPFM Fitting

Input: A pair of texture image M_t and depth image M_d, a DPFM D

Output: Landmark coordinate vector X

1. Preprocess the texture image M_t and depth image M_d to reduce noise as depicted in section III-A,
2. Create the initial shape instance X^0 with the vector of S_h, where b_s are all set to zeros and C^T, s, ρ are set to $(0, 0)^T$, 1, 0 respectively,
3. Extract local patches on the texture image M_t and the depth image M_r given X^0 as in section III-A,
4. Estimate texture and depth instances $\hat{G}^k$, $\hat{Z}^k$ as described in section III-C,
5. Compute the value of objective function f^k,
6. Predict S_h^{k+1} by the optimization algorithm,
7. Compute X^{k+1} and compare it with X^k to check the convergence. If not, $k = k+1$ and go to 2; else return X^{k+1}.

IV. AU Space for Facial Expression Recognition

Ekman and Friesen [46] defined the facial AUs as all visually discernible facial movements in terms of atomic facial actions. Over 7000 different AU combinations have been observed and some of these combinations are mapped into various affective states according to FACS affect interpretation database (FACSAID) [61]. For example, the occurrence of AU1, AU2, AU5, and AU26 can be mapped into the surprise while the occurrence of AU6 and AU12 can be mapped as happiness. These rules use the binarized AU detection results (AU occurrence: 1 or non-AU occurrence: 0) and normally it requires skilled experts to manually apply them. The proposed AU space is a multiple dimensional space, where a coordinate on an AU axis represents the belief that its AU occurs on a face. The coordinate ranges from 0 to 1 on all axes to facilitate the hyperplane construction later. Ideally the dimension of this space should match with the number of 44 basic AUs defined in [46]. However, some subtle AUs can only be detected on high-resolution facial data collected from professional devices. Thus, the AU space constructed in this paper is a subspace of the ideal AU space, which covers a set of N AUs available in the database under our investigation. This space is subsequently subdivided into regions of emotions or affective states using hyperplanes. Points located in a region share the same affective state.

AU space uses the continuous coordinates instead of binary values in FACSAID rules to interpret affective states. It does not require manual label from skilled experts and is more tolerant of AUs detecting results in margin area as depicted in Fig. 6, especially for those subtle AUs on which AU detectors are easier to obtain scores near the borderlines. Moreover, AU space is more flexible in recognizing emotions other than the

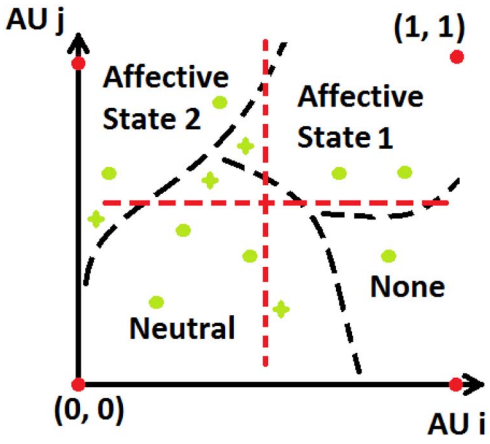


Fig. 6. Depiction of AU space conception. This figure demonstrates a conceptual 2-D AU space constructed just from AU_i and AU_j. The green dots and stars represent the coordinates computed for several 2.5-D faces using the proposed method, the red dots represent the four states of binary AU combinations, the black dashed lines represent the hyperplanes dividing the AU space, and the red dashed lines represent the threshold lines for classifying the binary outputs from AU detectors. According to the principle of the FACSAID rules, only four points (red dots) in the space may be mapped into affective states (note that scores from AU detectors can be binarized into these four points by thresholding). It can be seen that AU space-based emotion interpretation is more tolerant of AU detection results in margin areas, and thus reduces the misclassification of those coordinates near the borderlines (green stars).

six universal ones (happiness, sadness, surprise, fear, anger, and disgust) in comparison to the traditional judgement-based facial emotion recognition methods.

A. AU Detector

A coordinate on an AU axis describes how likely the corresponding AU(X) is displayed on a face. It can be computed as the probability of presence of facial features $\{\mathcal{F}\}$ knowing X

$$\mathcal{P}(\{\mathcal{F}\}|X) = \mathcal{P}\left(\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_{N_c}|X\right) = \prod_{l=1}^{N_c} \mathcal{P}\left(\mathcal{F}_l|X\right) \quad (8)$$

where N_c is the number of features we used for AU detection, $\mathcal{F}_l$ represents any type of facial feature in a vector. Based on landmarks localized manually or automatically by the DPSM, a set of global and local features can be extracted to comprehensively characterize facial activities. In total, a number of $N_c = 14$ features are extracted from three face representations (i.e., facial shape, facial texture, and facial geometry). These features include the displacements of landmarks $\mathcal{D}$ compared to a mean neutral face and distances between these landmarks $\mathcal{L}$ in the shape representation, the texture patches $\mathcal{G}$ around landmarks and the LBP at five scales $\mathcal{L}_t^{1-5}$ in the texture representation [62], the local depth patches $\mathcal{Z}$ in the vicinity of landmarks, and LBPs computed on the depth patches also at five scales L_d^{1-5} [62] in the depth representation. These features are selected since face deformations caused by AUs or expressions impact three face representations to different extents. For instance, AU7 and AU43 significantly change texture in the eye regions without moving eye corners. AU24 changes mostly the local depth and texture in the mouth region while having less influence on landmark locations. Therefore, it is essential to simultaneously extract features from these three different face representations.

In order to estimate $\mathcal{P}(\mathcal{F}_l|X)$ in (8), we adopt a feature model associated with the lth features in assuming that $\mathcal{P}(\mathcal{F}_l|X)$ follows the Gibbs–Boltzmann distribution. Given a set of features extracted from all training faces displaying a specific AU i_x, the (l, i_x)th feature model is trained to learn the person-independent AU occurrence distribution. The model is able to approximate a new feature by linearly combining the learnt variation modes, similar to the principle in DPSM. Specifically, for each feature $\mathcal{F}_l \in \{\mathcal{D}, \mathcal{L}, \mathcal{G}, \mathcal{L}_t^{1-5}, \mathcal{Z}, L_d^{1-5}\}$, PCA is applied to learn the variation modes of the features on the facial data displaying the AU i_x, preserving 95% major variations of the underlying feature. $\mathcal{F}_l^{i_x} = \bar{\mathcal{F}}_l^{i_x} + P_l^{i_x} b_l^{i_x}$, where $\bar{\mathcal{F}}_l^{i_x}$ is the feature mean, $P_l^{i_x}$ is the set of eigenvectors resulting from PCA, and $b_l^{i_x}$ is a set of parameters. We repeat the training for 14 features and thus obtain 14 sets of trained $\bar{\mathcal{F}}_l$, P_l, and σ_l for the i_xth AU. Then, we repeat this model building process for all (l, i_x) combinations.

B. Coordinate Computation

Given a new face κ, we first localize its landmarks and extract the aforementioned features $\mathcal{F}_\kappa$. The lth feature response $\hat{\mathcal{F}}_{l\kappa}^{i_x}$ from the i_xth AU detector can be generated by estimating the best parameter $b_l^{i_x} x = {P_l^{i_x}}^T (\mathcal{F}_{l\kappa} - \bar{\mathcal{F}}_l^{i_x})$. We set the boundary as $\pm 0.5\sigma_{lj}^{i_x}$ for the corresponding parameter in $b_{lj}^{i_x}$ to form $\hat{b}_{lj}^{i_x}$. Consequently, the deformations in the feature response are constrained so that separability between the genuine and nongenuine classes can be increased. Then the feature response $\hat{\mathcal{F}}_{l\kappa}^{i_x}$ is computed by inputting $\hat{b}_{lj}^{i_x}$ in its corresponding feature model.

Following the Gibbs–Boltzmann distribution, $\mathcal{P}(\mathcal{F}_l|X) \propto e^{A_l Q_l^{i_x}}$, where $Q_l^{i_x}$ is the match quality and A_l is a normalizing constant. $Q_l^{i_x}$ can be computed as the similarity between the observed feature $\mathcal{F}_{l\kappa}$ and its i_xth AU response $\hat{\mathcal{F}}_{l\kappa}$. Specifically, the similarity $Q_{l\kappa}$ is computed

$$Q\left(\mathcal{F}_{l\kappa}^{i_x}\right) = \left\langle \frac{\mathcal{F}_{l\kappa}}{\|\mathcal{F}_{l\kappa}\|}, \frac{\hat{\mathcal{F}}_{l\kappa}^{i_x}}{\left\|\hat{\mathcal{F}}_{l\kappa}^{i_x}\right\|} \right\rangle. \quad (9)$$

Inserting the Gibbs distribution into (8) and taking logarithm, the coordinate on the i_x AU axis for the face κ is thus obtained as

$$\mathcal{D}_\kappa^{i_x} = \sum_{l=1}^{N_c} A_l Q_{l\kappa}^{i_x}. \quad (10)$$

We repeat this process for all AU axes and obtain a vector of AU coordinates $\mathcal{D}_\kappa$

$$\mathcal{D}_\kappa = \left(\mathcal{D}_\kappa^1, \mathcal{D}_\kappa^2, \ldots, \mathcal{D}_\kappa^N\right)^T. \quad (11)$$

C. Affective State Interpretation

Coordinate computation is repeated for all N AU detectors and a N-dimensional coordinate vector $\mathcal{D}_\kappa$ can be obtained for the input face κ. We use this coordinate vector as input for hyperplane construction.

Support vector machine (SVM) [64] is a popular and powerful classifier which constructs a hyperplane or set of hyperplanes in high dimensional space. These hyperplanes achieve good separations by maximizing the distance to the nearest training data points of classes (support vectors) in the classification task. Though SVM is trained from the observed data, it forgets the individual observations after training and only saves the decision hyperplanes. The SVM is originally proposed for binary classification problem. Multiclass SVM has also been designed [65] which converts the single multiclass problem into multiple binary classification problems.

In our case, N dimensional coordinate vectors need to be classified into one of the M affective states γ. Thus, a multiclass SVM is adopted to construct these hyperplanes which divide AU space into M regions. The SVM follows the one-versus-all strategy. Specifically, M binary SVMs is constructed inside the multiclass SVM. The ith SVM is trained with all of the coordinate vectors in the ith class with positive labels, and the other coordinate vectors with negative labels. It solves the following problem:

$$\begin{aligned} \min_{w^i, v^i, \xi^i} \quad & \frac{1}{2}\left(w^i\right)^T w^i + C\sum_{j=1}^{\mathcal{T}} \xi_j^i \left(w^i\right)^T \\ & \left(w^i\right)^T \phi\left(\mathcal{D}_j\right) + v^i \geq 1 - \xi_j^i, \text{ if } \gamma_j = i \\ & \left(w^i\right)^T \phi\left(\mathcal{D}_j\right) + v^i \leq 1 - \xi_j^i, \text{ if } \gamma_j \neq i \\ & \zeta_j^i \geq 0, j = 1, \ldots, \mathcal{T} \end{aligned} \tag{12}$$

where i is the index of class and j is the index of a number of $\mathcal{T}$ training data. ϕ is a mapping function. M pairs of w, v are learnt as the hyperplane parameters along with the ξ. For a coordinate vector $\mathcal{D}_\kappa$ to classify, M decision scores are computed as $(w^i)^T\phi(\mathcal{D}_\kappa) + v^i$. The affective state of κ is hereby designated as the state with the highest values of the decision functions. The detailed elaboration on the process of multiclass SVM training and testing can be found in [65]. Since values in coordinate vectors are between zero and one, no further scaling process on their values is performed before feeding these vectors into the SVM.

V. Experimental Results

A. Databases

We used three publicly accessible face databases for evaluating the proposed landmarking and FER methods. The first is EURECOM Kinect face database [13], which consists of the RGB-D facial data from 52 subjects (14 females and 38 males) obtained by the Kinect camera. The data were captured in two sessions with an intersession period of about half month. In each session, the database provides the facial data in nine different facial expressions, lighting and occlusion conditions. The cohort employed in the evaluation consists of facial data displaying the neutral state, the open mouth and the smile expressions of the two sessions.

The second database is FRGC database [80]. It contains three sessions, where data from 4003 subject sessions were collected. In each session, a person's biometric data are collected and consist of four controlled still images, two uncontrolled still images, and one 3-D image. The 3-D image was taken under controlled illumination conditions. The 3-D images consist of both a depth image and a texture image. The 3-D images were acquired by a Minolta Vivid 900/910 series RGB-D sensor. The cohort we used includes all 1893 3-D images in FRGC v2 database collected in Fall 2003.

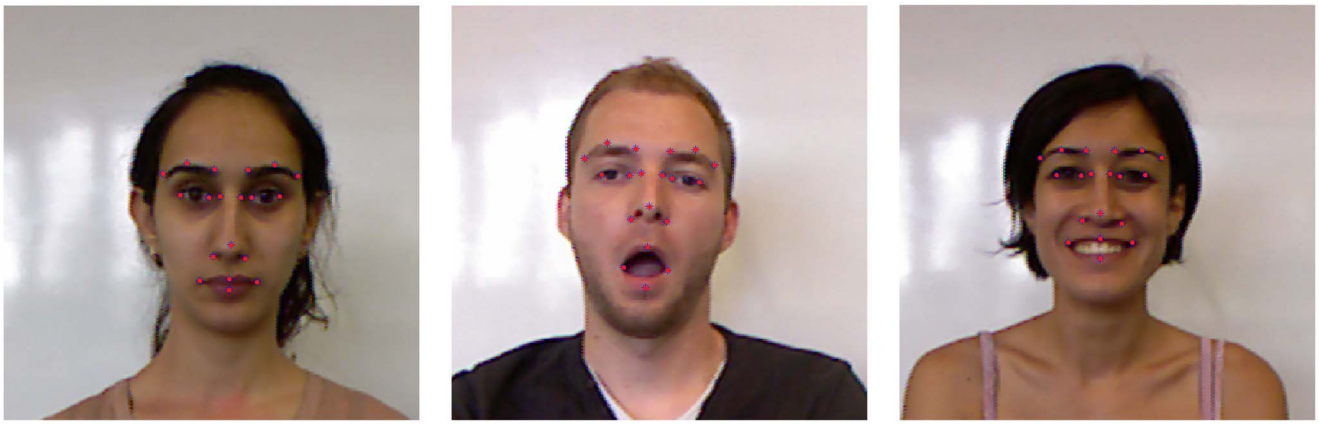
The third database is Bosphorus database [66]. It comprises of 105 subjects in various poses, expressions and occlusion conditions. The number of facial data from each subject varies from 31 to 54 and the number of total faces is 4652. The cohort employed in the evaluation consists of all facial data displaying neutral state, the six universal expressions and the AUs. Borphorus database contains the most various single occurred AUs with both texture and geometry information in the literature. It is necessary for us to obtain 2.5-D facial data and evaluate the AU space-based method on this database. Another relevant facial database is presented in [67]. However, it consists of only seven subjects and contains many occluded facial data caused by hand-over-face gestures. The data are not sufficient for us to conduct the subject-independent experiment procedure. Thus, we excluded this database from our tests. Note that the term "facial data" in this section specifically refers to a pair of facial texture and depth images captured simultaneously from a face using a RGB-D camera.

TABLE I
Landmark Localization Results on EURECOM Database

		1	2	3	4	5
a	mean	3.78%	3.69%	4.67%	4.33%	4.31%
	std	2.13%	2.15%	2.36%	3.01%	2.93%
b	mean	4.19%	4.24%	4.16%	5.33%	5.19%
	std	2.37%	2.30%	2.08%	3.20%	3.39%
c	mean	4.71%	4.69%	4.44%	6.66%	6.74%
	std	2.21%	1.89%	2.11%	3.26%	3.17%

B. Results on Landmarking

1) DPSM-Based Landmarking Evaluation on EURECOM Database: We trained a DPSM on the data from the first three male and the first three female subjects, so that the portion of the training data was roughly 10% of the whole cohort in EURECOM Kinect face database. Then the rest portion was used for testing. We have manually annotated the training set. Table I depicts the mean and standard deviation of localization errors, which were normalized by the interpupillary distance. The normalized error has been considered as a common evaluation criterion to compare the landmarking results across different face scales. Since the database provides manual annotations for six landmarks as the ground truth, but only five were overlapped with the landmarks in our set, we only reported localization errors on these five landmarks: 1) the left/right eye centers; 2) the nose tip; and 3) the left/right mouth corners. The first two rows depict the landmarking results on neutral faces; the second two rows depict the landmarking results on faces displaying smile expression; and the third two rows depict the results on faces with open mouth. Most of them had an average error less than 5%. Except for two mouth corners



(a) (b) (c)

Fig. 7. Depiction of 19 landmarks automatically localized by DPSM on faces in EURECOM database. Landmarks located on a (a) neutral face, (b) face with open mouth, and (c) face displaying a smile.

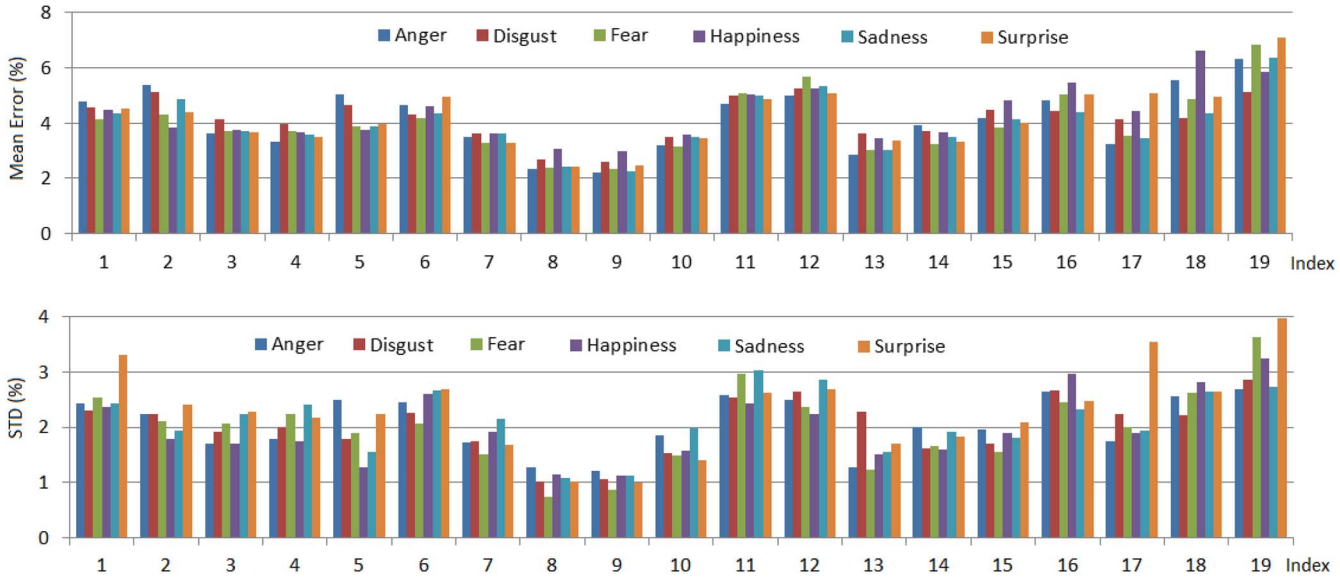


Fig. 8. Depiction of landmarking accuracy on different expressions from the Bosphorus database. The *y*-axis of the upper figure is the mean of normalized errors. 1: the left corner of the left eyebrow, 2: the middle of the left eyebrow, 3: the right corner of the left eyebrow, 4: the left corner of the right eyebrow, 5: the middle of the left eyebrow, 6: the right corner of the right eyebrow, 7: the left corner of the left eye, 8: the right corner of the left eye, 9: the left corner of the right eye, 10: the right corner of the right eye, 11: the left nose saddle, 12: the right nose saddle, 13: the left corner of the nose, 14: the nose tip, 15: the right nose saddle, 16: the left corner of the mouth, 17: the middle of the upper lip, 18: the right corner of the mouth, and 19: the middle of the lower lip.

whose errors were slightly increased in the open mouth expressions, landmarks were localized with consistent errors on all three expressions. The results demonstrated the robustness of our method on expression variations. A demonstration on landmarking examples is illustrated in Fig. 7.

2) DPSM-Based Landmarking Evaluation on Bosphorus Database: To evaluate landmarking accuracy of 19 landmarks and investigate impact of facial expressions on landmarking accuracy, we have conducted the second experiment on Bosphorus database. A cohort was utilized consisting of facial data displaying the six universal expressions from all available subjects. The texture and depth images from the first three male subjects and the first three female subjects were used to train another DPSM. The rest of the cohort was used for testing. Fig. 8 depicts the landmarking results in terms of mean and standard deviations of normalized errors on 19 landmarks. Landmarks with less deformation in expressions were relatively better localized, (i.e., eye corner, nose tip, and nose corner). The middle of the lower lip was detected with the worst accuracy in the surprise expression because of the large mouth displacement and ample deformation in this region. Fig. 9 depicts the accumulative error distributions for each expression. Most landmarks were localized consistently across different expressions with errors ranging from 2% to 8%.

3) DPSM-Based Landmarking Evaluation on FRGC v2 Database: To evaluate the contribution of texture and depth information, respectively, we have also evaluated DPSM on

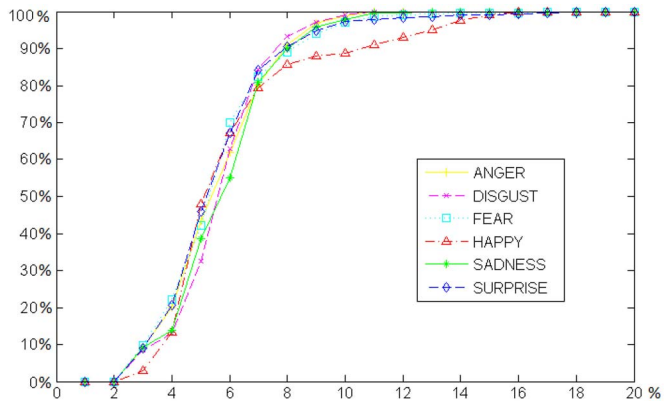


Fig. 9. Depiction of accumulated error distribution on landmarking over different expressions in Bosphorus database. The *y*-axis is the accumulated error percentage and the *x*-axis is the landmarking error.

TABLE II
LANDMARK RESULT COMPARISON ON FRGC V2 DATABASE

%	DPSM-D		DPSM-T		DPSM		CLM-Z [75]	
	Mean	STD	MEAN	STD	MEAN	STD	MEAN	STD
1	6.57	8.19	6.95	9.37	6.31	8.15	**6.12**	9.87
2	5.45	8.13	5.45	8.89	5.4	7.54	**4.48**	8.72
3	5.29	8.69	5.20	9.74	**4.88**	8.47	5.39	8.06
4	5.33	7.89	5.34	9.63	**4.92**	7.55	5.31	7.30
5	5.03	8.89	5.00	9.60	4.67	6.45	**4.21**	6.52
6	5.78	7.17	6.24	7.60	**5.71**	6.42	5.80	7.18
7	5.17	9.00	5.34	9.06	4.84	6.67	**3.22**	9.22
8	3.83	9.94	3.80	9.21	**3.57**	7.15	5.73	8.44
9	3.84	7.19	3.70	8.29	**3.35**	6.95	5.14	7.69
10	4.93	7.80	4.62	7.50	4.61	7.24	**3.16**	6.95
11	3.99	7.10	4.38	7.72	3.59	6.57	x	x
12	4.24	7.98	4.14	7.63	4.01	7.55	x	x
13	3.72	5.94	3.59	7.55	**3.50**	5.38	4.43	9.37
14	5.13	7.40	5.02	8.06	4.72	6.87	x	x
15	3.84	7.92	3.96	8.83	**3.80**	6.78	5.23	8.78
16	4.58	9.42	4.90	9.56	**4.51**	7.32	5.03	10.23
17	4.90	11.65	5.08	10.94	4.81	8.85	**3.84**	9.70
18	5.51	10.27	6.07	10.28	5.31	7.13	**4.38**	9.36
19	5.32	9.90	5.64	10.10	4.90	8.87	**4.16**	10.51
Avg.	4.87	8.45	4.97	8.92	4.60	7.26	4.73	8.62

FRGC v2 database. The results are depicted in Table II. In this test, the first 10% of the cohort were used in training and the rest 90% of the cohort were used in testing. Nineteen landmarks have been located using texture images, depth images and both (columns DPSM-D, DPSM-T, and DPSM). The average error using texture information is 4.87%, while the average error when using depth information only is 4.97%. When using both types of information, an average error of 4.60% has been achieved. It can be observed that texture images are more informative in landmarking compared to depth images. We further compared our method to the constrained local model-Z (CLM-Z) method [75] in column CLM-Z. The average landmarking error for CLM-Z is 4.73% for the 16 overlapped landmarks. Both methods have comparable results in mean accuracy, while the proposed method has a lower standard deviation. This is probably because of the object function and adoption of the simplex optimization method, which outputs localization results with fewer outliers and more consistency.

C. Results on Facial Expression Recognition

1) AU Coordinate Computation on Bosphorus Database: We have tested AU coordinate computation and depicted results in Table III. There is no RGB-D facial database containing all 44 types of AUs, thus we constructed the AU space using 25 types of single occurred AU data in Bosphorus database. The cohort utilized in this test includes all facial data displaying a single AU. We have carried out a tenfold person-independent cross-validation approach, where subjects in training and testing sets are mutually exclusive. Subjects were partitioned into two subsets in each round (totally ten rounds): 1) one with 90% subjects for training the AU detectors and 2) the other with 10% subjects for testing. To exclude the impact from landmarking errors, we have used manual landmarks in this test. For each AU class, the test data from the same AU were labeled as "genuine" and the test data from the different AU were labeled as "imposter." In the ideal AU space, the coordinates of the genuine data should be as high as possible while the coordinates of the imposter data should be as low as possible. In our test, values in genuine columns are the mean coordinates computed for the genuine data and values in imposter columns are the mean coordinates computed for imposter data. The average coordinates for the genuine data were close to one in Table III, which indicates accurate detection on genuine AU data. The average coordinates for the imposter data varied from 0.1249 at minimum to 0.5925 at maximum, and 19 out of 25 were below 0.5. Distinguishable gaps between two types of scores exist for all AUs. This validated the rationality and feasibility of the proposed coordinate computation for AU space.

TABLE III
AU COORDINATE COMPUTATION

Label	Genuine	Imposter	Label	Genuine	Imposter
AU1	0.9840	0.4236	AU17	0.9787	0.5813
AU2	0.9905	0.5073	AU18	0.9934	0.3792
AU4	0.9827	0.3638	AU20	0.9875	0.4336
AU5	0.9918	0.4326	AU22	0.9909	0.4439
AU6	0.9938	0.3769	AU23	0.9788	0.1513
AU7	0.9911	0.5636	AU24	0.9832	0.5925
AU9	0.9803	0.4929	AU25	0.9801	0.2769
AU10	0.9942	0.5384	AU26	0.9938	0.4907
AU11	0.9923	0.2768	AU27	0.9942	0.1249
AU12	0.9895	0.4197	AU28	0.9904	0.3326
AU14	0.9873	0.5222	AU34	0.9853	0.4755
AU15	0.9973	0.4813	AU43	0.9986	0.3432
AU16	0.9702	0.3142			

2) AU Detection Evaluation and Comparison: We have also compared the accuracy of AU coordinate computation with the literature in Fig. 10. This comparison followed AU verification scheme, where the coordinates along an axis were used to verify the occurrence of the target AUs. The receiver operating characteristic (ROC) area under the curve (AUC) were calculated from the AU coordinates computed previously in Table III. Since all the methods in comparison followed a tenfold person independent cross-validation approach, their results are directly comparable. Compared to the 3-D local binary pattern (3DLBP) method in [69] and the CS-3DLBP method in [70], the proposed AU coordinate computation achieved the highest AUC in 14 out of 26 AUs. These 14 AUs are mostly related to the eyebrows and the mouth, whose landmarks

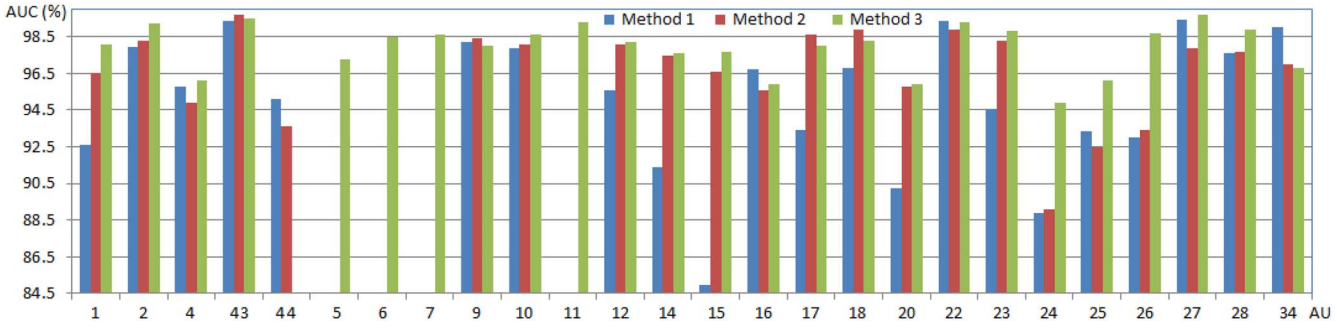


Fig. 10. ROC AUC achieved with different methods. Method 1 refers to the 3DLBP method in [69], method 2 refers to the CS-3DLBP method in [70], and method 3 is referred to AU coordinate computation proposed in this paper. AUC values are missing from the methods 1 and 2 for AU 5–AU 7 and AU 11, while AUC values are missing from the method 3 for AU 44. In 14 out of 26 AUs, the our coordinate computation method achieved the highest AUC.

TABLE IV
CONFUSION MATRIX OF THE EXPRESSION RECOGNITION ON BOSPHORUS DATABASE

Input \ Output	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutral
Anger	87.3/83.6%	9.4/10.7%	0.0/1.2%	0.0/0.0%	2.0/2.6%	0.0/1.3%	1.3/0.6%
Disgust	0.0/2.3%	92.6/85.7%	1.7/4.1%	4.5/4.5%	1.2/0.0%	0.0/1.7%	0.0/1.7%
Fear	0.0/2.6%	2.0/3.1%	83.2/78.8%	0.0/1.5%	3.6/2.8%	6.7/8.4%	4.5/2.8%
Happiness	0.0/0.0%	0.0/1.7%	0.0/0.0%	100.0/96.6%	0.0/0.0%	0.0/1.7%	0.0/0.0%
Sadness	2.3/4.1%	2.7/1.7%	0.0/0.0%	0.0/1.2%	91.4/85.2%	0.0/3.3 %	3.6/4.5%
Surprise	0.0/1.7%	0.0/0.0%	8.0/9.2%	0.0/0.0%	0.0/0.0%	92.0/89.1%	0.0/0.0%
Neutral	2.4/2.8%	0.5/1.2%	6.7/8.4%	0.0/0.0%	4.5/5.8%	2.3/1.7%	83.6/80.1%

Left values in each cell are the results obtained by using manual landmarks and right values are the results by using automatic landmarks.

often have significant displacements. Compared to the other two holistic methods, our features are extracted depending on facial landmarks and thus they can better capture the appearance changes when landmark displacement is relatively large, especially in the eye and mouth regions. It can also be observed that AUCs for all AUs have a smaller variance achieved by our method compared to the other two methods (1.3 by the proposed AU space method versus 3.8 by 3DLBP [69] and 2.5 by CS-3DLBP [70]), indicating the robustness of our method in computing coordinate.

3) Evaluation on FER Using AU Space: To test the effectiveness of interpreting facial affective states as AU space coordinates, we have conducted an experiment on a cohort of the Bosphorus database and depicted the results in Table IV. Facial data displaying the neutral state and the six universal expressions from all available subjects were selected to constitute this cohort. A 25 dimensional coordinate vector was computed for each face. A tenfold person-independent cross-validation approach was performed. In each fold, AU coordinates from the training facial data were used to determine the hyperplanes segmenting the AU space into seven regions corresponding to the neutral and the six universal emotions. The test AU coordinates were used to test if the emotions can be classified correctly. Each row in Table IV represents average rates across tenfolds which recognizes the facial expression in an input class into output classes. The values on the diagonal are the recognition rates, the values outside the diagonal are misclassified rates. Left values in each cell are the results when AU detectors used manual landmarks to extract facial features on this cohort data and right values are the results when AU detectors used automatic landmarks. Note that subjects used for training the DPSM were excluded in the FER testing when using automatic landmarks to avoid bias. The method achieved the best recognition rate 100%/96.6% when recognizing happiness for both cases, while achieved the lowest rate 83.2%/78.8% when recognizing fear. The average recognition rate for all seven expressions was 90.0% when applying the manual landmarks in AU detection, while the average recognition rate was 85.6% in the automatic case. Around 5% decrease in terms of the average recognition rate was caused by automatic landmarking errors. Also tested on Bosphorus database, Zernike-moments-based automatic FER in [71] achieved a 60.5% recognition rate using depth images only. Compared to it, the proposed AU space-based FER method has achieved a better recognition rate. The reasons for the improvement were probably because we used AU coordinates as an interpretation of facial expressions. Also we have used both texture and depth information in recognition.

To further evaluate the AU space-based emotion interpretation, we have compared the expression recognition rates between the AU space-based method and the binarization-based method in Fig. 11. In the binarization-based method, we set thresholds at their equal error rate in AU verification (referring to the AU verification experiment), resulting in a 25-D vector with only binary values for each face. These vectors can be further mapped to emotions using the FACSAID rules. We obtained these binary vectors from the AU coordinates computed with the manual landmark setup in the previous FER experiment. The mean recognition rate for binarization-based method on the six emotions is 88.6%, compared to 91.1% achieved by the AU space-based method.

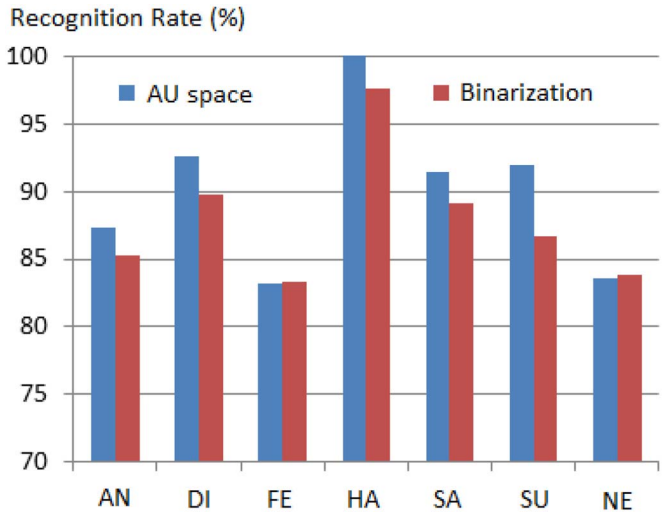


Fig. 11. Comparison on recognition rates for the six universal emotions between the AU space-based method and the binarization-based method. The x-axis represents the index to the six universal emotions: anger, disgust, fear, happiness, sadness, and surprise and the y-axis represents the recognition rates (RR). The blue bars represent the RR achieved by the AU space-based method and the red bars represent the RR achieved by the binarization-based method. The average RR for the former method is 91.1% compared to 88.6% achieved by the latter method.

This result demonstrates the AU space's tolerance of AU detection in margins, which resulted in a reduced misclassification rate of the emotions.

4) Evaluation on Feature Set Impact to AU Detection and FER: We have also compared the performance using different feature subsets in Table V. The AU coordinates and the hyperplanes were computed using the same setup as in Tables III and IV except the feature sets were different. In 3DLBP, multiscale 3DLBP features were extracted globally on depth maps in five scales. In LBP, multiscale LBP features were extracted globally on texture maps in five scales. L_d^{1-5} and L_t^{1-5} were the multiscale LBP feature subsets extracted locally from depth and texture patches in five different scales, as used in the original feature set. All represents all 14 features used in the original feature set. "All+3DLBP" represents all 14 features in the original feature set and multiscale 3DLBP features. All of these feature sets were incorporated into the proposed AU space method, respectively, and thus their results are directly comparable. Compared to local extracted features (L_t^{1-5} and L_d^{1-5}), global features (3DLBP and LBP), have achieved a better performance on both ROC AUC and recognition rate. Meanwhile, texture-based features (LBP and L_t^{1-5}) achieved better results than depth-based features (3DLBP and L_d^{1-5}). The best result in terms of AUC has been achieved using the all original features combined with 3DLBP feature. The best result in terms of RR has been achieved using the all original features.

5) Evaluation on Cross Database FER: To test the effectiveness of AU space-based FER method across different databases, we have conducted experiments to recognize three facial states on EURECOM Kinect face database with the AU detector trained from Bosphorus database. We used the facial data cohort and their automatic annotated landmarks in the first experiment, which is around 90% portion of the cohort. A tenfold person-independent cross-validation approach was performed. In each fold, AU coordinates from the training facial data were used to train the hyperplanes, while coordinates from the test facial data were used for recognizing their states. The classification results were depicted as the confusion matrix in Table VI. Facial data with neutral and open mouth states had recognition rates around 94% while facial data with the smile state were easier to confuse with the other two states, achieved only 86.6% recognition rate. The overall performance with an average recognition rate of 91.6% was satisfying. Moreover, the good recognition rate achieved on the nonuniversal expression state (i.e., open mouth) demonstrated the flexibility of the AU space being applied on facial states other than the six universal emotions.

TABLE V
PERFORMANCE COMPARISON ON DIFFERENT FEATURE SETS

	$3DLBP$	LBP	L_d^{1-5}	L_t^{1-5}	All	$All+3DLBP$
AuC (%)	88.7	90.5	86.4	87.6	97.3	97.4
RR (%)	85.5	88.3	83.2	87.8	91.1	90.8

AUC is the average Area under Curve over 25 AUs in the AU detection, and RR is the average Recognition Rate over the six universal expressions in facial expression recognition.

TABLE VI
CONFUSION MATRIX OF FER ON EURECOM DATABASE

	Neutral	Smile	Open Mouth
Neutral	94.3%	4.2%	1.5%
Smile	4.6%	86.6%	8.8%
Mouth Open	3.7%	2.5%	93.8%

VI. DISCUSSION ON FUTURE APPLICATIONS

A. Wearable Social Assistant Device

Wearable social assistant devices can be developed with support from augmented reality techniques. A hardware design of such devices include a RGB-D camera, a computation unit (e.g., a smart phone, a tablet, or a computerized wearable devices) and an user interface (a transparent head-mounted display (HMD) and/or a earphone), as depicted in Fig. 12. Enabled by this platform, the wearable device will analyze egocentric visual content in users' daily lives and annotate facial affective states on others' faces using the proposed landmarking and FER methods. The annotations can be provided to people with LV, ASD, and AD conditions by either overlapping a virtual text box on the actual faces through the transparent HMDs or reading out the states from the headphone.

B. 2.5-D Video Chatting

As announced in consumer electronics show (CES) 2014, tablets and laptops with integrated RGB-D camera modules will emerge in the commercial market in the middle of 2014. This may bring a revolution on video chatting applications, such as FaceTime or Skype. 2.5-D facial data can be captured and transmitted during chatting sessions. Based on such platforms, the proposed algorithms can provide affective computing functionality for these applications. During chatting sessions, people with LV, ASD, and AD conditions can benefit

Fig. 12. Demonstration on the hardware of the social assistant devices [68].

from the emotion annotations on other chatting participants, which can either be delivered as augmented visual signals on the screen or translated into signals in other sensory channels, such as voice or touch.

VII. Conclusion

To address the needs on automatic emotion annotation of people with AD, LV, and ASD conditions, we have proposed a 2.5-D facial landmarking method and an AU space-based FER method. These two methods work as a chain to automatic annotating visual data collected from RGB-D cameras. In landmarking method, DPFM is built and fitted to new faces to localize the facial landmarks. Facial features are then extracted by AU detectors based on these landmarks and further fed into coordinate computation in the proposed AU space. These coordinates are employed to recognize the emotions displayed on the face. Evaluated on three publicly accessible databases, the landmarking and facial emotion recognition methods have achieved satisfactory results. We further discussed two kinds of emotion AT which possibly can adopt the proposed algorithms to help people in their daily social lives.

References

[1] R. Adolphs, L. Sears, and J. Piven, "Abnormal processing of social information from faces in autism," *J. Cogn. Neurosci.*, vol. 13, no. 2, pp. 232–240, 2001.

[2] R. Hargrave, R. J. Maddock, and V. Stone, "Impaired recognition of facial expressions of emotion in Alzheimer's disease," *J. Neuropsych. Clin. Neurosci.*, vol. 14, no. 1, pp. 64–71, 2002.

[3] S. Krishna, D. Colbry, J. Black, V. Balasubramanian, and S. Panchanathan, "A systematic requirements analysis and development of an assistive device to enhance the social interaction of people who are blind or visually impaired," in *Proc. Workshop Comput. Vis. Appl. Vis. Impair.*, Anchorage, AK, USA, Jun. 2008, pp. 1–6.

[4] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognit.*, vol. 36, no. 1, pp. 259–275, Jan. 2003.

[5] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[6] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, Southampton, U.K., Apr. 2006, pp. 211–216.

[7] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, Amsterdam, The Netherlands, Sep. 2008, pp. 1–6.

[8] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, May 2009.

[9] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris, "3D facial expression recognition: A perspective on promises and challenges," in *Proc. Int. Conf. Workshops Autom. Face Gesture Recognit.*, Santa Barbara, CA, USA, Mar. 2011, pp. 603–610.

[10] (2014). *Creative Senz3D*. [Online]. Available: http://us.creative.com/p/web-cameras/creative-senz3D

[11] (2014). *Intel Realsense RGB-D Camera*. [Online]. Available: http://itersnews.com/?p=63992

[12] (2010). *Prototype of Wearable RGB-D Camera Glass*. [Online]. Available: http://www.ubergizmo.com/2010/06/sony-prototypes-eye-tracking-glasses-for-lifeblogging/

[13] T. Huynh, R. Min, and J.-L. Dugelay, "An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data," in *Proc. Asian Conf. Comput. Vis. Workshops*, Daejeon, Korea, Nov. 2012, pp. 133–145.

[14] X. Zhao, E. Dellandrea, L. Chen, and I. A. Kakadiaris, "Accurate landmarking of three-dimensional facial data in the presence of facial expressions and occlusions using a three-dimensional statistical facial feature model," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 5, pp. 1417–1428, Oct. 2011.

[15] C. B. Peterson, N. R. Prasad, and R. Prasad, "The future of assistive technologies for dementia," *Gerontechnology*, vol. 11, no. 2, pp. 195–204, 2012.

[16] S. Krishna, D. Colbry, J. Black, V. Balasubramanian, and S. Panchanathan, "A systematic requirements analysis and development of an assistive device to enhance the social interaction of people who are blind or visually impaired," in *Proc. Workshop Comput. Vis. Appl. Vis. Impair.*, Marseille, France, Oct. 2008, pp. 1–6.

[17] S. Krishna and S. Panchanathan, "Assistive technologies as effective mediators in interpersonal social interactions for persons with visual disability," in *Proc. Comput. Help. People Spec. Needs*, Vienna, Austria, Jul. 2010, pp. 316–323.

[18] C. Liu, K. Conn, N. Sarkar, and W. Stone, "Physiology-based affect recognition for computer-assisted intervention of children with autism spectrum disorder," *Int. J. Human-Comput. Stud.*, vol. 66, no. 9, pp. 662–677, Sep. 2008.

[19] A. L. Wainer and B. R. Ingersoll, "The use of innovative computer technology for teaching social communication to individuals with autism spectrum disorders," *Res. Autism Spectr. Disord.*, vol. 5, no. 1, pp. 96–107, Jan./Mar. 2011.

[20] L. Cruz, D. Lucio, and L. Velho, "Kinect and RGBD images: Challenges and applications," in *Proc. 25th SIBGRAPI Conf. Graph. Patterns Images Tuts.*, Ouro Preto, Brazil, Sep. 2012, pp. 36–49.

[21] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1318–1334, Oct. 2013.

[22] S. Izadi *et al.*, "KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera," in *Proc. 24th ACM Symp. User Interf. Softw. Technol.*, Santa Barbara, CA, USA, Sep. 2011, pp. 559–568.

[23] Z. Ren, J. Meng, J. Yuan, and Z. Zhang, "Robust hand gesture recognition with Kinect sensor," in *Proc. 19th ACM Int. Conf. Multimedia*, Scottsdale, AZ, USA, Nov. 2011, pp. 759–760.

[24] Y.-L. Chen, H.-T. Wu, F. Shi, X. Tong, and J. Chai, "Accurate and robust 3D facial capture using a single RGBD camera," in *Proc. Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 3615–3622.

[25] T. Weise, S. Bouaziz, H. Li, and M. Pauly, "Realtime performance-based facial animation," *ACM Trans. Graph.*, vol. 30, no. 4, p. 77, 2011.

[26] X. Zhang *et al.*, "A high-resolution spontaneous 3D dynamic facial expression database," in *Proc. Int. Conf. Workshops Autom. Face Gesture Recognit.*, Shanghai, China, Apr. 2013, pp. 1–6.

[27] M. Blum, J. T. Springenberg, J. Wulfing, and M. Riedmiller, "A learned feature descriptor for object recognition in RGB-D data," in *Proc. IEEE Int. Conf. Robot. Autom.*, St. Paul, MN, USA, May 2012, pp. 1298–1303.

[28] M. Szwoch, "FEEDB: A multimodal database of facial expressions and emotions," in *Proc. 6th Int. Conf. Human Syst. Interact.*, Sopot, Poland, Jun. 2013, pp. 524–531.

[29] R. I. Hg *et al.*, "An RGB-D database using Microsoft's Kinect for windows for face detection," in *Proc. Int. Conf. Signal Image Technol. Internet Based Syst.*, Naples, Italy, Nov. 2012, pp. 42–46.

[30] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *Proc. IEEE Int. Conf. Robot. Autom.*, Shanghai, China, May 2011, pp. 1817–1824.

[31] B. Ni, G. Wang, and P. Moulin, "RGBD-HuDaAct: A color-depth video database for human daily activity recognition," in *Consumer Depth Cameras for Computer Vision*, London, U.K.: Springer, Dec. 2013, pp. 193–208.

[32] I. Kakadiaris *et al.*, "3D face recognition in the presence of facial expressions: An annotated deformable model approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 640–649, Apr. 2007.

[33] C. Xua, T. Tana, Y. Wang, and L. Quanc, "Combining local features for robust nose location in 3D facial data," *Pattern Recognit. Lett.*, vol. 27, no. 13, pp. 1487–1494, Oct. 2006.

[34] J. D'House, J. Colineau, C. Bichon, and B. Dorizzi, "Precise localization of landmarks on 3D faces using Gabor wavelets," in *Proc. Int. Conf. Biometr. Theory Appl. Syst.*, Arlington, VA, USA, Sep. 2007, pp. 1–6.

[35] V. Bevilacqua, P. Casorio, and G. Mastronardi, "Extending Hough transform to a points cloud for 3D-face nose-tip detection," in *Proc. Int. Conf. Adv. Intell. Comput. Theor. Appl.*, Shanghai, China, Sep. 2008, pp. 1200–1209.

[36] H. Dibeklioglu, A. A. Salah, and L. Akarun, "3D facial landmarking under expression, pose, and occlusion variations," in *Proc. Int. Conf. Biometr. Theory Appl. Syst.*, Arlington, VA, USA, Sep. 2008, pp. 1–6.

[37] P. Szeptycki, M. Ardabilian, and L. Chen, "A coarse-to-fine curvature analysis-based rotation invariant 3D face landmarking," in *Proc. IEEE 3rd Int. Conf. Biometr. Theory Appl. Syst.*, Washington, DC, USA, 2009, pp. 1–6.

[38] X. Lu, A. K. Jain, and D. Colbry, "Matching 2.5D face scans to 3D models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 31–43, Jan. 2006.

[39] P. Nair and A. Cavallaro, "3-D face detection, landmark localization, and registration using a point distribution model," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 611–623, Jun. 2009.

[40] S. Berretti, A. D. Bimbo, P. Pala, B. B. Amor, and M. Daoudi, "A set of selected SIFT features for 3D facial expression recognition," in *Proc. Int. Conf. Pattern Recognit.*, Istanbul, Turkey, Aug. 2010, pp. 4125–4128.

[41] H. Tang and T. S. Huang, "3D facial expression recognition based on automatically selected features," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit. Workshops*, Anchorage, AK, USA, Jun. 2008, pp. 1–8.

[42] J. Wang, L. Yin, X. Wei, and Y. Sun, "3D facial expression recognition based on primitive surface feature distribution," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2006, pp. 1399–1406.

[43] I. Mpiperis, S. Malassiotis, and M. Strintzis, "Bilinear models for 3-D face and facial expression recognition," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 3, pp. 498–511, Sep. 2008.

[44] H. Soyel and H. Demirel, "3D facial expression recognition with geometrically localized facial features," in *Proc. Symp. Comput. Sci. Inf. Technol.*, Istanbul, Turkey, Oct. 2008, pp. 1–4.

[45] S. Ramanathan, A. Kassim, Y. V. Venkatesh, and W. S. Wah, "Human facial expression recognition using a 3D morphable model," in *Proc. IEEE Int. Conf. Image Process.*, Atlanta, GA, USA, Oct. 2006, pp. 661–664.

[46] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA, USA: Consult. Psychol., 1978.

[47] (1983). *Emotion Facs (emfacs)*. [Online]. Available: http://face-and-emotion.com/dataface/general/homepage.jsp

[48] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1683–1699, Oct. 2007.

[49] A. Savran and B. Sankur, "Automatic detection of facial actions from 3D data," in *Proc. Int. Conf. Comput. Vis. Workshop Human Comput. Interact.*, Kyoto, Japan, Sep./Oct. 2009, pp. 1993–2000.

[50] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1940–1954, Nov. 2010.

[51] K.-T. Song and S.-C. Chien, "Facial expression recognition based on mixture of basic expressions and intensities," in *Proc. Int. Conf. Syst. Man Cybern.*, Seoul, Korea, Oct. 2012, pp. 3123–3128.

[52] G. Sandbach, S. Zafeiriou, and M. Pantic, "Binary pattern analysis for 3D facial action unit detection," in *Proc. Brit. Mach. Vis. Conf.*, Guildford, U.K., Jun. 2012, pp. 1–12.

[53] A. Savran, B. Sankur, and M. T. Bilge, "Regression-based intensity estimation of facial action units," *Image Vis. Comput.*, vol. 30, no. 10, pp. 774–784, Oct. 2012.

[54] P. Soille, *Morphological Image Analysis: Principles and Applications*. Berlin, Germany: Springer, 1999, pp. 173–174.

[55] Pittsburgh Pattern Recognition. (Mar. 2011). *PittPatt Face Recognition Software Development Kit (PittPatt SDK) v5.2*. [Online]. Available: http://www.pittpatt.com

[56] X. Zhao, S. K. Shah, and I. A. Kakadiaris, "Illumination alignment using lighting ratio: Application to 3D-2D face recognition," in *Proc. Int. Conf. Workshops Autom. Face Gesture Recognit.*, Shanghai, China, Apr. 2013, pp. 1–6.

[57] T. F. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models—Their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, Jan. 1995.

[58] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, Sep. 2003.

[59] J. P. Lewis, "Fast normalized cross-correlation," *Vis. Interf.*, vol. 10, no. 1, pp. 120–123, 1995.

[60] J. Nelder and R. Mead, "A simplex method for function minimization," *Comput. J.*, vol. 7, no. 4, pp. 308–313, 1965.

[61] P. Ekman, E. Rosenberg, and J. Hager. (1998). *Facial Action Coding System Affect Interpretation Database (FACSAID)*. [Online]. Available: http://nirc.com/Expression/FACSAID/facsaid.html

[62] C. Shan and T. Gritti, "Learning discriminative LBP-histogram bins for facial expression recognition," in *Proc. Brit. Mach. Vis. Conf.*, Leeds, U.K., Sep. 2008, pp. 1–10.

[63] C. Dorai and A. Jain, "COSMOS—A representation scheme for 3D free-form objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 10, pp. 1115–1130, Oct. 1997.

[64] C. C. Chang and C. J. Lin. (2001). *LIBSVM: A Library for Support Vector Machines*. [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm

[65] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[66] A. Savran *et al.*, "Bosphorus database for 3D face analysis," in *Proc. 1st COST 2101 Workshop Biometr. Identity Manage.*, Roskilde, Denmark, 2008, pp. 47–56.

[67] M. Mahmoud, T. Baltrušaitis, P. Robinson, and L. D. Riek, "3D corpus of spontaneous complex mental states," in *Proc. Affect. Comput. Intell. Interact.*, Memphis, TN, USA, Sep. 2011, pp. 205–214.

[68] (2014). *Spaceglass, Meta.01 Developer Edition*. [Online]. Available: https://www.spaceglasses.com/buy

[69] G. Sandbach, S. Zafeiriou, and M. Pantic, "Local normal binary patterns for 3D facial action unit detection," in *Proc. Int. Conf. Image Process.*, Orlando, FL, USA, Sep./Oct. 2012, pp. 1813–1816.

[70] N. Bayramoglu, G. Zhao, and M. Pietikainen, "CS-3DLBP and geometry based person independent 3D facial action unit detection," in *Proc. Int. Conf. Biometr.*, Madrid, Spain, Jun. 2013, pp. 1–6.

[71] N. Vretos, N. Nikolaidis, and I. Pitas, "3D facial expression recognition using Zernike moments on depth images," in *Proc. Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 773–776.

[72] G. Tzimiropoulos and M. Pantic, "Gauss-Newton deformable part models for face alignment in-the-wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 1851–1858.

[73] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 FPS via regressing local binary features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 1685–1692.

[74] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 1867–1874.

[75] T. Baltrusaitis, P. Robinson, and L. Morency, "3D constrained local model for rigid and non-rigid facial tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, Jun. 2012, pp. 2610–2617.

[76] G. Fanelli, M. Dantone, and L. Van Gool, "Real time 3D face alignment with random forests-based active appearance models," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Shanghai, China, Apr. 2013, pp. 1–8.

[77] Q. Cai, D. Gallup, C. Zhang, and Z. Zhang, "3D deformable face tracking with a commodity depth camera," in *Proc. Eur. Conf. Comput. Vis.*, Crete, Greece, Sep. 2010, pp. 229–242.

[78] E. Antonakos, J. Alabort-Medina, G. Tzimiropoulos, and S. Zafeiriou, "HOG active appearance models," in *Proc. IEEE Int. Conf. Image Process.*, Paris, France, Oct. 2014, pp. 224–228.

[79] S. Cheng, S. Zafeiriou, A. Asthana, and M. Pantic, "3D facial geometric features for constrained local models," in *Proc. Eur. Conf. Comput. Vis.*, Crete, Greece, Sep. 2010, pp. 1–6.

[80] P. J. Phillips *et al.*, "Overview of the face recognition grand challenge," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 947–954.

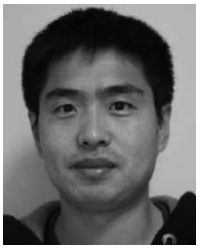
Xi Zhao (S'08–M'10) received the Ph.D. (Hons.) degree in computer science from Ecole Centrale de Lyon, Lyon, France, in 2010.

He conducted research in the fields of biometrics, face analysis, and pattern recognition, as a Research Assistant Professor with the Department of Computer Science, University of Houston, USA. He is currently an Associate Professor with Xi'an Jiaotong University, Xi'an, China. His current research interests include biometrics, affective computing, data analysis, mobile computing, computer vision, healthcare computing, and assistive technology.

Dr. Zhao currently serves as a Reviewer of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON INFORMATION, FORENSICS AND SECURITY, *Image and Vision Computing*, the IEEE International Conference on Automatic Face and Gesture Recognition, the IEEE International Conference on Advanced Video and Signal-Based Surveillance, and the International Conference on Pattern Recognition. He has served as the Co-Chair of the International Conference on Biometrics: Theory, Applications and Systems in 2013, the Program Committee of Workshop on 3-D Face Biometrics in 2013, and the International Conference on Affective Computing and Intelligent Interaction in 2015.

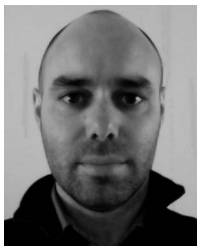
Jianhua Zou (M'10) received the Ph.D. degree with the Institute of Plasma Physics, Chinese Academy of Sciences, Beijing, China, in 1991

He is the Vice Dean with the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China. He conducted the post-doctoral research with the Huazhong University of Science and Technology, Wuhan, China, and Xi'an Jiaotong University in 1991–1993 and 1993–1995, respectively. He became a Professor with the System Engineering Institute, Xi'an Jiaotong University. He was the Principle Investigator of two projects funded by the National Science Foundation of China, about ten key projects supported by industry and military. As a Senior Visiting Scholar, he has established partnerships with Eindhoven University, Eindhoven, The Netherlands, and Philips Company, Amsterdam, The Netherlands. His current research interests include intelligent control, computer vision and pattern recognition, data mining, and knowledge discovery. He has published over 60 academic papers.



Huibin Li (S'09) received the B.S. degree in mathematics from Shaanxi Normal University, Xi'an, China, in 2006, the M.S. degree in mathematics from Xi'an Jiaotong University, Xi'an, in 2009, and the Ph.D. degree in computer science from Ecole Centrale de Lyon, Lyon, France, in 2013.

He jointed the Institute for Information and System Sciences, School of Mathematics and Statistics, Xi'an Jiaotong University. His current research interests include affective computing, biometrics, and 3-D shape analysis and recognition.



Emmanuel Dellandréa (M'06) received the master's and engineer degrees and the Ph.D. degree from Université de Tours, Tours, France, in 2000 and 2003, respectively, all in computer science.

He joined Ecole Centrale de Lyon, Lyon, France, as an Associate Professor in 2004. His current research interests include multimedia analysis, affective computing, and particularly affect recognition both in image and audio signals as well as facial expression analysis.

Ioannis A. Kakadiaris (SM'09) received the B.Sc. degree in physics from the University of Athens, Athens, Greece, the M.Sc. degree in computer science from Northeastern University, Boston, MA, USA, and the Ph.D. degree in computer science from the University of Pennsylvania, Philadelphia, PA, USA.

He was a Post-Doctoral Fellow with the University of Pennsylvania. He joined the University of Houston (UH), Houston, TX, USA, in 1997, where he is a Hugh Roy and Lillie Cranz Cullen University Professor of Computer Science, Electrical and Computer Engineering, and Biomedical Engineering and the Founder of the Computational Biomedicine Laboratory. His current research interests include multimodal biometrics, face recognition, computer vision, pattern recognition, biomedical image analysis, and predictive analytics.

Dr. Kakadiaris was a recipient of a number of awards, including the NSF Early Career Development Award, the Schlumberger Technical Foundation Award, the UH Computer Science Research Excellence Award, the UH Enron Teaching Excellence Award, and the James Muller Vulnerable Plaque Young Investigator Prize. His research has been featured on The Discovery Channel, National Public Radio, KPRC NBC News, KTRH ABC News, and KHOU CBS News. He held selected professional service leadership positions, such as the General Co-Chair of the 2013 Biometrics: Theory, Applications and Systems Conference and the 2014 SPIE Biometric and Surveillance Technology for Human and Activity Identification, the Program Co-Chair of the 2015 International Conference on Automatic Face and Gesture Recognition Conference, and the Vice-President for Technical Activities of the IEEE Biometrics Council.

Liming Chen (SM'14) received the joint B.Sc. degree in mathematics and computer science from the University of Nantes, Nantes, France, in 1984, and the master's and Ph.D. degrees in computer science from the University of Paris 6, Paris, France, in 1986 and 1989, respectively.

He first served as an Associate Professor with the Université de Technologie de Compiègne, Compiègne, France, then joined Ecole Centrale de Lyon (ECL), Lyon, France, as a Professor in 1998, where he leads an Advanced Research Team of Multimedia Computing and Pattern Recognition. He has been the Head of the Department of Mathematics and Computer Science, ECL since 2007. His current research interests include computer vision and multimedia, in particular 2-D/3-D face analysis and recognition, image and video analysis and categorization, and affective computing both in image and audio and video.