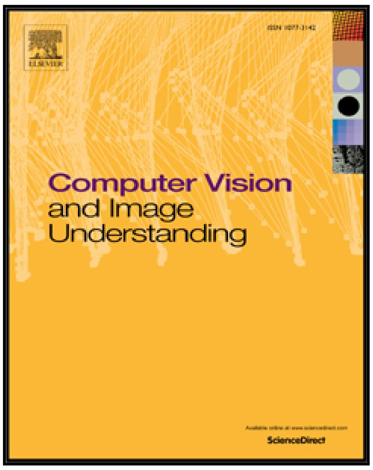
An Efficient Multimodal 2D + 3D Feature-based Approach to Automatic Facial Expression Recognition



An Efficient Multimodal 2D + 3D Feature-based Approach to Automatic Facial Expression Recognition

Huibin Li, Huaxiong Ding, Di Huang, Yunhong Wang, Xi Zhao,
Jean-Marie Morvan, Liming Chen

PII: S1077-3142(15)00158-7
DOI: 10.1016/j.cviu.2015.07.005
Reference: YCVIU 2286

To appear in: *Computer Vision and Image Understanding*

Received date: 14 June 2014
Revised date: 5 July 2015
Accepted date: 9 July 2015

Please cite this article as: Huibin Li, Huaxiong Ding, Di Huang, Yunhong Wang, Xi Zhao, Jean-Marie Morvan, Liming Chen, An Efficient Multimodal 2D + 3D Feature-based Approach to Automatic Facial Expression Recognition, *Computer Vision and Image Understanding* (2015), doi: 10.1016/j.cviu.2015.07.005

Highlights

- We propose a feature-based 2D+3D multimodal facial expression recognition method.
- It is fully automatic benefit from a large set of automatically detected landmarks.
- The complementarities between 2D and 3D features are comprehensively demonstrated.
- Our method achieves the best accuracy on the BU-3DFE database so far.
- A good generalization ability is shown on the Bosphorus database.

An Efficient Multimodal 2D + 3D Feature-based Approach to Automatic Facial Expression Recognition

Huibin Li[a], Huaxiong Ding[b], Di Huang[c,*], Yunhong Wang[c], Xi Zhao[d], Jean-Marie Morvan[e,f], Liming Chen[b]

[a]*School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China*
[b]*Ecole Centrale de Lyon, LIRIS UMR5205, Lyon, France*
[c]*State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing, China*
[d]*School of Management, Xi'an Jiaotong University, Xi'an, China*
[e]*Université Lyon 1, Institut Camille Jordan, Lyon, France*
[f]*GMSV Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia*

Abstract

We present a fully automatic multimodal 2D + 3D feature-based facial expression recognition approach and demonstrate its performance on the BU-3DFE database. Our approach combines multi-order gradient-based local texture and shape descriptors in order to achieve efficiency and robustness. First, a large set of fiducial facial landmarks of 2D face images along with their 3D face scans are localized using a novel algorithm namely incremental Parallel Cascade of Linear Regression (iPar-CLR). Then, a novel Histogram of Second Order Gradients (HSOG) based local image descriptor in conjunction with the widely used first-order gradient based SIFT descriptor are used to describe the local texture around each 2D landmark. Similarly, the local geometry around each 3D landmark is described by two novel local shape descriptors constructed using the first-order and the second-order surface differential geometry quantities, i.e., Histogram of mesh Gradients (meshHOG) and Histogram of mesh Shape index (curvature quantization, meshHOS). Fi-

[*]Corresponding author
Email addresses: huibinli@mail.xjtu.edu.cn (Huibin Li), huaxiong.ding@ec-lyon.fr (Huaxiong Ding), dhuang@buaa.edu.cn (Di Huang), yhwang@buaa.edu.cn (Yunhong Wang), zhaoxi1@mail.xjtu.edu.cn (Xi Zhao), morvan@math.univ-lyon1.fr (Jean-Marie Morvan), liming.chen@ec-lyon.fr (Liming Chen)

nally, the Support Vector Machine (SVM) based recognition results of all 2D and 3D descriptors are fused at both feature-level and score-level to further improve the accuracy. Comprehensive experimental results demonstrate that there exist impressive complementary characteristics between the 2D and 3D descriptors. We use the BU-3DFE benchmark to compare our approach to the state-of-the-art ones. Our multimodal feature-based approach outperforms the others by achieving an average recognition accuracy of 86.32%. Moreover, a good generalization ability is shown on the Bosphorus database.

Keywords: facial expression recognition, local texture descriptor, local shape descriptor, multimodal fusion

1. Introduction

Affect recognition aims to determine an individual's emotion by detecting and measuring the emotion related physiological (e.g., bodily symptoms), psychological (e.g., feelings) or behavioral (e.g., facial expression) characteristics [1], [2]. As an easily detectable, collectible, and measurable emotion component, facial expression is ideal for affect recognition and for human-computer interaction related applications [3]. However, Facial Expression Recognition (FER) is a very challenging problem mainly because of the diversity and hybridity of human expressions among different subjects in different cultures, genders and contexts.

In the past decades, a large number of FER approaches have been proposed. They can be categorized from three perspectives, namely the data modality, expression granularity, and temporal dynamics. From the first perspective, they are classified into 1) 2D FER (which uses 2D gray or color face images), 2) 3D FER (which uses 3D range images, point clouds, or meshes of faces), and 3) multimodal 2D + 3D FER (which uses both 2D and 3D facial data). From the second perspective, they are divided into 1) six basic facial expression (i.e., anger, disgust, fear, happiness, sadness, and surprise) recognition, 2) facial Action Unit (AU, e.g., brow raiser, lip tightener, and mouth stretch) detection and recognition. From the third perspective, they are categorized into static (still images) and dynamic (image sequences) FER. In this paper, we focus on the problem of recognizing the six basic facial expressions using multimodal 2D + 3D static images.

Appearance-based 2D FER has been widely investigated since 1990s [3]. The main research topics lie in three aspects: face detection, expression related feature extraction and classification. Comprehensive surveys of 2D FER approaches are given in [4], [3]. They are mainly classified into two categories, i.e., template-based and feature-based [4]. Template-based approaches usually fit a holistic face model to the input image or track it in the input image sequence. Active appearance model [5], point distribution model [6], mixture of probabilistic PCA [7], and topographic modeling [8] are some typical examples. Feature-based approaches generally localize the features of an analytic face model in the input image or track them in the input sequence. Gabor wavelets [9] and Local Binary Patterns (LBP) [10] based face representations are two popular representatives. Although considerable advancements have been achieved, 2D FER is still very challenging mainly due to its sensitivity to illumination, pose variations, and possible occlusions [4], [3].

Recently, with the rapid development of 3D imaging and scanning technologies, it becomes more and more popular to capture 3D face scans. Comparing with 2D face images, 3D face scans contain precise geometric shape information of facial surfaces, which is robust to illumination and pose variations, but more sensitive to facial expression changes. Thus, shape-based 3D FER has attracted increasing attentions. Similar to 2D, 3D FER approaches can also be categorized into template-based and feature-based. Template-based approaches usually build a parametric deformable face model first, and then extract the model parameters as expression features for recognition. 3D morphable model [11], bilinear deformable model [12], shape deformation model [13], and statistical feature model [14] are some famous examples. The main drawback of template-based approaches lies in that they require to establish one-to-one correspondence between 3D face scans, which is still a very challenging issue. Meanwhile, time consuming procedures like dense 3D face registration and model fitting are indispensable. Feature-based approaches generally extract 3D expression cues around facial landmarks using different facial surface geometric or differential quantities. For example, the distances between 3D facial landmarks are widely used in [15], [16], [17], and [18]. Moreover, 3D facial curves [19], facial geometry images and normal maps [20], [21] facial conformal images [22], facial surface normal [23], [24] and curvatures [23], [25], [24], and local depth-SIFT features [26] are some popular expression features. Feature-based approaches generally perform better than template-based ones. However, the bottleneck of feature-based approaches lies in accurate and robust 3D facial landmark localization, which is still a very difficult task [27]. More detailed surveys of 3D facial expression recognition are given in [28], [29].

Although the effectiveness of multimodal 2D + 3D face recognition has been well presented as in [30], [31], the investigation of multimodal 2D + 3D FER is very limited. Wang et al. [25] compared the FER accuracy of 3D primitive surface feature distribution based approach with 2D Gabor-wavelet and Topographic Context based ones on the BU-3DFE database, and found that 3D shape based approach is superior to 2D ones, especially for non-frontal faces. However, the effectiveness of combing 3D and 2D approaches was not discussed. Zhao et al. [14] used both 2D features (RGB values and LBP) and 3D features (3D coordinates and shape index values) in the 3D statistical feature model for prototypical expression recognition. But the results using only 2D features or 3D features were not reported, and thus the complementarity between 2D and 3D features was also not studied. In [32], the authors used both 2D and 3D dynamic data for real-time facial action and expression recognition. More precisely, they first extended the active shape model to handle 3D data for facial feature tracking. Then, they extracted numerous geometric measurements (e.g., the distances between landmarks and the boundary shape of lips) and surface deformation measurements (e.g., image gradient and surface curvature descriptors). Finally, the Rule Classifier was used for recognizing a subset of 11 important AUs and 4 facial expressions (i.e., happy, sad, surprise, disgust) on a dataset consisting of 832 sequences of 52 participants. Their experimental results demonstrated that the proposed 2D+3D algorithm performed much better than the 2D appearance-based algorithm (i.e., 2D ASM + Gabor filters + LDA) for recognizing the four facial expressions. This is a very illuminating approach for 2D+3D multimodal FER. However, they did not report the performance of each modality under their own framework. The importance of each modality is still unclear. Savran et al. [33] utilized multimodal 2D + 3D face data for facial AU detection. They found that 3D data generally perform better than 2D data, especially for lower AUs. Moreover, the fusion of two modalities can improve the detection rates from 93.5% (2D) and 95.4% (3D) to 97.1% (2D+3D). Except for facial AU detection and expression recognition, Wang et al. [34] quantified facial expression abnormality in Schizophrenia by combining 2D and 3D features. Their experimental results demonstrated that the combined features better characterized facial expressions than either individual 3D geometric or 2D texture features.

The above studies have preliminarily proved the fact that the combination of 2D and 3D data is better than either of the single 2D or 3D modality for expression characterization and AU detection, but deep analysis of the superiority for multimodal 2D+3D FER is still missing. An advantage of using 2D data is that it can be used to accurately localize a large set of facial landmarks on 2D face images and further on their 3D face scans due to the 2D-3D correspondence, which is the first contribution of this paper. More precisely, we propose to explore the incremental Parallel Cascade of Linear Regression (iPar-CLR) algorithm [35] to automatically localize 49 landmarks for each 2D face image and its corresponding 3D mesh scan. This large set of expression related landmarks are then used for extracting local texture and shape descriptors for expression classification. To the best of our knowledge, this is the first work which uses such large number of automatically detected landmarks for 2D and 3D multimodal FER. In contrast, the majority of existing feature-based 3D FER approaches reported their results on the BU-3DFE benchmark based on a large set of (typically 83) 3D facial landmarks manually localized by the database providers [15], [16], [17], [18], [19], [23], [25], [26]. Therefore, the proposed framework presents a promising way to these landmark-based approaches so that they can be made automatic using the iPar-CLR algorithm in 2D and 3D multimodal face space.

The second contribution of this paper is that a novel second-order image gradient based local texture descriptor (HSOG), a novel first-order mesh gradient (i.e., surface normal) based local shape descriptor (meshHOG), as well as a second-order mesh gradient (i.e., surface curvature) based local shape descriptor (meshHOS) are adapted in FER to comprehensively encode the expression variations in both the 2D and 3D modalities. According to our previous work [36], most of existing popular local image descriptors, such as HOG, LBP, and SIFT, only employ the first-order gradient information related to the slope and the elasticity, i.e., length, area, etc. when the image is regarded as a surface, and thereby partially characterize its geometric properties. By contrast, HSOG captures the curvature related cues of the surface, i.e., cliffs, ridges, summits, valleys, basins, and so on. Thus, HSOG can be applied to describe facial expression deformations (e.g., mouth stretch, lip stretcher, brow raiser). Moreover, in that paper, it was also demonstrated that HSOG outperformed the first-order gradient based local image descriptors (i.e. HOG, LBP, SIFT) when there were not severe scale variations, as in the applications of local image matching and scene classification. In this paper, we give another evidence of the effectiveness and generalization ability of HSOG for FER. Similarly, as general local shape descriptors, meshHOG and meshHOS provide a compact description of the facial surface normal and curvature information, and they have proved very efficient for 3D face identification in our previous works [37], [38]. In this paper, we interested in exploring their generalization abilities in 3D FER.

During the FER stage, both the early fusion (i.e., feature-level) and late fusion (i.e., score-level) strategies of 2D descriptors, 3D descriptors, as well as 2D and 3D descriptors are comprehensively demonstrated and their complementary characteristics are well revealed, which is our third contribution. The important findings behind the fusion results can be summarized as: 1) The second-order gradient based local texture or shape descriptor (HSOG or meshHOS) generally have stronger discriminative power than the first-order gradient based ones (SIFT or meshHOG). Moreover, different order 2D or 3D descriptors are complementary in encoding local texture or shape cues. 2) There exist large complementary characteristics between 2D and 3D descriptors of the same order (SIFT and meshHOG, HSOG and meshHOS), different order (SIFT and meshHOS, HSOG and meshHOG), as well as multiple orders (all four 2D and 3D descriptors).

Overall, we present an efficient multimodal (2D and 3D) and multiple-order (first and second) feature-based fully automatic FER approach, and validate it trough comprehensive experiments on the BU-3DFE database. Considerable complementary characteristics between the features of different orders and different modalities are highlighted either by early fusion or late fusion of 2D, 3D, as well as 2D and 3D descriptors. The generalization capability of our approach is further evaluated on the Bosphorus database.

This paper is an extension of our work presented in [23] and is organized as follows. Section 2 introduces the iPar-CLR based 2D and 3D facial landmark localization procedure. Section 3 and 4 give the construction details of the HSOG, meshHOG and meshHOS descriptors. Section 5 lists and compares the accuracies of each single 2D and 3D descriptor, and the ones of their fusion. The generalization capability of the proposed approach is discussed in Section 6. Finally, we conclude the paper and point out the limitations and future directions.

2. Joint 2D and 3D facial landmark localization

To extract expression related features, a set of key landmarks are required. In this paper, we introduce the incremental Parallel Cascade of Linear Regression (iPar-CLR) [35] for face landmarking in the 2D modality. iPar-CLR is an incremental and parallel version of of the Sequential Cascade of Linear

Regression (Seq-CLR) algorithm [39]. Given a set of training face images I_i associated with p 2D landmarks $\mathbf{x}^i \in \mathbb{R}^{2p\times1}$. f is a feature extraction function (e.g., SIFT) and $f(\mathbf{x}^i) \in \mathbb{R}^{128p\times1}$ in the case of extracting SIFT features. During training, one assumes that p corrected landmarks are known for each I_i, and denoted as $\mathbf{x}^i_*$. To reproduce the testing scenario, one runs the face detector on the training images to provide an initial configuration of the p landmarks $\mathbf{x}^i_0$, which corresponds to an average shape. In this setting, the Seq-CLR algorithm is formulated as:

$$\arg\min_{W_k,b_k} \sum_{I_i} \sum_{\mathbf{x}^i_k} \|\mathbf{x}^i_* - \mathbf{x}^i_k - W_k f(\mathbf{x}^i_k) - b_k\|^2. \tag{1}$$

In practice, W_0 and b_0 are first estimated using $\mathbf{x}^i_*$, $\mathbf{x}^i_0$, and $f(\mathbf{x}^i_0)$. Then, a sequence of regressions are computed to update $\mathbf{x}^i_k$ and make it converge to $\mathbf{x}^i_*$ step by step. iPar-CLR improves Seq-CLR by introducing a parametric 3D shape model for the configuration of p landmarks, and solving Eq. (1) in the parameter space. By assuming that the distribution of the perturbations of shape parameters is Gaussian, iPar-CLR is well suited for the task of incremental update. That is, it can incrementally update the pre-trained shape model according to the newly added face images.

When used for joint 2D and 3D facial landmark localization, the texture map is projected from each textured 3D face scan into a 2D regular grid domain using the interpolation techniques. Then, we apply iPar-CLR [40] to each projected 2D texture face image, outputting 49 2D landmarks (see Fig.1). These 2D landmarks are then transferred to 3D texture face space by the inverse of the above projection. Note that since all these 2D landmarks are located at the frontal part of the projected 2D face texture, the one-to-one correspondence between 3D texture and 2D texture can be approximately preserved during the projection mapping. Finally, the corresponding 3D landmarks are directly determined by the one-to-one correspondence between 3D texture and 3D geometry of the 3D face model. We evaluate iPar-CLR on the whole BU-3DFE database and the expressive samples in Bosphorus, and find that it can precisely localize all the pre-defined 49 facial landmarks for all samples even with variations in expression, ethnicity, gender and age etc. (see Fig.1 for some sampled results).

3. Construction of local 2D texture descriptors

3.1. First-order gradient based local texture descriptor: SIFT

We extract the SIFT [41] descriptor of each projected 2D texture face image at the locations of the detected 2D landmarks within 16 × 16 patches.

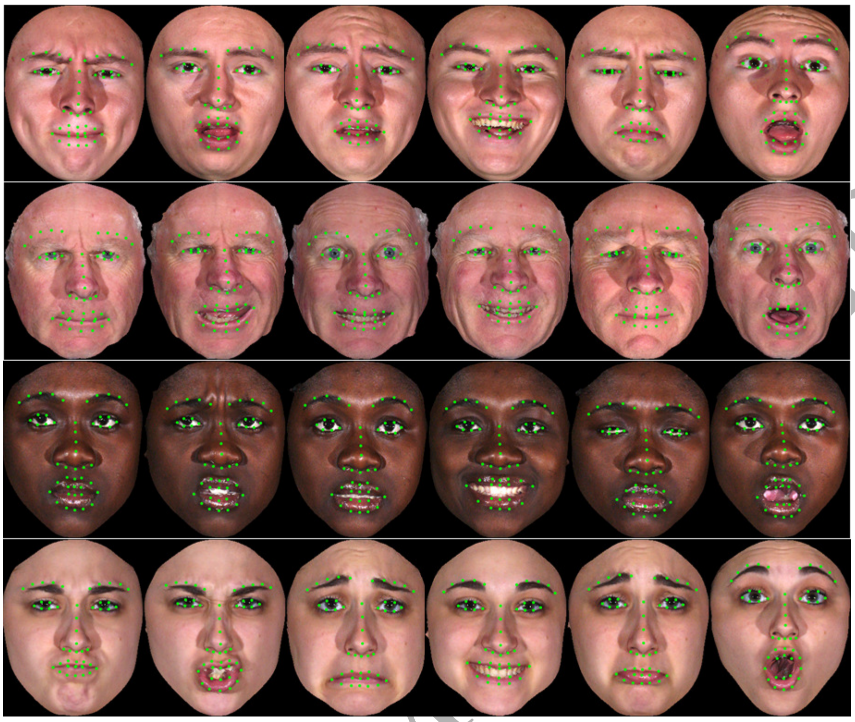


Figure 1: iPar-CLR based 2D landmark localization. 49 2D landmarks are localized on the projected 2D texture face images of the BU-3DFE database with different genders, ethnicities, ages, and expressions (from left to right, anger, disgust, fear, happiness, sadness and surprise).

The SIFT feature based facial representation of a 2D texture image is generated by concatenating all the SIFT features at the 49 landmarks according to the pre-defined order, resulting in a $128 \times 49 = 6,272$ dimensional feature vector. This vector is further normalized to the unit length for the following processing.

3.2. Second-order gradient based local texture descriptor: HSOG

The HSOG descriptor was originally proposed in [36] and proved very efficient for local image matching, object categorization, and scene classification. In this paper, we explore HSOG for 2D facial expression description. The construction of HSOG is composed of three steps:

(1) *Computation of the first order Oriented Gradient Maps (OGMs)*: The input of HSOG is a $R \times R$ image patch around each localized 2D facial landmark. For each image patch $\mathbf{I}(x, y)$, it outputs a number of Oriented Gradient

Maps (OGMs) $\{\mathbf{J}_o(x,y)\}_{o=1}^{L}$ by computing the Gaussian convolution of the positive orientation gradient maps, described as:

$$\mathbf{J}_o(x,y) = \mathbf{G}_{\Sigma} * \max\left(\frac{\partial \mathbf{I}(x,y)}{\partial o}, 0\right), o = 1, 2, ..., L, \tag{2}$$

where o represents a quantized direction, and $\mathbf{G}_{\Sigma}$ is a Gaussian kernel with standard deviation Σ, which is proportional to the size of image patch R.

(2) *Computation of the second order gradients*: Once these first order OGMs of all quantized directions are generated, they are used as the inputs for computing the second order gradients. Precisely, for each OGM $\mathbf{J}_o(x,y)$, we calculate its gradient magnitude $mag_o(x,y)$ and orientation $\theta_o(x,y)$ at every pixel location. The orientation value $\theta_o(x,y)$ is then re-scaled from the range of $[-\pi/2, \pi/2]$ to $[0, 2\pi]$, and quantized into L dominant orientations. After quantization, the entry n_o of each orientation θ_o is calculated as:

$$n_{\theta_o}(x,y) = \mod\left(\left\lfloor \frac{\theta_o(x,y)}{2\pi/L} + \frac{1}{2} \right\rfloor, L\right), o = 1, 2, \cdots, L. \tag{3}$$

(3) *Spatial pooling*: Daisy-style spatial pooling strategy is used in HSOG as illustrated in Fig.2. It is easy to find that there are four parameters that determine the HSOG descriptor, i.e., the size of the patch (R); the number of quantized orientations (L); the number of concentric rings (CR); the number of circles on each ring (C). The total number of the divided circles can be calculated as $T = CR \times C + 1$. Within each circle CIR_j, and for each

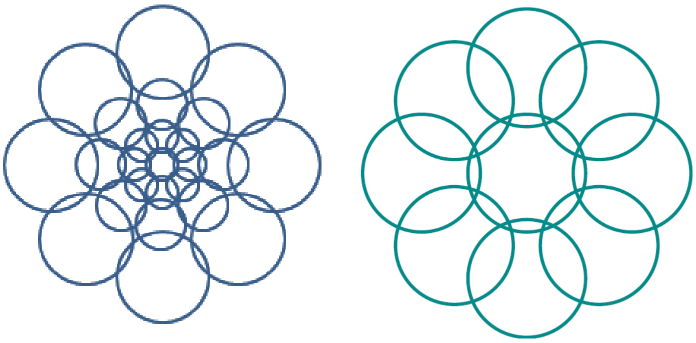


Figure 2: The daisy-style spatial pooling. Left: three concentric rings, and each with eight circles in HSOG. Right: one concentric ring with eight circles in meshHOG and meshHOS.

OGM $\mathbf{J}_o$, a second order gradient histogram is constructed by accumulating the gradient magnitudes mag_o of all the pixels with the same quantized orientation entry n_o.

$$\mathbf{h}_{oj}(i) = \sum_{(x,y)\in CIR_j} \delta(n_{\theta_o}(x,y) == i) * mag_o, \tag{4}$$

where $i = 0, 1, \cdots, L-1$; $o = 1, 2, \cdots, L$, $j = 1, 2, \cdots, T$, and δ is the characteristic function. Then, for each first order OGM $\mathbf{J}_o$, its second order gradient histogram h_o is generated by concatenating all the histograms from T circles:

$$\mathbf{h}_o = [\mathbf{h}_{o1}, \mathbf{h}_{o2}, \cdots, \mathbf{h}_{oT}]^T, \tag{5}$$

where $o = 1, 2, \cdots L$. Finally, the HSOG descriptor is obtained by concatenating all L histograms of the second order gradients as in Eq. (6). Each histogram $\mathbf{h}_o$ is normalized to a unit norm vector $\hat{\mathbf{h}}_o$ before concatenation.

$$HSOG = [\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \cdots, \hat{\mathbf{h}}_L]^T. \tag{6}$$

Similar to SIFT, the HSOG feature based expression representation of a 2D texture image is generated by concatenating all HSOG features of the localized landmarks and then normalized to the unit length. In this paper, we set $R = 25$, $L = 8$, $CR = 3$, $C = 4$ as in [36]. Thus, the dimension of the final HSOG feature vector for a face image is $T \times L \times 8 \times 49 = 13 \times 8 \times 8 \times 49 = 40,768$.

4. Construction of local 3D shape descriptors

The meshHOG and meshHOS descriptors were originally proposed in our previous work [37], [38] and proved efficient in 3D face identification. In this paper, we employ them in 3D FER. Similar to HSOG, meshHOG and meshHOS are built by the following three steps:

(i) *Computation of facial surface normal and curvature*: Each 3D facial surface is represented by a triangular mesh $\mathcal{T} = (\mathcal{F}, \mathcal{V})$, where $\mathcal{F}$ and $\mathcal{V}$ are the face and vertex sets. We compute the unit normal vector of each face by the cross product of its two edge vectors. Then the unit normal of each vertex $\mathbf{n}^v = [n_x^v, n_y^v, n_z^v]^T$ is achieved by averaging the normal vectors of its one-ring faces. The mesh gradient magnitude mag_v and orientation θ_v at each vertex are calculated as:

$$mag_v = \sqrt{(n_x^v/n_z^v)^2 + (n_y^v/n_z^v)^2}, \quad \theta_v = \arctan(n_y^v/n_x^v). \tag{7}$$

According to [42], the principal curvatures $k_{\max}$ and $k_{\min}$ are computed by fitting a cubic-order surface:

$$f(x,y) = \frac{A}{2}x^2 + Bxy + \frac{C}{2}y^2 + Dx^3 + Ex^2y + Fxy^2 + Gy^3 \tag{8}$$

and its normal vectors $(f_x(x;y), f_y(x;y), -1)$ using both the 3D coordinates and the normal vectors of the associated local neighbor points (two-ring). Once we have two principle curvatures, the shape index values, which describe different shape classes by a single number ranging from 0 to 1, is calculated as:

$$SI = \frac{1}{2} - \frac{1}{\pi}\arctan\left(\frac{k_{\max} + k_{\min}}{k_{\max} - k_{\min}}\right). \tag{9}$$

Figure 3 shows the shape index maps of sampled 3D faces with six prototypical expressions.

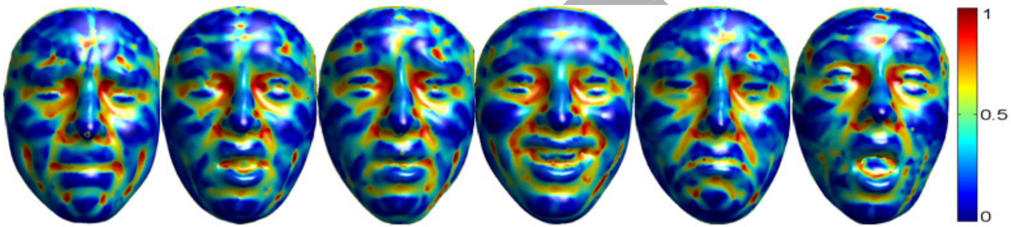


Figure 3: The shape index maps of sampled 3D faces with six prototypical expressions (from left to right, anger, disgust, fear, happiness, sadness, and surprise).

(ii) *Canonical orientation(s) assignment*: Similar to the SIFT feature, to achieve rotation invariance, one or more local coordinate systems (i.e., canonical orientations) should be determined at each localized 3D landmark. This can be accomplished by the following three steps: First, we build an initial local coordinate system, where the landmark v and its normal $\mathbf{n}^v$ are the origin and the positive z axis, respectively. And two perpendicular vectors in the tangent plane of v are randomly chosen as the x axis and y axis, respectively. Then, the gradient magnitudes and orientations of the vertices around the landmark with a given geodesic distance r_0 are computed, Gaussian-weighted by their corresponding gradient magnitudes, and put in a histogram of 360 bins. Dominant gradient orientations, that is, peaks in the histogram, are used to assign one or more canonical orientations to the landmark. Finally, the initial local coordinate system is rotated in the local tangent plane, making each canonical orientation as new x axis, and the new

y axis is computed by the cross product of the z axis (i.e., normal vector $\mathbf{n}^v$) and the new x axis (i.e., canonical orientation). Once the canonical orientations are determined, all the neighbor vertices and their normal vectors are transformed to the new local coordinate system for the following processes.

(iii) *Spatial pooling*: Similar to the HSOG feature, a simplified daisy-style spatial pooling strategy is also used for meshHOG and meshHOS. However, the pooling strategy here is performed on the tangent plane of each 3D landmark and on the local coordinate system determined by the assigned canonical orientations. As illustrated in Fig. 2, for the 3D descriptors, there is only one concentric ring associated with eight circles, resulting in nine sequential circles. Within each circle CIR_j, a mesh gradient histogram and a shape index histogram are constructed respectively. The histogram of mesh gradient is constructed by accumulating the gradient magnitudes mag_v of all vertices with the same quantized orientation entry $n_\theta(v)$ as:

$$\mathbf{hog}_j(i) = \sum_{v \in CIR_j} \delta(n_\theta(v) == i) * mag_v, \tag{10}$$

where $i = 0, 1, \cdots, 7$; $j = 1, 2, \cdots, 9$, $n_\theta(v)$ is entry of the quantized gradient orientation computed the same as $n_{\theta_o}(x, y)$ in (3). The histogram of shape index is constructed by accumulating the Gaussian weights $\mathbf{G}_\Sigma(v)$ of all vertices with the same quantized shape index value $n_{SI}(v)$

$$\mathbf{hos}_j(i) = \sum_{v \in CIR_j} \delta(n_{SI}(v) == i) * \mathbf{G}_\Sigma(v), \tag{11}$$

where $i = 0, 1, \cdots, 7$; $j = 1, 2, \cdots, 9$, $n_{SI}(v)$ is the quantized shape index values. Then, for each 3D landmark, its 3D descriptors are generated by concatenating all the histograms from nine circles in a clockwise direction,

$$HOG = [\mathbf{hog}_1, \mathbf{hog}_2, \cdots, \mathbf{hog}_9]^T, HOS = [\mathbf{hos}_1, \mathbf{hos}_2, \cdots, \mathbf{hos}_9]^T. \tag{12}$$

Each sub-histogram ($\mathbf{hog}_i$ or $\mathbf{hos}_i$) is normalized to the unit length before concatenation to eliminate the influence of non-uniform mesh sampling. Note that, intuitively, HOG describes the point-level bending pattern of the local shape around a landmark while HOS indicates the distribution of different shape categories. The expression representation (based on meshHOG or meshHOS) of a 3D face surface is generated by concatenating all HOG or HOS features of the localized 3D landmarks and then normalized to the unit length. Following [37], the geodesic radius r_0 is set to 22.50 mm, the radius of each circle is set to 10 mm, and the distance between the center of the centric circle and the one of each rounding circle is set to 15 mm. As a result, the dimension of the final meshHOG or meshHOS feature is $9 \times 8 \times 49 = 3,528$.

5. Experimental results

5.1. The BU-3DFE database

We make use of the widely used BU-3DFE database [43] to evaluate the proposed multimodal 2D + 3D local feature-based FER approach. This database consists of 2,500 textured 3D face scans of 100 persons in different expression, gender, race, and age. Six prototypical facial expressions (anger, disgust, fear, happiness, sadness, and surprise) with four intensity levels plus a neutral expression are displayed for each person. Examples of some projected 2D texture face images in BU-3DFE database are shown in Fig.1.

5.2. Experimental setup

To fairly conduct the identity-independent FER, we use the evaluation protocol in [13]. More precisely, we randomly select 60 persons, and keep the samples with the six prototypical facial expressions of two highest intensity levels. That is, $60 \times 6 \times 2 = 720$ samples are used for training and testing in total. Then, 648 samples of 54 persons (90%) and 72 of 6 persons (10%) are randomly divided for the training and testing data partition. To achieve stable recognition accuracy, this kind of 10-fold subject-independent cross-validation is conducted 100 rounds for all of our experiments. Based on these data partition strategies and the constructed 2D and 3D features, we utilize the SVM classifier with the Radial Basis Function (RBF) kernel for expression classification. The parameters for SVMs are tuned according to the 10-fold cross-validation in the training sets. To find the complementary characteristics between 2D descriptors, 3D descriptors, as well as 2D and 3D descriptors, we conduct both the early fusion (feature-level) and late fusion (score-level). For early fusion, the fused feature is generated by simply concatenating different descriptors. For late fusion, the mean of the recognition accuracies of different descriptors are used as the final accuracy.

5.3. *Performance evaluation*

5.3.1. *Local 2D texture descriptors and their fusion*

Table 1 shows the average expression recognition accuracies achieved using the single 2D descriptors and their fusion. From this table, we can see that: i) The average accuracies of the HSOG descriptor are much better than the ones of SIFT for anger and sadness, and comparable for the other expressions. ii) Early fusion largely improves the average accuracies of anger and sadness for SIFT, but also largely impairs the one of sadness for HSOG. iii) Late fusion generally performs better than early fusion, especially for the fear and sadness expressions. iv) Overall, the average accuracy of HSOG is 84.49%, which is better than SIFT (81.85%), and even slightly better than the ones of early fusion (82.85%) and late fusion (84.29%). We can conclude that the second-order gradient based local texture descriptor (HSOG) has more powerful discriminative ability than the popular first-order gradient based one (SIFT) for local texture-based FER. Moreover, there also exists some complementarity between different order descriptors for some specific expressions (e.g., anger and fear).

Table 1: Average confusion matrices of SIFT, HSOG, and their early and late fusions on BU-3DFE database.

	SIFT (**81.85**)						HSOG (**84.49**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**76.55**	6.53	2.83	0	14.09	0	**83.09**	4.24	3.21	0	9.37	0.08
DI	5.16	**84.51**	2.37	1.28	2.39	4.29	5.75	**85.00**	2.14	2.41	2.46	2.24
FE	3.64	6.41	**72.61**	5.73	8.85	2.76	1.06	5.63	**72.41**	9.48	8.35	3.07
HA	0	0.98	8.77	**89.37**	0	0	0.79	2.45	6.02	**89.82**	0.03	0.90
SA	20.25	1.43	7.29	0	**70.71**	0.32	12.82	3.42	3.57	0	**80.20**	0
SU	0.01	0.04	1.12	0.64	0.82	**97.38**	0	0.23	1.74	0.75	0.82	**96.47**
	early fusion: SIFT + HSOG (**82.86**)						late fusion: SIFT + HSOG (**84.29**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**83.91**	4.29	2.86	0	8.87	0.08	**82.16**	5.05	2.92	0	9.87	0
DI	6.12	**83.27**	2.87	1.68	1.66	4.42	5.81	**83.87**	2.56	1.64	2.22	3.9
FE	1.42	7.73	**70.82**	8.60	8.35	3.07	1.16	6.48	**74.40**	7.38	7.80	2.77
HA	0.03	2.59	8.28	**88.20**	0	0.90	0.02	1.70	7.52	**89.86**	0	0.90
SA	19.06	2.00	5.03	0	**73.91**	0	14.67	2.45	4.48	0	**78.42**	0
SU	0	0	2.13	0.01	0.82	**97.05**	0	0.07	1.33	0.72	0.82	**97.07**

5.3.2. *Local 3D shape descriptors and their fusion*

Table 2 shows the average expression recognition accuracies achieved using the single 3D descriptors and their fusion. From this table, we can find that: i) Except anger expression, meshHOS achieves better results than meshHOG, especially for happiness. ii) Early fusion and late fusion generally improve the accuracies of both 3D descriptors for all expressions except happiness with a slight drop in early fusion. iii) Overall, the average recognition accuracy of meshHOS is 80.55%, which is better than meshHOG (77.62%), and late fusion (82.70%) is superior to early fusion (81.23%). We can conclude that the second-order surface gradient-based local shape descriptor (meshHOS) has stronger discriminative capability than the first-order surface gradient-based one (meshHOG). Moreover, they also contain some complementary information when classifying some specific expressions (e.g., sadness and surprise).

Table 2: Average confusion matrices of meshHOG, meshHOS, and their early and late fusions on BU-3DFE database.

	meshHOG (**77.62**)						meshHOS (**80.55**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**79.75**	2.47	2.60	1.48	13.70	0	**77.96**	4.36	2.16	1.17	14.36	0
DI	4.11	**75.00**	8.89	4.92	2.17	4.91	3.34	**78.99**	6.78	3.83	3.31	3.74
FE	3.39	7.32	**66.23**	14.44	4.30	4.33	0.92	6.63	**69.50**	13.27	4.69	4.99
HA	0.77	0.59	15.52	**80.79**	0	2.33	0	0.37	9.85	**88.38**	1.40	0
SA	22.58	2.09	2.91	0	**72.32**	0.11	18.28	3.01	4.15	0	**74.52**	0.04
SU	0	1.01	7.38	0.01	0	**91.61**	0	2.08	3.98	0	0	**93.93**
	early fusion: meshHOG + meshHOS (**81.23**)						late fusion: meshHOG + meshHOS (**82.70**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**80.93**	2.56	2.56	1.45	12.50	0	**82.63**	2.48	2.56	1.48	10.85	0
DI	4.43	**80.12**	5.55	4.76	2.01	3.13	3.84	**80.97**	5.13	4.62	1.93	3.51
FE	1.29	6.18	**71.06**	13.08	3.29	5.10	1.67	5.72	**72.08**	12.22	3.44	4.87
HA	0	1.18	13.79	**85.03**	0	0	0	0.68	12.02	**87.21**	0.09	0
SA	19.13	1.38	2.82	0	**76.67**	0	15.78	1.57	3.91	0	**78.74**	0
SU	0	0.13	6.32	0.01	0	**93.54**	0	0.04	5.38	0	0	**94.57**

5.3.3. Local multimodal 2D + 3D descriptors and their fusion

In this section, we indicate that the local 2D texture and 3D shape descriptors contain strong complementary characteristics, and thus their fusion largely improves the expression recognition accuracies.

Table 3 lists the average expression recognition results of fusing the same order gradient-based 2D and 3D descriptors lead increase performance. Compared with the results in Table 1 and Table 2, we can see that both early fusion and late fusion of the same order gradient-based 2D and 3D descriptors are very efficient, especially for the case of late fusion, with an improvement up to 3% for SIFT, 7% for meshHOG, 2.3% for HSOG, and 6.3% for meshHOS in the average accuracy. Moreover, the improvement of sadness expression is up to 8% for meshHOG and 10% for SIFT. And the accuracies of happiness are improved about 5% for HSOG and 6% for meshHOS.

Table 3: The effectiveness of fusing the same order gradient-based 2D and 3D descriptors on BU-3DFE database.

	early fusion: SIFT + meshHOG (**83.68**)						late fusion: SIFT + meshHOG (**84.91**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**82.09**	5.25	1.83	0.08	10.75	0	**83.14**	4.07	1.59	0.56	10.63	0
DI	5.67	**84.75**	4.00	1.33	0.08	4.17	6.38	**82.65**	3.47	2.04	0.75	4.68
FE	1.50	6.00	**71.33**	11.00	6.67	3.50	0.20	6.63	**74.02**	8.06	7.85	3.24
HA	0	1.08	8.67	**89.50**	0	075	0	0.87	6.98	**91.74**	0	0.41
SA	17.33	1.08	3.42	0	**78.17**	0	15.28	0.91	3.36	0	**80.45**	0
SU	0	0.58	2.83	0	0.33	**96.25**	0	0.04	1.75	0	0.77	**97.43**
	early fusion: HSOG + meshHOS (**84.49**)						late fusion: HSOG + meshHOS (**86.80**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**83.50**	5.75	2.17	0	8.58	0	**86.10**	3.23	1.64	0.33	8.70	0
DI	5.42	**85.67**	1.42	2.17	2.33	3.00	4.96	**85.45**	1.66	2.29	2.48	3.17
FE	0.25	6.92	**70.58**	8.08	9.75	4.42	0.05	4.31	**75.87**	9.33	6.77	3.67
HA	0	2.08	6.25	**91.67**	0	0	0.02	1.08	3.47	**94.85**	0.59	0
SA	14.00	2.50	4.42	0	**79.08**	0	13.47	0.75	4.4	0	**81.38**	0
SU	0	0.08	3.17	0	0.33	**96.42**	0	0.01	1.96	0.07	0.82	**97.15**

Table 4 shows the average expression recognition results of fusing different order gradient-based 2D and 3D descriptors. Compared with the results in Table 1 and Table 2, we can find that the fusion of the different order gradient-based 2D and 3D descriptors is also very efficient except the case of early fusion of HSOG and meshHOG. Take the results of late fusion as an example, the average recognition accuracies are improved by 3.2% for SIFT, 5% for meshHOS, 1.3% for HSOG and 8.1% for meshHOG. In particular, the improvement of happiness expression is 5.5% in the case of fusing SIFT and meshHOS. And the accuracy of the sadness expression is improved up to 11% when lately fusing HSOG and meshHOG.

As reported in Table 5, when considering the fusion of all the first-order and second-order gradient-based local 2D texture and 3D shape descriptors, our approach achieves an average recognition accuracy of 85.92% for early fusion and 86.32% for late fusion. These scores largely outperform the ones achieved by only fusing 2D descriptors (82.86% and 84.29%) in Table 1 or 3D descriptors (81.23% and 82.70%) in Table 2. More precisely, the confusion matrices of these scores indicate that the 2D descriptors and 3D descriptors have strong complementary characteristics for all the six prototypical facial expressions.

5.3.4. Comparison with other methods

To validate the effectiveness of the proposed method in FER, we compare it with the state-of-the-art methods on the BU-3DFE dataset. To give a comprehensive analysis, four aspects, including the data modality, facial landmark, expression classifier, and recognition accuracy are compared.

Table 4: The effectiveness of fusing different order gradient-based 2D and 3D descriptors on BU-3DFE database.

	early fusion: SIFT + meshHOS (**85.15**)						late fusion: SIFT + meshHOS (**85.07**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**85.42**	5.75	2.25	0	6.58	0	**80.67**	5.08	1.83	0.08	12.33	0
DI	5.00	**86.92**	0.92	1.75	1.50	3.92	4.67	**86.33**	2.17	0.83	2.50	3.50
FE	0	6.25	**73.67**	6.42	9.58	4.08	0	6.92	**75.50**	7.08	7.33	3.17
HA	0	1.00	7.33	**91.67**	0	0	0	1.00	5.08	**93.92**	0	0
SA	16.83	1.25	5.75	0	**76.17**	0	15.58	0.83	6.92	0	**76.67**	0
SU	0	0.25	1.92	0	0.75	**97.08**	0	0	1.83	0	0.83	**97.33**
	early fusion: HSOG + meshHOG (**83.17**)						late fusion: HSOG + meshHOG (**85.75**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**81.00**	4.33	1.83	1.00	11.83	0	**82.58**	3.92	1.92	1.50	10.08	0
DI	4.75	**85.83**	3.67	2.58	1.17	2.00	5.33	**86.67**	1.67	2.50	1.33	2.50
FE	1.17	5.92	**72.75**	10.25	6.25	3.67	0	6.58	**74.83**	8.25	7.00	3.33
HA	0.08	1.83	9.58	**87.83**	0	0.67	0	2.42	6.00	**90.17**	0	1.42
SA	19.17	1.08	3.50	0	**76.25**	0	11.42	1.17	4.17	0	**83.25**	0
SU	0	1.08	3.58	0	0	**95.33**	0	0	2.08	0	0.92	**97.00**

Table 5: The effectiveness of fusing all four gradient-based 2D and 3D descriptors on BU-3DFE database.

	early fusion: all features (**85.92**)						late fusion: all features (**86.32**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**86.33**	3.91	1.58	0.06	8.13	0	**85.56**	3.88	1.58	0.17	8.81	0
DI	5.78	**84.27**	2.19	2.32	0.98	4.47	5.17	**84.35**	2.00	2.50	2.03	3.95
FE	0.02	4.06	**75.03**	10.93	6.41	3.56	0	4.64	**75.77**	9.26	7.02	3.30
HA	0	0.99	7.15	**91.86**	0	0	0	0.92	5.62	**93.42**	0	0.05
SA	14.69	0.43	3.52	0	**81.36**	0	14.28	0.92	3.55	0	**81.26**	0
SU	0	0	2.52	0	0.82	**96.67**	0	0	1.63	0	0.82	**97.55**

From Table 6, we find that all previous methods (except [25] and [14]) reported their FER accuracies on BU-3DFE using only 3D modality data. As mentioned in Section 1, the results of 2D and 3D data are separately reported in [25] and jointly reported in [14]. Complementarity analysis of 2D and 3D data in FER is missing. On facial landmark, early studies such as [25], [16], [18], [17], and [26] rely on a large number of manual landmarks. Recent studies try to avoid this impractical framework by utilizing global registration algorithms (e.g., [12], [13], [21], [24]), or building general face models (e.g., [14], [44]). Our method solves this problem by exploring the iPar-CLR algorithm to jointly detect a large number of 2D and 3D landmarks. For expression classification, SVM is the most popular classifier compared with the others such as Neutral Networks (NN), Sparse Representation-based Classifier (SRC), Bayesian Belief Net (BBN), and Multiple Kernel Learning (MKL).

Table 6: Performance comparison with the state-of-the-art methods on BU-3DFE database.

Method	Modality	Landmark	Classifier	Accuracy in protocol (%)		
				I	II	III
Wang et al. (2006) [25]	2D/3D	64 manual	LDA	83.60	61.79	-
Soyel et al. (2007)[15]	3D mesh	11 manual	NN	91.30	67.52	-
Soyel et al. (2008) [16]	3D mesh	83 manual	NN	93.72	-	-
Tang et al. (2008)[18]	3D mesh	83 manual	LDA	95.10	74.51	-
Tang et al. (2008) [17]	3D mesh	83 manual	SVM	87.10	-	-
Mpiperis et al. (2008) [12]	3D mesh	global registration	ML	90.50	-	-
Gong et al. (2009)[13]	3D depth	global registration	SVM	-	76.22	-
Zhao et al. (2010)[14]	2D+3D	19 automatic	BBN	82.30	-	-
Berretti et al. (2010) [26]	3D depth	27 manual	SVM	-	-	77.54
Lemaire et al. (2011) [44]	3D mesh	21 automatic	SVM	-	75.76	-
Li et al. (2012) [21]	3D depth	global registration	MKL	-	-	80.14
Zeng et al. (2013) [22]	3D depth	3 automatic	SRC	-	-	70.93
Zhen et al. (2015) [24]	3D mesh	global registration	SVM	-	84.50	83.20
Yang et al. (2015) [45]	3D mesh	global registration	SVM	-	84.80	82.73
Ours (2D+3D features)	2D+3D	49 automatic	SVM	-	**86.32**	-

In the literature, there are three FER protocols on BU-3DFE. Early tasks (e.g., [14], [16], [17], [18], [25]) chose 60 subjects and average the accuracies of one or two rounds of 10-fold cross-validation, totally with 10 or 20 times of train and test sessions (denoted by protocol I). This protocol has proved very sensitive to the identity variations of training and testing samples [13]. Gong et al. [13] later suggested to choose 60 subjects and average the accuracies of 100 rounds of 10-fold cross-validation, resulting in 1000 times of train and test sessions in total (i.e., protocol II). A similar protocol (i.e., protocol III) [26], randomly chose 60 subjects in each round of 10-fold cross-validation and average the accuracies of 100 rounds. From Table 6, we can find that the accuracies of the same methods [25], [15], [18] dropped more than 20% from protocol I to protocol II. Moreover, the accuracies of the same method achieved by protocol II and protocol III were close to each other as shown in [24] and [45]. Our proposed multimodal 2D+3D local feature-based approach reaches the highest average accuracy (86.32%) in protocol II.

6. Discussion

6.1. Generalization capability on Bosphorus database

In this section, we study the generalization capability of our proposed approach on the Bosphorus database. This database contains 4666 textured 3D face models of 105 subjects in various facial expressions, action units, poses and occlusions. To fairly conduct the identity-independent facial expression recognition, we still use the experimental protocol in [13] (i.e., protocol II). That is, we randomly select 60 persons who display all the six prototypical facial expressions. Totally, there are $60 \times 6 = 360$ samples used for training and testing. And 324 samples of 54 persons (90%) and 36 of 6 persons (10%) are randomly divided for the training and testing data partition. This kind of 10-fold cross-validation is conducted 100 rounds to achieve stable recognition accuracies, and the results are listed in Table 7.

Table 7: The average accuracies and confusion matrices (in %) on Bosphorus database.

	SIFT (**82.89**)			HSOG (**80.31**)			meshHOG (**65.39**)			meshHOS (**74.94**)		
	late fusion of SIFT + meshHOS (**84.44**)						late fusion of HSOG + meshHOS (**83.56**)					
	early fusion of 2D+3D descriptors (**84.33**)						late fusion of 2D+3D descriptors (**84.72**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**83.00**	5.83	0.17	0	11.00	0	**82.33**	6.67	0	0	11.00	0
DI	5.83	**82.50**	5.67	1.33	4.67	0	5.83	**82.83**	5.17	1.33	4.67	0.17
FE	1.17	1.67	**69.83**	1.50	4.00	21.83	0.17	3.17	**72.33**	2.17	3.67	18.50
HA	0	0.17	0	**99.83**	0	0	0	0	0	**100**	0	0
SA	3.83	8.00	0	0	**88.17**	0	4.33	6.67	0	0	**89.00**	0
SU	0	0	17.33	0	0	**82.67**	0	0	18.17	0	0	**81.83**

Compare the results in Table 7 with the ones achieved on the BU-3DFE dataset, we can see that: 1) except the SIFT descriptor, the accuracies of other 2D and 3D descriptors are decreased. For example, the performance of meshHOG is dropped from 77.62% on BU-3DFE to 65.39% on Bosphorus. 2) The fusion of 2D and 3D descriptors is still efficient such as the late fusion of SIFT and meshHOG, HSOG and meshHOS. 3) Comparable expression recognition accuracies (86.32% vs. 84.72%) are achieved on the two datasets when fusing all the 2D and 3D local descriptors. 4) Compare with the results in Table 5, the accuracies for happiness and sadness are much better on Bosphorus, while the ones for surprise are much better on BU-3DFE. The possible reasons resulting in 1) and 4) come from the large expression variations of different persons when they displaying the same expression. Noted that all the persons in Bosphorus are professional actors or actress, while the subjects in BU-3DFE are ordinary people such as the university students. As shown in Fig.4, sadness and anger look very similar for some people, and fear is always with month opening, which makes fear and surprise are largely confused with each other. Moreover, the disgust expression is very special and diversiform, which makes it confusing with sadness, anger and fear. It is

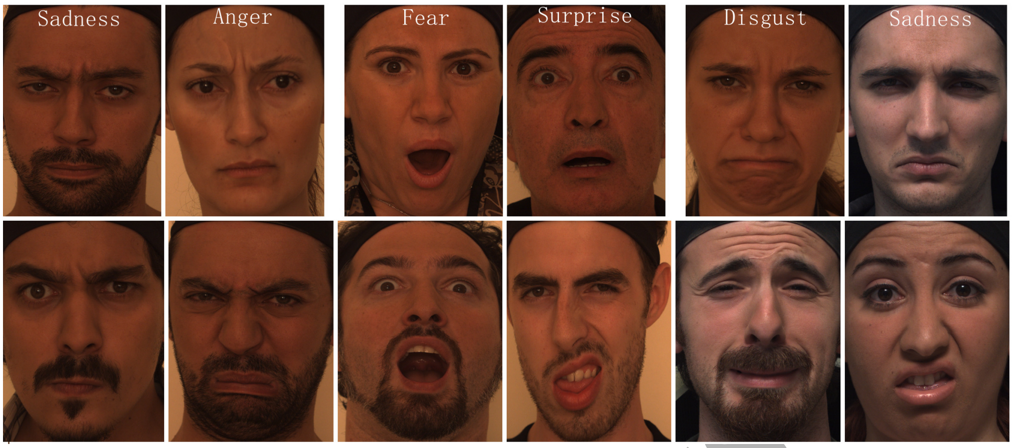


Figure 4: Examples of expression pairs with similar expression configurations but different expression labels in the Bosphorus database. The expression labels of the bottom three pairs are: anger and disgust, fear and disgust, sadness and disgust.

probably the reason that most anger samples are misclassified into sadness, and most surprise samples are misclassified to fear and vice versa as shown in the average confusion matrices in Table 7.

6.2. Complementarity analysis between 2D and 3D descriptors

To illustrate the complementary characteristics between 2D and 3D multimodal descriptors, we perform the Gentle AdaBoost algorithm [46] on the HSOG and meshHOS descriptors to select the most discriminative 2D and 3D facial landmarks (i.e., local regions used to compute HSOG or meshHOS) on BU-3DEF. More precisely, in each iteration of the Gentle AdaBoost algorithm, each landmark associated descriptor is first fed into a logistic regression weak classier, and the one with the lowest error rate is chosen as the most discriminative one in current iteration. Then, the weights of all the samples (landmarks) are updated, making the algorithm pay more attention on the misclassified samples. Finally, the algorithm stops when the top N discriminative landmarks are selected. Figure 5 shows the top 15 most discriminative landmarks automatically selected by this algorithm. From this figure, it is not difficult to find that the distributions of the top 15 most discriminative 2D and 3D facial landmarks are largely different from each other for all the six sampled facial expressions. This finding once again indicates that our proposed 2D and 3D multimodal local texture and shape descriptors indeed have strong complementary characteristics.

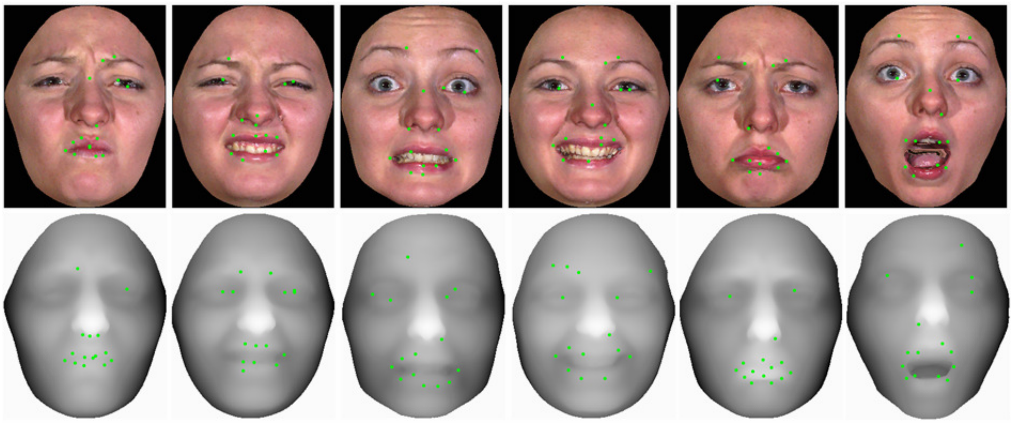


Figure 5: Illustration of the complementary characteristics between the 2D and 3D local descriptors. The top 15 most discriminative 2D and 3D landmarks are automatically selected by the Gentle AdaBoost algorithm from 2D and 3D face samples with different expressions (i.e., anger, disgust, fear, happiness, sadness, and surprise) on BU-3DFE. Note that depth images, instead of 3D meshes, are displayed for better visualization.

7. Conclusion and future work

In this paper, we present an efficient multimodal 2D + 3D feature-based approach for automatic FER. Based on the iPAR-CLR algorithm, we automatically localize 49 2D facial landmarks, and their corresponding 3D facial landmarks. Around each landmark, the HSOG based local texture descriptor and the SIFT descriptor are integrated for local 2D facial texture description. Furthermore, two mesh-based local shape descriptors, which consider both the first-order (surface normal) and the second-order (curvatures) surface gradients, are introduced to describe local 3D facial shapes. Both early fusion and late fusion of 2D, 3D, as well as 2D and 3D descriptors are comprehensively evaluated on the BU-3DFE database. All the experimental results demonstrate the effectiveness of integrating the 2D and 3D descriptors for expression recognition. Furthermore, we also analyze the generalization capability of the proposed approach on Bosphorus, and illustrate the complementary characteristics between the 2D and 3D descriptors.

Considering the limitation of current approach, in the future, we will go deeply in the following directions: i) The iPar-CLR based joint 2D and 3D facial landmark localization algorithm may fail with large pose variations and data missing. To solve this problem, we will investigate more robust algorithms such as [47]. ii) In current work, we use the simplest early and late fusion schemes. To find more intrinsic complementary characteristics between 2D and 3D modalities, we are going to explore better strategies. iii) In this paper, we focus on recognizing six basic expressions using multimodal 2D+3D static images. Following [48], [49], this work will also be extended to the problem of 3D action unit recognition and to dynamic 3D face spaces.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 11401464, No. 61202237 and No. 61273263; the China Postdoctoral Science Foundation (No. 2014M560785); the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20121102120016); the Research Program of State Key Laboratory of Software Development Environment (SKLSDE-2015ZX-30); the French research agency, Agence Nationale de la Recherche (ANR) under Grant ANR-07-SESU-004, ANR-2010-INTB-0301-01 and ANR-13-INSE-0004-02; the joint project by the LIA 2MCSI lab between the group of Ecoles Centrales and Beihang University; and the Fundamental Research Funds for the Central Universities.

References

[1] K. R. Scherer, What are emotions? and how can they be measured?, Social Science Information 44 (2005) 693–727.

[2] Z. Zeng, M. Pantic, G. I. Roisman, T. S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (1) (2009) 39–58.

[3] V. Bettadapura, Face expression recognition and analysis: the state of the art, arXiv preprint arXiv:1203.6722.

[4] M. Pantic, L. J. M. Rothkrantz, Automatic analysis of facial expressions: The state of the art, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1424–1445.

[5] B. Abboud, F. Davoine, M. Dang, Facial expression recognition and synthesis based on an appearance model, Signal Processing: Image Communication 19 (8) (2004) 723–740.

[6] A. Lanitis, C. J. Taylor, T. F. Cootes, Automatic interpretation and coding of face images using flexible models, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 743–756.

[7] L. Zalewski, S. Gong, Synthesis and recognition of facial expressions in virtual 3d views, in: IEEE international conference on automatic face and gesture recognition (FG), 2004, pp. 493–498.

[8] J. Wang, L. Yin, Static topographic modeling for facial expression recognition and analysis, Computer Vision and Image Understanding 108 (1) (2007) 19–34.

[9] M. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, Coding facial expressions with gabor wavelets, in: IEEE international conference on automatic face and gesture recognition (FG), 1998, pp. 200–205.

[10] C. Shan, S. Gong, P. W. McOwan, Facial expression recognition based on local binary patterns: A comprehensive study, Image and Vision Computing 27 (6) (2009) 803–816.

[11] S. Ramanathan, A. Kassim, Y. Venkatesh, W. S. Wah, Human facial expression recognition using a 3d morphable model, in: IEEE International Conference on Image Processing (ICIP), 2006, pp. 661–664.

[12] I. Mpiperis, S. Malassiotis, M. G. Strintzis, Bilinear models for 3-d face and facial expression recognition, IEEE Transactions on Information Forensics and Security 3 (3) (2008) 498–511.

[13] B. Gong, Y. Wang, J. Liu, X. Tang, Automatic facial expression recognition on a single 3d face by exploring shape deformation, in: Proceedings of the 17th ACM international conference on Multimedia, 2009, pp. 569–572.

[14] X. Zhao, D. Huang, E. Dellandréa, L. Chen, Automatic 3d facial expression recognition based on a bayesian belief net and a statistical facial feature model, in: 20th International Conference on Pattern Recognition (ICPR), 2010, pp. 3724–3727.

[15] H. Soyel, H. Demirel, Facial expression recognition using 3d facial feature distances, Image Analysis and Recognition (2007) 831–838.

[16] H. Soyel, H. Demirel, 3d facial expression recognition with geometrically localized facial features, in: 23rd International Symposium on Computer and Information Sciences (ISCIS), 2008, pp. 1–4.

[17] H. Tang, T. S. Huang, 3d facial expression recognition based on properties of line segments connecting facial feature points, in: IEEE international conference on automatic face and gesture recognition and workshops (FG), 2008, pp. 1–6.

[18] H. Tang, T. S. Huang, 3d facial expression recognition based on automatically selected features, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2008, pp. 1–8.

[19] A. Maalej, B. B. Amor, M. Daoudi, A. Srivastava, S. Berretti, et al., Local 3d shape analysis for facial expression recognition, in: 20th International Conference on Pattern Recognition (ICPR), 2010, pp. 4129–4132.

[20] O. Ocegueda, T. Fang, S. K. Shah, I. A. Kakadiaris, Expressive maps for 3d facial expression recognition, in: IEEE International Conference on Computer Vision Workshops (ICCV), 2011, pp. 1270–1275.

[21] H. Li, L. Chen, D. Huang, Y. Wang, J.-M. Morvan, 3d facial expression recognition via multiple kernel learning of multi-scale local normal patterns, in: 21st International Conference on Pattern Recognition (ICPR), 2012, pp. 2577–2580.

[22] W. Zeng, H. Li, L. Chen, J.-M. Morvan, X. D. Gu, An automatic 3d expression recognition framework based on sparse representation of conformal images, in: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013, pp. 1–8.

[23] H. Li, J.-M. Morvan, L. Chen, 3d facial expression recognition based on histograms of surface differential quantities, Advances Concepts for Intelligent Vision Systems (2011) 483–494.

[24] Q. Zhen, D. Huang, Y. Wang, L. Chen, Muscular movement model based automatic 3d facial expression recognition, in: MultiMedia Modeling, Springer, 2015, pp. 522–533.

[25] J. Wang, L. Yin, X. Wei, Y. Sun, 3d facial expression recognition based on primitive surface feature distribution, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, 2006, pp. 1399–1406.

[26] S. Berretti, A. D. Bimbo, P. Pala, B. B. Amor, M. Daoudi, A set of selected sift features for 3d facial expression recognition, in: 20th International Conference on Pattern Recognition (ICPR), 2010, pp. 4125–4128.

[27] M. Song, D. Tao, S. Sun, C. Chen, S. Maybank, Robust 3d face landmark localization based on local coordinate coding, Image Processing, IEEE Transactions on 23 (12) (2014) 5108–5122.

[28] T. Fang, X. Zhao, O. Ocegueda, S. Shah, I. Kakadiaris, 3d facial expression recognition: A perspective on promises and challenges, in: IEEE international conference on automatic face and gesture recognition and workshops (FG), 2011, pp. 603–610.

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

[30] K. Chang, K. W. Bowyer, P. Flynn, An evaluation of multimodal 2d+3d face biometrics, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (4) (2005) 619–624.

[31] A. S. Mian, M. Bennamoun, R. Owens, An efficient multimodal 2d-3d hybrid approach to automatic face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (11) (2007) 1927–1943.

[32] F. Tsalakanidou, S. Malassiotis, Real-time 2d+3d facial action and expression recognition, Pattern Recognition 43 (5) (2010) 1763–1775.

[33] A. Savran, B. Sankur, M. T. Bilge, Facial action unit detection: 3d versus 2d modality, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, pp. 71–78.

[34] P. Wang, C. Kohler, F. Barrett, R. Gur, R. Verma, Quantifying facial expression abnormality in schizophrenia by combining 2d and 3d features, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1–8.

[35] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Incremental face alignment in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1859–1866.

[36] D. Huang, C. Zhu, Y. Wang, L. Chen, Hsog: A novel local image descriptor based on histograms of the second order gradients, IEEE Transactions on Image Processing 23 (11) 4680–4695.

[37] H. Li, D. Huang, P. Lemaire, J.-M. Morvan, L. Chen, Expression robust 3d face recognition via mesh-based histograms of multiple order surface differential quantities, in: IEEE International Conference on Image Processing (ICIP), IEEE, 2011, pp. 3053–3056.

[38] H. Li, D. Huang, J.-M. Morvan, Y. Wang, L. Chen, Towards 3d face recognition in the real: A registration-free approach using fine-grained matching of 3d keypoint descriptors, International Journal of Computer Vision 113 (2) (2015) 128–142.

[39] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[40] https://sites.google.com/site/chehrahome/.

[41] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[42] J. Goldfeather, V. Interrante, A novel cubic-order algorithm for approximating principal direction vectors, ACM Transactions on Graphics 23 (1) (2004) 45–63.

[43] L. Yin, X. Wei, Y. Sun, J. Wang, M. J. Rosato, A 3d facial expression database for facial behavior research, in: IEEE international conference on automatic face and gesture recognition (FG), 2006, pp. 211–216.

[44] P. Lemaire, B. Ben Amor, M. Ardabilian, L. Chen, M. Daoudi, Fully automatic 3d facial expression recognition using a region-based approach, in: Proceedings of the joint ACM workshop on Human gesture and behavior understanding, 2011, pp. 53–58.

[45] X. Yang, D. Huang, Y. Wang, L. Chen, Automatic 3d facial expression recognition using geometric scattering representation, in: IEEE International Conference on Automatic Face Gesture Recognition (FG), 2015.

[46] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, Annals of Statistics 28 (2) (2000) 337–407.

[47] S. Z. Gilani, F. Shafait, A. Mian, Shape-based automatic detection of a large number of 3d facial landmarks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[48] Y. Sun, M. Reale, L. Yin, Recognizing partial facial action units based on 3d dynamic range data for facial expression recognition, in: IEEE International Conference on Automatic Face Gesture Recognition (FG), 2008, pp. 1–8.

[49] M. Reale, X. Zhang, L. Yin, Nebula feature: A space-time feature for posed and spontaneous 4d facial behavior analysis, in: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013, pp. 1–8.