Multimodal 2D+3D Facial Expression Recognition With Deep Fusion Convolutional Neural Network

Huibin Li, *Student Member, IEEE*, Jian Sun, *Member, IEEE*, Zongben Xu, *Member, IEEE*, and Liming Chen, *Senior Member, IEEE*

***Abstract*—This paper presents a novel and efficient deep fusion convolutional neural network (DF-CNN) for multimodal 2D+3D facial expression recognition (FER). DF-CNN comprises a feature extraction subnet, a feature fusion subnet, and a softmax layer. In particular, each textured three-dimensional (3D) face scan is represented as six types of 2D facial attribute maps (i.e., geometry map, three normal maps, curvature map, and texture map), all of which are jointly fed into DF-CNN for feature learning and fusion learning, resulting in a highly concentrated facial representation (32-dimensional). Expression prediction is performed by two ways: 1) learning linear support vector machine classifiers using the 32-dimensional fused deep features, or 2) directly performing softmax prediction using the six-dimensional expression probability vectors. Different from existing 3D FER methods, DF-CNN combines feature learning and fusion learning into a single end-to-end training framework. To demonstrate the effectiveness of DF-CNN, we conducted comprehensive experiments to compare the performance of DF-CNN with handcrafted features, pre-trained deep features, fine-tuned deep features, and state-of-the-art methods on three 3D face datasets (i.e., BU-3DFE Subset I, BU-3DFE Subset II, and Bosphorus Subset). In all cases, DF-CNN consistently achieved the best results. To the best of our knowledge, this is the first work of introducing deep CNN to 3D FER and deep learning-based feature-level fusion for multimodal 2D+3D FER.**

***Index Terms*—Deep fusion convolutional neural network (DF-CNN), facial expression recognition (FER), multimodal, textured three-dimensional (3D) face scan.**

Manuscript received April 1, 2016; revised September 20, 2016 and April 7, 2017; accepted June 2, 2017. Date of publication June 8, 2017; date of current version November 15, 2017. The work of H. Li was supported in part by the NSFC under Grant 11401464, in part by the Chinese Postdoctoral Science Foundation under Grant 2014M560785, in part by the International Exchange Foundation of China NSFC, and in part by the United Kingdom RS under Grant 61711530242. The work of J. Sun was supported in part by the NSFC under Grant 61472313 and Grant 11622106. The work of L. Chen was supported in part by the French Research Agency, l'Agence Nationale de Recherche, through the Jemime project under Contract ANR-13-CORD-0004-02 and the Biofence project under Contract ANR-13-INSE-0004-02, and in part by the PUF 4D Vision project funded by the Partner University Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sen-Ching Samson Cheung. *(Corresponding author: Jian Sun.)*

H. Li, J. Sun, and Z. Xu are with the Institute for Information and System Sciences, School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: huibinli@mail.xjtu.edu.cn; jiansun@mail.xjtu.edu.cn; zbxu@mail.xjtu.edu.cn).

L. Chen is with the LIRIS UMR 5205, Department of Mathematics and Informatics, Ecole Centrale de Lyon, Lyon 69134, France (e-mail: liming.chen@ec-lyon.fr).

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/TMM.2017.2713408

I. Introduction

Facial expressions, as a form of nonverbal communication, and a primary means of conveying social information among humans, are ideal for human emotion measurement, computation, and interpretation. Therefore, machine-based automatic facial expression recognition (FER) has a wide range of applications in human-computer interaction, facial animation, entertainment, and psychology study [3], [27], [43], [58], etc. It has been extensively investigated over the past decades in the fields of multimedia, affective computing, and computer vision [6], [12], [37], [40].

Existing FER methods generally can be classified from three perspectives, namely the data modality, expression granularity, and temporal dynamics [12], [37], [40]. From the first perspective, they are classified into: 2D FER (which uses 2D face images), 3D FER (which uses 3D face shape models), and 2D+3D multi-modal FER (which uses both 2D and 3D face data). From the second perspective, they are divided into: 1) recognition of prototypical facial expressios (i.e., anger, disgust, fear, happiness, sadness and surprise), 2) detection and recognition of facial Action Units (AU, e.g., brow raiser, lip tightener, and mouth stretch). From the third perspective, they are categorized into static (still images) or dynamic (image sequences) FER [40], [68]. In this paper, we focus on the problem of recognizing the six prototypical facial expressions using multi-modal 2D and 3D static face data (i.e., textured 3D face scans).

In the literature of FER, the majority of methods are based on 2D face images or videos (e.g., [5], [6], [8], [18], [37], [49], [55], [56], [58], [61], [62]). Despite significant advances have been achieved, 2D methods still fail to solve the challenging problems of illumination and pose variations [37]. Designing FER systems using infrared facial images is a beneficial attempt to solve the illumination issue [54], [55]. But infrared images are usually fail to capture subtle facial deformations, e.g., skin wrinkles [18], and also sensitive to the effect of wearing glasses, which is often occur in uncontrolled condition. With the fast development of 3D imaging and scanning technologies, FER using 3D face scans has attracted more and more attentions [12], [13], [16], [40]. This is mainly due to that 3D face scans are naturally robust to lighting and pose variations. Moreover, 3D facial shape deformations caused by facial muscle movements contain important cues to distinguish different expressions. To meet the requirements of real applications, FER based on multi-modality data (e.g., visual and audio [49], visible and infrared face images [54], [55]), especially using

TABLE I
MOTIVATIONS: A FEW EXAMPLES OF CURRENT FER RESEARCH

	Handcrafted features	Learned Features
2D FER	HOG: Hu *et al.* [7] LBP: Zhao *et al.* [65] Gabor: Zhang *et al.* [64]	Deep CNN: Yu and Zhang [60] DBN: Kahou *et al.* [19] Auto-Encoder: Rifai *et al.* [39]
3D FER	Depth-SIFT: Berretti *et al.* [1] Normal-LBP: Li *et al.* [23] Curvature-HOG: Lemaire *et al.* [22]	Learned feature for 3D FER?
2D+3D FER	Handcrafted feature-level fusion Handcrafted score-level fusion Savran *et al.* [42], Li *et al.* [24]	Learning-based fusion for 2D+3D FER?

both 2D face images and 3D face models [16], [24] [42], [50], is becoming a promising research direction due to that there exist large complementarity among different modalities.

This paper is a new attempt along this promising direction, which dedicates to exploring multi-modal 2D+3D FER method by combing the advantages of both 2D and 3D face data. The main challenges of such combination involve the following two issues: *1) how to find a unified framework to generate discriminative facial representations for both 2D and 3D face data? 2) how to optimally combine the facial representations of 2D and 3D face data for expression prediction?*

As illustrated in Table I, handcrafted features such as HOG [7], LBP [65], and Gabor [64] have been widely used for facial representations in 2D FER. Similarly, these handcrafted features have also been widely employed in 3D FER, which are used to describe 3D facial shape information by coding different types of geometric maps like depth-SIFT [1], normal-LBP [23], and curvature-HOG [22]. Recently, with the significant breakthrough of deep learning, such kind of handcrafted features have been proven to be suboptimal. Thanks to the continuous updating and releasing of large 2D expression datasets (e.g., Acted Facial Expressions in the Wild (AFEW) [10] and Static Facial Expressions in the Wild (SFEW) [9]), leaning facial representations using deep learning is becoming the mainstream in 2D FER. For example, following the Emotion Recognition in the Wild (EmotiW) Grand Challenge, a large number of deep learning based approaches, such as deep convolutional neural network (CNN) [60], deep belief network (DBN) [19], and auto-encoder [39] have been successfully used in 2D FER as shown at the right side of Table I.

However, to the best of our knowledge, deep learning has never been used to learn 3D facial representations in 3D FER. This motivates us to fill this gap although a very limited number of 3D face scans with expression labels are available. Inspired by the fact that the *off-the-shelf* pre-trained deep CNN models have surprising and consistent good generalization ability for various visual recognition tasks [11], [38], A promising way is using transfer learning method that fine tunes a pre-trained deep CNN model using as many as possible 3D face data.

Deep CNN can provide a unified framework to learn facial representations for both 2D and 3D face data. Then, how to find a strategy to optimally combine these learned 2D and 3D facial representations is becoming the key issue. As illustrated in Table I, the suboptimal handcrafted feature-level fusion and score-level fusion are widely used in current multi-modal 2D+3D FER methods. The importance weights of 2D and 3D facial features have not be well explored. This motivates us to design a learning-base fusion strategy, i.e., a novel deep fusion network, which can automatically learn sophisticated fusion weights of 2D and 3D facial representations for multi-modal 2D+3D FER. Overall, this paper presents a unified end-to-end learning framework (i.e., Deep Fusion CNN or DF-CNN), which can deal with both feature learning and fusion learning for multi-modal 2D+3D FER. Therefore, the main novelties and contributions of this paper can be summarized as follows:

1) This is the first work of introducing deep CNN to 3D FER and using learned features to describe 3D facial expressions. To overcome the issue that training 3D faces are far from enough, we propose to use multiple types of facial attribute maps to learn facial representations by fine tuning pre-trained deep CNN models trained from large-scale image dataset for generic visual tasks.
2) This paper proposes to use a deep fusion net (i.e., a learning-based feature-level fusion) to learn the optimal combination weights of 2D and 3D facial representations for multi-modal 2D+3D FER. This is totally different from the suboptimal handcrafted feature-level fusion and score-level fusion used in existing 2D+3D FER.
3) This paper presents a Deep Fusion CNN, which combines feature learning and fusion learning into a unified end-to-end training framework, and consistently outperforms the handcrafted features, pre-trained deep features, fine-tuned deep features, and state-of-the-art 3D FER methods on three 3D face datasets.

The remainder of this paper is organized as follows. Related works for 2D, 3D and 2D+3D FER are introduced in Section II. Section III gives an overview of the proposed approach. Section IV introduces the computational details of generating different facial attribute maps. Section V describes our DF-CNN in detail, involving net architecture, training strategy, and visualization. Experimental results are shown in Section VI, and Section VII concludes the paper.

II. RELATED WORKS

A. Related Works on 3D and 2D+3D FER

Current 3D FER approaches are mainly *model-based* or *feature-based* [12]. *Model-based* methods generally employ dense rigid registration and non-rigid fitting techniques to get the one-to-one point correspondence among face scans. This generates a generic expression deformable model, which can be used to fit unknown face scans, and the fitting parameters are finally used as expression features. For example, Mpiperis *et al.* [35] proposed to build a novel bilinear facial deformable model to characterize the behaviors of facial non-rigid deformations. Given a new 3D face model, its expression and identity parameters can be estimated using the well-trained bilinear model. These parameters are then used as expression features and fed into the Maximum Likelihood classifier for expression prediction. Similarly, Gong *et al.* [15] suggested to learn a model to decompose the shape of an expressive face into a neutral-style basic facial shape component (BFSC) and an expression shape component (ESC). The ESC is then used to design expression features. Zhao *et al.* [66] proposed to build a statistical facial feature model (SFAM) for automatic facial landmarking, both 3D shape and 2D texture features are extracted around these landmarks for expression recognition. *Feature-based* methods generally extract local expression features around facial landmarks based on surface geometric attributes or differential quantities. For example, 3D landmark distances [44], [45], [46], [47], local surface patch distances [24] [32] [33], geometry and normal maps [36], conformal images [63], surface normal [26] and curvatures [26], [53] are some popular features use for 3D FER. As a typical local feature-based method, Maalej *et al.* [32] [33] proposed to extract local surface patches around 70 facial landmarks in the 3D mesh. These patches were then parameterized by a set of closed iso-level curves at the landmarks. The distance between two patches was computed by the geodesic distance of deforming their corresponding iso-level curves in the Riemannian shape analysis space. Finally, multi-boosting and support vector machines (SVM) classifiers were used to classify the six prototypical facial expressions. By combing the advantages of both feature-based and model-based methods, Zhen *et al.* [67], [68] proposed to study 3D FER problem from the perspective of facial muscular movement model. Their method first automatically segments 3D face shapes into several facial regions according to the muscular movement model. Then, each region is described by a set of geometric features. The weights of different regions are learned by genetic algorithm, and SVM classifier with score-level fusion is used for expression prediction. Savran *et al.* [42] utilized multi-modal 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). Li *et al.* [24] proposed a fully automatic multi-modal 2D+3D feature based FER approach. Both 2D texture descriptors and 3D geometry descriptors are used to describe the appearances and geometric deformations of local facial patches around automatically detected 2D and 3D facial landmarks. The complementarity between 2D descriptors, 3D descriptors, and 2D+3D descriptors are demonstrated in their experiments based on both feature-level and score-level fusion strategies of the SVM classifier.

The main weakness of model-based methods lie in that they require to establish dense correspondence among face scans, which is still a challenging issue. Moreover, time consuming procedures like dense 3D face registration and model fitting are usually indispensable in practice. *Feature-based* methods generally perform better than *model-based* ones. However, their performances are largely dependent on the accuracy of 3D facial landmarking, which is also a challenging task [12]. FER based on 2D+3D multi-modal data is becoming a promising research direction due to that there exist large complementarity among different modalities. Giving a complete survey for 3D FER is out the scope of this paper, readers are strongly suggested to refer to the comprehensive survey [40] for the issues of 3D and 4D face acquisition, dense correspondence, alignment, tracking, available databases, as well as the details of feature extraction, selection, classification, and temporal modeling for static and dynamic 3D facial expression recognition.

B. Related Works on 2D FER

Rifai *et al.* [39] designed a multi-scale contractive convolutional network to learn hierarchical expression features which are robust to the variations of factors like pose, identity, morphology of the face. Tang [48] demonstrated the advantages of replacing the softmax loss function of a deep CNN by a linear SVM loss for 2D FER. Liu *et al.* [30] proposed a unified Boosted Deep Belief Network framework to iteratively optimizing the expression training process of feature learning, feature selection, and classifier construction. Burkert *et al.* [2] proposed a convolutional neural network (CNN) architecture for 2D FER and claimed that it outperforms the earlier proposed CNN based approaches. Liu *et al.* [29] designed a 3D CNN incorporating a deformable parts learning component for dynamic expression analysis. The authors also proposed the action unit inspired deep networks for 2D FER [28]. Khorrami *et al.* [20] showed both qualitatively and quantitatively that CNNs can learn facial action units when doing expression recognition, and their method achieved state-of-the-art performance on the extended Cohn-Kanade (CK+) and the Toronto Face Dataset (TFD). Kahou *et al.* [19] developed a deep learning approach for emotion recognition in video. Their method respectively trained a CNN for video and a deep belief net for audio. "Bag of mouths" features are also extracted to further improve the performance. To fusion different models, the ensemble weights are determined with random search. The idea of ensemble multiple deep models has also been used in Kim *et al.* [21]. This work trained 216 deep CNNs by varying network architectures, input normalization, and weight initialization and by adopting several learning strategies. Then, the valid-accuracy-based exponentially-weighted decision fusion method was proposed to ensemble different CNNs.

The work by Yu and Zhang [60] is probably the most related work to ours. This method proposed to independently train multiple differently initialized CNNs and output their training responses. To combine multiple CNN models, they proposed to learn the ensemble weighs of the network responses by minimizing the log likelihood loss or hinge loss. Despite with the same spirit of fusing deep models, our proposed learning strategy differs from [60] significantly. First, they trained multiple CNNs by varying the network initialization, while we only need to train a single CNN for different facial attribute maps. As shown in our experiments (Section VI-D), this kind of single network training can largely reduce both computate time and memory consumption, while still preserve the accuracy. Second, their method learned different weights for different networks, thus corresponding to a learning-based score-level fusion strategy, while ours corresponds to a learning-based feature-level fusion strategy.

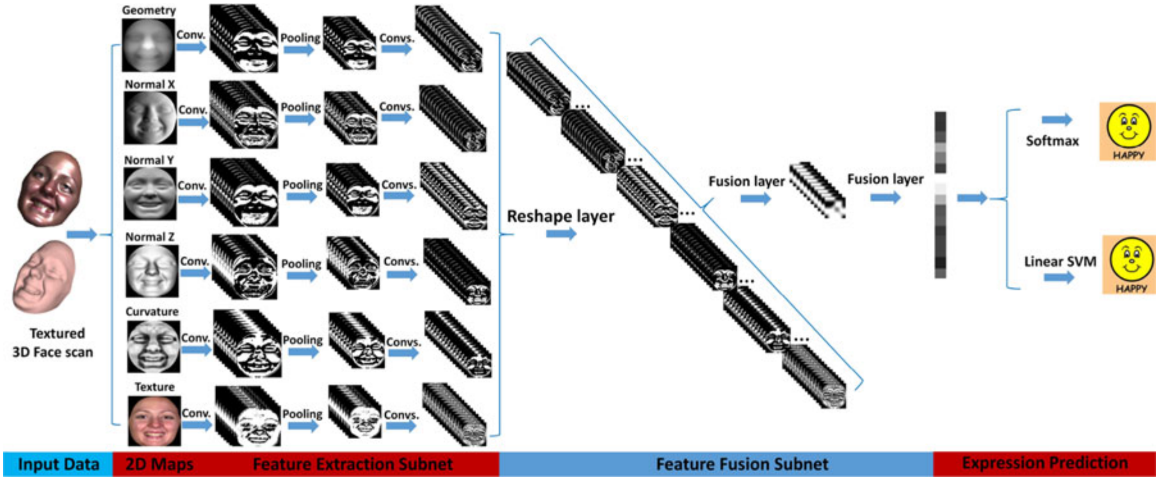


Fig. 1. Pipeline of the proposed DF-CNN-based multimodal 2D+3D FER approach. Each textured 3D face scan is represented as six types of 2D facial geometric and photometric attribute maps (i.e., 3D coordinates based geometry map, normal vectors based normal maps, principle curvatures based curvature map, and texture map). These attribute maps are jointly fed into the feature extraction subnet of DF-CNN with sharing parameters, generating hundreds of multi-channel feature maps. All these feature maps are then fed into the feature fusion subnet (including a reshape and two fusion layers) of DF-CNN, resulting in a highly concentrated facial representation (32-dimensional fused deep feature). Finally, the softmax-loss layer is followed for network training (see Section V-A for details). The final expression label prediction is performed by two ways: learning linear SVM classifiers using the 32-dimensional fused deep features or directly performing softmax prediction based on the six-dimensional probability vectors.

III. Overview of the Proposed Approach

Fig. 1 illustrates the pipeline of the proposed DF-CNN approach for 2D+3D FER. Given a set of preprocessed textured 3D face scans with different expressions, each of which is first represented as six types of 2D facial attribute maps (see Section IV), including geometry map (3D coordinates), three normal component maps (normal vectors), normalized curvature map (principle curvatures), and texture map. Then, these six facial attribute maps of each textured 3D face scan are jointly fed into the feature extraction subnet (repetitions of convolution, ReLU, and pooling layers) with sharing parameters, resulting in several hundreds of multi-channel feature maps. All these feature maps are fed into the following feature fusion subnet (including a reshape and two feature fusion layers), leading to a highly concentrated facial representation (32-dimensional fused deep feature). Finally, the softmax-loss or softmax layer is followed for network training or expression prediction (see Section V-A for details).

For DF-CNN training, considering that there are very limited numbers of textured 3D face scans with expression labels, the feature extraction subnet is initialized using the *off-the-shelf* convolutional layers of a pre-trained deep model (e.g., *vgg-net-m*). This kind of pre-trained deep models have been proven to have a good generalization ability for generic visual recognition tasks [11], [38]. The feature fusion subnet is randomly initialized, and the whole net is trained by the back-prorogation algorithm using the softmax-loss function and the stochastic gradient descent (SGD) algorithm.

For DF-CNN testing, six facial attribute maps of each textured 3D face scan are jointly fed into the feature extraction and feature fusion subnets, generating a highly concentrated facial representation (32-dimensional fused deep feature). This deep feature is further transformed into a 6-dimensional expression probability vector by the final softmax layer. Expression label prediction is preformed by training linear SVM classifiers using the 32-dimensional fused deep features (i.e., DF-CNN$_{\text{svm}}$) or directly performing softmax prediction based on the 6-dimensional probability vectors (i.e., DF-CNN$_{\text{softmax}}$).

IV. Attribute Maps of a Textured 3D Face

To comprehensively describe the geometric and photometric attributes of a textured 3D face scan, six types of 2D facial attribute maps, namely the geometry map, texture map, three normal maps, as well as normalized curvature map are employed. Given a raw textured 3D face scan, we first run the preprocessing pipeline algorithm (see Section VI-A) to generate a 2D texture map I_t and a geometry map I_g. The coordinates information of each geometry map are then used to estimate the surface normals and curvatures, resulting in three normal component maps I_n^x, I_n^y, and I_n^z, and one normalized curvature (i.e. shape index) map I_c. Finally, a textured 3D face scan I can be described by six types of 2D facial attribute maps: $I = \{I_g, I_n^x, I_n^y, I_n^z, I_c, I_t\}$, as shown in Fig. 2. The details for generation of normal maps and curvature map are introduced as follows.

A. Normal Maps

Given a normalized facial geometry map I_g represented by a $m \times n \times 3$ matrix

$$I_g = [p_{ij}(x,y,z)]_{m\times n} = [p_{ijk}]_{m\times n\times\{x,y,z\}} \tag{1}$$

where $p_{ij}(x,y,z) = (p_{ijx}, p_{ijy}, p_{ijz})^T, (1 \leq i \leq m, 1 \leq j \leq n, i,j \in \mathbb{Z})$ represents the 3D coordinates of point p_{ij}. Let its unit normal vector matrix $(m \times n \times 3)$ be

$$I_n = [n(p_{ij}(x,y,z))]_{m\times n} = [n_{ijk}]_{m\times n\times\{x,y,z\}} \tag{2}$$

where $n(p_{ij}(x,y,z)) = (n_{ijx}, n_{ijy}, n_{ijz})^T, (1 \leq i \leq m, 1 \leq j \leq n, i,j \in \mathbb{Z})$ denotes the unit normal vector of p_{ij}. In this paper, we utilize the local plane fitting method [17] to estimate I_n. That is to say, for each point $p_{ij} \in I_g$, its normal vector

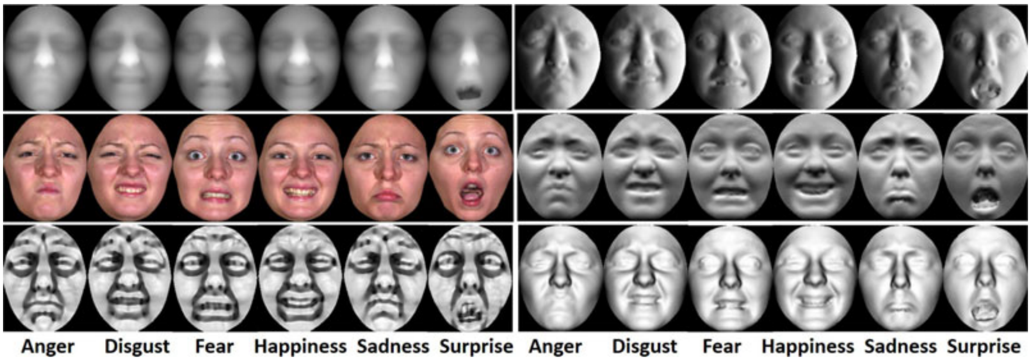


Fig. 2. Illustration of the six types of 2D geometric and photometric facial attribute maps of six textured 3D face scans (subject F0001 in the BU-3DFE dataset) with six prototypical facial expressions (i.e., anger, disgust, fear, happiness, sadness, and surprise). The left hand column shows: the geometry maps, texture maps, and curvature maps, and the three normal maps (components x, y, and z) are shown at the right hand column.

$n(p_{ij})$ can be estimated as the normal vector of the following local fitted plane:

$$S_{ij} : n_{ijx}q_{ijx} + n_{ijy}q_{ijy} + n_{ijz}q_{ijz} = d \tag{3}$$

where $(q_{ijx}, q_{ijy}, q_{ijz})^T$ represents any point within the local neighborhood of point p_{ij} and $d = n_{ijx}p_{ijx} + n_{ijy}p_{ijy} + n_{ijz}p_{ijz}$. In this work, a neighborhood of 5 × 5 window is used. To simplify, each normal component in (2) can be represented by an $m \times n$ matrix

$$I_n = \begin{cases} I_n^x = [n_{ij}^x]_{m\times n}, \\ I_n^y = [n_{ij}^y]_{m\times n}, \\ I_n^z = [n_{ij}^z]_{m\times n} \end{cases} \tag{4}$$

where $\|(n_{ij}^x, n_{ij}^y, n_{ij}^z)^T\|_2 = 1$.

B. Curvature Map

Similar to the local plane fitting method used for normal estimation, we explored the local cubic fitting method [14] to estimate the principle curvatures. This method assumes that the local geometry of a surface is approximated by a cubic surface patch. For robustly solving the local fitting problem, both the 3D coordinates and the normal vectors of the neighboring points of the point $p_{ij} \in I_g$ to be estimated are used. That is, we are fitting the following equations:

$$\begin{cases} z(x,y) = \frac{a}{2}x^2 + bxy + \frac{c}{2}y^2 + dx^3 + ex^2y + fxy^2 + gy^3 \\ z_x = ax + by + 3dx^2 + 2exy + fy^2 \\ z_y = bx + cy + 3gy^2 + 2fxy + ex^2. \end{cases} \tag{5}$$

These equations can be solved by the least squares regression, and the shape operator $\mathbf{S}$ can be computed as

$$\mathbf{S} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

Then, the eignvalues of $\mathbf{S}$ give the two principle curvatures κ_1 and κ_2 at point $p_{ij} \in I_g$. The normalized curvatures (i.e., shape index value) at this point is defined by

$$\frac{1}{2} - \frac{1}{\pi}\arctan\left(\frac{\kappa_1 + \kappa_2}{\kappa_1 - \kappa_2}\right). \tag{6}$$

Fig. 2 shows six types of 2D geometric and photometric facial attribute maps of six textured face scans with six prototypical facial expressions of subject F0001 in the BU-3DFE database.

V. Deep Fusion Convolutional Neural Network

This section first describes the architecture and training details of DF-CNN. To intuitively highlight the discriminative ability of DF-CNN, both the highly concentrated 32-dimensional fused deep features and the expression-specific saliency maps are visualized.

A. DF-CNN: Architecture and Training

The architecture of DF-CNN is formed by a feature extraction subnet, a feature fusion subnet, and a softmax layer. The feature extraction subnet is used to generate hierarchical and over-completed facial representations (i.e., feature maps) for each type of attribute maps. And the feature fusion subnet is used to combine hundreds of feature maps from different types of attribute maps into a highly concentrated deep feature. The main building blocks of feature extraction subnet include the convolutional layers and ReLU nonlinearity, while the reshape layer and fusion layers are main components of feature fusion subnet. The details of these components are introduced as follows:

Convolutional layer and ReLU nonlinearity: A convolutional layer transforms a 3D volume of activation maps (i.e., feature maps) to another through a set of learnable 3D filters. In particular, input a volume of activation maps of the previous layer $\mathcal{Y}^{l-1} \in \mathbb{R}^{W_{l-1}\times H_{l-1}\times D_{l-1}}$, and K_l 3D filters $\{\mathcal{W}_k^l\}_{k=1}^{K_l}$, each with size $W_f^l \times H_f^l \times D_{l-1}$, it outputs a 3D volume of activation maps $\mathcal{Y}^l \in \mathbb{R}^{W_l\times H_l\times D_l}$ at layer l. Let the convolutional stride be S, and the amount of zero padding be P, then we have $W_l = (W_{l-1} - W_f^l + 2P)/S + 1$, $H_l = (H_{l-1} - H_f^l + 2P)/S + 1$, and $D_l = K_l$. The k-th 2D activation map $\mathcal{Y}_k^l$ is denoted by

$$\mathcal{Y}_k^l = \varphi(\mathcal{W}_k^l * \mathcal{Y}^{l-1} + b_k^l) \tag{7}$$

where $b_k^l \in \mathbb{R}$ denotes the bias term of k-th filter $\mathcal{W}_k^l$, $*$ is the convolution operator, and φ is the rectified linear units (ReLU): $\varphi(x) = \max(0, x)$.

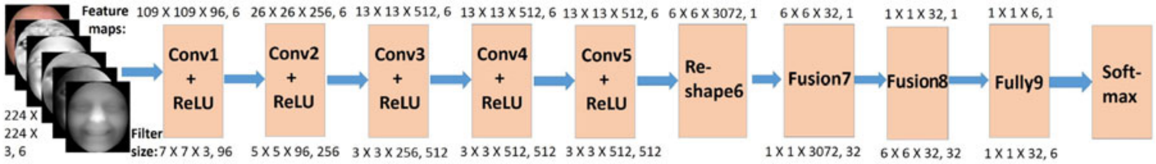


Fig. 3. Architecture of the proposed deep fusion convolutional neural network (DF-CNN). Six types of facial attribute maps of a textured 3D face model are jointly fed into five feature convolutional layers (convolution + ReLU + Pooling), a reshape layer, a feature channel fusion layer, a spatial dimension fusion layer, and a final softmax layer. The sizes and numbers of input data, feature maps and filters are listed for each layer.

Reshape Layer: This layer is used to concatenate all the 3D volumes of activation maps produced from all types of 2D facial attribute maps. Suppose DF-CNN has L convolutional layers in total, and acts on N different types of facial attribute maps, then the reshape layer operation is defined as

$$\mathcal{Y}^L_{Re} = \texttt{Reshape}(\{\mathcal{Y}^L(I_i)\}_{i=1}^N) = [\mathcal{Y}^L(I_1)|, \cdots, |\mathcal{Y}^L(I_N)] \in \mathbb{R}^{W_L \times H_L \times (K_L \times N)} \tag{8}$$

where I_i is i-th type of facial attribute maps, and the notation $[\cdot|, \cdots, |\cdot]$ denotes the concatenation of 3D matrices along feature channel dimension.

Feature channel fusion layer: This is a fully connected layer, which is used to fuse all the activation volumes extracted from all types of facial attribute maps in feature channel dimension. Let this feature channel fusion layer be the $(L+1)$-th layer, and its input be the output of the reshape layer $\mathcal{Y}^L_{Re}$, which is fully connected with K_{L+1} 3D filters $\{\mathcal{W}^{L+1}_k\}_{k=1}^{K_{L+1}}$, each with size $1 \times 1 \times (K_L \times N)$, then the output of this layer is $\mathcal{Y}^{L+1} \in \mathbb{R}^{W_{L+1} \times H_{L+1} \times D_{L+1}}$. Here $W_{L+1} = W_L$, $H_{L+1} = H_L$, and $D_{L+1} = K_{L+1}$. The k-th 2D activation map $\mathcal{Y}^{L+1}_k$ is denoted by

$$\mathcal{Y}^{L+1}_k = \varphi(\mathcal{W}^{L+1}_k * \mathcal{Y}^L_{Re} + b^{L+1}_k) \tag{9}$$

where $b^{L+1}_k \in \mathbb{R}$ denotes the bias term of k-th filter $\mathcal{W}^{L+1}_k$. That is to say, to achieve an activation volume $\mathcal{Y}^{L+1}$ with much smaller number of feature channels, the number of filters K_{L+1} should be much smaller than the number of feature channels in the previous activation volume $\mathcal{Y}^L_{Re}$.

Spatial dimension fusion layer: This is also a fully connected layer, which is used to fuse the activation volume $\mathcal{Y}^{L+1}$ in the height-width spatial dimension. Let this fusion layer be the $(L+2)$-th layer, and its input be the output volume of feature channel fusion layer $\mathcal{Y}^{L+1}$, which is fully connected with K_{L+2} 3D filters $\{\mathcal{W}^{L+2}_k\}_{k=1}^{K_{L+2}}$, each with size of $\mathcal{W}^{L+2}_k \in \mathbb{R}^{W_{L+1} \times H_{L+1} \times K_{L+1}}$, then the output activation feature of this layer is $\mathcal{Y}^{L+2} \in \mathbb{R}^{1 \times 1 \times K_{L+2}}$. The k-th 2D value $\mathcal{Y}^{L+2}_k$ is denoted by

$$\mathcal{Y}^{L+2}_k = \varphi(\mathcal{W}^{L+2}_k * \mathcal{Y}^{L+1} + b^{L+2}_k) \tag{10}$$

where $b^{L+2}_k \in \mathbb{R}$ denotes the bias term of k-th filter $\mathcal{W}^{L+2}_k$.

Softmax layer: Given K possible expression classes, the softmax layer has K nodes denoted by p_i, where $i = 1, 2, \cdots, K$. p_i specifies a discrete probability distribution of expressions, therefore, $\sum_{i=1}^K p_i = 1$. Let $\mathcal{Y}^{L+2}$ be the output of spatial dimension fusion layer, and $\{\mathcal{W}^{L+3}_k \in \mathbb{R}^1 \times 1 \times K_{L+2}\}_{k=1}^K$ be K weights fully connecting spatial dimesion fusion layer to softmax layer. Then the total input into a softmax layer, denoted by $\mathcal{Y}^{L+3}$, is

$$\mathcal{Y}^{L+3}_k = \mathcal{W}^{L+3}_k \mathcal{Y}^{L+2} + b^{L+3}_k \in \mathbb{R} \tag{11}$$

then we have

$$p_i = \frac{\texttt{exp}(\mathcal{Y}^{L+3}_k)}{\sum_j^6 \texttt{exp}(\mathcal{Y}^{L+3}_j)}. \tag{12}$$

The predicted expression class $\hat{i}$ would be

$$\hat{i} = \arg\max_i p_i. \tag{13}$$

In practice, considering that there are very limited numbers of 3D face scans with expression labels, we use the convolutional architecture and parameters of a pre-trained deep CNN model to build and initialize the convolutional layers. In particular, we choose *vgg-net-m* [4] as the pre-trained deep model since it performs well and involves moderate amount of parameters. In principle, other pre-trained deep CNN models or newly designed deep CNN models are also possible to be used if enough numbers of training samples are available. The parameters of fusion layers and softmax layer are randomly initialized. The detailed architecture of DF-CNN, including the sizes and numbers of filters and activation maps for each layer, is illustrated in Fig. 3.

As shown in Fig. 3, DF-CNN comprises five convolutional layers, a reshape layer, two fusion layers, and a softmax layer. Moreover, ReLU neuron is used after all convolutional layers and feature fusion layers. The max pooling layer is used following the first, second, and the fifth convolutional layers. And Local Response Normalization (LRN) layer is used before the first and second pooling layers.

Each 2D facial attribute map is converted to color scale and resized to 224 × 224 × 3, and then all six types of attribute maps of each textured 3D face scan are jointly fed into feature extraction subnet of DF-CNN, generating six activation volumes, each with size of 6 × 6 × 512. These six activation volumes are concatenated and reshaped into size of 6 × 6 × 3,072 by reshape layer (Reshape6 in Fig. 3). The reshaped activation volumes are fused by the following feature channel fusion layer (Fusion7 in Fig. 3), resulting in an activation volume with size of 6 × 6 × 32. This activation volume is further fused by spatial dimension fusion layer (Fusion8 in Fig. 3), generating a highly concentrated facial representation (i.e., 32-dimensional fused deep feature). This fusion layer is followed by another fully connected layer, which outputs a 6-dimensional expression probability vector. Finally, a softmax loss layer is used to train all the parameters of DF-CNN based on the back-propagation algorithm.

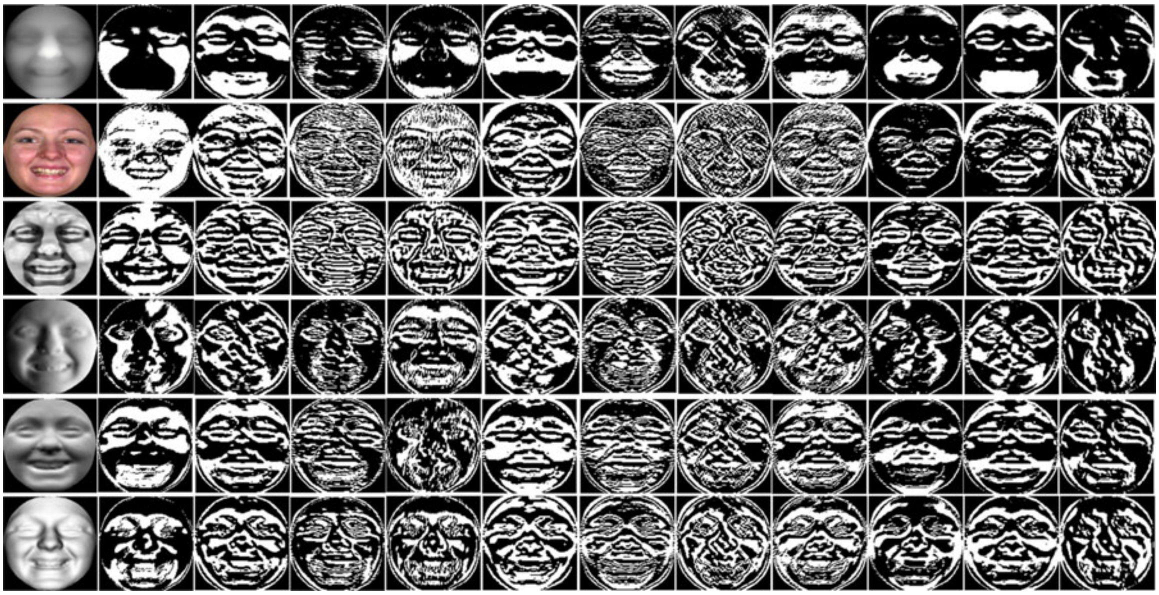


Fig. 4. Visualization of 11 typical feature maps of geometric and photometric facial attribute maps extracted from the first convolution layer of DF-CNN. From top to bottom are the feature maps for the geometry map, texture map, curvature map, and normal maps with components x, y, and z.

During training, the weight decay parameter is set to 5e-4. The learning rate and momentum parameters are set to 1e-4 and 0.9, respectively. The open source implementation MatConvNet[1] is used to build DF-CNN. During testing, expression label of a textured 3D face scan is predicted by two ways: 1) training linear SVM classifiers using the highly concentrated 32-dimensional deep features (i.e., DF-CNN$_{\text{svm}}$). 2) performing softmax prediction based on the 6-dimensional vectors of expression probabilities (i.e., DF-CNN$_{\text{softmax}}$).

B. DF-CNN: Deep Feature Visualization

To have an intuitive impression and gain insight into the discriminative ability of DF-CNN, we visualize both the "low-level" and "high-level" deep features extracted from the first convolution layer and the last fusion layer of DF-CNN, respectively. Fig. 3 shows that there are totally 96 3D filters in the first convolution layer, thus we can generate 96 "low-level" feature maps for each type of facial attribute maps. Fig. 4 illustrates 11 typical feature maps for each type of facial attribute maps of a textured 3D face scan with happiness expression. From this figure, we can see that diverse feature maps can be extracted from DF-CNN using different filters and different attribute maps. Moreover, each feature map looks similar to conventional gradient-like facial maps extracted from the shadow handcrafted features (e.g., LBP and Gabor face maps in [25]). Such a large number of feature maps can comprehensively capture various expression-related facial shape or texture deformations, of course with very high dimensions. Therefore, how to combine such a large number of over-completed and redundant deep representations into a single compact facial representation becomes the key issue to be solved. Fortunately, DF-CNN is designed to handle this problem, and providing us a high-level, low-dimensional, and high discriminative facial representation.

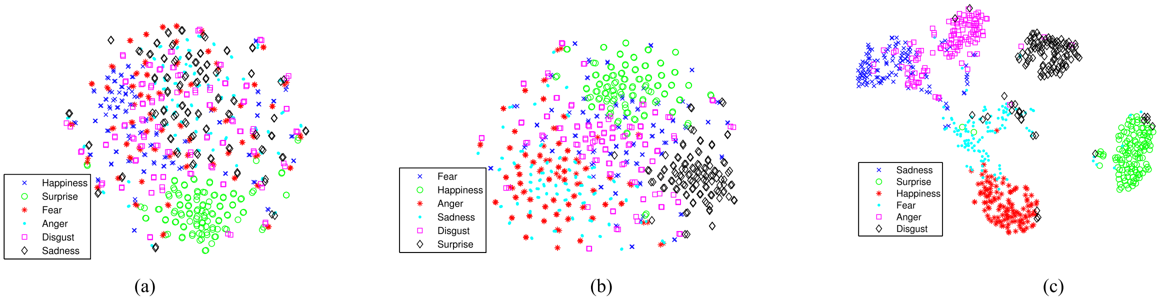
To highlight the high discriminative property DF-CNN, Fig. 5 visualizes the clustering structures of t-SNE [51] based 2-dimensional embedding of handcrafted feature, pre-trained deep feature, and 32-dimensional fused deep feature associated with six prototypical facial expressions. In particular, the same features (i.e., Gabor, *vgg-net-m-conv5*, and 32-dimensional fused deep feature) are used as those in Section VI-B. Notice that for Gabor and *vgg-net-m-conv5*, the features of different attribute maps are concatenated together (i.e., feature-level fusion) to generate a single high-dimensional representation of each textured 3D face scan. The feature dimensions of Gabor and *vgg-net-m-conv5* are 40,320 (6 attribute maps, each one is described as a 6,720-dimensional Gabor feature) and 110,592 (6 attribute maps, each one is described as feature maps with size $6 \times 6 \times 512$), respectively. Fig. 5 shows that the 32-dimensional fused deep feature has an obvious clustering structure for different expression categories, while other two types of features demonstrate large category-wised overlapping. This clearly indicates that the 32-dimensional fused deep features learned by DF-CNN has more discriminative power to distinguish different expressions than handcrafted feature Gabor and pre-trained deep feature *vgg-net-m-conv5*.

C. DF-CNN: Saliency Map Visualization

Since different facial expressions relate to different ways of local facial shape deformations, the importance weights of different facial parts are generally quite different for expression predicting as shown in [31], [69]. In this section, we show that the importance weights can be revealed by pixel-level expression related saliency maps of DF-CNN.

To this end, we visualize the importance of each pixel for its final discrimination ability of different facial

[1] [Online]. Available: http://www.vlfeat.org/matconvnet/



(a) (b) (c)

Fig. 5. Comparison of the clustering structures of t-SNE-based two-dimensional embedding of the handcrafted feature (i.e., Gabor), pre-trained deep feature (i.e., vgg-net-m-conv5), and 32-dimensional fused deep feature learned by DF-CNN associated with six prototypical facial expressions.

expressions. For example, for "happiness", we visualize the saliency map for a textured 3D face by the importance of each image pixel contributing to the final discrimination of "happiness". To compute the saliency map of a textured 3D face scan $I_\Lambda = \{I_g, I_n^x, I_n^y, I_n^z, I_c, I_t\}$ w.r.t. an expression indexed by e, we construct a score function for assigning this face to expression e by

$$S(I_\Lambda|e, \Theta) = {w_e}^T f(I_\Lambda, \Theta) \tag{14}$$

where Θ is the set of learned parameters (i.e., filters and biases) of DF-CNN, $f(I_\Lambda, \Theta)$ is the 32-dimensional fused deep feature of I_Λ, and w_e is the weight of a trained SVM classifier for expression e using $f(I_\Lambda, \Theta)$. Obviously, the higher value of ${w_e}^T f(I_\Lambda, \Theta)$ implies higher confidence in labeling this textured 3D face as expression e. We next compute the gradient of score function in (14) w.r.t. the input pixels

$$G(x|I_\Lambda, e, \Theta) = \sum_{I \in I_\Lambda} {w_e}^T \frac{\partial f(I_\Lambda, \Theta)}{\partial I(x)} \tag{15}$$

where x denotes any pixel of an attribute map. $\frac{\partial f(I_\Lambda, \Theta)}{\partial I(x)}$ is the gradient of fused deep feature w.r.t. the attribute map I at pixel x, which can be computed by the back-propagation algorithm of DF-CNN from the spatial dimension fusion layer. Its absolute value $|G(x|I_\Lambda, e, \Theta)|$ measures the importance of pixel x in labeling I as expression e. We call this term computed over all pixels of all facial attribute maps of a textured 3D face scan as *saliency map*.

Fig. 6 visualizes some examples of saliency maps for different expressions. The saliency map is re-scaled to [0, 1]. We visualize it by fusing the face texture map with a dark blue background using the saliency map as weights. The less important pixels are shown in dark blue in these maps. We observe some interesting phenomena from these maps. First, mouth is the most salient facial part for discriminating all these expressions of interest, particularly for sadness and surprise. Second, the distributions of those salient maps for all expressions are approximately consistent with the patterns of facial shape deformations, which may spread over the whole faces with different importance. These observations indicate that the proposed DF-CNN can provide a discriminative facial representation and can distinguish facial expressions using the discriminative facial parts.

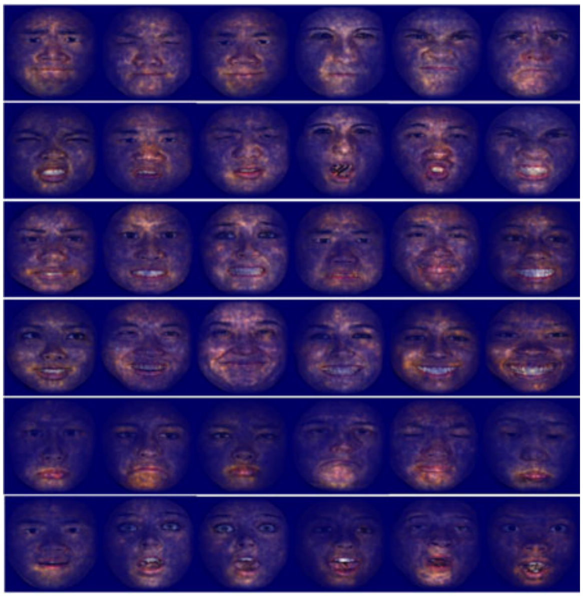


Fig. 6. Visualization of the DF-CNN based facial expression saliency maps. From top to bottom rows: saliency maps for anger, disgust, fear, happiness, sadness, and surprise. The less important pixels are shown in dark blue.

VI. Experimental Evaluation

To evaluate the effectiveness of DF-CNN for multi-modal 2D+3D FER, we will compare its performance with popular handcrafted features, pre-trained deep features, fine-tuned deep features, and state-of-the-art methods over three expression subsets of two 3D face datasets (i.e., BU-3DFE and Bosphorus). Finally, we will discuss the issues of feature extraction with or without parameter sharing, effectiveness of learning-based fusion, and optimality of linear SVM based expression prediction.

A. Databases and Preprocessing

BU-3DFE database: The BU-3DFE (Binghamton University 3D Facial Expression) Database [59] has been the benchmarking for static 3D FER [12]. It includes 100 subjects (56 females and 44 males), with age ranging from 18 to 70 years old, and with a variety of racial ancestries (e.g., White, Black, East-Asian). Each subject has 25 samples of seven expressions: one sample for neutral, and other 24 samples for six prototypical

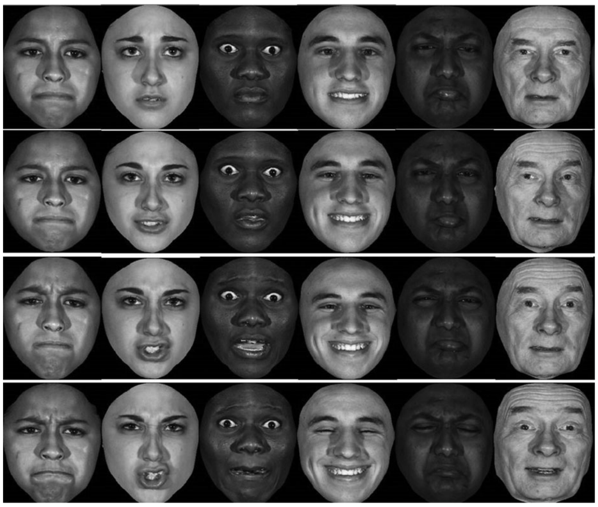


Fig. 7. Samples of 2D texture maps of BU-3DFE database with different genders, ethnicities, ages, expressions (from left to right, anger, disgust, fear, happiness, sadness, and surprise), and levels of expression intensity (from top to bottom: level 1 to level 4).

expressions (anger, disgust, fear, happiness, sadness, and surprise), each includes four levels of intensity (see Fig. 7). As a result, this database consists of 2,500 2D texture images and 2,500 geometric shape models. To fairly compare DF-CNN with state-of-the-art methods, and to validate the effectiveness of DF-CNN for samples with lower levels of expression intensity, the following two subsets are used.

1) *BU-3DFE Subset I: This subset is the standard dataset used for 3D FER.* It contains 1,200 2D and 3D face pairs (i.e., 7,200 2D facial attribute maps) of 100 subjects with 6 prototypical expressions and two higher levels of expression intensity.
2) *BU-3DFE Subset II: This subset includes all samples of BU-3DFE except the 100 neutral samples.* It contains 2,400 2D and 3D face pairs (i.e., 14,400 2D facial attribute maps) of 100 subjects with 6 prototypical expressions of four levels of intensity. To our knowledge, the samples with lower levels of expression intensity have not been used for 3D FER.

Bosphorus 3D Face Database: The Bosphorus 3D Face Database [41] has been widely used for 3D face recognition under adverse conditions, 3D facial action unit detection, 3D facial landmarking, etc. It contains 105 subjects and 4,666 pairs of 3D face models and 2D face images with different action units, facial expressions, poses and occlusions. In this dataset, there are totally 65 subjects performing the six prototypical expressions with near frontal view. Each person has only one 2D (or 3D) sample for each expression, resulting in 390 2D and 3D face pairs. To better partition, we use the following subset for experimental evaluations.

1) *Bosphorus Subset:* It contains 360 2D and 3D face pairs (i.e., 2,160 facial attribute maps) of 60 subjects with 6 prototypical expressions.

Preprocessing: We performed similar preprocessing for both BU-3DFE subsets and Bosphorus subset. First, we used the Iterative Closest Point algorithm for 3D face registration. Then, we performed nose detection, face cropping, re-sampling, and projection procedures using the 3D face normalization method proposed in [34]. Finally, we achieved the normalized 2D range images (i.e., geometry maps) with x, y, and z coordinates. Once we have geometry maps, other geometric facial attribute maps can be estimated according to the method introduced in Section IV. The 2D texture maps of BU-3DFE dataset are generated by projecting 3D texture images with linear interpolation. Samples of preprocessed facial attribute maps of BU-3DFE database are shown in Fig. 2. And Fig. 8 illustrates some samples of 2D texture images of Bosphorus subset.

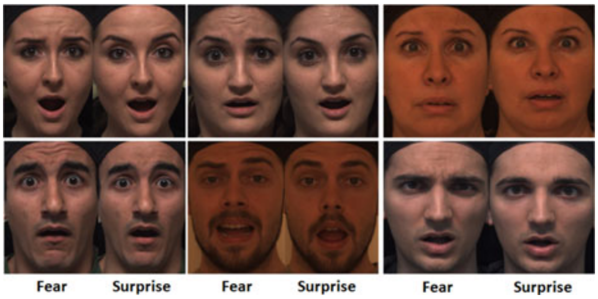


Fig. 8. Six pairs of 2D texture images with fear and surprise expressions of Bosphorus database. It's not easy even for humans to distinguish these fear and surprise pairs illustrated in this figure.

B. Evaluation and Comparison on BU-3DFE Subset I

Experimental protocol: This experimental protocol is firstly used in [15] and has been proven to be more stable than the one used in [53]. In this protocol, 60 subjects, each with 12 samples (i.e., 6 prototypical expressions with two higher levels of intensity) are randomly selected from the BU-3DFE subset I. That is to say, 720 textured 3D face scans (i.e., 4,320 2D facial attribute maps) are used. To achieve stable results, 1,000 times random and independent 54-versus-6-subject-partition experiments (1,000 times train and test sessions in total) are performed. For each partition, 648 textured 3D face scans of 54 subjects are used for training and 72 textured 3D face scans of 6 subjects are used for testing. Different partitions are independently trained and tested, and the average expression recognition accuracy of all the 1,000 test sessions across all 6 prototypical expressions are reported for the final evaluation.

In particular, we use the remaining 40 subjects (i.e., 2,880 2D facial attribute maps) of BU-3DFE Subset I to train our DF-CNN. Once DF-CNN is trained, it is then used to extract the 32-dimensional fused deep features of the other 60 subjects. These fused deep features are then used to train linear SVM classifiers for expression prediction using above 1,000 times 54-versus-6-subject-partition experiments (i.e., DF-CNN$_{\text{svm}}$). Alternatively, expression labels of the other 60 subjects are also predicted directly by the softmax layer of the trained DF-CNN (i.e., DF-CNN$_{\text{Softmax}}$). It's important to note that result of this one-time prediction is very close to (86.20% vs. 86.25%) the one achieved by predicting expression label using maximum value of the 6-dimensional expression probabilities with the same 1,000 times 54-versus-6 experimental protocol.

TABLE II
COMPARISON OF THE AVERAGE ACCURACIES WITH HANDCRAFTED FEATURES ON BU-3DFE SUBSET I

Method	I_g	I_n^x	I_n^y	I_n^z	I_c	I_t	*All*
MS-LBP	76.47	76.77	77.87	76.41	77.70	71.65	81.74
dense-SIFT	80.29	79.97	**82.35**	80.95	80.28	75.56	83.16
HOG	**81.89**	**82.09**	80.58	**81.81**	77.95	78.11	83.74
Gabor	77.95	78.80	81.97	81.10	**81.65**	**80.36**	**84.72**
DF-CNN$_{\text{svm}}$	–	–	–	–	–	–	**86.86**
DF-CNN$_{\text{softmax}}$	–	–	–	–	–	–	86.20

TABLE III
COMPARISON OF THE AVERAGE ACCURACIES WITH PRE-TRAINED DEEP FEATURES ON BU-3DFE SUBSET I

Method	I_g	I_n^x	I_n^y	I_n^z	I_c	I_t	*All*
caffe-alex-conv5	77.53	78.87	81.50	78.71	**80.83**	81.40	83.74
vgg-net-m-conv5	80.38	**80.37**	81.68	81.23	79.23	**82.14**	**84.22**
vgg-net-16-conv5-3	**81.72**	78.55	**83.06**	**81.25**	76.95	78.46	83.78
caffe-alex-full7	68.64	73.43	76.64	75.72	**74.52**	**74.45**	**82.56**
vgg-net-m-full7	73.34	**74.99**	**77.51**	**76.77**	68.81	70.93	81.56
vgg-net-16-full7	**76.71**	72.22	73.87	74.61	64.35	67.03	82.45
DF-CNN$_{\text{svm}}$	–	–	–	–	–	–	**86.86**
DF-CNN$_{\text{softmax}}$	–	–	–	–	–	–	86.20

TABLE IV
COMPARISON OF THE AVERAGE CONFUSION MATRICES WITH GABOR AND PRE-TRAINED DEEP FEATURE FOR ALL FACIAL ATTRIBUTE MAPS ON BU-3DFE SUBSET I

Gabor (average accuracy = 84.72)						
%	AN	DI	FE	HA	SA	SU
AN	**85.53**	1.67	0.93	0	11.88	0
DI	1.63	**84.48**	6.19	4.35	0	3.36
FE	3.56	6.24	**65.98**	12.70	4.34	7.18
HA	0	0.83	3.03	**96.14**	0	0
SA	18.70	0	1.20	0.95	**79.15**	0
SU	0	1.26	1.67	0	0.04	**97.03**
vgg-net-m-conv5 (average accuracy = 84.22)						
%	AN	DI	FE	HA	SA	SU
AN	**86.96**	1.68	0.83	0	10.53	0
DI	1.88	**80.43**	8.47	4.29	0	4.93
FE	3.23	9.53	**66.41**	12.83	2.10	5.91
HA	0	0.25	3.48	**96.27**	0	0
SA	19.82	0	2.83	0.39	**76.96**	0
SU	0	0.04	1.67	0	0.01	**98.28**
DF-CNN$_{\text{svm}}$ (average accuracy = 86.86)						
%	AN	DI	FE	HA	SA	SU
AN	**82.08**	3.60	2.42	0	11.90	0
DI	3.27	**84.94**	5.70	2.50	0	3.59
FE	1.84	5.28	**79.24**	8.33	0.81	4.50
HA	0	0	3.74	**96.26**	0	0
SA	12.63	0.10	5.56	0.53	**81.18**	0
SU	0	0.07	1.67	0	0.83	**97.43**

1) Comparison With Handcrafted Features: This paragraph compares the performance of DF-CNN with the ones achieved by using handcrafted features. Four classical handcrafted image features: MS-LBP, dense-SIFT, HOG, and Gabor, which have been proven to be quite efficient for both 2D and 3D facial expression analysis, are employed for comparisons. Please refer to [25], [52], [22], and [25], respectively for the implementations of these features. When used for multi-modal 2D+3D FER, these features are first extracted from each type of facial attribute maps, then respectively fed into linear SVM[2] classifier with default parameter of C. To achieve final results, score-level fusion of SVM scores with sum rule is used.

Table II shows the average expression recognition accuracies across all six expressions of four handcrafted features, and the proposed DF-CNN on BU-3DFE subset I. From Table II, we can conclude that: 1) Gabor and HOG generally perform better than dense-SIFT and MS-LBP. In particular, Gabor achieves the highest fusion accuracy of 84.72%, which outperforms HOG, dense-SIFT, and MS-LBP by 0.98%, 1.56%, and 2.98%, respectively. 2) For different facial attribute maps, normal maps (I_n^x, I_n^y, and I_n^z) generally perform better than others, and the fusion of all six attribute maps (i.e., All) achieves the best performance. These results indicate that different facial attribute maps indeed contain large complementary information for multi-modal 2D+3D FER. 3) DF-CNN$_{\text{svm}}$ and DF-CNN$_{\text{softmax}}$ achieves similar and much better results (86.86% vs. 86.20%) than handcrafted features.

2) Comparison With Pre-trained Deep Features: This paragraph compares the performance of DF-CNN with the ones achieved by using deep features extracted from three deep models (i.e., *caffe-alex*, *vgg-net-m*, and *vgg-net-16*) pre-trained on the ImageNet database [4]. Notice that the convolutional layers of *vgg-net-m* is used to initialize our DF-CNN. Similar to the case of handcrafted features, each type of facial attribute maps are separately fed into these pre-trained models to extract deep features, and then linear SVM classifiers are trained for expression classification. The final fusion results are achieved by performing score-level fusion of SVM scores with sum rule. For comparisons, deep features extracted from the 5th convolutional layer (*net-conv5*) and the penultimate fully connected layer (*net-full7*) are used for each type of facial attribute maps.

Table III shows the average expression recognition accuracies of pre-trained deep features and DF-CNN on BU-3DFE subset I. From Table III, we can find that: 1) Different pre-trained deep features have different superiorities associated with different facial attribute maps. For example, *vgg-net-16-conv5-3* achieves the best score for I_n^y, while *vgg-net-m-conv5* performs best for I_n^x. 2) For the fusion scores, *vgg-net-m-conv5* and *caffe-alex-full7* achieve slightly better results than others among pre-trained deep features. 3) The deep features extracted from convolutional layers (i.e., conv5) of pre-trained deep models generally perform much better than the ones extracted from fully connected layers (i.e., full7). 4) Our method achieves consistently better results than all pre-trained deep features. Notice that the dimension of *vgg-net-m-conv5* for one type of facial attribute maps is 18,432, which is much higher than the 32-dimensional fused deep feature produced by DF-CNN.

Table IV compares the average confusion matrices achieved by Gabor feature, *vgg-net-m-conv5* and DF-CNN$_{\text{svm}}$. It can be seen that DF-CNN$_{\text{svm}}$ outperforms Gabor for all expressions ex-

[2] [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/liblinear/

TABLE V
COMPARISON OF THE AVERAGE ACCURACIES WITH FINE-TUNED DEEP FEATURES ON BU-3DFE SUBSET I

Method	I_g	I_n^x	I_n^y	I_n^z	I_c	I_t	*All*
caffe-alex-ft-full7$_{\text{svm}}$	79.44	79.84	80.51	79.50	79.46	80.83	84.05
vgg-net-m-ft-full7$_{\text{svm}}$	79.68	**82.85**	**82.15**	80.30	**82.01**	81.62	84.85
vgg-net-16-ft-full7$_{\text{svm}}$	**80.21**	82.30	82.04	**80.43**	80.87	**84.10**	**86.01**
caffe-alex-ft$_{\text{softmax}}$	78.19	80.96	81.94	78.75	78.89	80.83	83.61
vgg-net-m-ft$_{\text{softmax}}$	**78.33**	**83.06**	**82.78**	**81.11**	**81.11**	80.42	85.00
vgg-net-16-ft$_{\text{softmax}}$	78.33	82.08	80.69	79.19	79.31	**84.17**	**85.14**
DF-CNN$_{\text{svm}}$	–	–	–	–	–	–	**86.86**
DF-CNN$_{\text{softmax}}$	–	–	–	–	–	–	86.20

TABLE VI
COMPARISON OF THE AVERAGE CONFUSION MATRICES WITH FINE-TUNED DEEP MODEL FOR ALL FACIAL ATTRIBUTE MAPS ON BU-3DFE SUBSET I

vgg-net-m-ft-full7$_{\text{svm}}$ (average accuracy = 84.85)						
%	AN	DI	FE	HA	SA	SU
AN	**81.48**	1.92	0.83	0	15.77	0
DI	1.95	**81.91**	8.58	3.14	0	4.42
FE	3.42	6.96	**73.51**	11.57	1.99	2.56
HA	0	0.77	3.48	**95.74**	0	0
SA	15.88	0.19	4.18	0	**79.75**	0
SU	0	0.83	1.67	0	0.78	**96.72**
vgg-net-16-ft-full7$_{\text{svm}}$ (average accuracy = 86.01)						
%	AN	DI	FE	HA	SA	SU
AN	**86.19**	2.52	0.83	0	10.45	0
DI	1.97	**82.00**	9.59	2.27	0	4.17
FE	2.73	7.42	**74.38**	12.99	0.79	1.68
HA	0	0.83	3.47	**95.69**	0	0
SA	16.21	0	4.29	0	**79.50**	0
SU	0	0	1.67	0	0.06	**98.27**
DF-CNN$_{\text{svm}}$ (average accuracy = 86.86)						
%	AN	DI	FE	HA	SA	SU
AN	**82.08**	3.60	2.42	0	11.90	0
DI	3.27	**84.94**	5.70	2.50	0	3.59
FE	1.84	5.28	**79.24**	8.33	0.81	4.50
HA	0	0	3.74	**96.26**	0	0
SA	12.63	0.10	5.56	0.53	**81.18**	0
SU	0	0.07	1.67	0	0.83	**97.43**

cept anger (with a difference of 3.45%). It is worth noting that DF-CNN$_{\text{svm}}$ has more powerful discriminative ability to distinguish fear expression, promoting the accuracy upto 13.26% and 12.83% for Gabor feature and *vgg-net-m-conv5*.

3) Comparison With Fine-Tuned Deep Models: To further demonstrate the effectiveness of DF-CNN, we also compared it with fine-tuned deep models. The same pre-trained deep models: *caffe-alex*, *vgg-net-m*, and *vgg-net-16* are used for fine-tuning. For each pre-trained deep model, we keep the net architecture of all layers and parameters unchanged except the final fully connected layer. In particular, since we have six expression classes, the filter weight with size of $1 \times 1 \times 4096 \times 1000$ is changed to $1 \times 1 \times 4096 \times 6$ and randomly initialized, and the corresponding 1000-dimensional bias vector is also replaced by a 6-dimensional zero vector. Then, we separately fine-tune the pre-trained deep models using different facial attribute maps, resulting in six fine-tuned deep models for each pre-trained deep model. Finally, testing data associated with each kind of attribute maps are fed into the corresponding fine-tuned deep model for feature extraction. Similar to DF-CNN, expression prediction for each kind of attribute maps is achieved by the following two ways: 1) learning linear SVM classifiers using the 4,096-dimensional fine-tuned deep features (e.g., *vgg-net-m-ft-full7*$_{\text{svm}}$); 2) performing softmax prediction using the 6-dimensional fine-tuned deep features of expression probabilities (e.g., *vgg-net-m-ft*$_{\text{softmax}}$). To achieve fusion results, score-level fusion with sum rule are used for both cases.

Table V shows the average expression recognition accuracies of fine-tuned deep features and DF-CNN on BU-3DFE subset I. From Table V, we can find that: 1) Fusion of multiple facial attribute maps can also significantly improve the accuracies for all fine-tuned deep features. 2) Fine-tuned deep feature *vgg-net-16* achieves significantly better results for texture maps, and also achieves the highest accuracies (86.01% and 85.14%) for both two prediction ways. This conclusion of deeper net performs better is consistent with the one in [4]. 3) Our DF-CNN initialized by *vgg-net-m* still achieves the best results. It's necessary to compare the results of Table III and Table V. It's easy to find that significant improvements have been achieved from pre-trained to fine-tuned deep features, particularly for the case of 4096-dimensional deep features extracted from the penultimate fully connected layer (i.e., full7). For example, the improvements are upto 16.52% for curvature maps and 17.07% for texture maps when considering the pre-trained and fine-tuned *vgg-net-16-full7*.

Table VI compares the average confusion matrices achieved by two fine-tuned deep features: *vgg-net-m-ft-full7*$_{\text{svm}}$, *vgg-net-16-ft-full7*$_{\text{svm}}$, and our DF-CNN$_{\text{svm}}$. It's not difficult to see that DF-CNN$_{\text{svm}}$ achieves consistent better results than *vgg-net-m-ft-full7*$_{\text{svm}}$ for all six expressions. It even achieves better results than *vgg-net-16-ft-full7*$_{\text{svm}}$, which is fine-tuned from a much deeper pre-trained deep model. In particular, the superiority for fear expression is upto 4.86% .

4) Comparison With Other Methods: To comprehensively evaluate the effectiveness of DF-CNN, we compared it with 18 state-of-the-art methods on BU-3DFE subset I. To give a thoroughly analysis, four aspects, including the data modality, expression feature, expression classifier, and recognition accuracy are compared in Table VII.

1) *For data modality*, we can see that all previous methods reported their results using only 3D data exception of [24] and [66]. It is worth noting that Li *et al.* [24] proposed a local feature-based multimodal 2D+3D FER method, and studied the complementarity between 2D and 3D features. However, their fusion results were produced by handcrafted feature-level and score-level fusion schemes. In contrast, our method can automatically combine different 3D geometric and 2D photometric maps into a single 32-dimensional fused deep feature.

2) *For expression feature*, one way is directly building histograms of surface geometric quantities, such as coordinates (e.g., [66], [67]), normals (e.g., [24], [26], [67]), and curvatures (e.g., [24], [26], [53], [66], [67]). Another way is extracting popular handcrafted features (e.g., HOG, SIFT, LBP, DWT) from

TABLE VII
COMPARISON OF EXPRESSION FEATURES, CLASSIFIERS, AND ACCURACIES WITH THE STATE-OF-THE-ART ON BU-3DFE SUBSET I (NOTICE THAT THE ACCURACIES IN THE LEFT COLUMN ARE ACHIEVED BY AVERAGING 100 ROUND INDEPENDENT 10-FOLD CROSS-VALIDATION TESTS, WHILE THE ONES IN THE RIGHT COLUMN ARE ACHIEVED BY AVERAGING ONLY ONE OR TWO ROUND 10-FOLD CROSS-VALIDATION TESTS)

Methods	Data	Feature	Classifier	Accuracy	
Wang *et al.* [53]	3D	curvatures/hist.	LDA	61.79	83.60
Soyel *et al.* [44]	3D	points/distance	NN	67.52	91.30
Soyel *et al.* [45]	3D	points/distance	NN	–	93.72
Tang *et al.* [46]	3D	points/distance	LDA	74.51	95.10
Tang *et al.* [47]	3D	slopes, distance	SVM	–	87.10
Mpiperis [35]	3D	deformable model	ML	–	90.50
Gong *et al.* [15]	3D	depth/PAC	SVM	76.22	–
Berretti *et al.* [1]	3D	depth/SIFT	SVM	77.54	–
Maalej *et al.* [33]	3D	facial curves	muiti-boosting	–	98.81 92.75
Li *et al.* [26]	3D	normals, curv./hist.	SVM	82.01	–
Li *et al.* [23]	3D	normals/LBP	MKL	80.14	–
Lemaire [22]	3D	curvature/HOG	SVM	76.61	–
Ocegueda [36]	3D	coordinates, normals curvatures/DWT	Logistic Reg.	–	90.40
Zeng *et al.* [63]	3D	curvatures/LBP	SRC	70.93	–
Zhen *et al.* [67]	3D	coordinates, normals, shape index	SVM	84.50	–
Yang *et al.* [57]	3D	depth, normals, curv./scattering	SVM	84.80	–
Zhao *et al.* [66]	2D+3D	intensity,coordinates, shape index/LBP	BBN	–	82.30
Li *et al.* [24]	2D+3D	meshHOG/SIFT meshHOS/HSOG	SVM	86.32	–
DF-CNN$_{\text{svm}}$	2D+3D	32-D deep feature	SVM	**86.86**	–
DF-CNN$_{\text{softmax}}$	2D+3D	6-D deep feature	Softmax	86.20	–

depth maps (e.g., [1], [15], [36], [57]), normal maps (e.g., [23], [36], [57], [22]), or curvature maps (e.g., [36], [57], [63]). As mentioned in our introduction section, all these state-of-the-art works for 3D-FER are based on handcrafted expression features. In contrast, our method can learn highly concentrated and discriminative facial representation (only 32-dimensional) from six types of facial attribute maps.

3) *For expression classifier*, SVM (e.g., [1], [15], [26]) is the most popular classifier compared with others such as Neural Networks (NN), Maximal Likelihood (ML), Bayesian Belief Net (BBN), multi-boosting, and Sparse Representation-based Classifier (SRC). It is worth noting that a majority of methods are based on SVM classifier with non-linear RBF kernel (e.g., [26], [57], [63], [67]) or using multiple kernel learning [23] to combine multiple high-dimensional features (e.g., normal-LBP in [23]), while our results are based on linear SVM classifier with default parameter.

4) *For recognition accuracy*, benefiting from the end-to-end training framework of DF-CNN, the fused deep features produced by DF-CNN have strong discriminative ability to distinguish different expressions. In particular, our method (DF-CNN$_{\text{svm}}$) achieves the highest accuracy of 86.86% compared with all state-of-the-art methods using the same (or very similar [1]) experimental protocol. Notice that both the experimental protocol used in our paper [15] and the similar one used in [1] have been proved stable since the scores are achieved by averaging 100 times independent 10-fold cross-validation tests. It should be pointed out that directly comparisons of the two accuracy columns in Table VII are far from fair since the results listed in the second column were achieved based on an unstable experimental protocol (i.e., 10-fold or 20-fold cross-validation) firstly used in [53]. For example, the accuracy of [33] is reduced from 98.81% to 92.75% when using 20-fold instead of 10-fold cross-validation. As produced by Gong *et al.* [15], the accuracies of [53], [44], [46] were dropped significantly (more than 20%) when using a more stable experimental protocol. Overall, different from state-of-the-art methods, the proposed DF-CNN combines feature learning and fusion learning into a single end-to-end training framework, and achieves the best accuracy for multimodal 2D+3D FER under the more stable experimental protocol.

TABLE VIII
COMPARISON OF THE AVERAGE ACCURACIES WITH HANDCRAFTED FEATURES, PRE-TRAINED DEEP FEATURES, AND FINE-TUNED DEEP FEATURES ON BU-3DFE SUBSET II

Method	I_g	I_n^x	I_n^y	I_n^z	I_c	I_t	*All*
MS-LBP	73.50	74.58	73.54	73.21	73.37	66.08	77.75
dense-SIFT	**76.25**	75.79	77.42	76.58	75.88	71.79	79.42
HOG	76.25	**76.88**	76.29	**77.75**	76.29	72.04	79.71
Gabor	73.04	75.00	**78.29**	76.42	**76.33**	**75.86**	**80.00**
vgg-net-m-conv5	**76.17**	75.04	76.92	**76.54**	75.54	76.42	**79.75**
vgg-net-m-full7	70.21	69.71	72.67	70.67	67.00	66.83	77.38
vgg-net-m-ft-full7$_{\text{svm}}$	75.17	**76.62**	**77.08**	75.83	**78.12**	**78.67**	**81.08**
vgg-net-m-ft$_{\text{softmax}}$	74.62	75.33	76.96	75.79	77.88	78.54	80.71
DF-CNN$_{\text{svm}}$	–	–	–	–	–	–	81.04
DF-CNN$_{\text{softmax}}$	–	–	–	–	–	–	**81.33**

C. Evaluation and Comparison on Other Datasets

This section will show more experimental results evaluated on BU-3DFE subset II and Bosphorus subset.

Experimental protocol: To get more training data and to reduce the effect of data bias for DF-CNN training, we used the standard 10-fold cross-validation (10 train and test sessions) experimental setting. That is, different DF-CNNs should be trained for different sessions, and the average recognition accuracies of 10 different DF-CNNs across all six prototypical expressions are reported for evaluations and comparisons. In particular, for BU-3DFE subset II, 100 subjects are randomly divided into 10 subsets, and for each session, 12,960 attribute maps of 90 subjects are used for training and the remaining 1,440 attribute maps of 10 subjects are used for testing. Similarly, for Bosphorus subset, 60 subjects are randomly divided into 10 subsets, and for each session, 1,944 attribute maps of 54 subjects are used for training and the remaining 216 attribute maps of 6 subjects are used for testing.

1) Results on BU-3DFE Subset II: Table VIII reports the performance comparisons of the proposed DF-CNN with handcrafted features, pre-trained deep features, and fine-tuned deep features on BU-3DFE Subset II. From this table, we can conclude that: 1) As before, Gabor feature still achieves the best results among handcrafted features. It even slightly outperforms the pre-trained deep feature *vgg-net-m-conv5* (80% vs. 79.75%).

TABLE IX
COMPARISON OF THE AVERAGE ACCURACIES WITH HANDCRAFTED FEATURES PRE-TRAINED DEEP FEATURES, AND FINE-TUNED DEEP FEATURES ON BOSPHORUS SUBSET

Method	I_g	I_n^x	I_n^y	I_n^z	I_c	I_t	*All*
MS-LBP	71.11	69.44	70.56	66.67	62.78	62.50	73.33
dense-SIFT	70.28	73.89	72.78	73.89	72.50	65.56	76.39
HOG	**72.50**	**74.22**	73.89	**74.72**	71.94	**71.94**	77.22
Gabor	67.78	73.61	**75.83**	71.61	**75.56**	70.56	**77.50**
vgg-net-m-conv5	**71.94**	72.50	73.61	71.67	72.78	73.06	**79.72**
vgg-net-m-full7	61.11	63.33	63.89	65.83	60.56	61.94	75.56
vgg-net-m-ft-full7$_{\text{svm}}$	71.67	**72.78**	74.72	**76.11**	71.94	**73.61**	79.17
vgg-net-m-ft$_{\text{softmax}}$	71.39	72.78	**75.28**	75.00	**73.33**	73.61	**79.72**
DF-CNN$_{\text{svm}}$	–	–	–	–	–	–	**80.28**
DF-CNN$_{\text{softmax}}$	–	–	–	–	–	–	80.00

2) Fine-tuned deep features achieve significantly better results than pre-trained deep features, e.g., 81.08% for *vgg-net-m-ft-full7* vs. 77.38% for *vgg-net-m-full7*. 3) Our DF-CNN based methods achieve comparable (81.04% vs. 81.08%) or slightly better (81.33% vs. 81.08%) results compared with fine-tuned deep features. It is worth noting that for each train and test session, DF-CNN only needs to train a single CNN for both feature learning and feature fusion, while fine-tuned deep feature based method needs to respectively train different deep models for different types of facial attribute maps, and respectively extract fine-tuned deep features from different deep models and combine all scores by hand. This leads to much more consumptions of training time and parameter space compared with DF-CNN. Notice that the results of two BU-3DFE subsets clearly indicate that the samples with lower levels of expression intensity are indeed much more difficult to be recognized than the higher level ones.

2) Results on Bosphorus Subset: Table IX reports the performance comparisons of the proposed DF-CNN with handcrafted features, pre-trained deep features, and fine-tuned deep features on Bosphorus subset. Similar to the conclusions achieved on BU-3DFE subset I and subset II, we have: 1) Gabor feature achieves the highest accuracy of 77.50% among handcrafted features. 2) Fine-tuned deep feature (i.e., *vgg-net-m-ft-full7*) also significantly outperforms the pre-trained one (i.e., *vgg-net-m-full7*). Note that although the pre-trained deep feature *vgg-net-m-conv5* achieves the same accuracy of 79.72% as the fine-tuned deep feature vgg-net-m-ft$_{\text{softmax}}$, the feature dimension is much higher (18,432 $\times$ 6 vs. 6). 3) Our DF-CNN based methods achieve slightly better results compared with the fine-tuned deep features. Overall, Bosphorus subset is the most difficult dataset among the three subsets used in this paper.

3) Comparison With Other Methods: To compare the performance of the proposed DF-CNN with other methods on BU-3DFE subset II and Bosphorus subset, we reproduced three state-of-the-art methods (i.e., [23], [24], and [57]) on these two datasets using the same experimental protocol (i.e., 10-fold cross-validation with the same subjects for training and testing in each train and test session) as DF-CNN. In particular, [23] and [24] are two of our previous methods. Results of [57] were reproduced using the code shared by the authors. Notice that multiple kernel learning was used in [23], non-linear SVM was used in [24] and [57] for expression prediction, respectively. For fair comparison, the non-linear SVM was replaced by linear SVM classifier, and the sum rule based score-level fusion was used for [24] and [57].

TABLE X
COMPARISON WITH THE STATE-OF-THE-ART ON THE BU-3DFE SUBSET II AND BOSPHORUS SUBSET

Method	BU-3DFE Subset II	Bosphorus subset
Li *et al.* (2012) [23]	78.50	75.83
Li *et al.* (2015) [24]	80.42	79.72
Yang *et al.* (2015) [57]	80.46	77.50
DF-CNN$_{\text{svm}}$	**81.04**	**80.28**
DF-CNN$_{\text{softmax}}$	**81.33**	**80.00**

Table X reports the performance comparisons of DF-CNN with state-of-the-art methods [23], [24] and [57] on both BU-3DFE subset II and Bosphorus subset. From this table, we can see that method [23] achieves the lowest accuracy on both subsets. Methods [24] and [57] achieve very similar results (80.42% vs. 80.46%) on BU-3DFE subset II, while method [24] performs better by 2.22% on Bosphorus subset. Our DF-CNN achieves the best results on both two subsets. Similar to the case on BU-3DFE subset I, DF-CNN has significant superiority to distinguish fear expression. For example, on the Bosphorus subset, DF-CNN$_{\text{svm}}$ achieves an average recognition rate of 65% for fear expression, which is much higher than the results of 36.67%, 51.67%, and 43.33% achieved by [23], [24], and [57], respectively. It is worth noting that distinguishing the samples of Bosphorus subset with fear expression and surprise expression is a very difficult task even for humans as illustrated in Fig. 8. From this figure, we can see that there only exist very subtle differences between fear and surprise pairs of the same person.

Overall, the proposed DF-CNN unifies feature learning and fusion learning into a single end-to-end training framework, and performs better than handcrafted features, pre-trained deep features, fine-tuned deep features, and state-of-the-art methods, resulting in a good generalization ability on BU-3DFE subset II and Bosphorus subset for multimodal 2D+3D FER.

D. Discussion

To further validate the effectiveness of DF-CNN, three issues: feature extraction with or without parameter sharing, effectiveness of learning-based fusion, and optimality of linear SVM based expression prediction are discussed in this paragraph. Noting that all the following discussions are based on BU-3DFE subset I and the corresponding experimental protocol introduced in Section VI-B.

1) Feature Extraction With or Without Parameter Sharing: As shown in Fig. 1, the CNN parameters are shared for different types of facial attribute maps in the feature extraction subnet of DF-CNN. Alternatively, different attribute maps can also been separately fed into different CNNs for feature fusion, then adding the following feature fusion and expression prediction layers. Clearly, the latter one (namely DF-CNN^a) needs to learn more parameters and thus perhaps performs better. In

TABLE XI
COMPARISON OF DF-CNN AND DF-CNN^a (WITHOUT PARAMETER SHARING, I.E., DIFFERENT ATTRIBUTE MAPS CORRESPONDING TO DIFFERENT FEATURE EXTRACTION SUBNETS) ON BU-3DFE SUBSET I

Method	Parameter	Time/epoch	Accuracy
DF-CNN$_{\text{svm}}$	$\simeq$ 50 MB	7.4 Hz	**86.86**
DF-CNN$_{\text{softmax}}$	$\simeq$ 50 MB	7.4 Hz	**86.20**
DF-CNN$^a_{\text{svm}}$	$\simeq$ 300 MB	4.1 Hz	86.48
DF-CNN$^a_{\text{softmax}}$	$\simeq$ 300 MB	4.1 Hz	85.97

TABLE XII
COMPARISON OF THE LEARNING-BASED FUSION STRATEGY (DF-CNN) WITH OTHERS ON BU-SUBSET I

Feature and fusion	linear SVM (score-level)	MKL (kernel)	DF-CNN$_{\text{svm}}$ (learning-based)	DF-CNN$_{\text{softmax}}$ (learning-based)
DF-CNN-in-conv5	84.22	85.07	84.79	83.90
DF-CNN-ft-conv5	84.17	**85.73**	**86.86**	**86.20**

TABLE XIII
COMPARISON OF DIFFERENT CLASSIFIERS OVER THE 32-DIMENSIONAL DEEP FEATURES EXTRACTED FROM DF-CNN ON BU-3DFE SUBSET I

Classifier	Logistic Regres.	k-Nearest Neighbor	Naive Bayes	Random Forests	kernel SVM	linear SVM
Accuracy (%)	81.03	85.84	85.90	85.18	86.76	**86.86**

Table XI, we compared the parameter quantity, compute time, and accuracy between DF-CNN and DF-CNN^a. We can see that, comparing with DF-CNN, DF-CNN^a has much more parameters (50M vs. 300M) and thus runs more slowly (7.4 Hz vs. 4.1 Hz). However, DF-CNN still achieves slightly better results than DF-CNN^a. This might be due to that we used very limited number of training samples to train DF-CNN and DF-CNN^a. Therefore, we guess that if one has sufficient training samples available, DF-CNN^a still has a large potential to outperform DF-CNN in general but needs to learn more parameters, and to take more training time.

2) Effectiveness of Learning-Based Fusion: To show the effectiveness of learning-based fusion, we compared DF-CNN with two popular classifiers: linear SVM with score-level fusion and multiple kernel learning (MKL) with kernel-level fusion. Deep CNN features *DF-CNN-in-conv5* and *DF-CNN-ft-conv5* are respectively extracted from the feature extraction subnet of DF-CNN with initialized and fine-tuned CNN parameters. From Table XII, we can see that MKL with kernel-level fusion achieves better results than liner SVM with score-level fusion in both cases. Our fine-tuned DF-CNN, which combines feature learning and fusion learning in a single end-to-end training framework, achieves significant better results than liner SVM and MKL.

Moreover, to see the effect of feature fusion subnet, we fixed all the initialized CNN parameters of the feature extraction subnet, and only learned the parameters of the following feature fusion subnet. This is equivalent to learn hierarchical fusion weights to combine the high-dimensional pre-trained deep features. From Table XII, we can see that this kind of pure fusion learning-based DF-CNN can achieve slightly better results (84.79% vs. 84.22%) than linear SVM with handcrafted score-level fusion, but significantly worse than the proposed DF-CNN. This indicates that the combination of feature learning and fusion learning into a single end-to-end training framework is very important for the proposed DF-CNN.

3) Optimality of Linear SVM-Based Prediction: To validate the optimality of using linear SVM classifier for expression prediction, we compared it with five popular classifiers: logistic regression, k-Nearest Neighbor, naive bayes, random forests, and rbf-kernel SVM. All experiments were carried out on BU-3DFE subset I based on the 1,000 times 54-vs-6 experimental setting, and using the 32-dimensional fused deep features produced by DF-CNN. The hyper-parameters of these classifiers (e.g., the value of k in k-Nearest Neighbor, number of trees in random forests, and γ in rbf-kernel SVM) were carefully selected by cross-validation on the training set of each train session. In contrast, the parameter C in linear SVM was set to be the default value 1 for all 1,000 times train sessions.

Table XIII reports the comparison results. We can see that: 1) All classifiers achieve comparable results except logistic regression, which indicates again that the 32-dimensional fused deep feature is very discriminative. 2) Among all classifiers, linear SVM has obvious advantages in both accuracy and speed (without parameter tuning). Therefore, linear SVM is generally the best candidate classifier for expression prediction using fused deep features produced by DF-CNN.

Finally, it is worth noting that we have also studied the issue of optimal dimension for the fused deep feature produced by DF-CNN. Our experimental results indicate that the 32-dimensional fused deep feature can achieve slightly better results than both 16-dimensional and 64-dimensional fused deep features.

VII. CONCLUSION AND FUTURE WORK

This paper presents a novel deep fusion convolution neural network (DF-CNN) for subject-independent multi-modal 2D+3D FER. DF-CNN comprises a feature extraction subnet, a feature fusion subnet, and a softmax-loss layer. Each textured 3D face scan is firstly represented as six types of facial attribute maps, all of which are then jointly fed into DF-CNN for feature extraction and feature fusion, resulting in a highly concentrated facial representation. Expression prediction is performed by two ways: 1) learning linear SVM classifiers using the 32-dimensional fused deep features; 2) directly performing softmax prediction using the 6-dimensional expression probabilities. Different from existing methods for 3D FER, DF-CNN combines feature learning and fusion learning into a single end-to-end training framework. To demonstrate the effectiveness of DF-CNN, we conducted comprehensive experiments to compare the performance of DF-CNN with handcrafted features, pre-trained deep features, fine-tuned deep features, and the state-of-the-art methods on three subsets of two popular 3D face datasets (i.e., BU-3DFE and

Bosphorus). In all cases, DF-CNN consistently achieves the best results. Both visualization and quantification results indicate that the 32-dimensional fused deep feature of DF-CNN has strong discriminative ability to distinguish different facial expressions.

In the future, some other issues of DF-CNN such as how to choose the optimal pre-trained deep CNN for initialization, and the optimal loss function for training will be studied. Moreover, we will also study to extend current DF-CNN framework to multi-modal 2D+3D video based facial expression recognition, or other multi-modal facial emotion prediction problems such as action unit detection and expression intensity estimation.

References

[1] S. Berretti, A. Bimbo, P. Pala, B. Amor, and M. Daoudi, "A set of selected sift features for 3d facial expression recognition," in *Proc. 20th Int. Conf. Pattern Recog.*, 2010, pp. 4125–4128.

[2] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "Dexpression: Deep convolutional neural network for expression recognition," *CoRR*, 2015. [Online]. Available: http://arxiv.org/abs/abs/1509.05371

[3] R. A. Calix, S. A. Mallepudi, B. Chen, and G. M. Knapp, "Emotion recognition in text for 3-d facial expression rendering," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 544–551, Oct. 2010.

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. British Mach. Vis. Conf.*, 2014, pp. 1–12.

[5] W. S. Chu, F. de la Torre, and J. Cohn, "Selective transfer machine for personalized facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 529–545, Mar. 2017.

[6] C. A. Corneanu, M. Oliu, J. F. Cohn, and S. Escalera, "Survey on RGB, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1548–1568, Aug. 2016.

[7] M. Dahmane and J. Meunier, "Emotion recognition using dynamic grid-based hog features," in *Proc. IEEE Int. Conf., Automat. Face Gesture Recog. Workshops*, Mar. 2011, pp. 884–888.

[8] M. Dahmane and J. Meunier, "Prototype-based modeling for facial expression analysis," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1574–1584, Oct. 2014.

[9] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. IEEE Int. Conf., Comput. Vis. Workshops*, Nov. 2011, pp. 2106–2112.

[10] A. Dhall, S. Member, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia Mag.*, vol. 19, no. 3, pp. 34–41, Jul. 2012.

[11] J. Donahue *et al.*, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proc. 31st Int. Conf. Mach. Learn.*, vol. 32, 2014, pp. I-647–I-655.

[12] T. Fang, X. Zhao, O. Ocegueda, S. Shah, and I. Kakadiaris, "3D facial expression recognition: A perspective on promises and challenges," in *Proc. IEEE Int. Conf. Automat. Face Gesture Recog. Workshops*, Mar. 2011, pp. 603–610.

[13] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris, "3d/4d facial expression analysis: An advanced annotated face model approach," *Image Vis. Comput.*, vol. 30, no. 10, pp. 738–749, 2012.

[14] J. Goldfeather and V. Interrante, "A novel cubic-order algorithm for approximating principal direction vectors," *ACM Trans. Graph.*, vol. 23, no. 1, pp. 45–63, 2004.

[15] B. Gong, Y. Wang, J. Liu, and X. Tang, "Automatic facial expression recognition on a single 3D face by exploring shape deformation," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 569–572.

[16] M. Hayat and M. Bennamoun, "An automatic framework for textured 3D video-based facial expression recognition," *IEEE Trans. Affective Comput.*, vol. 5, no. 3, pp. 301–313, Jul. 2014.

[17] R. Hoffman and A. K. Jain, "Segmentation and classification of range images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, no. 5, pp. 608–620, Sep. 1987.

[18] Y. Huang, Y. Li, and N. Fan, "Robust symbolic dual-view facial expression recognition with skin wrinkles: Local versus global approach," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 536–543, Oct. 2010.

[19] S. E. Kahou *et al.*, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[20] P. Khorrami, T. L. Paine, and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2015, pp. 19–27.

[21] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee, "Hierarchical committee of deep CNNs with exponentially-weighted decision fusion for static facial expression recognition," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 427–434.

[22] P. Lemaire, L. Chen, M. Ardabilian, and M. Daoudi, "Fully automatic 3D facial expression recognition using differential mean curvature maps and histograms of oriented gradients," in *Proc. IEEE Automat. Facial Gesture Recog., Workshop 3D Face Biometrics*, Apr. 2013, pp. 1–7.

[23] H. Li, L. Chen, D. Huang, Y. Wang, and J.-M. Morvan, "3D facial expression recognition via multiple kernel learning of multi-scale local normal patterns," in *Proc. 21st Int. Conf. Pattern Recog.*, 2012, pp. 2577–2580.

[24] H. Li *et al.*, "An efficient multimodal 2d + 3d feature-based approach to automatic facial expression recognition," *Comput. Vis. Image Understand.*, vol. 140, pp. 83–92, 2015.

[25] H. Li, D. Huang, L. Chen, and Y. Wang, "A group of facial normal descriptors for recognizing 3D identical twins," in *Proc. IEEE 5th Int. Conf. Biometrics: Theory, Appl. Syst.*, Sep. 2012, pp. 271–277.

[26] H. Li, J.-M. Morvan, and L. Chen, "3D facial expression recognition based on histograms of surface differential quantities," in *Proc. Advances Concepts Intell. Vis. Syst.*, 2011, pp. 483–494.

[27] K. Li *et al.*, "A data-driven approach for facial expression retargeting in video," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 299–310, Feb. 2014.

[28] M. Liu, S. Li, S. Shan, and X. Chen, "Au-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, 2015.

[29] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, *Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis*. Cham, Switzerland: Springer, 2015, pp. 143–157.

[30] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. IEEE Conf., Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1805–1812.

[31] P. Liu *et al.*, "Feature disentangling machine—A novel approach of feature selection and disentangling in facial expression analysis," in *ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*. Cham, Switzerland: Springer, 2014, pp. 151–166.

[32] A. Maalej, B. B. Amor, M. Daoudi, A. Srivastava, and S. Berretti, "Local 3D shape analysis for facial expression recognition," in *Proc. 20th Int. Conf. Pattern Recog.*, Aug. 2010, pp. 4129–4132.

[33] A. Maalej, B. B. Amor, M. Daoudi, A. Srivastava, and S. Berretti, "Shape analysis of local facial patches for 3D facial expression recognition," *Pattern Recog.*, vol. 44, no. 8, pp. 1581–1589, 2011.

[34] A. Mian, M. Bennamoun, and R. Owens, "Automatic 3D face detection, normalization and recognition," in *Proc. 3D Data Process., Vis. Transmiss.*, 2006, pp. 735–742.

[35] 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.

[36] O. Ocegueda, T. Fang, S. K. Shah, and I. A. Kakadiaris, "Expressive maps for 3D facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Nov. 2011, pp. 1270–1275.

[37] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.

[38] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 2014, pp. 512–519.

[39] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *Computer Vis. ECCV 2012* (Lecture Notes Comput. Sci. 7577). Berlin, Germany: Springer, 2012, pp. 808–822.

[40] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image Vis. Comput.*, vol. 30, no. 10, pp. 683–697, 2012.

[41] A. Savran *et al.*, *Bosphorus Database for 3D Face Analysis*. Berlin, Germany: Springer, 2008, pp. 47–56.

[42] A. Savran, B. Sankur, and M. T. Bilge, “Facial action unit detection: 3D versus 2D modality,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2010, pp. 71–78.

[43] M. Song *et al.*, “A generic framework for efficient 2-D and 3-D facial expression analogy,” *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1384–1395, Nov. 2007.

[44] H. Soyel and H. Demirel, “Facial expression recognition using 3D facial feature distances,” in *Image Analysis and Recognition*, (Lecture Notes Comput. Sci. 4633). Berlin, Germany: Springer, 2007, pp. 831–838.

[45] H. Soyel and H. Demirel, “3D facial expression recognition with geometrically localized facial features,” in *Proc. 23rd Int. Symp. Comput. Inf. Sci.*, 2008, pp. 1–4.

[46] H. Tang and T. Huang, “3D facial expression recognition based on automatically selected features,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2008, pp. 1–8.

[47] H. Tang and T. Huang, “3D facial expression recognition based on properties of line segments connecting facial feature points,” in *Proc. 8th IEEE Int. Conf. Automat. Face Gesture Recog.*, Sep. 2008, pp. 1–6.

[48] Y. Tang, “Deep learning using support vector machines,” in *Proc. Workshop Representational Learning*, 2013, pp. 1–6.

[49] A. Tawari and M. M. Trivedi, “Face expression recognition by cross modal data association,” *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1543–1552, Nov. 2013.

[50] F. Tsalakanidou and S. Malassiotis, “Real-time 2d+3d facial action and expression recognition,” *Pattern Recog.*, vol. 43, no. 5, pp. 1763–1775, 2010.

[51] L. van der Maaten and G. Hinton, “Visualizing data using t-sne,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[52] A. Vedaldi and B. Fulkerson, “VLFeat: An open and portable library of computer vision algorithms,” in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1469–1472.

[53] J. Wang, L. Yin, X. Wei, and Y. Sun, “3D facial expression recognition based on primitive surface feature distribution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 1399–1406.

[54] S. Wang *et al.*, “A natural visible and infrared facial expression database for expression recognition and emotion inference,” *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 682–691, Nov. 2010.

[55] S. Wang *et al.*, “Analyses of a multimodal spontaneous facial expression database,” *IEEE Trans. Affective Comput.*, vol. 4, no. 1, pp. 34–46, Jan. 2013.

[56] C. H. Wu, W. L. Wei, J. C. Lin, and W. Y. Lee, “Speaking effect removal on emotion recognition from facial expressions based on eigenface conversion,” *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1732–1744, Dec. 2013.

[57] X. Yang, D. Huang, Y. Wang, and L. Chen, “Automatic 3D facial expression recognition using geometric scattering representation,” in *Proc. IEEE Int. Conf. Automat. Face Gesture Recog.*, May 2015, vol. 1, pp. 1–6.

[58] M. Yeasin, B. Bullot, and R. Sharma, “Recognition of facial expressions and measurement of levels of interest from video,” *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 500–508, Jun. 2006.

[59] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, “A 3D facial expression database for facial behavior research,” *Proc. IEEE Int. Conf. Automat. Face Gesture Recog.*, Apr. 2006, pp. 211–216.

[60] Z. Yu and C. Zhang, “Image based static facial expression recognition with multiple deep network learning,” in *Proc. ACM Int. Conf. Multimodal Interaction*, IEEE, Nov. 2015, pp. 435–442.

[61] S. Zafeiriou and I. Pitas, “Discriminant graph structures for facial expression recognition,” *IEEE Trans. Multimedia*, vol. 10, no. 8, pp. 1528–1540, Dec. 2008.

[62] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe, “Learning personalized models for facial expression analysis and gesture recognition,” *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 775–788, Apr. 2016.

[63] W. Zeng, H. Li, L. Chen, J.-M. Morvan, and X. D. Gu, “An automatic 3D expression recognition framework based on sparse representation of conformal images,” in *Proc. 10th IEEE Int. Conf. Workshops Automat. Face Gesture Recog.*, Apr. 2013, pp. 1–8.

[64] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, “Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron,” in *Proc. 3rd IEEE Int. Conf. Workshops Automat. Face Gesture Recog.*, Apr. 1998, pp. 454–459.

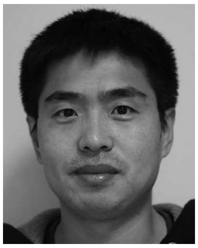
[65] G. Zhao and M. Pietikainen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[66] X. Zhao, D. Huang, E. Dellandrea, and L. Chen, “Automatic 3D facial expression recognition based on a Bayesian belief net and a statistical facial feature model,” in *Proc. Int. Conf. Pattern Recog.*, 2010, pp. 3724–3727.

[67] Q. Zhen, D. Huang, Y. Wang, and L. Chen, “Muscular movement model based automatic 3D facial expression recognition,” in *MultiMedia Modeling* (Lecture Notes Comput. Sci. 8935). Berlin, Germany: Springer, 2015, pp. 522–533.

[68] Q. Zhen, D. Huang, Y. Wang, and L. Chen, “Muscular movement model based automatic 3d/4d facial expression recognition,” *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1438–1450, Jul. 2016.

[69] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, “Learning multiscale active facial patches for expression analysis,” *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499–1510, Aug. 2015.



Huibin Li (S’11) 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, China, in 2009, and the Ph.D. degree in mathematics and computer science from the Université de Lyon, CNRS, Ecole Centrale de Lyon, LIRIS, Lyon, France, in 2013.

He is currently an Assistant Professor with the School of Mathematics and Statistics, Xi’an Jiaotong University. His research interests include 3D shape analysis, 3D face recognition, 3D facial expression analysis, discrete differential geometry, and geometric data analysis, modeling and learning.

Jian Sun (S’08–M’10) received the Ph.D. degree in applied mathematics from Xi’an Jiaotong University, Xi’an, China, in 2009.

He was a visiting student with Microsoft Research Asia, Beijing, China (Nov. 2005–Mar. 2008), a Postdoctoral Researcher with the University of Central Florida, Orlando, FL, USA (Aug. 2009–Apr. 2010), and a Postdoctoral Researcher with the Willow Team of Ècole Normale Supérieure de Paris/INRIA, Paris, France (Sep. 2012–Aug. 2014). He is now a Professor with the School of Mathematics and Statistics, Xi’an Jiaotong University. His research interests include the mathematics and machine learning-based approaches for image processing/recognition, and medical image analysis.

Zongben Xu (M’15) received the Ph.D. degree in mathematics from Xi’an Jiaotong University, Xi’an, China, in 1987.

He is currently the Chief Scientist of the National Basic Research Program of China (973 Project), and the Director of the Institute for Information and System Sciences, Xi’an Jiaotong University. His current research interests include intelligent information processing and applied mathematics.

Dr. Xu was elected as a Member of the Chinese Academy of Sciences in 2011. He was the recipient of the National Natural Science Award of China in 2007. He was the recipient the CSIAM Su Buchin Applied Mathematics Prize in 2008. He delivered a 45 minute talk at the International Congress of Mathematicians 2010.

Liming Chen (A’04–M’06–SM’14) received the joint B.Sc. degree in mathematics and computer science from the University of Nantes, Nantes, France, in 1984, and the M.S. degree and the Ph.D. degree 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, and then in 1998 joined, as a Professor, Ecole Centrale de Lyon, where he leads an advanced research team in multimedia computing and pattern recognition. He has been the Head of the Department of Mathematics and Computer Science from 2007. His current research interests include multimedia processing, discrete differential geometry, and statistical learning, with applications in particular to 2D/3D face analysis and recognition, image and video analysis and categorization.