IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 38, No. 2, April, 2021, pp. 321-330
Journal homepage: http://iieta.org/journals/ts

Facial Expression Recognition Using 3D Points Aware Deep Neural Network

Imen Hamrouni Trimech[1*], Ahmed Maalej[2], Najoua Essoukri Ben Amara[1]

[1] Université de Sousse, Ecole Nationale d'Ingénieurs de Sousse, LATIS-Laboratory of Advanced Technology and Intelligent Systems, Sousse 4023, Tunisia
[2] Université de Kairouan, Institut Supérieur de Mathématiques Appliquées et d'informatique de Kairouan, Kairouan 3100, Tunisia

Corresponding Author Email: najoua.benamara@eniso.rnu.tn

https://doi.org/10.18280/ts.380209

Received: 29 November 2020
Accepted: 23 March 2021

Keywords:
3D facial expression recognition (3D FER), facial expression synthesis, facial surface representation, 3D point-based deep neural network (DNN)

ABSTRACT

Point cloud-based Deep Neural Networks (DNNs) have gained increasing attention as an insightful solution in the study field of geometric deep learning. Point set aware DNNs have proven capable of dealing with the unstructured data type and successful in 3D data applications such as 3D object classification, segmentation and recognition. On the other hand, two major challenges remain understudied when it comes to the use of point cloud-based DNNs for 3D facial expression (FE) recognition. The first challenge is the lack of large labelled 3D facial data. The second is how to obtain a point-based discriminative representation of 3D faces. To address the first issue, we suggest to enlarge the used dataset by generating synthetic 3D FEs. For the second one, we propose to apply a level-curve based sampling strategy in order to exploit crucial geometric information. The conducted experiments show promising results reaching 97.23% on the enlarged BU-3DFE dataset.

1. INTRODUCTION

Facial Expressions (FEs) perform a crucial role in the nonverbal communication. They afford indeed, perceptual indicators of the intention, the impression and the emotional state of persons. Managing and identifying one's emotions and others' is considered as a sign of high emotional intelligence.

The work of Ekman [1] is considered as one of the valuable studies on FE understanding and analysis. Ekman and Frisen [2] developed the facial action coding system to encode FEs as a combination of 44 Action Units (AUs). They also defined six universal categories of emotions, referred to as the basic emotions: *Angry, Disgust, Fear, Happy, Sad,* and *Surprise*.

A considerable amount of work on FE Recognition (FER) has been conducted on both research and industrial areas. Large and complex projects, such as mental state and behavioural studies, emotion understanding and classification, facial animation and rigging, stress signal and fatigue identification, are few examples of the widespread applications related to different FER involving different fields, such as psychology, affective computing, human-computer interaction and automatic surveillance.

FER has been extensively studied in the 2D domain (i.e. 2D images and image sequences) and various approaches have been developed [3]. Despite their valuable enhancement in terms of accuracy and expression recognition rates, 2D-based FER methods show some weaknesses when facing problems like illumination condition changes, head pose variation, and occlusion of facial appearances (e.g. hair, beard and glasses).

Therefore, 3D data have been presented as an alternative [4] to alleviate the aforementioned inherent problems with 2D data.

Due to their efficiency in analyzing the FER problem, conventional FER approaches have been prominent techniques in the last two decades. However, these handcrafted-based techniques suffer several limitations. One of the first challenging problems is how to decide which feature extraction algorithm is suitable for a given problem. There is no a priori way of knowing about the quality of the feature to be designed. The efficiency of a feature extraction method can only be established posteriori, through the statistical methods of the feature quality assessment. Moreover, most feature extraction methods induce very high feature space dimensions. Therefore, feature selection techniques such PCA and LDA are used to reduce the feature space dimension. Such techniques allow excluding highly correlated features and irrelevant ones so as to achieve higher classification accuracy and lower computational cost. Nevertheless, discarding features and reducing massive amounts of data samples can cause data loss and data restriction issues. Such operations hinder the exploitation and mining of data to the fullest extent, especially when dealing with large scale real-world data. These issues can become even more difficult to handle for challenging FER applications, such as real time 2D/3D FER or AU recognition, where the number of expressions goes beyond six basic universal emotions.

Furthermore, feature engineering methods are, in general, labor-intensive, complex and error-prone. Hence, trying to develop and compare these handcrafted features is problematic and reveals critical issues like dataset dependency, replicability and reproducibility of FER approaches.

Nowadays, Deep Learning (DL) has become the basis of most state-of-the-art techniques used in the machine learning field. Motivated by the high inference achievement demonstrated by Deep Neural Network (DNN) architectures, recent FER approaches have employed DNNs to classify FEs. Promising results in terms of FER accuracy have been achieved [5-8]. Most of the previous work on 3D FER has focused on transforming 3D faces into 2D representations in order to allow for a straightforward use of CNN-based architectures. However, such representations induce resolution-degradation and information-loss problems. Therefore, previous 3D FER approaches remain suboptimal and unable to fully describe subtle and complex details of shape cues.

The problem of 3D FER remains understudied when it comes to the use of unstructured aware DNN frameworks. Two main reasons account for this: The first is how to get the optimal 3D face representation that capture most of the facial expression cues, and the second is the lack of large facial databases.

In this work, we propose two major contributions to tackle both issues. First, along with exploiting 3D point sets extracted from curve-based representation as an input to the used DNN architecture, we propose to study the impact of the number of 3D points on the obtained accuracy in order to select the adequate one. Second, we suggest to increase the size of original data by generating realistic 3D FEs using a classic non-rigid registration technique. Thorough experiments are conducted and promising 3D FER results are obtained using the static BU-3DFE database.

The remainder of the paper is organized as follows: Section 2 reviews the related work on 3D FER. Section 3 introduces an overview of our proposed 3D FER approach. Section 4 presents the conducted experiments and obtained results. In section 5, we delve into discussion. Finally, section 6 concludes the paper and presents some future work.

2. RELATED WORK

In this paper, we are interested in recognizing six prototypical FEs from 3D static face scans using a DL paradigm. Most DL-based 3D FER methods, as presented in Table 1, have two principal characteristics in common. First, they are multi-modal data-oriented approaches, and second they apply off-the-shelf pre-trained CNN models. The way these approaches conduct multi-modal (2D+3D) studies is through the use of 2D RGB/gray images and 2D representations of 3D face scans. Range images (i.e. depth maps) and derivative maps (i.e. Gaussian, mean, shape index, etc.) are examples of 2D representations. Several benefits can be depicted from 2D + 3D multimodality. First, it allows Data Augmentation (DA) and alleviates, to some extent, the labelled data scarcity problem. Second, the multi-modality aspect is, in general, coupled with different fusion strategies (e.g. data fusion and classifier fusion) which can be applied on different levels. A multi-modal recognition system is indeed capable of exploiting the complementary property shared between different image data attributes (or modes), which contributes to the robustness enhancement of this type of approaches.

Furthermore, using 2D representations of 3D models assures the consistency of the data format with the input layer of CNN architectures. Most FER approaches exploiting 3D samples typically transform the input samples into a regular 2D format for a prompt use of CNN models. Moreover, this allows the use of well-established pre-trained CNN models. Within the context of transfer learning, where relevant features learned from natural images can benefit the FE recognition task, comes with significant gains in terms of execution time, with better initialization options and enhanced recognition accuracy with appropriate fine-tuning. Overall, the reuse of pre-trained CNN models is dictated by the very limited labelled FE datasets.

The first work towards automatic 3D FER using DL was proposed by Li et al. [9]. They pre-processed the 3D face scans to extract several 2D images from the BU-3DFE database. Six types of 2D facial attributes were utilized, namely three normal maps (X, Y, Z), a geometry map (2D range image), a curvature map (principal curvatures) and a provided texture map. They applied pre-trained CNN models, such as Caffe-Net and VGG-Net, to extract what they referred to as *"deep representation"*. They investigated each convolutional layer and computed per-layer FER accuracy. Based on their study of efficiency versus accuracy, they suggested that the feature maps of the fifth convolutional layer were the best to represent deeply learned features. Rather than using a standard six-node softmax layer to classify the input into one of the six basic emotions, the authors applied linear SVM classifiers at the top of the pre-trained CNN models. This choice might find its support in an earlier study of Tang [10], where he revealed a small, but consistent, gain of substituting the softmax layer with an SVM classifier. Detailed comparisons with handcrafted feature-based 3D FER and their deep representation generated using pre trained CNN models were reported to highlight the performance of their approach. Another gap-filling study was put forward by Huynh et al. [7]. They combined two CNN architectures. The first CNN was used with textured data, which enabled handling both gray and colored images. The second CNN handled depth images. The approach was applied on the BU-3DFE database and an overall recognition rate of 92% was reported. Li et al. [11] suggested an extension of their work [9] with a Deep Fusion CNN (DF-CNN) architecture composed of two subnets; a feature subnet and a fusion one. The fusion subnet was proposed to enhance their original CNN model while fusing the learned deep features with the softmax activation function. Convolutional, ReLU and pooling layers were the main building blocks of the feature extraction subnet. As for the fusion subnet, it was built on reshaping and fusion layers. To show the effectiveness of the DF-CNN, exhaustive evaluations and comparisons against previous 3D FER approaches were reported. Indeed, 86.8% accuracy was achieved overcoming the state-of-the-art methods.

In a simple, yet effective, study of Yang and Yin [6], 3D landmarks were used to generate a mask configuration around salient facial regions. Clipping window blocks were centered around these landmarks to enable the rendering of relevant face regions (mouth, eyes and eyebrows). In addition to the curvature and depth maps, a mask based on the regions of interest was used in the third input channel of the CNN model so that the focus of the DL model could be activated on specific areas that were more likely involved in the FE changes. Both BU-4DFE and BU-3DFE databases were used for training and validation/testing. Clearly, 75.9% and 69% accuracy were attained on BU-4DFE and BU-3DFE, respectively. Jan et al. [12] pursued further research on the individual contribution of distinct facial regions to perform an expression. They suggested to exploit the eyes, eyebrows, mouth and nose regions. These four relevant parts were extracted from both depth and texture images. Then, they were propagated individually through a pre-trained CNN model.

Table 1. State-of-the art DNN based 3D FER approaches

Reference	Database	The used DNN based approach		Recognition rate
		Input type	DNN architecture /model	
Yang and Yin [6]	BU-3DFE BU-4DFE	- Curvature map - Depth map - Generated masking configuration	- Application of VGG-Face model	69% 75.9%
Huynh et al. [7]		- Texture image - Depth image	- Combination of two CNN architectures dealing with both input types (texture and depth images)	92%
Li et al. [9]	BU-3DFE	- Texture map - Geometry map - Three normal maps - Normalized curvature map	- Application of pre-trained CNN architectures (Caffe-Net and VGG-Net)	84.87%
Li et al. [11]		-Texture map - Geometry map -Three normal maps - Normalized curvature map	- Extension of [9]. - Enhancement of DL network. - Exploitation of fusion subnet to fuse learned deep features	86.8%
Jan et al. [12]	BU-3DFE	- Facial key parts cropped out from texture and depth images (mouth, nose, eyes and eyebrows)	- Exploitation of a fusion subnet - Fusion of different feature maps learned from each facial part	88.54%

The feature maps learned from each facial part were fused using a fusion subnet. After the assessment of the deep feature representation of each region, the mouth was considered as the facial region that brought the most value to FEs, compared with other facial parts. Although the 2D representation encoded important cues of the original 3D shape, such a representation showed some limitations as for capturing complex and detailed shape information.

In general, significant loss of geometric information is entailed while using 2D planes of projected 3D data. Thus, the aforementioned 3D FER approaches remain suboptimal for encoding fine and subtle shape changes due to FEs.

3. PROPOSED APPROACH

We propose to address the 3D FER problem from a pure geometric view point. Among different representation types of 3D objects (i.e. point cloud, polygonal/triangular mesh, voxel grid and implicit surface), particularly for the 3D face model, we are interested in the point cloud representation.

A point cloud is a set of points $\{P_{i|i=1,\dots,n}\}$ where each one is defined by its (x, y, z) coordinate attributes. Fundamentally, these spatial attributes hold fundamental facial geometric details and allow for capturing local shape information. Point clouds, known as unified structures, have the advantage of being compact and accurately encoding highly complex objects. Our goal is to develop a DL-based 3D FER approach that obviates irregularities and connectivity complexities of meshes, surfaces or volumes, and to use only the 3D point cloud instead, as presented in Figure 1. In this work, we consider treating the deep 3D FER problem as a classification one. More explicitly, we propose to apply a DNN model which accepts a set of 3D points as an input and outputs K scores for all K candidate classes, where K is the number of FEs to be recognized. These FEs are the six basic emotions. On the other hand, the main faced challenge in the 3D-point learning process is related to the fact that point clouds are unordered. Therefore, the model needs to be invariant to N! permutations.

Indeed, 3D points should create a meaningful subset allowing the model to detain combinatorial interactions among neighboring points. A symmetric function defined by 'Maxpooling' is consequently used.

To construct a set of symmetric functions f defining the DNN, the following observation is used: f is symmetric as long as g is symmetric.

$$f(x_1, x_2, \dots, x_n) = \delta \circ g(h(x_1), \dots, h(x_n))$$

First, each 3D input point is transformed independently and identically by a small network h defined by five convolutional layers, as presented in Figure 5. Then, 3D point features are aggregated using the Maxpooling layer, which is defined by the symmetric function g. Afterwards, the aggregated information goes through a subnetwork δ, mainly composed of a set of fully connected layers.

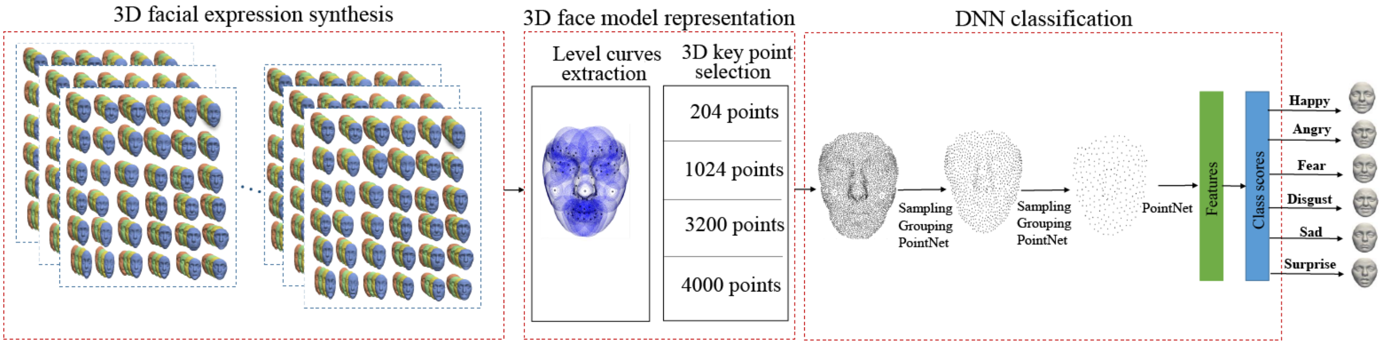


Figure 1. Overview of proposed approach

In fact, to deal with the geometric information of 3D shapes, these shapes are typically modeled as Riemannian manifolds.

The manifolds that are equipped with a Riemannian metric are usually discretized as point clouds or meshes. Hence, the neighbourhoods of local points are defined with the distance metric and a 3D shape locally resembles to a Euclidian space. Such a standard practice allows intrinsically working with geometric shapes, i.e. to locally apply metrics and compute measurements near each point. Thus, the generalization of convolution operations, as main building blocks of classic CNN models to 3D shapes which are considered locally as Euclidean objects, becomes intuitive. The convolution filters of the new geometric CNN are deformation-invariant by construction. Pointnet ++ uses indeed the distance metric to define local regions by partitioning the set of points. Afterwards, local features are extracted from the neighbourhoods of points and then aggregated to generate higher-level features. These steps are repeated until obtaining the whole 3D facial points features.

3.1 3D FE synthesis

One major requirement for the success of DL models is the availability and abundance of labelled data to learn from. This is mainly affordable in the 2D domain especially with natural images. When acquiring such data, we only need a simple camera sensor embedded on some mobile device. Such a facility is unavailable for 3D data, especially for real data where there is a need for special 3D sensors and techniques for data acquisition. For instance, existing 3D FE databases are small-scale ones among others. For the BU-3DFE dataset, it does only contain 2500 FE labeled samples. The available datasets are still far from being enough to train DL models. Thus, there is a need to generate more ground truth data for training. One common practice to alleviate the lack of large-scale training datasets is to generate synthetic data. In order to generate 3D face expressive models, two strategies exist: 1) reconstruct 3D faces from 2D facial images, and 2) employ a statistical 3D face model and manipulate the model distribution parameters to generate synthetic 3D face samples.

The first strategy, commonly referred to as Monocular 3D Facial Shape Reconstruction (MFSR), takes advantage of the abundance of 2D face image data. Different techniques have been proposed to reconstruct 3D faces from 2D images, starting from the pioneer work of Blanz and Vetter [13], who put forward an approach to model 3D human faces from single and/or multiple facial images. This was the first work to introduce 3D Morphable Model (3DMM) for facial shape reconstruction using a PCA-based linear subspace that captures shape variations in human faces. With the emergence of DL techniques, more advanced MFSR approaches have been developed, such as RSNIEF [14], UH-E2FAR [15], Ganfit [16], MMFace [17], and AvatarMe [18].

The challenge of synthesizing 3D faces not only requires inferring accurate, smooth and fine-grained geometry information from 2D face images, but also renders high-resolution and detailed texture information (i.e. skin reflectance). Capturing this information within the constraints of arbitrary poses, lighting conditions and occlusion makes the task even more challenging. State-of-the-art MFSR approaches such as GANFIT and AvatarMe exploit the power of genarative-adversial-network models to produce high-resolution photorealistic 3D faces from *in-the-wild* images.

As for the second strategy, synthetic 3D FE models can be generated using 3D face morphable models, like 3DMM [13], BFM [19], and AFM [20]. Such generic 3D human-like faces are parameterized models that are controlled by a set of geometric and photo-metric parameters. Usually, they are explicitly formulated as a linear combination of; a set of shapes and blend-shape bases, which involves both identity and expression parameter vectors. Therefore, it is possible to create a large number of 3D faces. Gilani and Mian [21] proposed to synthesize new identities from 3D models by simultaneously interpolating between the facial identity and FE spaces.

Another solution consists in using 3D-face databases to compute a statistical 3DMM that likely encodes both facial identities and expressions. Samples of synthetic 3D FEs can be generated from the random sampling of the 3DMM. Even though the above methods are effective in synthesizing 3D faces, most of them fail to reconstruct highly detailed cues of geometric and texture information. The 3D face geometry reconstruction approach from 2D facial images [13] is restricted due to the various faced challenges by the used algorithm, such as illumination conditions and diversified FEs. In addition, exploiting the 3DMM [21, 22] fails to accurately depict the complicated structure of facial details [18].

According to the emotion universality hypothesis, all humans convey six prototypical expressions (anger, disgust, sadness, fear, surprise and happiness) with closely similar facial region movements.

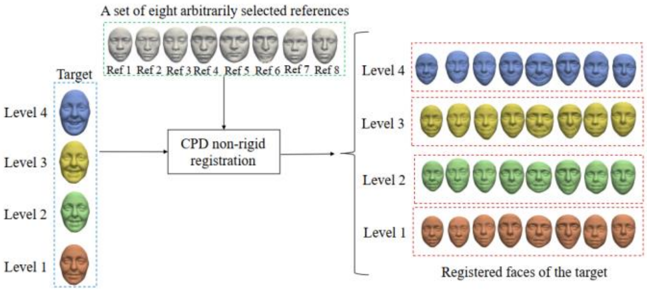

A set of eight arbitrarily selected references
Ref 1
Ref 2
Ref 3
Ref 4
Ref 5
Ref 6
Ref 7
Ref 8
Target
Level 4
Level 3
Level 2
Level 1
CPD non-rigid registration
Level 4
Level 3
Level 2
Level 1
Registered faces of the target

Figure 2. Illustration of proposed DA for BU-3DFE while using eight randomly picked references to enhance DNN training of Fes

Based on this hypothesis, we propose to generate additional realistic 3D FEs by utilizing the available BU-3DFE dataset. We apply a Coherent Point Drift (CPD) non-rigid registration to reproduce the facial expression of an expressive face (target) to a neutral one (source). Compared to the state-of-the-art 3D FE synthesis methods, this technique is simple, yet effective. The size of the used database is consequently increased by reusing the same samples of the BU-3DFE database.

As presented in Figure 7, the synthetized FEs are realistic and different from the original ones.

Let S define the initial dataset and T specify the increased one. DA can be represented as:

$$\phi: S \rightarrow T$$

The augmented dataset is therefore represented as:

$$\hat{S} = S \cup T$$

where, $\hat{S}$ encloses the original training set S and the new generated data T using transformation functions ϕ. Here, ϕ is the non-rigid transformation resulting from the CPD technique between two different point sets, *X* and *Y*.

The CPD algorithm defines the registration by a probability density estimation where the first point set $X = (x_1, x_2, \ldots, x_N)^T$ is the reference, and the second point set $Y = (y_1, y_2, \ldots, y_M)^T$ is the target. The latter is represented by Gaussian Mixture Model (GMM) centroids.

The GMM probability density is given by the following equation:

$$P(x) = \sum_{m=1}^{M+1} P(m)p(x|m)$$

with

$$P(x=m) = \begin{cases} (1-\omega)\frac{1}{M} & if\ m \neq (M+1) \\ \omega\frac{1}{N} & otherwise \end{cases}$$

and $p(x|m) = \frac{1}{(\pi\sigma^2)^{\frac{D}{2}}} e^{-\frac{(x-y_m)^2}{2\sigma^2}}$.

where, σ^2 defines the GMM variance, (*m*=1, …, *M*)), *D* presents the dimension of point sets, and w defines the (*M*+1) uniform weight distribution comprised between 0 and 1. The extra (*M*+1) refers to the assignment outliers.

3.2 3D face model representation

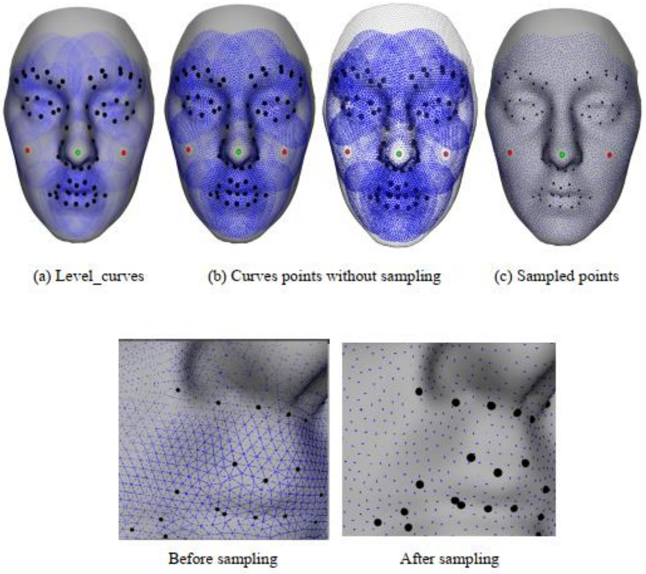
Most 3D DL-based FER research studies have transformed 3D meshes into regular representations such as multi-view images and 2D maps in order to allow for straightforward use of CNN-based architectures. FE data need to be sampled in order to focus on relevant face parts. Therefore, we suggest a preprocessing step.

Our proposed representation consists in summarizing each 3D face shape by a set of level curves, from which we get a set of key points using a sampling technique. Figure 3 and Figure 4 illustrate the used point cloud representation resulting from a uniform curve sampling. We start first by selecting a set of 68 landmarks characterizing the informative 3D facial regions defined by the mouth, nose, eyebrows and eyes regions. Afterwards, we extract two extra points around the cheeks, which are obtained through computing the midpoints of the geodesic paths connecting both the mouth and the outside eye corners. Thus, 70 landmarks are considered as reference points, and each 3D face model is then represented by a set of level curves centered on these points.

Let *S* be a facial surface, and *l* a finite set of $n \in \mathbb{N}$ landmarks defining fixed anatomical points on *S*, denoted by $l = \{l_1, l_2, , \ldots, l_n\}$. We employ the 3D Cartesian coordinates of these anatomical landmarks, represented by *n* ordered triplets $\left(l_{i_x}, l_{i_y}, l_{i_z},\right)_{i=\{1,2,\ldots,n\}}$, to extract patches $\{P_1 \ldots P_n\}$ centered on each landmark *i*. Each patch P_i is an indexed collection of level curves $\mathsf{C}^j_{\lambda_1<\lambda<\lambda_0}$, where λ refers to a constant value of the distance function between the landmark point $l_{i,}$, taken as a reference point, and all points of the curve C^j. In addition, λ_0 defines to the maximum considered distance value. For the curve extraction, we select the Euclidean distance function $||l_i - p||$, which is sensitive to deformation. We exploit it as a function characterizing the length between a reference point r_i and any point *P* on surface *S*, as shown in the following equation:

$$||r_i - P||: \mathsf{C}^i_\lambda = \{P \in S | ||r_i - P|| = \lambda\} \subset S, \lambda \in [0, \lambda_0]$$

where, λ_0 defines the maximum value of λ, and C^i_λ refers to a closed curve comprising a group of points *P* placed with an identical distance λ from the reference point r_i. Uniform sampling consists in finding a good approximation by choosing *n* equally spaced sample points $\{t_1, t_2, \ldots, t_N\}$ defining vertices $\{v_1, v_2, \ldots, v_n\}$, where $v_i = \mathsf{C}^i_\lambda$, while keeping n small. This data representation enables accurately capturing facial surface local deformation.


(a) Level_curves (b) Curves points without sampling (c) Sampled points

Before sampling After sampling

Figure 3. Level curves extraction and 3D point set sampling

3.3 3D FER

3.3.1 DL on point clouds

Recent years have witnessed an increasing interest in developing deep net architectures capable of reasoning about data types, whose underlying structure lies in a non-Euclidean space. Significant effort has been directed towards extending conventional DL models to the non-Euclidean domain. A number of techniques have been proposed to generalize DNNs and adapt convolutional operations to process non-Euclidean representations of 3D objects (i.e. point cloud, mesh surface, graph). Thanks to the availability of 3D CAD models for training and the efficiency of these techniques has been proven on common 3D object applications involving classification, segmentation and recognition.

More in-depth information about the advances in 3D DL architectures and their applicability on various 3D data representations, while categorizing these representations into Euclidean and non-Euclidean ones, can be found in the interesting survey by Ahmed et al. [22]. We are particularly interested in the point cloud representation, considered as one important type of 3D geometric data offering a simple, compact and unified structure. Compared to other representation types (mesh, graph, etc.), the point representation allows avoiding complexities and combinatorial irregularities of the connectivity properties.

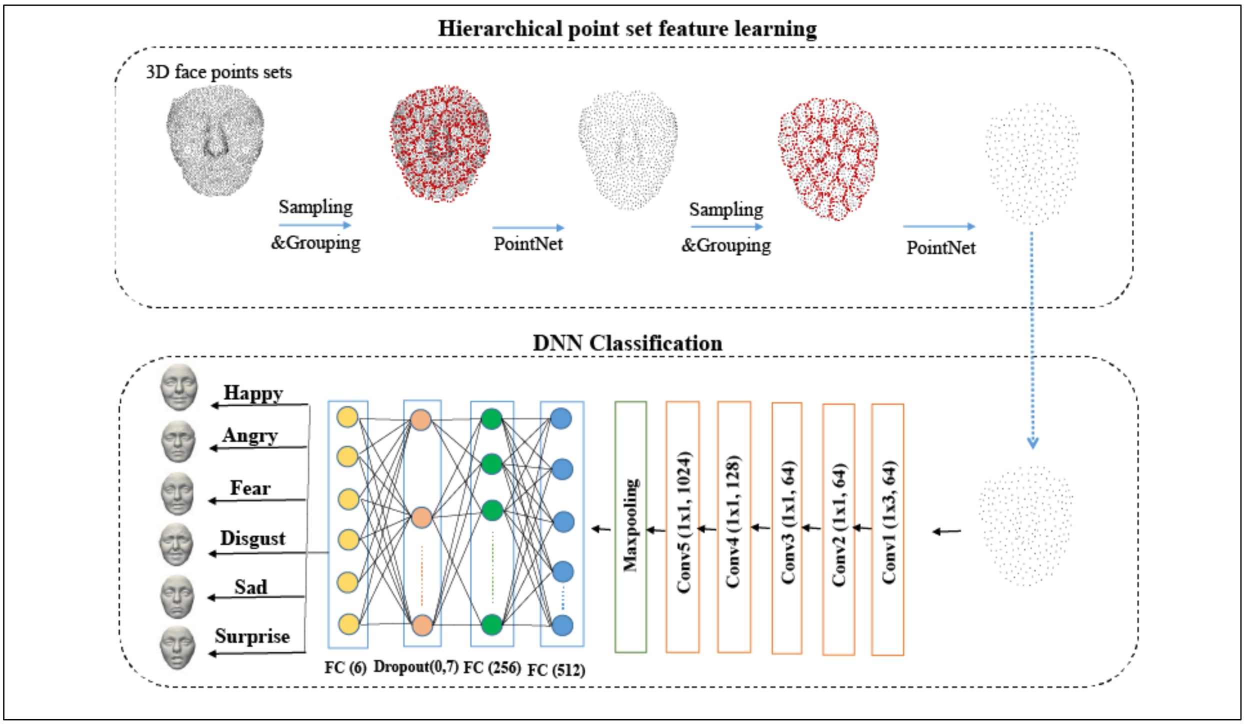


Figure 4. DNN architecture capturing local deformations for 3D FER

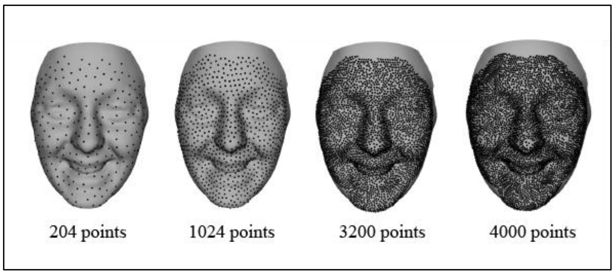


Figure 5. 3D face point cloud representation generated from different sampling of curve-based representation

Few prior approaches have studied DL on point clouds for 3D object classification and segmentation. PointNet of Qi et al. [23] was a pioneer work in this direction. It was the first DL model designed to directly process 3D point sets. Using point sets sampled from CAD models of ModelNet40, ShapeNet, and simulated kinect scans for training, evaluation and testing, the model could output, per hole and per point, labels for the inputted point set. Thus, it allowed object classification, part segmentation, and scene semantic parsing. Though simple, PointNet achieved high performance on the latter tasks, with a set of fully connected layers, max pooling, average pooling and an attention-based weighted sum. Although invariant to member permutation and rigid transformation, PointNet, by design, did not capture local features induced by point neighboring. Therefore, a second version was proposed by the same authors, PointNet++, which applied PointNet recursively and captured local features by encoding fine grained structures from small point neighborhoods. SpiderCNN [24] extended convolutional operations from regular point structures (i.e. image pixels) to irregular point sets. By parameterizing a family of convolutional filters, they introduced SpiderConvs as the new convolutional layers for point clouds.

In this work, we proceed to recursively partition the input facial 3D point sets into local overlapping regions to capture local features from neighboring points, and then group them to produce a higher-level representation. The local feature learner is defined by the original PointNet architecture.

3.3.2 DNN architecture

In this work, we exploit a 3D points aware DNN architecture where point sets are defined as vectors $\{P_{i|i=1,\ldots,n}\}$ of (x, y, z) coordinates. Such an architecture involves a shared MLP network with different output layer sizes (64, 128, 1024 respectively). In addition, it employs a max pooling layer followed by two connected layers with different output layer sizes: 512 and 256. A batch normalization as well as ReLU are applied to all layers. Then, we consider the registered 3D point sets as an input for our classifier. The feature learning process in this work is based on three abstraction levels. The first one, denoted by sampling layer, allows the selection of the centroids in local areas. In fact, we select a set $\{x_{i_1}, x_{i_2}, \ldots, x_{i_m}\}$ from the input points $\{x_1, x_2, \ldots, x_m\}$ using the iterative farthest point sampling [25].

The second one, called the grouping layer, retrieves the centroid neighboring points to generate several sub-point sets. The input point set size of this layer is $M \times (b + S)$, where M , b and S respectively denote the number, coordinates and features of points. The output is a collection of point sets of size $M' \times K \times (b + S')$, where M' represents the subsampled points, K is the number of centroid neighbouring points, and S' is the new feature vector that encodes the local regions. The third one is a PointNet layer exploiting the same architecture presented by Qi et al. [23].

4. EXPERIMENTATION AND RESULTS

In this section, we carry out a set of experiments on the BU-3DFE database to substantiate the effectiveness of using 3D point sets as a DNN input for FE classification, and to evaluate the robustness of our proposed DA approach in FER using DL.

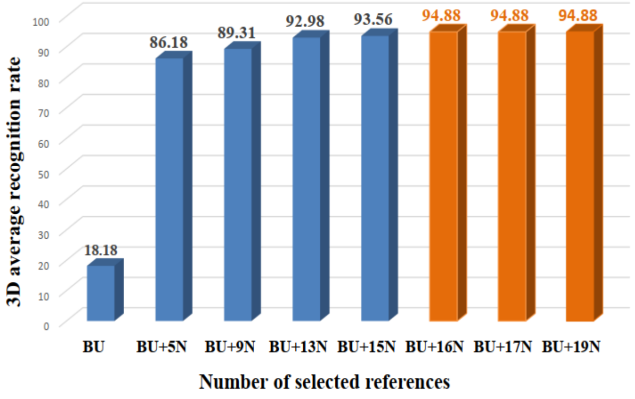


Figure 6. 3D FER improvement while varying the number of selected references

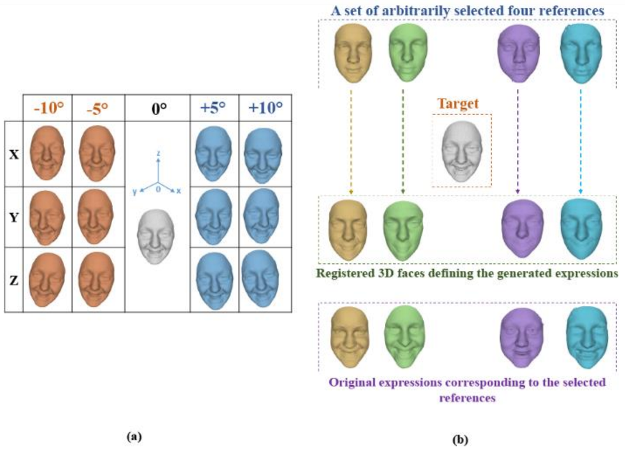

-10°
-5°
0°
+5°
+10°
X
Y
Z
A set of arbitrarily selected four references
Target
Registered 3D faces defining the generated expressions
Original expressions corresponding to the selected references
(a)
(b)

Figure 7. DA: (a) DA based on rotation with different degrees for one happy face, (b) DA exploiting non-rigid CPD registration

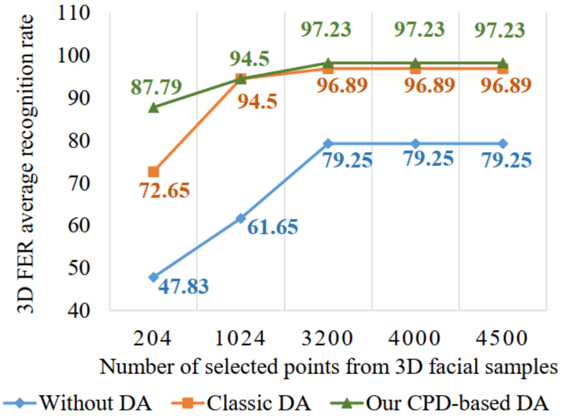


Figure 8. Comparison between different obtained 3D FER average rates

4.1 Dataset description

BU-3DFE [26] is considered as the first created dataset for static FER. It contains 100 face models (56 female and 44 male) with a diversification in ethnic and racial ancestries (white, black, Asian, Indian, Latino). It involves the six prototypical expressions, namely: Anger (AN), Happiness (HA), Disgust (DI), Sadness (SA), Fear (FE), and Surprise (SU) with four intensity levels (low (1), middle (2), high (3) and highest (4)). Thus, each subject possesses 25 FE samples (24 expressions + 1 neutral). The BU-3DFE database is annotated with 83 landmarks located in salient facial regions.

4.2 Experimental results

We propose to increase the BU-3DFE database size by using the non-rigid CPD registration. Given the neutral face of one person, we can generate additional resembling facial samples of its own six prototypical expressions. In fact, the non-rigid registration is between a set of arbitrarily chosen neutral 3D faces of different people, defining the references, and the 3D expressive faces of the whole dataset, named targets.

Figure 2 illustrates an example of the augmented dataset using eight arbitrarily selected references. We end up with a total number of 21600 (8 x 2400 + 2400 (original data)) samples, which is seven times larger than the original size. The number of selected references is varied until reaching 40800 samples relevant to created 16-reference-dependent 3D data. In order to optimize the calculation time for the choice of the number of references, we present the 3D faces by their 68 landmarks, discarding 15 landmarks of the border, as presented by Trimech et al. [27]. Then, we progressively augment the number of selected references to increase the database size. FE accuracy stabilizes when the reference number is equal or higher to 16, as shown in Figure 6.

Thus, we limit our augmented 3D data to 40800 samples relevant to the choice of 16 references for all our experiments.

Compared to conventional DA (Figure 7) based on the rotation transformation, the non-rigid registration using the CPD has a greater impact on the FER rate, which improves by reaching 97.3% (Figure 8).

4.3 Implementation details

We implement the used DNN architecture by using the TensorFlow framework on a PC loading Intel Xeon(R) CPU E5-2650 0 @ 2.00GHz x 16, TITAN X (Pascal)/PCIe/SSE2. The learning process using the augmented dataset lasts about 4h per one epoch. The input 3D facial models are sampled using a set of extracted points from level curves. The batch size is set to 32. We use a learning rate equal to 0.001, and Adam is the used optimizer to train the model.

For the data split, we use 80% for training and 20% for testing.

5. DISCUSSION

5.1 Classic DA

Unlike 2D massive face databases such as FaceNet [28] and VGGFace [29], 3D annotated FE databases are still limited. Consequently, the 3D data variability is insufficient, which may lead to the problem of overfitting. Most state-of-the-art 3D FER work has essentially exploited rotation as a main transformation to augment the size of the used dataset [6, 30]. We adopt a similar transformation as [6] and apply rotation on each 3D face model with different rotation angles (-10°, -5°, $+5^{\circ}$, $+10^{\circ}$) along three directions (x, y, z) named raw, roll and pitch directions. The database size is consequently increased to reach 31200 samples ((2400 x3) x 4 + 2400 (original data)).

Table 2. Average confusion matrix (percentage values) before DA exploiting 204 points for each 3D face model

	AN	DI	FE	HA	SA	SU
AN	**61.84**	14.47	7.89	1.31	14.47	0
DI	29.76	**38.09**	17.85	7.14	7.14	0
FE	14.77	5.68	**44.31**	17.04	15.90	2.27
HA	6.84	2.73	32.87	**47.94**	6.84	2.73
SA	36.14	6.02	15.66	3.61	**36.14**	2.40
SU	1.33	0	29.33	8.0	2.66	**58.66**

Table 3. Average confusion matrix (percentage values) with classic DA exploiting 204 points for each 3D face model

	AN	DI	FE	HA	SA	SU
AN	**78.85**	2.36	1.84	0.30	16.32	0.30
DI	19.49	**60.48**	7.86	3.24	8.49	0.41
FE	10.03	2.42	**56.81**	5.80	22.17	2.47
HA	2.68	1.59	14.01	**68.98**	11.33	1.39
SA	8.45	0.82	2.16	0.72	**87.21**	0.61
SU	0.66	0	8.18	1.32	6.19	**83.62**

Table 4. Average confusion matrix (percentage values) before DA exploiting 3200 points for each face model

	AN	DI	FE	HA	SA	SU
AN	**70.73**	13.41	2.43	2.43	0	10.97
DI	6.41	**88.46**	0	0	0	5.12
FE	6.25	1.25	**80.0**	2.5	1.25	8.75
HA	1.23	1.23	1.23	**95.06**	0	1.23
SA	5.79	0	1.44	0	**53.62**	39.13
SU	1.12	0	6.74	3.37	1.12	**87.64**

Table 5. Average confusion matrix (percentage values) with classic DA exploiting 3200 points for each face model

	AN	DI	FE	HA	SA	SU
AN	**98.34**	0.41	0.10	0.31	0.72	0.10
DI	2.10	**96.52**	0.84	0.10	0	0.42
FE	1.19	0.97	**94.20**	2.79	0.29	0.69
HA	0.10	0	0.42	**99.47**	0	0
SA	4.72	0.31	0.73	0.20	**94.01**	0
SU	0.10	0.10	0.74	0.21	0	**98.2**

Table 2 and Table 4 present the resulting confusion matrices before DA using respectively 204 points and 3200 points. We notice an enhancement of accuracies values while applying classic DA (Table 3 and Table 5).

We reach an average recognition rate equal to 96.89%, while presenting each 3D face by 3200 points (Table 5).

Furthermore, we notice that by increasing the number of point sets, accuracy is significantly enhanced by 24.24% and stabilizes at 96.89%. Indeed, we notice a significant reduction of FE misclassification rates. The highest FE accuracy is mainly relative to *Happy, Angry, Disgust* and *Sad* expressions.

5.2 CPD-based DA

Applying CPD-based DA allows generating new realistic facial expressions while being different from the original ones of the BU dataset (Figure 7).

In Table 6 and Table 7, we respectively present confusion matrices using 204 points and 3200 points after our CPD-based DA.

The accuracy of 97.23% is reached, which is higher than the one obtained with classic DA (Table 7).

The misclassification rates decrease with a 6.51% in the case of *Angry* and *Sad* expressions and with 3.65% for the case of *Fear* and *Disgust* ones. The resulting misclassifications rates between the different expressions are mainly due to the FEs similarity pattern as presented in Figure 9.

Table 6. Average confusion matrix (percentage values) for our CPD-based DA with 204 points selection

	AN	DI	FE	HA	SA	SU
AN	**86.04**	3.15	1.83	0.14	8.29	0.51
DI	4.43	**88.56**	2.85	0.90	1.27	1.95
FE	2.18	4.43	**79.50**	7.48	2.90	3.48
HA	0.21	0.93	7.07	**90.54**	0.64	0.57
SA	9.51	0.81	1.93	0.37	**87.21**	0.14
SU	0.36	0.58	2.42	0.80	0.88	**94.92**

Table 7. Average confusion matrix (percentage values) for our CPD-based DA with 3200 points selection

	AN	DI	FE	HA	SA	SU
AN	**96.58**	0.96	0.59	0	1.78	0.07
DI	1.02	**97.59**	1.02	0.07	0.07	0.21
FE	0.29	0.78	**96.77**	1.28	0.35	0.50
HA	0.07	0.07	2.52	**97.17**	0	0.14
SA	2.67	0.07	0.81	0	**96.29**	0.14
SU	0.14	0.07	0.59	0.14	0	**99.03**

Table 8. Comparison between DL-based 3D FER works on the same BU-3DFE dataset

Works	Features	Accuracy (%)
Yang and Yin [6]	Depth + curvature + mask	75.9
Huynh et al. [7]	Texture + facial shape	92
Li et al. [9]	Geometry + curvature + texture + normal maps	86.86
Jan et al. [12]	Texture + depth maps	88.54
Wei et al. [31]	Geometry + curvature + texture + normal maps	88.08
Oyedotun et al. [32]	Fused RGB+ depth map	89.31
Zhu et al. [33]	Depth+ shape index + normal maps	87.69
Our Work	3D point sets	97.23

Comparing the two confusion matrices after DA, using 3200 points, presented in Table 5 and Table 7 for both classic and our CPD-based DA, we notice that for some cases the misclassification after CPD-based DA shows a small increase.

For instance, the misclassification between *Sad* and *Surprise* expressions augments from 0 to 0.14. Besides, *Sad* and *Fear* expressions are misclassified with 0.81%. Overall, these small variations do not influence the rise in most obtained accuracy after CPD-based DA.

Table 8 presents a comparison between different 3D FER studies using the same database (BU-3DFE). In fact, BU-3DFE is considered as the most used database in DL-based 3D FER work. This is mainly due to the database diversification in age, gender and cultural backgrounds, as well as the large FE variations related to emotional states or external environments.

Most DL-based 3D FER work has exploited 2D representations as texture, depth, curvature and normal maps, which have been generated from 3D face models.

These 2D representations facilitate the use of 2D input based DL architectures and increase the size of the used database by considering wide techniques of 2D image generation from 3D models. Despite their performance, using 2D representations for FER can degrade the useful 3D information that may be relevant for 3D FER. In fact, 2D representations are not optimal to detect complicated shape deformation. This is mainly caused by the projection of 3D data into 2D planes, which leads to the natural and significant loss of geometric information.

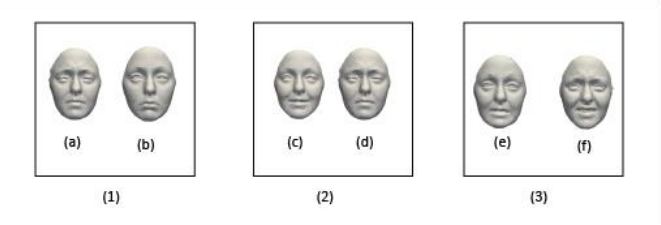


Figure 9. Similarity patterns between different expressions (1): Level 1 of respectively the (a) Anger and (b) Sad expressions, (2): Level 1 of respectively the (c) Happiness and (d) Anger expressions, (3): Level 1 of (e) Fear expression and level 2 of (f) disgust expression

As presented in Table 8, using 3D point sets improves the recognition accuracy up to 97.23%. In fact, focusing on the direct use of 3D point sets and exploiting the inherently homogeneous and compact representation of 3D shapes ease the learning process of shape features [34]. Being depicted by its coordinate attributes (x, y, z) facilitates indeed capturing the fine-grained local patterns in the region-based local context.

Furthermore, most literature work, as in Ref. [7, 9, 12, 31-34], has exploited only the two highest levels of the BU-3DFE database due to the fact that weak expressions (lower levels) are hardly recognized. On the other hand, in our approach we exploit the four different levels of BU-3DFE, which improves the recognition of weak expressions. This is due to the use of representations based on 3D point sets. Moreover, exploiting the DL architecture allows capturing critical and subtle FE details.

6. CONCLUSION

In this study, we present a novel approach for 3D FER based on the use of a DNN architecture exploiting 3D point sets as an input. In addition, we suggest a novel DA strategy based on the non rigid CPD registration, hence generating new additional realistic 3D facial expressions that allow us to augment the initial size of the BU-3DFE database. We have reported an encouraging average recognition rate of 97.23% overcoming most of *state-of-art* works [30, 32-34].

As future work, we aim to test the robustness of our proposed approach with Bosphorus dataset. We are currently studying the generalization of our CPD-based DA for different detected 3D points.

REFERENCES

[1] Ekman, P. (1992). Facial expressions of emotion: New findings, new questions. Psychological Science, 3(1): 34-38. https://doi.org/10.1111/j.1467-9280.1992.tb00253.x

[2] Ekman, P., Friesen, W. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto: Consulting Psychologists Press.

[3] Pantic, M., Rothkrantz, L.J.M. (2000). Automatic analysis of facial expressions: The state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12): 1424-1445. https://doi.org/10.1109/34.895976

[4] Medioni, G., Waupotitsch, R. (2003). Face modeling and recognition in 3-D. Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures, Washington, pp. 232-233. https://doi.org/10.1109/AMFG.2003.1240848

[5] Zhang, T., Zheng, W., Cui, Z., Zong, Y., Yan, J., Yan, K. (2016). A deep neural network-driven feature learning method for multi-view facial expression recognition. IEEE Transactions on Multimedia, 18(12): 2528-2536. https://doi.org/10.1109/TMM.2016.2598092

[6] Yang, H., Yin, L. (2017). CNN based 3D facial expression recognition using masking and landmark features. 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 556-560. https://doi.org/10.1109/ACII.2017.8273654

[7] Huynh, X.P., Tran, T.D., Kim, Y.G. (2016). Convolutional neural network models for facial expression recognition using BU-3DFE database. In: Kim, K., Joukov, N. (eds) Information Science and Applications (ICISA) 2016. Lecture Notes in Electrical Engineering, vol 376. Springer, Singapore. https://doi.org/10.1007/978-981-10-0557-2_44

[8] Liang, D., Liang, H., Yu, Z., Zhang, Y. (2019). Deep convolutional BiLSTM fusion network for facial expression recognition. The Visual Computer, 36: 499-508. https://doi.org/10.1007/s00371-019-01636-3

[9] Li, H., Sun, J., Wang, D., Xu, Z., Chen, L. (2015). Deep representation of facial geometric and photometric attributes for automatic 3D facial expression recognition. CoRR, vol. abs/1511.03015.

[10] Tang, Y. (2013). Deep learning using support vector machines. CoRR, vol. abs/1306.0239, 2013.

[11] Li, H., Sun, J., Xu, Z., Chen, L. (2017). Multimodal 2D+3D facial expression recognition with deep fusion convolutional neural network. IEEE Transactions on Multimedia, 19(12): 2816-2831. https://doi.org/10.1109/TMM.2017.2713408

[12] Jan, H., Ding, H., Meng, L., Chen, Li, H. (2018). Accurate facial parts localization and deep learning for 3D facial expression recognition. 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), pp. 466-472. https://doi.org/10.1109/FG.2018.00075

[13] Blanz, V., Vetter, T. (1999). A morphable model for the synthesis of 3D faces. Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, New York, NY, USA, pp. 187-194. https://doi.org/10.1145/311535.311556

[14] Richardson, E., Sela, M., Kimmel, R. (2016). 3D face reconstruction by learning from synthetic data. 2016 Fourth International Conference on 3D Vision (3DV), pp. 460-469. https://doi.org/10.1109/3DV.2016.56

[15] Dou, P., Shah, S.K., Kakadiaris, I.A. (2017). End-to-End 3D face reconstruction with deep neural networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1503-1512. https://doi.org/10.1109/CVPR.2017.164

[16] Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S. (2019). GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1155-1164. https://doi.org/10.1109/CVPR.2019.00125

[17] Yi, H., Li, C., Cao, Q., Shen, X., Li, S., Wang, G., Tai, Y.W. (2019). MMFace: A multi-metric regression network for unconstrained face reconstruction. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7655-7664. https://doi.org/10.1109/CVPR.2019.00785

[18] Lattas, A., Moschoglou, S., Gecer, B., Ploumpis, S., Triantafyllou, V., Ghosh, A., Zafeiriou, S. (2020). AvatarMe: realistically renderable 3D facial reconstruction "in-the-wild". 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[19] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp 296-301. https://doi.org/10.1109/AVSS.2009.58

[20] Kakadiaris, A., Passalis, G., Toderici, G., Murtuza, M.N., Lu, Y., Karampatziakis, N., Theoharis, T. (2007). Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(4): 640-649. https://doi.org/10.1109/TPAMI.2007.1017

[21] Gilani, S.Z., Mian, A. (2018). Learning from millions of 3D scans for large-scale 3D face recognition. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 896-1905. https://doi.org/10.1109/CVPR.2018.00203

[22] Ahmed, E., Saint, A., Shabayek, A.E.R., Cherenkova, K., Das, R., Gusev, G., Aouada, D., Ottersten, B.E. (2018). Deep learning advances on different 3D data representations: A survey. CoRR, vol. abs/1808.01462.

[23] Qi, C.R., Su, H., Mo, K., Guibas, L.J. (2016). PointNet: Deep learning on point sets for 3D classification and segmentation. CoRR, vol. abs/1612.00593.

[24] Xu, Y., Fan, T., Xu, M., Zeng, L., Qiao, Y. (2018). SpiderCNN: Deep learning on point sets with parameterized convolutional filters. CoRR, vol. abs/1803.11527.

[25] Yan, D.M., Guo, J., Jia, X., Zhang, X., Wonka, P. (2014). Blue-noise remeshing with farthest point optimization. chez Computer Graphics Forum, 33(5): 167-176. https://doi.org/10.1111/cgf.12442

[26] Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J. (2006). A 3D facial expression database for facial behavior research. Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, Washington, pp. 211-216. https://doi.org/10.1109/FGR.2006.6

[27] Hamrouni Trimech, I., Maalej, A., Essoukri Ben Amara, N. (2019). Data augmentation using non-rigid CPD registration for 3D facial expression recognition. chez 2019 16th International Multi-Conference on Systems, Signals & Devices (SSD), pp. 164-169. https://doi.org/10.1109/SSD.2019.8893278

[28] Schroff, F., Kalenichenko, D., Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 815-823. https://doi.org/10.1109/CVPR.2015.7298682

[29] Parkhi, M., Vedaldi, A., Zisserman, A. (2005). Deep face recognition. Department of Engineering Science, University of Oxford.

[30] Wang, X., Wang, K., Lian, S. (2019). A survey on face data augmentation. arXiv:1904.11685v1.

[31] Wei, X., Li, H., Sun, J., Chen, L. (2018). Unsupervised domain adaptation with regularized optimal transport for multimodal 2d+ 3d facial expression recognition. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 31-37. https://doi.org/10.1109/FG.2018.00015

[32] Oyedotun, O.K., Demisse, G., Shabayek, A.E.R., Aouada, D., Ottersten, B. (2017). Facial expression recognition via joint deep learning of RGB-depth map latent representations. Proceedings of the IEEE International Conference on Computer Vision, pp. 3161-3168. https://doi.org/10.1109/ICCVW.2017.374

[33] Zhu, K., Du, Z., Li, W., Huang, D., Wang, Y., Chen, L. (2019). Discriminative attention-based convolutional neural network for 3D facial expression recognition. 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pp. 1-8. https://doi.org/10.1109/FG.2019.8756524

[34] Hamrouni Trimech, I., Maalej, A., Essoukri Ben Amara, N. (2020). Point-based deep neural network for 3D facial expression recognition. 2020 International Conference on Cyberworlds (CW), Cean, France, pp. 164-171. https://doi.org/10.1109/CW49994.2020.00035