3D Facial Expression Recognition: A Perspective on Promises and Challenges

T. Fang, X. Zhao, O. Ocegueda, S.K. Shah and I.A. Kakadiaris*

Abstract— **This survey focuses on discrete expression classification and facial action unit recognition performed using 3D face data, possibly including a corresponding 2D texture image. Research trends to date are summarized and the limitations of current methods are discussed. The challenges towards the development of more accurate and automated 3D facial expression recognition methods are identified. We also call for standardized experimental protocols in order to draw fair and meaningful comparisons between different systems.**

I. INTRODUCTION

Facial expression analysis/recognition[1] has interested many researchers due to its various purposes and applications. It plays a key role in emotion recognition and thus contributes to the development of human-computer interaction systems. It can also reinforce face recognition systems by providing prior knowledge on the facial motions and facial feature deformations. This is particularly intriguing considering that the mouth area contains significant amount of discriminative information [1], yet it is where most of the facial deformations take place. Other applications include but are not limited to psychological studies, tiredness detection, facial animation, robotics as well as virtual reality.

Facial expressions are generated by facial muscle contractions which result in temporary facial deformations in both facial geometry and texture. In the past, the main focus of expression analysis has been the 2D domain due to the prevalence of data in the relevant modalities (i.e., images and videos). Comprehensive surveys in this area include those by Fasel and Luettin [2], Pantic *et al.* [3] and Zeng *et al.* [4]. While these 2D facial expression recognition (FER) systems have achieved remarkable performance, existing challenges in 2D face recognition still present themselves in 2D expression analysis (i.e., illumination and pose variations). Three-dimensional data, on the other hand, are invariant to such changes and are information-rich by nature. Recent successes in 3D face recognition ([5], [6]) can naturally be exploited for expression recognition. To the best of our knowledge, no survey has been performed on the topic of 3D FER. This area of research has drawn much attention since the BU-3DFE database [7] was made publicly available in 2006, and we can thus expect more effort in that direction. Therefore, it would be beneficial to summarize past work and provide potential future directions.

This paper is structured as follows: Section II contains background information in expression recognition. Previous work in 3D FER is presented in categories and research directions are identified in Sections III and IV, respectively. The survey is then concluded in Section V.

II. BACKGROUND

A. Basic Emotions vs. Action Units

The two main streams of facial expression analysis have been message-based approaches and sign-based approaches [8]. The message-based approaches focus on interpretation of specific facial patterns and classify expressions into a predefined number of discrete categories, in which the most commonly used are the six basic emotions (anger, disgust, fear, happiness, sadness and surprise) [9]. The sign-based approaches, on the other hand, provide descriptions for facial deformations at an abstract level in an objective manner and defer the decision making process to other high-level algorithms or human experts. To completely describe all perceptible changes, the Facial Action Coding System (FACS) has been proposed by Ekman and Friesen [10].

The prototypical emotion categories and their characteristic facial expressions have been proved to be perceived by humans in the same way regardless of cultural differences. Hence, most of the studies on vision-based facial expression analysis rely on this categorization of expressions. However, they fall short in covering the whole range of emotions that people may experience in everyday lives, when more subtle emotions (e.g., anxiety) and combination of emotions often occur. Action units (AUs) of the FACS, on the other hand, are independent of interpretation and thus can be used as the input to decision making processes based on high-level rules or interpretations, such as Emotional FACS (EMFACS) [11] and FACS Affect Interpretation Database (FACSAID) [12]. Furthermore, AUs are more flexible as thousands of anatomically possible facial expressions can be described by combinations of merely 27 basic AUs and a number of AU descriptors. Hence, they are more suitable for describing spontaneous facial behaviors [4].

Albeit an increasing number of studies are based on automatic AU recognition, basic emotion classification still remains the most popular research topic. This trend is also reflected in the 3D domain, where the majority of existing works attempt to classify the basic emotions ([13], [14], [15], [16], [17], [18], [19], [20]) while only a few try to recognize AUs ([21], [22], [23]). One reason behind this is the lack of

*This research was funded in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL) and by the University of Houston (UH) Eckhard Pfeiffer Endowment Fund. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, the U.S. Government, or UH.

[1] We use the term *expression analysis* and *expression recognition* interchangeably in the rest of the paper.

FACS-coded databases. Sun *et al.* [21] manually labeled 8 AUs in the BU-4DFE database [24] for their work in partial AU recognition but the annotations are not publicly available. The Bosphorus database [25] has enabled the community to investigate AU recognition in 3D and it remains the only 3D database that provides facial action coding.

B. Static vs. Dynamic

In vision-based expression analysis, it is suggested that the dynamics of facial expressions provide important cues about the underlying emotions that are not available in static images [24]. This is especially true concerning spontaneous affective behaviors. Schmidt and Cohn [26] have shown that spontaneous smiles reach onsets faster than posed smiles and can have multiple rises of the mouth corners. Moreover, they are accompanied by other muscle activities that appear either simultaneously with mouth corner rises or follow them within 1 *s*. Generally an expression process can be segmented into four steps: neutral, onset, apex, and offset. The duration of typical muscle activities varies from 250 *ms* to 5 *s*. Thus, taking into consideration the temporal dynamics is of importance when evaluating expression intensity level and categorizing facial expressions or muscle activities.

Nevertheless, while most of the works in 2D FER use video as input, only a few attempts have been made to analyze facial behaviors in 3D videos ([27], [14], [21], [28]). This is partly because only relatively recently has a database of 3D dynamic sequences been made publicly available (i.e., the BU-4DFE database [24]). Chang *et al.* [27] have created an expression database of 3D videos using an acquisition system based on a camera-projector pair and active stereo, but it was never publicly released. Furthermore, Yin *et al.* [13] used key-frame interpolation to obtain intermediate frames between the expression models of four different intensities in the BU-3DFE database, in order to incorporate temporal information when classifying expressions.

C. Expression Databases

To date, there are three publicly available 3D databases designed *specifically* for expression analysis (i.e., the BU-3DFE, the BU-4DFE and the Bosphorus databases). A summary of these databases is given in Table I. Note that we consider a database dedicated to expression analysis only when it contains datasets displaying 6 basic emotions or different AUs of the FACS. Other databases such as FRGC v2 [29] and GavabDB [30] are rarely used for expression analysis purposes although they contain expression variations, due to an incomplete expression set and/or an irregular distribution of these variations.

Most of the existing 3D FER systems were evaluated on the BU-3DFE database [7], not only because it was the first one to become publicly available but also because of the manually annotated dense landmark set provided with the release. Among the 100 subjects, 56 are female and 44 are male. Using 3D sensors from 3dMD [31], each subject was captured performing the expressions corresponding to the six basic emotions at four different intensity levels along with a neutral expression. The database contains the raw 3D scans with associated texture images as well as the face models that are cropped from the original scans, which greatly facilitate research in human face analysis. In the BU-4DFE database [24], 3D videos were captured using the Di3D (Dimensional Imaging [32]) dynamic face capturing system. Each of the six universal expressions was performed gradually from neutral, onset, apex, offset and then back to neutral in approximately four seconds. There are 58 female and 43 male subjects, with a variety of ethnicities. The datasets are distributed in individual frames, as a result there are roughly 60,600 3D models with associated texture images in total. The Bosphorus database [25] contains 4,652 face scans, captured using a structured-light based 3D digitizer device [33]. Each scan has been manually labeled with 24 facial landmarks. There are 45 female and 60 male subjects, among which 27 are professional actors/actresses. Apart from containing many samples of the six universal facial expressions, it is also the only publicly available database to date that has dedicated scans for AUs in 3D. There are also 3D face scans with different poses, realistic occlusions (e.g., glasses, hands around mouth and eye rubbing) and facial hair (mustache and beard).

TABLE I

3D Facial Expression Databases

Name	Datasets[a]	Exp.[b]	Lmk.[c]	Elicited?[d]
BU-3DFE [7]	2,500 (100)	6	83	N
BU-4DFE [24]	606 videos (101)	6	N/A	N
Bosphorus [25]	4,652 (105)	6/28 AUs	24	N

a) All 3D data have associated 2D textures (number of subjects in parentheses); b) expression types and action units based on the provided labels; c) manually labeled landmarks as the ground truth; d) denotes whether the expressions were genuinely elicited or posed as instructed.

D. An Ideal FER System

Pantic *et al.* [3] listed the general properties of an ideal facial expression analyzer. Such an ideal system will benefit from solutions to multiple computer vision problems (e.g., face detection, landmark localization and illumination normalization). Moreover, depending on the modality of the input, the approaches as well as associated challenges to achieve these goals may be different. For instance, with 3D data it becomes possible not only to deal with a large extent of rigid head motions, but also to extract geometrically invariant features (thus bypassing the problem of pose estimation). Three-dimensional data are also illumination invariant by nature. Hence, a module designated to deal with variations in lighting is no longer necessary. Nevertheless, when a fully automatic FER system is desired, one should choose the appropriate module or provide fusion schemes for these tasks when more than one modality is available (i.e., 3D data with corresponding texture).

As will be discussed in Section III, many of the existing works only tackle the core problems of 3D FER (i.e., feature computation and expression classification) rather than developing a full-fledged automatic system. A possible explanation is that the BU-3DFE database on which the approaches were evaluated offers cropped 3D models (so that no face detection needs to be done) and manually annotated facial landmarks (based on which the features will be extracted, and thus skipping one extra step). Nevertheless, there have been a few attempts to also provide automatic landmark localization in their FER systems ([19], [17], [20]) and some model-based approaches ([17]) have the potential to extract the face region of the input data.

III. CURRENT APPROACHES

Existing approaches in 3D FER and their key properties are summarized in Table II. They can be broadly divided into two categories: feature-based and model-based. Feature-based 3D FER methods focus on the extraction of facial features directly from the input scan while model-based approaches usually involve a generic face model as an intermediate to bring input scans into correspondence by means of registration and/or deformation.

A. *Feature-based Methods*

Wang *et al.* [34] used cubic-order polynomial functions to approximate the continuous surface at each vertex of the input mesh [48]. The estimated coefficients of the polynomial function at a particular vertex $\mathbf{v}$ then form the Weingarten matrix for the local surface patch. The eigenvalues and eigenvectors of this matrix, along with the gradient magnitude that can be derived by the normal direction at $\mathbf{v}$, form a feature set that is used to assign to $\mathbf{v}$ a unique primitive 3D surface label based on a set of classification rules. To overcome the lack of correspondence between the meshes, the authors define seven facial regions using the 64 facial landmarks. The histograms of surface labels are computed for each region and normalized by the number of vertices in the region. This treatment introduces a sense of correspondence that facilitates the subsequent classification step, where the best performance is obtained using Linear Discriminant Analysis (LDA). Note that due to the geometrical invariance of curvature-based features, no rigid transformation is required to achieve this correspondence.

Soyel and Demirel [15] selected six distance measures among a pool of landmarks that maximize the differences of facial expressions to form the feature vectors. The intuition behind this selection comes from the definition of the fundamental facial expressions by the MPEG-4 facial animation parameters (FAPs) [49]. The authors argued that among all 84 feature points specified by MPEG-4, only a small set are not static due to the contraction and expansion of facial muscles when one of the universal expressions is displayed. However, the authors did not specify how to identify this set of feature points. By utilizing facial symmetry, they are able to trim down the number of facial features to merely 11 points, from which the six characteristic distances are extracted. One of the distances, which is essentially the width of the face contour, is used to normalize other distances as an attempt to make the feature scale-invariant. Subsequently, a neural network trained using a backpropagation algorithm is used to classify the expressions. Note that this distance-based feature is also invariant against rigid transformations. In a follow-up work [38], the authors used a similar framework but this time based on the FACS, the set of feature points are different and the distances are computed using several feature points instead of just two, which in the authors' view can cancel out individual variations. That work has been extended further in [16] and [41], where an automatic feature selection mechanism was introduced. Distances between all possible pairs of the 83 manual annotations of the BU-3DFE database are enumerated and normalized. Principal Component Analysis (PCA) is performed on this feature space to reduce its dimensionality and then LDA is applied to find the optimal subspace that preserves the most discriminant information. Realizing that some expression classes are close to each other in the subspace, the authors proposed to regroup these classes into clusters and perform the subspace projection followed by neural network classification in a hierarchical manner. Under the same framework, but without the coarse-to-fine scheme, Tekgüç *et al.* [42] adopted the Non-dominated Sorted Genetic Algorithm II for feature selection and obtained a slightly lower recognition rate.

Tang and Huang [39] explored similar distance features. They proposed an automatic feature selection method based on maximizing the average relative entropy of marginalized class-conditional feature distributions. Less than 30 "best" features were automatically selected using this method from the pool of all possible line segments between the 83 landmarks. Regularized AdaBoost algorithm with three weak classifiers (i.e., Nearest Neighbor (NN), Naive Bayes (NB) and LDA) was used for classification. As a preprocessing step, the feature distances on the neutral scan of a subject are subtracted from the features of his/her expressive scan. In another work [40] they took a manual approach and carefully devised a set of 96 discriminative features that includes not only the normalized distances but also the slopes of the line segments connecting a subset of the 83 landmarks. The distances were normalized by the corresponding facial animation parameter units (FAPUs), which are used to scale the FAPs according to the MPEG-4 standard [49], and were defined as fractions of distances between certain feature points on a face model in its neutral state. This also implies the availability of the neutral scan of the input subject. In addition, the slope features are also normalized to unit vectors. A multi-class Support Vector Machine (SVM) using the *one*-against-*one* scheme is selected for classification.

In their automatic pipeline, Gong *et al.* [18] assume that an expressive 3D face scan is an approximate sum of the expressional shape component (ESC) and the basic facial shape component (BFSC). The BFSC is estimated from a group of aligned neutral scans as well as the input expressive scan. The alignment is done by using local shape difference boosting [50] and then depth images are obtained by sampling the projection of the 3D shapes. Estimation of BFSC is essentially an eigen decomposition and projection process. Based on [51], the authors state that facial expressions are represented orthogonal to the eigenvectors so that, by projecting the expressive scan onto the subspace, the corresponding neutral scan (i.e., BFSC) can be estimated. Then, an expression descriptor is computed by taking the difference between the depth maps of the original scan and its BFSC at selected regions, which eventually becomes the input to an SVM classifier.

Berretti *et al.* [44] compute SIFT features at selected

TABLE II

METHODS FOR 3D FACIAL EXPRESSION ANALYSIS

Reference	Category	2D Used?[a]	Dynamic?[b]	Landmarks	Subject Independent?[c]	Database[d]	Expression Types	Reported Performance (%)[e]
Chang [27]	Model	N	Y	22 semi-auto	Y	Private	6	N/A
Wang [34]	Feature	N	N	64 manual	Y	BU-3DFE	6	83.6
Yin [13]	Model	Y	Y	64 semi-auto	Y	BU-3DFE	6	80.2
Ramanathan [35]	Mode	Y	N	Not used	Y	Private	4	97.0
Soyel [15]	Feature	N	N	11 manual	Y	BU-3DFE	7	91.3
Wang [36]	Feature	Y	N	58 auto	Y	Private	4	83.0
Sun [21]	Model	Y	Y	83 auto	Y	BU-4DFE	6/8 AUs	80.9/87.1
Sun [14]	Model	Y	Y	83 auto	Y	BU-4DFE	6	90.4
Mpiperis [17], [37]	Model	N	N	auto	Y	BU-3DFE	6	90.5, 92.3
Soyel [38]	Feature	N	N	23 manual	Y	BU-3DFE	7	87.8
Tang [39], [40]	Feature	N	N	83 manual	N	BU-3DFE	6	95.1, 87.1
Rosato [28]	Model	Y	Y	22 auto	Y	BU-3DFE BU-4DFE	7/6	80.1/85.9
Venkatesh [19]	Feature	Y	N	68 auto	Y	BU-3DFE	6	81.7
Soyel [16], [41]	Feature	N	N	83 manual	Y	BU-3DFE	7	93.7
Gong [18]	Feature	N	N	auto	Y	BU-3DFE	6	76.2
Tekgüç [42]	Feature	N	N	83 manual	Y	BU-3DFE	7	88.2
Savran [43]	Model	N	N	auto	Y	Bosphorus	22 AUs	91.4
Zhao [23]	Feature	Y	N	19 manual	Y	Bosphorus	7 AUs/16 AUs	94.2/85.6
Savran [22]	Feature	Y	N	auto	Y	Bosphorus	25 AUs	97.1
Berretti [44]	Feature	N	N	27 manual	Y	BU-3DFE	6	77.5
Zhao [20]	Feature	Y	N	19 auto	Y	BU-3DFE	6	82.3
Maalej [45]	Feature	N	N	24 manual	Y	BU-3DFE	6	N/A
Venkatesh [46]	Feature	Y	N	Not used	Y	BU-3DFE	6	85.6
Tsalakanidou [47]	Model	Y	Y	81 auto	N	Private	5/11 AUs	85.0/83.6

a) Denotes whether the method makes use of the 2D texture associated with the 3D data; b) denotes whether the method uses temporal information from a sequence of 3D data; c) denotes whether the method requires a neutral scan or the identity of the subject; d) may be a subset of the listed database; e) the average recognition rates are listed only for reference not for comparison due to different experiment settings.

locations on the depth images, which are generated from sampling the 3D data. Feature selection is performed using the minimal-redundancy maximal-relevance model [52] and FER is accomplished using six *one*-vs-*all* SVM. Maaleg *et al.* [45] proposed a curve-based representation of face patches. The length of the geodesic path is used as a measure for the similarity between any two closed curves [53]. Consequently, the similarities between the corresponding level curves of two face patches are accumulated to indicate the similarity between them. The similarity scores from all patches form the final descriptor to the *binary* classifiers using SVM or AdaBoost.

Some of the feature-based approaches also make use of the 2D texture image associated with the 3D data. Wang *et al.* [36] demonstrated an application of FER in diagnosing Schizophrenia. They quantify the abnormality in facial expressions by combining 2D and 3D features. More specifically, AdaBoost is used for face detection [54] followed by Active Appearance Model (AAM) [55] to automatically locate the fiducial landmarks on the 2D images. From these landmarks, a set of 2D geometric features are extracted. On the 3D surface, Gaussian and mean curvatures are estimated from fitting a continuous surface similar to [34], but the authors follow the HK classification [56] instead to assign the four shape labels. To incorporate texture information, they compute moment invariants by the Gabor wavelets. A series of normalizations are performed to remove individual scale and topological differences as well as the influences of lighting and skin colors. Note that the subjects' neutral face is used when normalizing the geometric features. PCA is applied to reduce the feature dimensionality, followed by LDA to maximize separation between classes. A probabilistic K-NN method [57] is used for classification.

Savran *et al.* [22] evaluated AU recognition performance when using only 3D or 2D modality as well as the fusion of the two. To detect AUs from 2D images the authors deployed the method proposed by Barlett *et al.* [58]. Gabor wavelets are computed from the images. Subsequently, AdaBoost is used for feature selection and SVM is used for classification. For a fair comparison, the 3D data first undergo a series of preprocessing steps and then the mean curvature values are estimated from the smoothed surface. These curvature values are resampled in the 2D domain via orthographic projection and thus can be directly compared with texture images as an input to the same AU recognition pipeline. Fusion is achieved by concatenating the features selected by AdaBoost in both modalities.

Venkatesh *et al.* [19] used texture images associated with the 3D data for automatic feature point extraction. First the texture image is segmented into six regions (eyebrows, eyes, nose and mouth) [59]. Within each bounding box an active contour algorithm is used to find a contour, which is sampled uniformly to obtain the relevant interest points. A shape matrix is formed by retrieving the 3D positions of these points on the texture image. The so-called flow matrix is then generated by subtracting the matrix of a neutral scan from that of a expressive scan of a particular subject. A modified PCA method is performed on training sets of different expressions and six matrix templates are formed by the projected coefficient. During classification, a test scan is projected onto the subspaces and is assigned to the expression that yields the minimal difference in matrix norm between the its coefficients and the template. In another work [46], they proposed a resampling approach to establish a correspondence among the 3D scans. They choose the nose tip as the origin and recenter the 3D data, which are then uniformly sampled using the Qhull algorithm [60]. The color information from the associated texture is interpolated accordingly. This results in a matrix with six channels of color and geometry information. The flow matrix is computed and the final feature vector is a concatenation of the Fourier components of the flow matrix reduced to 1/4 of its original size. Finally K-means clustering and NN schemes are employed for classification.

Zhao *et al.* [20] recognized six basic emotions using the statistical facial feature model (SFAM) and the Bayesian Belief Network (BBN). With local texture and geometry information as well as the support from the global landmark configuration, 19 landmarks can be automatically localized by SFAM. During training, templates of different expressions are built for each of the nine features computed from landmark topology, facial texture and geometry. The BBN then outputs the expression state with the highest belief accumulated from matching each feature with its template. The authors also performed AU recognition using an extended SFAM and an extended feature set [23]. Similarly, templates are created for each AU during training. The final matching score is a weighted sum of the correlation response between the feature set and its template. The input scan is recognized to contain the AU that yields the highest score.

B. Model-based Methods

Yin *et al.* [13] extended the work in [34] by introducing a tracking model for estimating motion trajectories, which are used to construct a spatial-temporal descriptor. A facial expression label map (FELM) based tracking approach is proposed. The tracking model is first aligned to the 3D face scan, and then deformed to fit the target scan by minimizing an energy function. To create the sequential models from the BU-3DFE database, 40 intermediate frames are generated using the key-frame interpolation and synthesis approach based on the four models corresponding to the four intensity levels. The FELM vector and the motion vector are concatenated to form the descriptor, which becomes the input to an LDA classifier.

Mpiperis *et al.* [37] followed a similar deformable model approach as Kakadiaris *et al.* [5] to bring 3D face scans into correspondence. The feature is essentially the geometry (i.e., vertex positions) of the fitted model. PCA and LDA are applied sequentially to reduce the feature dimension and find the optimal subspace. Then particle swarm optimization (PSO) [61] is used to discover a set of rules for FER. In [17], the authors also established the deformation subspace with PCA, together with an automatic mouth boundary detection, an input 3D scan can be represented as a linear combination of the principal modes of this subspace. An asymmetric bilinear model is built with the fitted models and incorporated in a maximum-likelihood classification framework for FER.

Savran and Sankur [43] first generate mean curvature maps by parameterizing the 3D scans with least squares conformal mapping [62]. The curvature maps are brought into correspondence via a multi-resolution elastic registration. Based on the log-likelihood ratio test, a binary decision is made for a particular AU given the observed curvature features at the interior of a manually defined region and the likelihood estimated from the training samples.

Ramanathan *et al.* [35] adapted Shelton's algorithm [63] to obtain correspondence between textured 3D meshes. A Morphable Expression Model (MEM) was then created which incorporates expression-dependent face variations in terms of morphing vectors. An input scan can be represented in the subspace of MEM and the FER is based on the Euclidean distance between the morphing vector of each expression and that of the input scan.

Furthermore, there are also a few FER systems that can process 3D dynamic sequences (sometimes referred to as 3D videos or 4D data). Chang *et al.* [27] built a coarse mesh model and fit it manually to the initial frame of the range data. A 2D tracker was then employed and the model's projection was warped by the 22 tracked feature points. The depth of the vertex was recovered by minimizing the distance between the model and the range data. The generalized expression manifold was built on the facial deformations of the training frames w.r.t. a standard model using Lipschitz embedding [64]. FER was formulated as the estimation of the posterior probability for each expression category.

Rosato *et al.* [28] took an alternative approach to Yin *et al.* [13] for feature tracking with a generic model by automatically establishing vertex correspondences across input scans or dynamic sequences. A deformable template approach [65] was applied to extract 22 feature points on the 2D face texture. The 3D meshes were parameterized in a 2D plane by the circle pattern approach [66]. The proposed coarse-to-fine model adaptation approach between the planar representations was used and the correspondences are extrapolated back to the 3D meshes. The composition of the descriptor and the classifier are the same as in Yin *et al.* [13]. The FER performance was evaluated on both the BU-3DFE [7] and the BU-4DFE database [24].

Also working with the BU-4DFE database, Sun *et al.* [21] used an AAM to track feature points in the 2D texture frames and retrieve their 3D positions. Influence of rigid head motion was eliminated by registering each 3D frame to the initial neutral scan and displacement of the tracked points between the two was used as the feature vector. The statistical information and the temporal dynamics of the training data were learned by HMMs [67] and the Bayesian decision rule was used to classify query sequences given the trained models for either the AUs or the prototypical expressions. This approach was taken one step further in [14]. After the 3D positions of the feature points were identified, radial basis function (RBF) based interpolation was used to adapt the generic model to these feature points. Similar to [34], geometric surface label maps were generated from the adapted models. LDA was used to achieve optimal feature space transformation and the performance of various HMM-based classifiers are evaluated.

Tsalakanidou and Malassiotis [47] presented a fully automatic FER system which is capable of operating at 4-10 frames per second utilizing both 2D and 3D data acquired from a real-time 3D scanner. They build a Active Shape Model (ASM) [68] which is a PDM learned from

81 manually annotated 3D landmarks accompanied by a 2D gradient profile for each landmark. Given a new 2D-3D image pair they fit the ASM to the data using the gradient information in the neighborhood of each landmark. The feature vectors combine geometric information of the landmarks and the statistics on the density of edges and curvature around the landmarks. For classification, a specific set of rules is defined for each expression and each action unit based on the variation of each component of the feature vector w.r.t. a base feature vector computed from the data of a particular subject with neutral expression. Despite its subject-dependent nature, this is the first fully automatic, real-time 2D-3D FER system reported in the literature.

IV. RESEARCH DIRECTIONS

A. Database Challenges

As stated by Zeng *et al.* [4], context is very important for the interpretation of facial expressions. This is a virtually unexplored area in 3D FER, since most of the work either classifies the six basic emotions or recognize AUs based on certain manifestation of the face. However, to make these FER systems really useful in practical situations complex emotions must be taken into consideration. The context of an expression (e.g., location and current task of the expresser, relationship between the expresser and the receiver) goes a long way toward detecting the true affect state of the expresser at that moment. Nonetheless, the study of context dependency in 3D FER is hindered because there is no publicly available database that contains 3D dynamic sequences with the elicited (spontaneous) rather than posed expressions [69]. Furthermore, AUs are more flexible than the six prototypic facial expressions that thousands of anatomically possible facial expressions can be described by a small number of AUs and AU descriptors. If these promises are to be investigated in the 3D domain, there is a need for new databases of 3D dynamic sequences displaying spontaneous facial behaviors, along with the FACS coding done by certified annotators.

However, it is inherently difficult to acquire such kind of data in 3D due to limitations of the sensors. Yin *et al.* [7] have discussed the concerns regarding the elicitation of authentic expressions. Current 3D acquisition systems cannot be hidden as well as 2D cameras and inevitably influence the authenticity of the elicited expressions. In other words, the subjects may be aware of the capturing process and subconsciously pose the expression that they think the operators are expecting. Moreover, acquisition accessories such as high wattage lights and infrared projectors may distract the subject. Bowyer *et al.* [70] have pointed out current limitations of the 3D sensing technologies. They stated that while 3D shape per se is illumination independent, the sensing of 3D shape is not. Variations of the illumination in an acquisition setting can greatly affect the shape captured by the 3D sensor. While sensors that capture static 3D scans can provide their own illumination (i.e., flash) to overpower the ambient light, for dynamic sequences where only constant illumination can be used, conspicuous high wattage lighting equipment is usually a necessity. To date, there is yet no technology that can make 3D captures without drawing undesirable attention.

B. Algorithm Improvement

The BU-3DFE database has been the most frequently used database on which previous work was evaluated. We have to emphasize the fact that the way this database is presented to the research community brings great convenience in this area. More specifically, it provides cropped face regions of the raw 3D scan as well as a dense set of 83 manually labeled landmarks. However, it should not stop researchers from developing automated preprocessing steps to do the same. Many of the previous 3D FER approaches are semi-automatic, in the sense that they require more or less manual annotations to proceed with feature extraction. There has been some work in localizing landmarks using only 3D data, but they do not tackle facial expressions directly. For instance, the landmark set detected by Perakis *et al.* [71] lacks some key points for the purpose of FER and Nair and Cavallaro [72] do not consider the landmarks in the mouth region at all. In fact, most of the automatic approaches make use of the associated 2D texture image or video frames for landmark detection ([36], [19], [20]).

According to Pantic *et al.* [3] and Wang *et al.* [73], not only does each person have his/her own maximal intensity of displaying a particular facial expression, he/she may also have a style of displaying the expression that is unique on some level. This observation suggests that methods which model simultaneously identity and expression (e.g., bilinear models used by [17]) *may* have an inherent advantage. Other approaches either perform rudimentary normalization or assume the availability of the subject's neutral scan and use it for calibration, in the hope of minimizing the inter-subject difference. Furthermore, Wang *et al.* [73] also stated that the expression manifolds are actually nonlinear and linear methods will not be able to discover the underlying manifold. Among the surveyed work, none has investigated nonlinear embedding of the expression space except the use of nonlinear classifiers.

From a more general point of view, there are some other pending investigations in both 2D and 3D domain, which include but are not limited to the estimation of expression intensity, the impact of aging on the facial expression patterns and the interpretation of AU combinations into more complex human emotions. Computational complexity, although rarely mentioned, is crucial if 3D data are to be used in typical HCI scenarios, where real-time response is desired. To date only the system proposed by Tsalakanidou and Malassiotis [47] is able to come close in this regard.

C. Standardized Protocols

While the majority of the previous work in 3D FER has reported performance on the BU-3DFE, the experiment settings vary from group to group. Hence, a direct comparison of the methods developed in the past five years just by the claimed performance may not be fair. In order to develop a common ground for the evaluation of the 3D

FER methods, standardized protocols must be proposed to define experiments under different scenarios, much like the Face Recognition Grand Challenge [29]. A collection of databases (or a new one) with different modalities (i.e., 2D, 3D and 2D/3D over time) need to be made available to the participants who follow the protocols to evaluate their work, and it is then upon the participants' judgement which modalities to use. The file masks for generating training and testing cohorts need to be specified and be used consistently. These masks may present different levels of difficulty. The protocol must also specify a set of landmarks that is realistic to detect, should the methods require manual annotations. Two FER scenarios can be considered, subject-dependent and subject-independent (based on the assumption of the availability of the subject's neutral scan during testing).

V. CONCLUSIONS

In this paper, we surveyed the existing works in 3D FER, many of which have shown promising results in specific experimental conditions. However, the majority of the top performing methods still require manual annotations on the datasets. The robustness of these methods against landmark localization error has not yet been investigated. Furthermore, among the surveyed systems only a handful of them work with dynamic 3D data and/or the recognition of AUs instead of six basic emotion categories. We expect that more effort will be directed to this area now that the corresponding databases are available ([24], [25]). The fact that none of the proposed approaches addresses the challenges of spontaneous expressions also calls for specialized databases. In addition, real-time response is usually a favorable feature for any HCI system. Hence, how to reduce the computational complexity of 3D FER is yet another intriguing problem. Lastly, we propose to develop a set of standardized protocols, so that fair comparisons can be drawn between the experiment results.

REFERENCES

[1] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *Journal of the Optical Society of America A*, vol. 14, no. 8, pp. 1724–1733, August 1997.

[2] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 1999.

[3] M. Pantic and L. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, December 2000.

[4] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2007.

[5] I. Kakadiaris, G. Passalis, G. Toderici, M. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis, "Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 640–649, Apr. 2007.

[6] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 large-scale experimental results," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 831–846, 2010.

[7] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3D facial expression database for facial behavior research," in *Proc.* 7^{th} *International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, Apr. 10-12 2006, pp. 211–216.

[8] J. F. Cohn, "Foundations of human computing: facial expression and emotion," in *Proc.* 8^{th} *International Conference on Multimodal Interfaces*, Banff, Alberta, Canada, 2006, pp. 233–238.

[9] P. Ekman and W. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[10] ——, *Facial action coding system: A technique for the measurement of facial movement.* Palo Alto, CA: Consulting Psychologists Press, 1978.

[11] www.face-and-emotion.com/dataface/facs/emfacs.jsp.

[12] www.face-and-emotion.com/dataface/facsaid/description.jsp.

[13] L. Yin, X. Wei, P. Longo, and A. Bhuvanesh, "Analyzing facial expressions using intensity-variant 3D data for human computer interaction," in *Proc.* 18^{th} *International Conference on Pattern Recognition*, Atlanta, GA, Oct. 8-11 2006, pp. 1248 – 1251.

[14] Y. Sun and L. Yin, "Facial expression recognition based on 3D dynamic range model sequences," in *Proc.* 10^{th} *European Conference on Computer Vision*, Marseille, France, 2008, pp. 58–71.

[15] H. Soyel and H. Demirel, "Facial expression recognition using 3D facial feature distances," in *Proc. International Conference on Image Analysis and Recognition*, vol. 4633, Montreal, Canada, Aug. 22-24 2007, pp. 831–838.

[16] ——, "Optimal feature selection for 3D facial expression recognition with geometrically localized facial features," in *Proc. Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control*, Famagusta, Cyprus, September 2009.

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

[18] B. Gong, Y. Wang, J. Liu, and X. Tang, "Automatic facial expression recognition on a single 3D face by exploring shape deformation," in *Proc.* 17^{th} *ACM International Conference on Multimedia*, New York, NY, 2009, pp. 569–572.

[19] Y. Venkatesh, A. Kassim, and O. Murthy, "A novel approach to classification of facial expressions from 3D-mesh datasets using modified PCA," *Pattern Recognition Letters*, vol. 30, no. 12, pp. 1128–1137, 2009.

[20] X. Zhao, D. Huang, E. Dellandréa, and L. Chen, "Automatic 3D facial expression recognition based on a bayesian belief net and a statistical facial feature model," in *Proc.* 20^{th} *International Conference on Pattern Recognition*, Istanbul, Turkey, 2010, pp. 3724–3727.

[21] Y. Sun, M. Reale, and L. Yin, "Recognizing partial facial action units based on 3D dynamic range data for facial expression recognition," in *Proc.* 8^{th} *IEEE International Conference on Automatic Face and Gesture Recognition*, Amsterdam, The Netherlands, 2008.

[22] A. Savran, B. Sankur, and M. Bilge, "Facial action unit detection: 3D versus 2D modality," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, San Francisco, CA, June 2010, pp. 71–78.

[23] X. Zhao, E. Dellandréa, L. Chen, and D. Samaras, "AU recognition on 3D faces based on an extended statistical facial feature model," in *Proc.* 4^{th} *IEEE International Conference on Biometrics: Theory, Applications and Systems*, Washington, DC, 2010.

[24] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *Proc.* 8^{th} *IEEE International Conference on Automatic Face and Gesture Recognition*, Amsterdam, The Netherlands, Sep. 17-19 2008.

[25] A. Savran, N. Alyuz, H. Dibeklioglu, O. Celiktutan, B. Gokberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Proc. First COST 2101 Workshop on Biometrics and Identity Management*, Roskilde University, Denmark, May 7-9 2008, pp. 47–56.

[26] K. Schmidt and J. Cohn, "Dynamics of facial expression: normative characteristics and individual differences," in *Proc. IEEE International Conference on Multimedia and Expo*, Tokyo, Japan, August 2001, pp. 728–731.

[27] Y. Chang, M. Vieira, M. Turk, and L. Velho, "Automatic 3D facial expression analysis in videos," in *Proc. Analysis and Modeling of Faces and Gestures*, Beijing, China, 2005, pp. 293–307.

[28] M. Rosato, X. Chen, and L. Yin, "Automatic registration of vertex correspondences for 3D facial expression analysis," in *Proc.* 2^{nd} *IEEE International Conference on Biometrics: Theory, Applications and Systems*, Arlington, VA, Sep. 29 - Oct. 1 2008.

[29] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the

face recognition grand challenge," in *Proc. International Conference on Computer Vision and Pattern Recognition*, vol. 1, San Diego, CA, 2005, pp. 947–954.

[30] A. B. Moreno and A. Sánchez, "GavabDB: a 3D face database," in *Proc. 2^{nd} COST275 Workshop on Biometrics on the Internet*, Vigo, Spain, March 2004, pp. 77–85.

[31] www.3dmd.com.

[32] www.di3d.com.

[33] www.inspeck.com.

[34] J. Wang, L. Yin, X. Wei, and Y. Sun, "3D facial expression recognition based on primitive surface feature distribution," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, New York, NY, Jun. 17-22 2006, pp. 1399–1406.

[35] S. Ramanathan, A. Kassim, Y. Venkatesh, and W. S. Wah, "Human facial expression recognition using a 3D morphable model," in *Proc. IEEE International Conference on Image Processing*, Atlanta, GA, 2006, pp. 661–664.

[36] P. Wang, C. Kohler, F. Barrett, R. Gur, and R. Verma, "Quantifying facial expression abnormality in schizophrenia by combining 2D and 3D features," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, Jun. 17-22 2007.

[37] I. Mpiperis, S. Malassiotis, V. Petridis, and M. Strintzis, "3D facial expression recognition using swarm intelligence," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, Mar. 31 - Apr. 4 2008, pp. 2133–2136.

[38] H. Soyel and H. Demirel, "3D facial expression recognition with geometrically localized facial features," in *Proc. 23^{rd} International Symposium on Computer and Information Sciences*, Istanbul, Turkey, 2008.

[39] H. Tang and T. S. Huang, "3D facial expression recognition based on automatically selected features," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Anchorage, AK, June 2008.

[40] H. Tang and T. Huang, "3D facial expression recognition based on properties of line segments connecting facial feature points," in *Proc. 8^{th} IEEE International Conference on Automatic Face and Gesture Recognition*, Amsterdam, The Netherlands, 2008, pp. 1–6.

[41] H. Soyel and H. Demirel, "Optimal feature selection for 3D facial expression recognition using coarse-to-fine classification," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 18, no. 6, pp. 1031–1040, 2010.

[42] U. Tekgüç, H. Soyel, and H. Demirel, "Feature selection for person-independent 3D facial expression recognition using NSGA-II," in *Proc. 24^{rd} International Symposium on Computer and Information Sciences*, Guzelyurt, Turkey, 2009, pp. 35–38.

[43] A. Savran and B. Sankur, "Automatic detection of facial actions from 3D data," in *Proc. 12^{th} IEEE International Conference on Computer Vision Workshops*, Kyoto, Japan, 2009, pp. 1993–2000.

[44] S. Berretti, A. D. Bimbo, P. P. B. B. Amor, and M. Daoudi, "A set of selected SIFT features for 3D facial expression recognition," in *Proc. 20^{th} International Conference on Pattern Recognition*, Istanbul, Turkey, 2010, pp. 4125–4128.

[45] A. Maalej, B. Amor, M. Daoudi, A. Srivastava, and S. Berretti, "Local 3D shape analysis for facial expression recognition," in *Proc. 20^{th} International Conference on Pattern Recognition*, Istanbul, Turkey, 2010, pp. 4129–4132.

[46] Y. Venkatesh, A. Kassim, and O. Murthy, "Resampling approach to facial expression recognition using 3D meshes," in *Proc. 20^{th} International Conference on Pattern Recognition*, Istanbul, Turkey, 2010, pp. 3772–3775.

[47] F. Tsalakanidou and S. Malassiotis, "Real-time 2D+3D facial action and expression recognition," *Pattern Recognition*, vol. 43, no. 5, pp. 1763–1775, 2010.

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

[49] I. Pandzic and R. Forchheimer, *MPEG-4 facial animation: the standard, implementation and applications*. Chichester: John Wiley & Sons Ltd, 2002.

[50] Y. Wang, J. Liu, and X. Tang, "Robust 3D face recognition by local shape difference boosting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1858–1870, 2010.

[51] C. Padgett and G. Cottrell, "Representing face images for emotion classification," in *Proc. Advances in Neural Information Processing Systems*, vol. 9, 1997, pp. 894–900.

[52] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[53] S. H. Joshi, E. Klassen, A. Srivastava, and I. H. Jermyn, "A novel representation for efficient computation of geodesics between n-dimensional curves," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, 2007.

[54] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[55] T. Cootes, G. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[56] E. Trucc and R. Fisher, "Experiments in curvature-based segmentation of range data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 2, pp. 177–182, 1995.

[57] C. Holmes and N. Adams, "A probabilistic nearest neighbor method for statistical pattern recognition," *Journal of the Royal Sta- tistical Society: Series B (Statistical Methodology)*, vol. 64, pp. 295–306, 2002.

[58] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[59] P. Campadelli and R. Lanzarotti, "Fiducial point localization in color images of face foregrounds," *Image and Vision Computing*, vol. 22, no. 11, pp. 863–872, 2004.

[60] C. Barber, D. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Transactions on Mathematical Software*, vol. 22, no. 4, pp. 469–483, 1996.

[61] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE International Conference on Neural Networks*, Perth, Australia, 1995, pp. 1942–1948.

[62] B. Lévy, S. Petitjean, N. Ray, and J. Maillot, "Least squares conformal maps for automatic texture atlas generation," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 362–371, 2002.

[63] C. Shelton, "Morphable surface models," *International Journal of Computer Vision*, vol. 38, pp. 75–91, 2000.

[64] J. Bourgain, "On lipschitz embedding of finite metric spaces in hilbert space," *Israel Journal of Mathematics*, vol. 52, pp. 46–52, 1985.

[65] L. Yin and A. Basu, "Generating realistic facial expressions with wrinkles for model-based coding," *Computer Vision and Image Understanding*, vol. 84, no. 2, pp. 201–240, November 2001.

[66] L. Kharevych, B. Springborn, and P. Schröder, "Discrete conformal mappings via circle patterns," *ACM Transactions on Graphics*, vol. 25, no. 2, pp. 412–438, 2006.

[67] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[68] A. Lanitis, C. Taylor, and T. Cootes, "Automatic interpretation and coding of face images using flexible models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 743–756, 1997.

[69] Anitha C, M. Venkatesha, and B. S. Adiga, "A survey on facial expression databases," *International Journal of Engineering Science and Technology*, vol. 2, no. 10, pp. 5158–5174, 2010.

[70] K. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition," *Computer Vision and Image Understanding*, vol. 101, no. 1, pp. 1–15, Jan. 2006.

[71] P. Perakis, G. Passalis, T. Theoharis, G. Toderici, and I. Kakadiaris, "Partial matching of interpose 3D facial data for face recognition," in *Proc. 3^{rd} IEEE International Conference on Biometrics: Theory, Applications and Systems*, Arlington, VA, Sep. 28-30 2009, pp. 439–446.

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

[73] Y. Wang, X. Huang, C.-S. Lee, S. Zhang, Z. Li, D. Samaras, D. Metaxas, A. Elgammal, and P. Huang, "High resolution acquisition, learning and transfer of dynamic 3-D facial expressions," *Computer Graphics Forum*, vol. 23, no. 3, pp. 677–686, 2004.