A DYNAMIC GEOMETRY-BASED APPROACH FOR 4D FACIAL EXPRESSIONS RECOGNITION

Mohamed Daoudi, Hassen Drira, Boulbaba Ben Amor, Stefano Berretti

A DYNAMIC GEOMETRY-BASED APPROACH FOR 4D FACIAL EXPRESSIONS RECOGNITION

Mohamed Daoudi, Hassen Drira, Boulbaba Ben Amor *

Telecom Lille1, LIFL (UMR 8022)
France

Stefano Berretti

University of Firenze,
Italy

ABSTRACT

In this paper we present a fully automatic approach for identity-independent facial expression recognition from 3D video sequences. Towards that goal, we propose a novel approach to extract a scalar field that represents the deformations between faces conveying different expressions. We extract relevant features from this deformation field using LDA and then train a dynamic model on these features using HMM. Experiments conducted on BU-4DFE dataset following state-of-the-art settings show the effectiveness of the proposed approach.

Index Terms— 4D facial expressions recognition, Riemannian geometry, HMM

1. INTRODUCTION

Automatic recognition of facial expressions has emerged as an active research field with applications in several different areas, such as human-machine interaction, psychology, computer graphics, facial animation of 3D avatars, etc. The first systematic studies on facial expressions date back to the late 70s with the pioneering work of Ekman [1]. In these studies, it is evidenced that, apart the *neutral* expression, the *prototypical* facial expressions can be categorized into six classes, representing *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*.

This categorization of facial expressions has been proved to be consistent across different ethnicities and cultures, so that these expressions are in some sense "universally" recognized. In his studies, Ekman also evidenced that facial expressions can be coded through the movement of face points as described by a set of *action units*. These results inspired many researchers to analyze facial expressions in videos by tracking facial features and measuring the amount of facial movements in subsequent frames [2]. In fact, there is the awareness that facial expressions are highly dynamical processes and looking at sequences of face instances, rather than to still images, can help to improve the recognition performance. More properly, facial expressions can be seen as dynamical processes that involve the 3D space and the temporal dimension (3D plus time, referred to as 4D), rather than being just a static or dynamic 2D behavior. In addition, 3D face scans are expected to feature less sensitivity to lighting conditions and pose variations. These considerations motivated a progressive shift from 2D to 3D in performing facial shape analysis, with the research on 3D facial expression recognition gaining a great impulse thanks to the recent availability of new databases, like the *Binghamton University* BU-3DFE [3], and the *Bosphorus* database [4].

Now, the introduction of appropriate data sets, such as the BU-4DFE developed at *Binghamton University* [5], makes also possible the study and recognition of facial expressions from 4D data. This trend is also inspired by the revolution of inexpensive acquisition devices such as the consumer 3D cameras. In such conditions, the extension of traditional methods developed for expression recognition from 2D videos or from 3D static models can be not effective or even possible and new solutions are required.

According to this, the main goal of this work is to propose and experiment an innovative solution for 4D facial expression recognition.

1.1. Methodology and Contributions

In this paper, we propose a fully automatic facial expression recognition approach that exploits the motion extracted from 3D facial videos. An overview of the proposed approach is given in Fig. 1. In the preprocessing stage, the 3D mesh in each frame is first aligned to the previous one and then cropped. The 3D motion is then captured based on a dense scalar vector field that represents the 3D deformation between two successive frames. Then, we use Linear Discriminat Analysis (LDA) to transform derived feature space to an optimal compact space to better separate different expressions. Given the optimized features, the second stage is to learn one HMM for each expression. Temporal modeling of the full expression is performed via neutral-onset-apex-offset HMMs, in an unsupervised fashion. Expression classification is then performed by using HMMs trained with the time vari-

*This work has bee supported by the French research agency ANR through the 3D Face Analyzer project under the contract ANR 2010 INTB 0301 01.

ations of the extracted features. Experimental results show that the proposed approach is capable to improve state of the art performance on the BU-4DFE.

The present work includes the following main contributions:

1. A novel Scalar Vector Field (SVF) defined on radial curves of 3D faces; The SVF grounds on Riemannian shape analysis and is capable to accurately capture the deformations occurring between 3D faces represented as sets of radial curves;

2. A new approach for facial expression recognition from 3D dynamic sequences, which combines the high descriptiveness of SVF extracted from successive 3D faces of a sequence with the temporal modeling and classification provided by HMMs;

3. A through experimental evaluation that compares the proposed solution with state of the art methods on a common data set and setting; Results shows that our approach is capable to achieve state of the art results.

Finally, to the best of our knowledge, this is the first fully automatic approach for dynamic 3D facial expressions recognition.

The rest of the paper is organized as follows: In Sect. 2, a face representation model is proposed that captures facial features relevant to categorize expression variations in 3D dynamic sequences. In Sect. 3, the HMM based classification of the selected features is addressed. Experimental results and comparative evaluation obtained on the BU-4DFE are reported and discussed in Sect. 4. Finally, conclusions and future research directions are outlined in Sect. 5.

2. 3D SHAPE MOTION ANALYSIS

One basic idea to capture facial deformation across 3D video sequences is to track densely meshes' vertices along successive 3D frames. To do so, as the meshes resolutions vary across 3D video frames, establishing a dense matching on consecutive frames is necessary. Sun et al. [6] proposed to adapt a generic model (a tracking model) to each 3D frame. However, a set of 83 predefined key-points is required to control the adaptation based on radial basis function. The main limitation is that the 83 points are manually annotated in the first frame of each sequence. A second solution is presented by Sandbach et al. [7], where the authors used an existing non-rigid registration algorithm called Free Form Deformation (FFD) based on B-splines interpolation between a lattice of control points. The dense matching is a step of preprocessing stage to estimate a motion vector field between frames t and t-1. However, the results provided by the authors are limited to three facial expressions: *happy*, *angry* and *surprise*.

To address this problem, we propose to represent facial surfaces by a set of parameterized radial curves emanating from the tip of the nose. Such an approximation of facial surfaces by indexed collections of curves can be seen as solution to facial surface parameterizations which capture locally their shapes.

A parameterized curve on the face, $\beta : I \rightarrow \mathbb{R}^3$, where $I = [0, 1]$, is represented mathematically using the *square-root velocity function* [8], denoted by $q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}}$. This specific parametrization has the advantage of capturing the shape of the curve and provides simple calculus. Let define the space of such functions: $\mathcal{C} = \{q : I \rightarrow \mathbb{R}^3, \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^3)$, where $\|\cdot\|$ implies the $\mathbb{L}^2$ norm. With the $\mathbb{L}^2$ metric on its tangent spaces, $\mathcal{C}$ becomes a Riemannian manifold.

The short path between q_1 and q_2 is given by the following expression:

$$\psi^*(\tau) = \frac{1}{\sin(\theta)} \left(\sin((1-\tau)\theta)q_1 + \sin(\theta\tau)q_2\right) \ , \qquad (1)$$

and $\theta = d_{\mathcal{C}}(q_1, q_2) = cos^{-1}(\langle q_1, q_2 \rangle)$. We point out that $sin(\theta) = 0$ if the distance between the two curves is null, in other words $q_1 = q_2$. In this case, for each τ, $\psi^*(\tau) = q_1 = q_2$. The tangent vector field on this geodesic is then given as $\frac{d\psi^*}{d\tau} : [0, 1] \rightarrow T_\psi(\mathcal{C})$:

$$\frac{d\psi^*}{d\tau} = \frac{-\theta}{\sin(\theta)} \left(\cos((1-\tau)\theta)q_1 - \cos(\theta\tau)q_2\right) \ . \qquad (2)$$

$$\frac{d\psi^*}{d\tau}|_{\tau=0} = \frac{\theta}{\sin(\theta)} \left(q_2 - \cos(\theta)q_1\right) \ , \qquad (3)$$

with $\theta \neq 0$.

In the practice, the first step to capture the deformation between two given 3D faces F_1 and F_2 is to extract the radial curves. Let β^1_α and β^2_α denote the radial curves that make an angle α with a reference radial curve on faces F_1 and F_2, respectively. The reference curve is chosen to be the vertical curve as the faces have been rotated to the upright position during the preprocessing step. The tangent vector field $\dot{\psi_\alpha}^*$ that represents the energy E given in Eq. (1) needed to deform β^1_α to β^2_α is then calculated for each index α. We consider the magnitude of this vector field at each point k of the curve for building a scalar vector field on the facial surface $V^k_\alpha = ||\dot{\psi}^*|_{(\tau=0)}(k)||$, where α denotes the angle to the vertical radial curve and k denotes a point on this curve. This scalar vector field quantifies the local deformation between the faces F_1 and F_2. In practice, we consider 100 curves and 50 points on each curve. Fig. 2.a illustrates the proposed idea. A neutral mesh is reported on the left. The vector field is computed between the neutral face and apex frames of each expression. The values of the vector field are reported using range of colors. In particular, black colors represent the highest deformations whereas the lower values of the vector field are represented in blue. It can be observed, as the regions with high deformation lie in different parts of the face for different expressions. For example, as intuitively expected, the

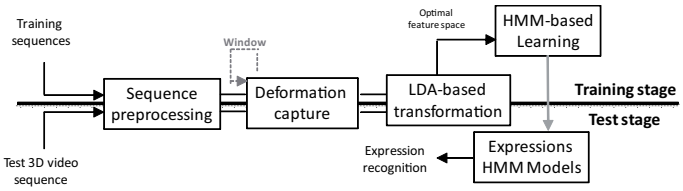

Fig. 1. Overview of the proposed approach in training and test stages, including preprocessing, 3D deformation capture, dimension reduction, and HMM-based classification

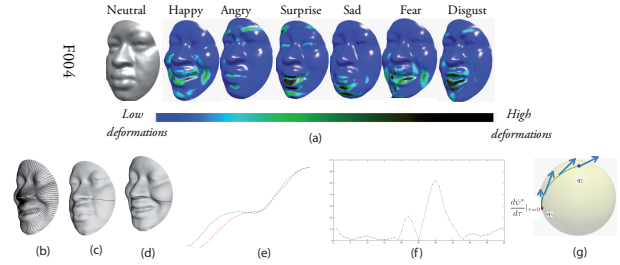

Fig. 2. (a) The first row reports deformation maps computed by the proposed method for the six emotional states computed from a neutral face. In the second row, we illustrate: (b) the extracted radial curves; (c) A radial curve on a neutral face, and (d) the correspondent radial curve on the face with happy expression; (e) the two curves are reported together. The plot of the trade-off between points on the curve and values of magnitude of $\frac{d\psi^*}{d\tau}|_{\tau=0}(k)$ is reported in (f) and an illustration of this parallel vector field across the geodesic between q_1 and q_2 in the space of curves $\mathcal{C}$ is reported in (g).

corners of the mouth and the cheeks are mainly deformed for happiness expression. In Fig. 2.b, we illustrate the face conveying happy expression with extracted radial curves. Fig. 2.c and Fig. 2.d illustrate two correspondent radial curves on neutral and happy faces respectively. These curves are reported together in Fig. 2.e, one can easily see the deformation between them although they lie at the same angle and belong to the same person. The amount of the deformation between the two curves is calculated using Eq. (3) and the plot of the magnitude of this vector at each point of the curve is reported in Fig. 2.f.

3. EXPRESSION CLASSIFICATION BASED ON HMMS

In our case, sequences of 3D frames constitute the temporal dynamics to be classified, and each prototypical expression is modeled by an HMM (a total of 6 HMMs λ^{expr} is required, with $expr \in \{an, di, fe, ha, sa, su\}$). Four states per HMM (N=4) are used to represent the temporal behavior of each expression. This corresponds to the idea that each sequence starts and ends with a neutral expression (state S_1); The frames that belong to the temporal intervals where the face changes from neutral to expressive and back from expressive to neutral are modeled by the *onset* (S_2) and *offset* (S_4) states, respectively; Finally, the frames corresponding to the highest intensity of the expression are captured by the apex state (S_3). Fig. 3 exemplifies the structure of the HMMs in our framework.

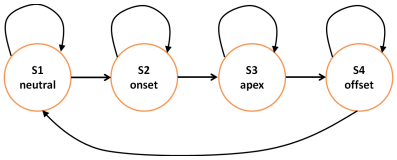

Fig. 3. Structure of the HMMs modeling a 3D facial sequence. The four states model, respectively, the *neutral*, *onset*, *apex* and *offset* frames of the sequence. As shown, from each state it is possible to remain in the state itself or move to the next one (*Bakis* or left-right HMM).

4. EXPERIMENTAL RESULTS

The proposed framework for facial expression recognition from dynamic sequences of 3D face scans has been experimented using the BU-4DFE database. Main characteristics of the database and results are reported in the following.

4.1. Expression Classification Performance

4D expression sequences of the BU-4DFE from *Binghamton University* [5] are affected by large outliers, mainly acquired in the hair, neck and shoulders regions.

After preprocessing operations (alignment, detection of the nose tip, etc..), data of 60 subjects have been selected to perform recognition experiments, whereas the remaining 41 subjects have been used for a preliminary tuning of the proposed algorithms. The 60 subjects are randomly partitioned into 10 sets, each containing 6 subjects, and 10-fold cross validation has been used for test, where at each round 9 of the 10 folds (54 subjects) are used for training while the remaining (6 subjects) are used for test. The recognition results of 10 rounds are then averaged to give a statistically significant performance measure of the proposed solution.

The proposed approach is able to correctly classify all the sequences with an accuracy of 100%. Indeed, the classification model is capable to correctly identify 3D dynamic expression sequences. This provides a measure of the overall capability to classify 3D frames sequences composed of around hundred frames and with a typical behavioral pattern.

In some contexts the classification of individual 3D frames is also relevant in that can permit an online analysis of a 3D video. Following the experimental protocol proposed in [6], this is obtained by the definition of a large set of very short subsequences extracted using a sliding window on the original expression sequences. The subsequences have been defined in [6] with a length of 6 frames with a sliding step of one frame from one subsequence to the following one. For example, with this approach, a sequence of 100 frames originates a set of $6 \times 95 = 570$ subsequences, each subsequences differing from one frame from the previous one. This accounts for the fact that, in general, the subjects can come into the system not necessarily starting with a neutral expression, but with a generic expression. Classification of these very short sequences is regarded as an indication of the capability of the expression recognition framework to identify individual expressions. According to this, for this experiment we retrained the HMMs on 6 frame subsequences constructed as discussed above. The 4-state structure of the HMMs still resulted adequate to model the subsequences. Also in this experiment, we performed 10-folds cross validation on the overall set of subsequences derived from the 60×6 sequences (31970 in total).

The results obtained by classifying individual 6-frames subsequences of the expression sequences (*frame-by-frame* experiment) are reported in the confusion matrix of Tab. 1. Values in the table have been obtained by using 6-frames subsequences as input to the 6 HMMs and using the maximum emission probability criterion as decision rule. It is evident that the proposed approach is capable to accurately classify very short sequences containing very different 3D frames, with an average accuracy of 93.83%. It can be noted that the lower recognition is obtained for the *disgust* expression (91.54%) which is mainly confused with the *angry* and *fear* expression. Interestingly, these three expressions capture negative emotive states of the subjects, so that similar facial muscles can be activated.

Table 1. Average confusion matrix for 6-frames subsequences (percentage values)

	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	**93.95**	1.44	1.79	0.28	2.0	0.54
Disgust	3.10	**91.54**	3.40	0.54	1.27	0.15
Fear	1.05	1.42	**94.55**	0.69	1.67	0.62
Happy	0.51	0.93	1.65	**94.58**	1.93	0.40
Sad	1.77	0.48	1.99	0.32	**94.84**	0.60
Surprise	0.57	0.38	3.25	0.38	1.85	**93.57**

4.2. Discussion and Comparative Evaluation

To the best of our knowledge, the only three works reporting results on expression recognition from dynamic sequences of 3D scans are [6], [7] and [9]. These works have been verified on the BU-4DFE dataset, but the testing protocols used in the experiments are quite different, so that a direct comparison of the results reported in these papers is not possible.

The approach in [6] is not completely automatic and also presents high computational cost. In fact, a generic model (i.e., tracking model) is adapted to each depth model of a 3D sequence. The adaptation is controlled by a set of 83 predefined keypoints that are manually identified and tracked in 2D. The person-independent expression recognition experiments were performed on 60 selected subjects out of the 101 subjects of the BU-4DFE database, by generating a set of 6-frame subsequences from each expression sequence to construct the training and testing sets. The process were repeated by shifting the starting index of the subsequence every one frame till the end of the sequence. The rationale used by the authors for this shifting was that a subject could come to the recognition system at any time, thus requiring the recognition process could start from any frame. Following a 10-fold cross-validation, an average recognition rates of 90.44% was reported. So, it results that expression recognition results are actually provided not on variable length sequences of 3D depth frames, but just on very short sequences with a predefined length of 6 frames.

The method proposed in [7] is fully automatic with respect to the processing of facial frames in the temporal sequences, but uses *supervised* learning to train a set of HMMs. Though performed offline, supervised learning requires manual annotation and counting on a consistent number of training sequences that can be a time consuming operation. In addition, a drawback of this solution is the computational cost due to Free-Form Deformations based on B-spline interpolation between a lattice of control points for nonrigid registration and motion capturing between frames. This hinders the possibility of the method to adhere to a real time protocol of use. Preliminary tests were reported on three expressions: *angry*, *happiness* and *surprise*. Authors motivated the choice of the happiness and anger expressions with the fact that they are at either ends of the valence expression spectrum, whereas

surprise was also chosen as it is at one extreme of the arousal expression spectrum. However, these experiments were carried out on a subset of subjects accurately selected as acting out the required expression. Verification of the classification system was performed using a 10-fold cross-validation testing. On this subset of expressions and subjects, an average expression recognition rate of 81.93% is reported.

In [9] a fully automatic method is also proposed, that uses an *unsupervised* learning solution to train a set of HMMs. In this solution, preprocessing is very important in that an accurate alignment of the 3D mesh of each frame is required in order to extract the level set curves that are used for face representation. This increases the computational cost of the approach making questionable its use where a real time constraint is required. Expression recognition is performed on 60 subjects from the BU-4DFE database for the expressions of *happiness*, *sadness* and *surprise*. Results of 10-fold cross-validation show an overall recognition accuracy of 92.22%, with the highest performance of 95% obtained for the happiness expression.

Tab. 2 summarizes the results scored by the above methods compared to those presented in our work. Considering the classification of the entire 4D sequences, our solution clearly outperforms those in [7] and [9], even working on six expressions instead of just three, evidencing the capability of the proposed face representation to capture salient features to discriminate between different expressions. With respect to the *frame-by-frame* classification experiment, our results are more than 3% better than those in [6], with the advantage of using a completely automatic approach and a simpler classification model using temporal HMMs with fewer states.

Table 2. Comparison to earlier work (values in percentage)

Av. RR	[6]	[6]	[7][1]	[9][2]	**This work**
Frame-by-frame	80.04	90.44	73.61	-	**93.83**
Sequence	-	-	81.93	92.22	**100**

[1] Performances provided for *happiness*, *anger* and *surprise* expressions.
[2] Performances provides for *happiness*, *sadness* and *surprise* expressions.

5. CONCLUSION AND FUTURE WORK

In this paper, we presented a fully automatic approach for identity-independent facial expression recognition from 3D video sequences. Through a facial shapes representation by collections of radial curves, a Riemannian shape analysis was applied to quantify dense deformations and extract motion from successive 3D frames. The resulting features are then used to train a HMM after LDA-based feature space transformation. The proposed approach outperforms previous ones; it is capable to accurately classify very short sequences containing very different 3D frames, with an average accuracy of 93.83% following state-of-the-art settings on the BU-4DFE dataset. A limitation of the approach is the nose tip detection in case of non frontal views and/or occlusion. The BU-4D contains frontal 3D faces without occlusion, however, in realistic scenario, more elaborated techniques should be applied to detect the nose tip.

6. REFERENCES

[1] P. Ekman, "Universals and cultural differences in facial expressions of emotion," in *Proc. Nebraska Symposium on Motivation*, Lincoln, NE, 1972, vol. 19, pp. 207–283.

[2] 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, Jan. 2009.

[3] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Southampton, UK, Apr. 2006, pp. 211–216.

[4] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Proc. First COST 2101 Workshop on Biometrics and Identity Management*, May 2008.

[5] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3d dynamic facial expression database," in *Int. Conf. on Automatic Face and Gesture Recognition (FGR08)*, Amsterdam, The Netherlands, Sept. 2008, pp. 1–6.

[6] Y. Sun and L. Yin, "Facial expression recognition based on 3D dynamic range model sequences," in *Proc. Eur. Conf. on Computer Vision*, Marseille, France, Oct. 2008, pp. 58–71.

[7] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Daniel Rueckert, "A dynamic approach to the recognition of 3d facial expressions and their temporal models," in *IEEE Conf. on Automatic Face and Gesture Recognition*, Santa Barbara, CA, Mar. 2011, pp. 406–413.

[8] A. Srivastava, E. Klassen, S. Joshi, and I. Jermyn, "Shape analysis of elastic curves in euclidean spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1415–1428, July 2011.

[9] Vuong Le, Hao Tang, and Thomas S. Huang, "Expression recognition from 3d dynamic faces using robust spatio-temporal shape features," in *IEEE Conf. on Automatic Face and Gesture Recognition*, Santa Barbara, CA, Mar. 2011, pp. 414–421.