Dynamic 3D Facial Expression Recognition Using Robust Shape Features

Ahmed Maalej[1], Hedi Tabia[2], and Halim Benhabiles[3]

[1] LSIS/ENSAM, Arts et Mtiers ParisTech, Lille, France
[2] ETIS/ENSEA, University of Cergy-Pontoise, CNRS, UMR 8051, France
[3] ESIGELEC/IRSEEM (EA 4353), Saint-Etienne du Rouvray, France

Abstract. In this paper we present a novel approach for dynamic facial expression recognition based on 3D geometric facial features. Geodesic distances between corresponding 3D open curves are computed and used as features to describe the facial changes across sequences of 3D face scans. Hidden Markov Models (HMMs) are exploited to learn the curves shape variation through a 3D frame sequences, and the trained models are used to classify six prototypic facial expressions. Our approach shows high performance, and an overall recognition rate of 94.45% is attained after a validation on the BU-4DFE database.

Keywords: Facial expression recognition, shape analysis, hidden Markov models.

1 Introduction

Facial expression recognition has increasingly gained interest of researchers in computer science field, leading to a continuous need to develop systems that are capable of understanding human emotions. Indeed, since machines are becoming more and more involved in everyday human life and take part in both his living and workspace, there is a tendency to embed these machines with intelligent modules that are able to analyse and recognize the human expressions. The facial expression recognition topic can find its applications in various domains such as psychology, medical care, security, etc.

The recent progress of 3D imaging systems (stereo vision cameras, laser/structured-light 3D scanners, time of flight and RGB-D cameras) has made the creation of facial range models simple and abundant. Three-dimensional data has emerged to provide an additional and valuable information which is the depth (z-value) information. Besides such data has shown the potential to alleviate problems encountered in 2D-based approaches: small pose variation can be handled and illumination differences can be avoided. 3D face databases become more and more available, providing the worldwide researchers of Face and Facial Expression Recognition community a large-scale data for training and evaluating their approaches. 3D facial expression recognition approaches can be categorized into two classes as well: static and dynamic. Different methods have been proposed in the static direction and

J.-K. Kämäräinen and M. Koskela (Eds.): SCIA 2013, LNCS 7944, pp. 309–318, 2013.

gave promising results. However, research in the dynamic direction is revealed to be more valuable as well as challenging. Since that facial expression is, by nature, a highly dynamical process, hence studying the dynamic cues while looking at sequences of expressive face frames can help improve the recognition performance. So far, there exists several approaches that exploit 3D facial expression dynamics. The first approach was proposed by Sun et al. [1] where they developed a rich spatio-temporal descriptor build through combination of a template and geometrical features. The template is a 3D generic deformable model constructed to estimate the physical process of facial changes due to expressions and vertex flow estimation is derived to compute vertex displacement from one frame to another. As for the geometrical feature they proposed to compute curvatures and developed an automatic surface labelling approach to classify the 3D primitive surface features into eight label basic categories. As a result each range model in a given frame sequence can be represented by a spatio-temporal feature vector that describes both shape and motion, hence providing a robust facial surface representation. Two-dimensional HMM models, spatial HMM (S-HMM) and a temporal HMM (T-HMM), were then used to conduct facial expression classification and a recognition rate of 83.7% was reached. Sandbach et al. [5] proposed to use HMM models for temporal modelling of the full expression sequence to be represented by 4 segments which are neutral-onset-apex-offset expression segments. They applied Free-Form Deformations for motion capture between frames, and extracted motion features using a quad-tree decomposition of several motion fields. Features selection is then derived using GentleBoost technique, and the obtained average recognition rate of three basic expressions(i.e., happy, angry and surprise) was 81.93%. Le et al. [2] proposed a level curve based approach to capture the shape of 3D facial models. The level curves are extracted using the arc-length parametrization and were partitioned into normalized segments. Then the Chamfer distance is applied to quantify the shape deformation between the corresponding segments. These measures are then used as spatio-temporal features to train HMMs. Using the BU-4DFE database to evaluate their approach, they reached an overall recognition accuracy of 92.22% for three prototypic expressions (i.e., happy, sad and surprise). Fang et al. [7] proposed a fully automatic pipeline to classify expression from 4D data. Their pipeline is set to start with a robust registration of each pair of consecutive frames of a given sequence. They developed a two-step technique to derive a mesh matching process; the first step consists in establishing vertex correspondence between two meshes with providing two alternative methods, one is based on spin images similarities and the other based on the Euclidean distance between MeshHOG descriptors. Then RANdom SAmple Consensus (RANSAC) is applied to alleviate the problem of nosy point correspondence due to outliers that might be generated by the vertex correspondence step. A deformable face model (AFM) is then applied to generate a fitted sequence from a 4D data set. Local Binary Patterns (LBP) descriptors are computed and flow image is estimated to represent the deformation vectors.

In the final stage of their proposed pipeline, they applied support vector machines for classification and outperformed previous work by achieving 95.75 % average classification rate on the BU-4DFE database.

2 Method

In this section we present a fully automatic pipeline for classifying six prototypical expressions from 4D data. First, a preprocessing step is applied to extract the face area and discard the non-informative part of the raw image. Second, a registration step is run using a global registration method applied on each pair of consecutive frames. Then, we adopt a sparse surface representation based on both contour and profile curves to approximate the 3D face model. These curves are employed to conduct shape analysis using a Riemannian framework, and extract temporal features. Finally, the temporal dynamics of the extracted features are learned using HMMs and the Bayesian decision rule is used to classify the query sequences given the trained models for the basic expressions.

2.1 Preprocessing

The raw data obtained from even most accurate 3D sensors is far from being perfect and clean for straightforward processing, as it may contain spikes, holes and noise. A preprocessing stage must be applied to remove these anomalies before any further operations can be performed. Thus preprocessing is important for any recognition system, especially when knowing that all the features will be extracted from the output of this stage. We developed an automatic preprocessing pipeline that is set to apply different tools and follow multiple steps. The first step is the nose detection step, the nose tip is a key point that is needed for preprocessing and also for facial surface representation. Exploiting the fact that in most 3D facial scans brought by publicly available databases, the nose is the closest facial region to the 3D acquisition systems, we simply detect the nose tip using horizontal and vertical slicing of the facial surface and a search for the maximum value of the z-coordinate of these curved profiles. This is done for the first frame in a sequence of frames, for the remaining ones, this detection technique is refined since and the search area in a current frame is reduced to a small sphere centered on the nose tip detected in the previous frame. The second step of data preprocessing is the cropping step, which consists of keeping the required facial area from the raw image and removing the irrelevant parts (i.e, neck, hair, face boundaries).

2.2 Facial Surface Representation

The primary concern for a surface representation in a facial expression recognition system, focus on extracting detailed information. This information needs to be accurate, concise and useful for statistical modeling. We chose to represent 3D facial shapes by the union of curves, profile and contour curves. This combination of curves are obtained from a curve extraction stage. For the contour

curves, that are iso-radius curves, are extracted from the intersection of the facial surface and a sphere defined by the the nose tip, as a center point, and a radius $r = \sqrt{x^2 + y^2 + z^2}$. These contour curves are closed curves and of various length, that is the number of points per curve. The bigger is the value of r, the larger is the number of points of the curve. As for the profile curves, they are open curves that are extracted using the intersection of the face model with a plane. The plane is also defined by a the noise tip and a normal vector parallel to the xy plane. Contrary to the contour curves, where the starting point is also an ending point of a curve, the starting point of all profile curves is set to be the nose tip and the ending point is the edge of face determined by the cropping step of the preprocessing stage. A collection of contour curves is obtained using a sphere defined by different radius values, and the set of profile curves are obtained by a plane with different rotation angles applied on its normal vector. Furthermore, we are interested in building features to quantify the shape deformation related to local regions of the facial surfaces. This led us to the idea of partitioning the contour curves into segments . Both profile curves and segments of contour curves are open curves, that represent the 3D facial surface in a concise manner Fig. 1. These curve are also reliable for computations, memory storage, can be efficiently displayed and suitable for deriving shape analysis.

2.3 Shape Analysis

In order to conduct shape analysis, we apply an efficient framework for analyzing the shape of these curves. Anuj et al. [3] introduced a square-root velocity function (SRVF) representation for analyzing shapes of closed curves in $\mathbb{R}^n$. We start by defining an open curve and the space of all parametrized open curves using differential geometry. Let β be an open curve with $\beta : I \to \mathbb{R}^3$, where $I = [0, 1]$ stands for the domain of parametrization and is set to allow focusing on curves of unit length that can be obtained through a scaling process. β is supposed to be continuous and whose derivative is $\dot{\beta}(t)$, exists almost everywhere and never vanishes: $\dot{\beta}(t) \neq 0, \forall t$.

β can be represented by the SRVF $q(t)$, given by:

$$q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}} \tag{1}$$

Note that q is already invariant to the translation of β in $\mathbb{R}^3$. However it is still dependent on rotation and the choice of parametrization.

Let $\mathcal{C}$ be the space of square-root velocity functions, or the space of all open curves, defined by: $\mathcal{C} = \{q : I \to \mathbb{R}^3, \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^3)$, where $\|.\|$ implies the $\mathbb{L}^2$ norm. Here the elements of $\mathcal{C}$ have a unit $\mathbb{L}^2$ norm, $\mathcal{C}$ is a Hypersphere in the Hilbert space $\mathbb{L}^2(I, \mathbb{R}^3)$. Given any two open curves β_1 and β_2 we can represent them respectively by q_1 and q_2 of the space $\mathcal{C}$, we would like to quantify the similarities and dissimilarities between their corresponding shapes. It is important to remind that these quantifications should not depend on the rotation,

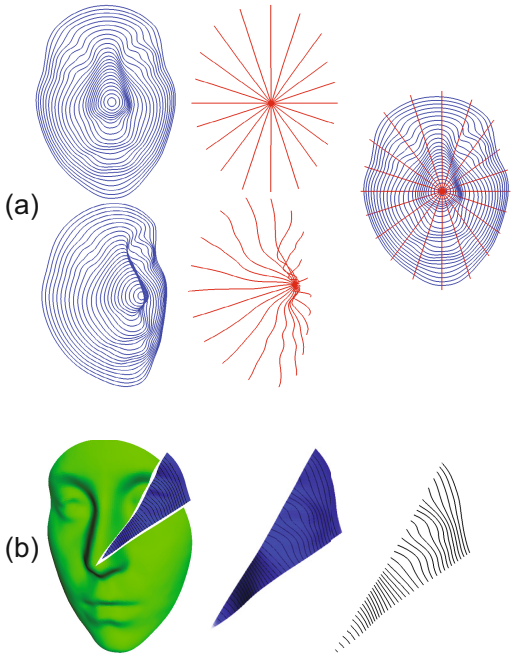

Fig. 1. Curve-based facial surface representation: (a) contour curves (80 curves), profile curves (10 curves) and their combination, (b) extraction of a sector area from the facial surface and generation of segments of curves

and re-parametrization that can change the curve but do not change its shape. As the elements of $\mathcal{C}$ have a unit $\mathbb{L}^2$ norm, $\mathcal{C}$ is a Hypersphere in the Hilbert space $\mathbb{L}^2(I, \mathbb{R}^3)$. Using the Riemannian structure we can write explicit forms for geodesics between any two open curves $q_1, q_2 \in \mathcal{C}$ and it is simply given by the minor arc of great circle connecting them on this Hypersphere. Let α denote the geodesic path between q_1, q_2 and that is defined by $\alpha : [0, 1] \rightarrow \mathcal{C}$:

$$\alpha(\tau) = \frac{1}{\sin(\theta)} \left(\sin((1-\tau)\theta)q_1 + \sin(\theta\tau)q_2\right) \ , \tag{2}$$

where $\tau \in [0, 1]$ and $\theta = \cos^{-1}\langle q_1, q_2\rangle$. And the length of the geodesic path (geodesic distance), that we denote it by $d_{\mathcal{C}}$, is given by:

$$d_{\mathcal{C}}(q_1, q_2) = \cos^{-1}\langle q_1, q_2\rangle \tag{3}$$

3 HMM-Based Classification

The characterization of the shape-based signature of facial open curves, through an acquired expression video sequence, is the core of the proposed approach. The goal is to characterize the real world facial expressions in terms of signal

models. The principle reason for applying signal models is to provide a basis for the theoretical description of the expression processing system. Besides, developing a signal model for a particular process, has been proven to be efficient, work extremely well in practice and enables us to build practical systems, such as prediction, recognition and identification systems. Thus, modeling a given expression by a signal can help considerably in building an expression recognition system in a valuable manner.

In this work we adopt a stochastic signal model, which is the HMM [4]. The underlying assumption of the use of HMM is that the facial expression can be well characterized as a parametric random process and that the parameters of the stochastic process can be estimated in a well defined manner.

Given a 3D dynamic sequence composed of T frames, from each frame of the 3D face model, we extract a combination of contour and profile curves. These curves are then partitioned to obtain a set of open curves $\{\beta_k\}_{1 \leqslant k \leqslant N}$ where each curve β_k characterizes a local shape of the facial surface. The SRVF representation q_k is applied to encode the shape information of β_k. The length of the geodesic distance separating the curve q_k, computed for a current frame, and q_k, computed for the next frame, is calculated according to Eq. 3. We denote by $s(q_k)$ the set of all geodesic distances calculated through the frame sequence for the k^{th} curve, $s(q_k) = \left\{d_{\mathcal{C}}(q_k^t, q_k^{t+1})\right\}_{1 \leqslant t \leqslant T-1}$. $s(q_k)$ encodes the temporal dynamics of facial expressions, and will be considered as the observed sequence for HMM application .

3.1 HMM-Based Signature of Facial Curves

An HMM is a is a temporal probabilistic model and a finite set of states that are not directly observable (hidden states), each state is characterized by a probability distribution function. To completely define an HMM, we need the following elements:

- $S = S_1, S_2, \ldots, S_m$, a set of m states, where each state can be associated with a particular shape information captured from an open curve.
- $A = \{a_{i,j}\}$, $1 \leqslant i, j \leqslant m$, is the transition matrix representing the probability of moving from state S_i to state S_j. So that, $a_{ij} = P\left[Q_{t+1} = S_j | Q_t = S_i\right]$, $1 \leqslant t \leqslant T$ with $a_{ij} \geqslant 0$, $\sum_{j=1}^{m} a_{ij} = 1$ and where Q_t represents the model state at time t. This matrix encodes how the curve shape deformation evolves through 3D image sequence.
- $B = \{b(o/S_i)\}$ is the emission matrix representing the emission probability of the observation o when system state is S_i. Where o stands for the geodesic distance information that can take values in $\mathbb{R}^+$. And $b(o/S_i)$ is a Gaussian probability density function.
- $\pi = \pi_i$, corresponds to the initial state probability distribution, representing probabilities of initial states $\pi_i = P\left[Q_1 = S_i\right]$, $1 \leqslant i \leqslant m$ with $\pi_i \geqslant 0$ and $\sum_{i=1}^{m} \pi_i = 1$

  In our work, given an observed sequence $s\,(q)$, the HMM parameters are learned, using the well-known Baum-Welch (BW) algorithm, which is able

to determine the parameters of the model λ_i by maximizing the likelihood $P\left(s\left(q_i\right)|\lambda_i\right)$. In this way, the HMM gives a statistical encoding of the facial curves deformation from a frame to another, taking into account the uncertainty in the data.

3.2 Training Stage

Let $O_j^{EXP} = [s(q_1), s(q_2), \ldots, s(q_N)]$ be the set of the extracted open curves for subject j having the expression $EXP \in \{HA, AN, FE, DI, SA, SU\}$. For expression recognition we collect this data from subjects with the same expression. Then, the face expression model $\Lambda^{EXP} = \left[\lambda_1^{EXP}, \ldots, \lambda_N^{EXP}\right]$ is generated by learning an HMM λ_i by taking into account all the subjects $j = 1, \ldots, M$ with that specific expression. Indeed, the process is repeated for all different expressions, by obtaining $\Lambda^{HA}, \Lambda^{AN}$ and so on. In our case, four states per HMM (N=4) are used to represent the temporal behaviour 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 that correspond to the highest intensity level of the expression are captured by the apex state (S_3). Fig. 2 exemplifies the structure of the HMMs in our framework.

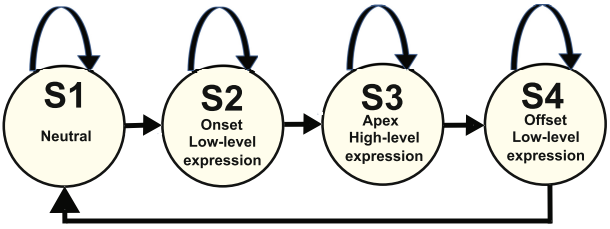

Fig. 2. Illustration of the 4 states structure of the left-right HMM model, respectively, the neutral, onset, apex and offset, applied to learn the sequential variations of the open curves shape

3.3 Testing Stage

Given a test sample of a subject with an arbitrary expression, the N open curves are extracted and the respective sequences are collected by defining $O_{test}^{UNK} = [\, s_{test}(q_1), \ldots, s_{test}(q_N)]$ (i.e., the sequence of the N open curves of the test subject having an unknown expression). The face expression recognition is computed first by adopting a maximum likelihood approach on each open curve sequence. In particular, for each open curve sequence $s(q_i)$, the following score is computed:

$$ml_i = \max_{EXP} logP(s_{test}(q_i)/\lambda_i^{EXP}) \tag{4}$$

Therefore, each curve sequence votes for a particular expression, and a majority criterion is used for classifying the expression of the test face.

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 [6]. The database contains videos of the six basic expressions that were captured for a total number of 101 subjects (58 female and 43 male). Each subject was asked to perform the six basic expressions, each expression was captured using a stereo acquisition technique, and 3D frames were produced according to a passive stereo-photogrammetry approach. Each acquired video lasts for almost 4 seconds, at a rate of 25 frames per second, resulting in an average number of 100 frames per video sequence. For each video sequence, we conduct the preprocessing pipeline as described previously, then we proceed with the curve-based representation Fig. 3. For each frame, we construct 80 contour curves and 10 profile curves, the profile curves define the boundaries of sector areas of the face model, and are used to generate segments of contour curves for more local representation. We end up with a total number of 80 × 20 (segments)+20 (radial curves starting from the nose tip) = 1620 (open curves).

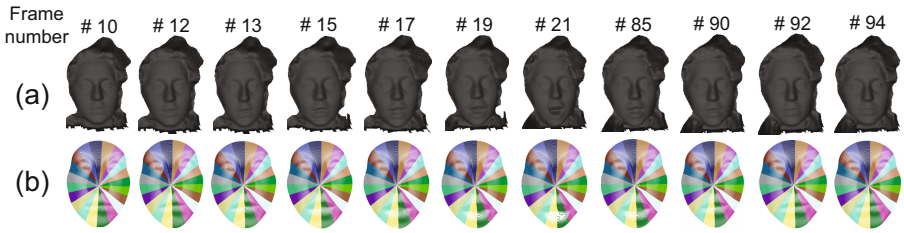

Fig. 3. Example of selected frames taken from the BU-4DFE database (female F017 surprise expression): (a) raw data showing the shape model (b) the corresponding preprocessed data showing the curve based representation considered for our study

Data of 100 subjects of the BU-4DFE database are considered to conduct facial expression recognition experiments. The subjects were partitioned into 10 sets, each containing 10 subjects, and 10-fold cross has been used for validation, where at each round 9 of the 10 folds (90 subjects) are used for training while the rest (10 subjects) are used for test. The recognition averaged on 10 rounds. The recognition results of 10 rounds are then averaged to give a statistically significant performance measure of the proposed solution.

Following the experimental protocol proposed in [1], 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

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 subsequences. Also in this experiment, we performed 10-folds cross validation, on the overall number of subsequences derived from the 100×6 sequences.

The results obtained by classifying individual 6-frames subsequences of the expression sequences 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 94.45%. It can be noted that the higher recognition rate is obtained for the surprise expression 96.57%, and the lower recognition is obtained for the angry expression 92.34% which is mainly confused with the disgust and fear expressions. Interestingly, these three expressions capture negative emotive states of the subjects, so that similar facial muscles can be activated.

Table 1. Average confusion matrix (percentage values)

	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	**92.34**	2.45	2.33	0.57	1.89	0.42
Disgust	1.82	**94.74**	1.75	0.44	0.93	0.32
Fear	1.65	1.29	**93.85**	0.79	1.97	0.45
Happy	0.56	0.84	0.68	**95.98**	0.36	1.58
Sad	1.98	1.24	2.33	0.88	**93.24**	0.33
Surprise	0.57	0.48	1.15	0.38	0.85	**96.57**

5 Conclusion

In this paper we propose to employ a curve-based representation of 3D facial surfaces for 4D facial expression recognition. A combination of contour and profile curves are extracted and partitioned to obtain local open curves. These curves are used to derive shape analysis and quantify their shape variation over time. The computed shape features are treated as a signal model and HMM is applied to learn the dynamics of facial expressions. The proposed approach is experimented on the BU-4DFE database and the obtained results are reported. Our method

shows an overall recognition accuracy as high as 94.45%. With these promising results, our future work will focus on testing the performance of our approach on real world data, such as RGB-D data, captured from low-end consumer devices. This will lead us to deal with multiple constraints, like optimizing the curve extraction process which is time consuming, and ameliorate our approach to handle challenges such as low resolution data and head pose variations.

References

1. Sun, Y., Yin, L.: Facial Expression Recognition Based on 3D Dynamic Range Model Sequences. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 58–71. Springer, Heidelberg (2008)
2. Le, V., Tang, H., Huang, T.S.: Expression recognition from 3D dynamic faces using robust spatio-temporal shape features. In: IEEE International Conference on Automatic Face Gesture Recognition and Workshops, FG (2011)
3. Srivastava, A., Klassen, E., Joshi, S.H., Jermyn, I.H.: Shape Analysis of Elastic Curves in Euclidean Spaces. IEEE Transactions on Pattern Analysis and Machine Intelligence (2010)
4. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE (1989)
5. Sandbach, G., Zafeiriou, S., Pantic, M., Rueckert, D.: Recognition of 3D facial expression dynamics. In: Image and Vision Computing (2012)
6. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3D dynamic facial expression database. In: 8th IEEE International Conference on Automatic Face Gesture Recognition, FG (2008)
7. Fang, T., Zhao, X., Shah, S.K., Kakadiaris, I.A.: 4D facial expression recognition. In: IEEE International Conference on Computer Vision Workshops (ICCV Workshops) (2011)