Eurographics Workshop on 3D Object Retrieval (2012)
M. Spagnuolo, M. Bronstein, A. Bronstein, and A. Ferreira (Editors)

Real-time Expression Recognition from Dynamic Sequences of 3D Facial Scans

Stefano Berretti, Alberto del Bimbo, and Pietro Pala

Dipartimento di Sistemi e Informatica, University of Firenze, Firenze, Italy

Abstract

In this paper, we address the problem of person-independent facial expression recognition in dynamic sequences of 3D face scans. To this end, an original approach is proposed that relies on automatically extracting a set of 3D facial points, and modeling their mutual distances along time. Training an Hidden Markov Model for every prototypical facial expression to be recognized, and combining them to form a multi-class classifier, an average recognition rate of 76.3% on the angry, happy and surprise expressions of the BU-4DFE database has been obtained. Comparison with competitor approaches on the same database shows that our solution is able to obtain effective results with the clear advantage of an implementation that fits to real-time constraints.

Categories and Subject Descriptors (according to ACM CCS): I.3.8 [Computer Graphics]: Applications— I.3.5 [Computer Graphics]: Computational Geometry and Object Modeling—Curve, surface, solid, and object representations

1. Introduction

In the last few years, 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, driver fatigue detection, etc. The first systematic studies on facial expressions date back to the late 70s with the pioneering work of Ekman [Ekm72]. 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. 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 [YWS*06], and the *Bosphorus* database [SAD*08]. Now, new challenges are also posed by the facial expression recognition in 4D, with the introduction of appropriate data sets, such as the BU-4DFE developed at *Binghamton University* [YCS*08] and the Hi4D-ADSIP [BM11] of the *University of Central Lancashire*. This trend is also inspired by the revolution of inexpensive acquisition devices such as the consumer 3D cameras [Kin10], that makes accessible 3D cameras to a large number of consumers. Hence, the quantity of 4D data is expected to grow very rapidly in the next years. However, the direct extension of traditional methods developed for 2D face recognition or expression recognition can be not effective or even possible with these new devices, so that new solutions are required. In order to motivate our approach to 4D facial expression recognition and relate it to the state of the art solutions, in the following we provide an overview of existing methods for 3D and 4D facial expression recognition.

DOI: 10.2312/3DOR/3DOR12/085-092

1.1. Related work

Most of the work on 3D facial expression recognition can be categorized as based on: *generic facial model* or *feature classification*. In the first category, a template face model is trained with some prior knowledge, such as feature points, shape and texture variations or local geometry labels. A dense correspondence between faces is usually required to build the general model, and facial landmarks are often used to this end. However, the requested precision in establishing dense correspondences, demands for manual annotation in many cases. Methods that fall in this category are those in [RKVW06, MMS08] and [GWLT09]. Approaches in the second category (i.e., methods that use feature classification), extract features from 3D scans and classify them into different expressions. Notable works in this category are either semi-automatic, in that rely on manually selected facial landmarks, like those presented in [WYWS06, SD07, TH08, MBD*11], or completely automatic as the solution proposed in [BdP*10]. Interestingly, all these works are experimented and compared on the BU-3DFE database.

There are a few works that use 4D data to perform facial expression recognition. In [SY08], a spatio-temporal expression analysis approach based on 3D dynamic geometric facial model sequences is proposed. The approach integrates a 3D facial surface descriptor and Hidden Markov Models (HMMs) to recognize facial expressions. Experiments were performed on the BU-4DFE. The main limit of this solution resides in the use of 83 manually annotated landmarks of the BU-4DFE that are not released for public use. In [SZPR11], a method that exploits 3D motion-based features between frames of 3D facial geometry sequences for dynamic facial expression recognition is proposed. An expressive sequence is modeled to contain an onset followed by an apex and an offset. Feature selection methods are applied in order to extract features for each of the onset and offset segments of the expression. These features are then used to train an HMM in order to model the full temporal dynamics of the expression. The system was tested on a subset of the BU-4DFE for the recognition of *anger*, *happiness* and *surprise*. In [LTH11], 3D facial shapes are compared using facial level curves. The pair- and segment-wise distances between the level curves comprise the spatiotemporal features for expression recognition from 3D dynamic faces. The paper further introduces universal background modeling and maximum a posteriori adaptation for HMMs, leading to a decision boundary focus classification algorithm. High recognition accuracy on the *happiness*, *sad* and *surprise* expressions of the BU-4DFE are reported. The work in [FZSK11] proposes a fully automatic 4D facial expression recognition approach with a particular emphasis on 4D data registration and dense correspondence between 3D meshes along the temporal line. The variant of the Local Binary Patterns (LBP) descriptor proposed in [ZP07], which computes LBP on three orthogonal planes, is used as face descriptor along the sequence. Results are provided on the BU-4DFE for all expressions and for the subsets of expressions used in [SZPR11] and [LTH11]. However, the need to perform 4D data registration and dense correspondence between subsequent 3D frames, makes the approach not suited for on-line processing of 4D sequences.

1.2. Method and contribution

From the analysis above, it emerges that the large part of existing works on 3D facial expression recognition rely on the presence of landmarks accurately identified on the face surface. The fact that several landmarks are not automatically detectable and the precision required for their positioning demand for manual annotation of train and test scans. This limits the applicability of many approaches and makes them difficult to be extended to the 4D case. At the same time, solutions specifically tailored for 4D are still preliminary, posing little or no attention to time constraints.

Motivated by these considerations, in this work we propose to use local descriptors to perform 4D facial expression recognition. Differently from existing 4D approaches, we exploit the local characteristics of the face by automatically extracting a small set of 3D *facial points* from depth maps of the face and computing mutual distances between them as face descriptors. Expression classification is then performed by using HMMs trained with the time variations of the extracted distance features. Experimental results show that the proposed approach is capable to achieve effective results on the BU-4DFE, also obeying to real time constraints.

In synthesis, three are the main contributions of this work:

- An original method to automatically detect an effective set of 3D facial points from depth images of the face;
- A simple and efficient modeling of the face based on mutual distances between 3D facial points;
- An effective and efficient temporal modeling of the 3D dynamic sequences based on HMM.

Finally, to the best of our knowledge, this is the first fully automatic approach for 4D facial expression recognition performing in real-time.

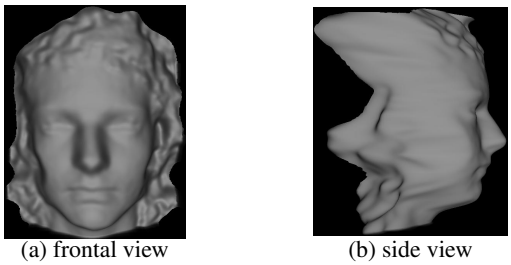
The rest of the paper is organized as follows: In Sect. 2, the main characteristics of the BU-4DFE and the preprocessing operations performed on the facial scans are described. In Sect. 3, a face representation model is proposed that captures facial features relevant to categorize expression variations in 3D dynamic sequences. In Sect. 4, the HMM based classification of the selected features is addressed. Experimental results and comparative evaluation are reported and discussed in Sect. 5. Finally, conclusions and future research directions are outlined in Sect. 6.

2. The BU-4DFE database

To investigate the usability and performance of 3D dynamic facial sequences for facial expression recognition, a dynamic

3D facial expression database has been recently created at *Binghamton University* [YCS*08]. The 3D scans have been constructed by capturing a sequence of stereo images and producing a depth map for each frame according to a passive stereo-photogrammetry approach. Each subject was requested to perform the six prototypic expressions (i.e., *angry*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*) separately. Each expression sequence contains neutral expressions in the beginning and the end, so that each expression was performed gradually from neutral appearance, low intensity, high intensity, and back to low intensity and neutral. Each 3D sequence captures one expression at a rate of 25 frames per second and each 3D sequence lasts approximately 4 seconds with about 35,000 vertices per scan (i.e., 3D *frame*). The database consists of 101 subjects (58 female and 43 male, with an age range of 18-45 years old) including 606 3D model sequences with 6 prototypic expressions and a variety of ethnic/racial ancestries (i.e., 28 Asian, 8 African-American, 3 Hispanic/Latino, and 62 Caucasian). More details on the BU-4DFE can be found in [YCS*08].

From a preliminary analysis, we note that the resolution of the individual scans of 3D sequences is not very high. In fact, the average number of vertices per scan is reasonable (about 35,000), but the number of vertices used to represent the face region is considerably lower due to the large outliers acquired in the hair and shoulder regions (see Fig. 1). The lack of facial geometric details makes the 3D sequences quite challenging to be used for facial expression recognition and face recognition. It can be also observed that the 3D frames present a near frontal pose with some slight changes occurring mainly in the azimuthal plane. This motivated us to perform expression recognition without requiring accurate pose normalization which is typically a time consuming operation. Based on these considerations, the preprocessing of the 3D frames is reduced to face cropping based on nose tip (see next Section), median filtering in the z-coordinate, holes filling using cubic interpolation and re-sampling on an uniform square grid at 0.7*mm* resolution.



(a) frontal view (b) side view

Figure 1: *BU-4DFE: A raw 3D frame before preprocessing: (a) Frontal view; (b) Side view.*

An example of a 3D dynamic facial sequence of a subject with "happy" expression is shown in Fig. 2, where 2D frames (not used in our solution), 3D frames and the depth maps extracted from the 3D frames after preprocessing are reported. For each row, five frames are given (out of the 98 total frames of the sequence). In particular, for each column in the figure: Frames in (a) and (e) represent the first and last frame of the sequence, respectively; Frames in (b) and (c) provide a sample of the extent of the facial expression in the *onset* and *offset* intervals of the sequence, respectively; Frames in (c) are taken from the interval of the sequence with the *apex* intensity of the expression.

3. Description of 3D dynamic sequences

Facial expressions are a dynamic process induced by spatiotemporal variations of facial muscles. The dynamic 3D face data provides both 3D geometric and motion information of such variations. In our approach, spatial variations due to expression changes in relevant regions of the face are captured by the facial distances computed between detected facial points. Instead, the temporal dynamic of the facial distances is learned by HMMs. This ensures a low computational cost of the approach and fits real-time constraints in the processing and classification of 3D dynamic face sequences. In particular, individual 3D frames are processed in three steps: *(i)* Identification of the tip and width of the nose; *(ii)* Detection of *facial points* in the mouth and eyes regions; *(iii)* Computation of distances between facial points.

Tip and width of the nose - The point with maximum gray value in the central region of the depth map of the face is used as initial estimate of the nose tip, and its position is then refined using the surface curvature [GMB10]. The *Gaussian* curvature (K) and the *mean* curvature (H) of the depth map are computed from the first and second derivatives of the surface. To reduce the effect of surface noise on second derivatives, the surface is smoothed with a Gaussian filter and approximated using a biquadratic polynomial. Following the observation that the region surrounding the nose tip is *convex* ($H < 0$), and has high *elliptic* Gaussian curvature ($K > 0$), the nose tip is determined in the convex part of the central region of the face as the point with a local maximum of the elliptic Gaussian curvature which is closest to the initial estimate of the nose tip. On the nose tip is then centered a sphere used to crop the 3D points than lie within a radius of 90*mm*.

The search of the two points that define the nose width is performed in a window of 50*mm* width and 42*mm* height centered on the nose tip [GMB10]. In this region, the edges of the depth image are identified using a *Laplacian of Gaussian* (LoG) edge detector with $\sigma = 3$ pixels. The edges of the left and right part boundaries of the nose are detected by traversing outwards horizontally in both directions from the tip of the nose and by retaining the first edges encountered. In order to compute the boundary curvature, the contour is coded counter-clockwise according to the *Freeman chain code*. Then, a *derivative of Gaussian* (doG) filter is applied to the chain code in order to smooth and derive it.

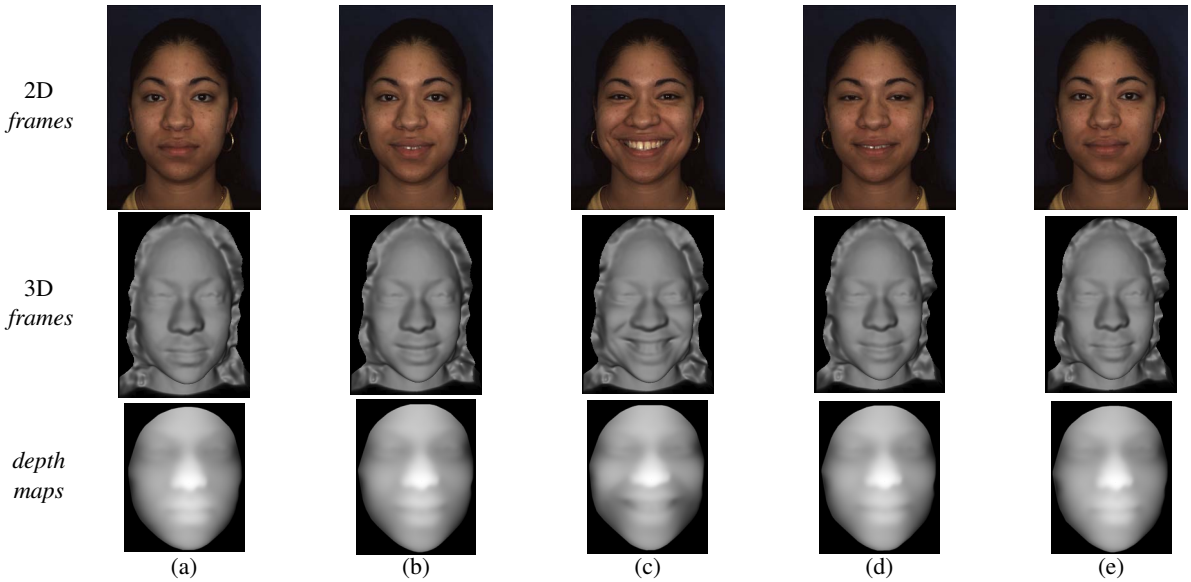


Figure 2: *BU-4DFE: 2D and 3D frames, and depth maps extracted from the 3D frames, of the dynamic sequence of subject F021 ("happy" expression). For each row, five frames are reported (out of 98 total frames): (a) frame #0, first frame of the sequence (neutral expression); (b) frame #18, start of the expression; (c) frame #47, apex of the expression; (d) frame #82, end of the expression; (e) frame #97, last frame of the sequence (neutral expression).*

Finally, the "critical" points are identified in correspondence to the local minima of the derivative that correspond to maxima of the curvature (i.e., the points along the nasal boundary with high negative curvature). The outer-left and outer-right critical points are selected as the points that determine the nose width (indicated as *leftnose* and *rightnose*).

Facial points in the mouth and eyes regions - The face regions that change maximally with facial expressions are those of the mouth and of the eyes. Following the studies of Farkas [FM87], the detection of these regions proceeds using the location of the nose tip and of the nose width points.

The vertical limits of the mouth are determined by detecting the upper and lower lip as the regions with elliptic Gaussian curvature below the tip of the nose. The points in these regions at the horizontal coordinate of the nose tip are used as the upper and lower points of the mouth. The nose width $nw_x = |leftnose_x - rightnose_x|$ is used to constrain horizontally the mouth region, being the left and right bounds, respectively, $leftnose_x + 0.7 \times nw_x$, and $rightnose_x - 0.7 \times nw_x$. The horizontal coordinate of the nose tip is also used as left bound for the left mouth region, and as right bound for the right mouth region. In these two regions, the SIFT *detector* [Low04] is run to identify the outer mouth points. In fact, SIFT works on 2D gray-scale images and the keypoints detected are mainly located at corner points of the image. When applied to depth maps, SIFT scale space extrema coincide with local depth variations that are preserved through multiple levels of resampling and smoothing. As result, detected keypoints are located in points that morphologically characterize the 3D shape. However, since several keypoints can be detected in each region, only the most stable keypoint (i.e., that detected at the largest scale) is retained as outer mouth point in the left and right mouth regions. The horizontal limits of the mouth, together with the vertical ones, are then used to sample two more points in the upper and lower lips in both the left and right regions of the mouth (uniform spacing and curvature information are used to locate these points). In summary, the mouth is sampled with 5 points in the upper lip and 5 points in the lower lip, and with the two outer mouth points, as shown in Fig. 3.

The eyes region is upper bounded by the eyebrows and lower bounded by the cheekbones. The movement of eyebrows and the eyelids convey significant information on the facial expressions. However, the eyes region with the eyelids is typically acquired with noise in 3D, so that it is difficult to extract effective features from it. Due to this, we analyze the eyes region and characterize its variations with facial expressions by using the movement of the eyebrows. To this end, the upper limit of the eyebrows and the cheekbones (i.e., the lower bound of the eyes cavities) are detected by computing the elliptic Gaussian curvature. As an example, Fig. 3 shows the points that limit the eyes region in two 3D frames.

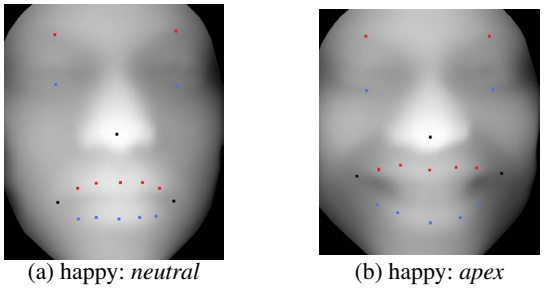
Distances between facial points - Using the detected facial points, the face in each 3D frame is represented by computing distances between points, as follows:

- *Mouth region* - The five distances between corresponding facial points detected in the upper and lower lips are used to model the vertical changes of the mouth due to different

expressions. The average value between these distances is used (feature f_0^t, *mouth*). In addition, the distance between the outer points of the mouth is used to capture horizontal variations of the mouth width (feature f_1^t, *mouth width*).

- *Nose region* - The nose tip is a stable point of the face across different expressions. The distance between this point and the upper point of the mouth is considered to relate absolute variations of the mouth with respect to other parts of the face (feature f_2^t, *mouth-nose*).
- *Eyes region* - The distances between the lower and upper points of the left and right eyes regions are computed and averaged (feature f_3^t, *eyebrows*).

For each feature f_k^t, the apex t indicates the dependence of the feature from the current 3D frame ($t = 1,\ldots,T$, being T the number of frames in a 3D sequence).



(a) happy: *neutral* (b) happy: *apex*

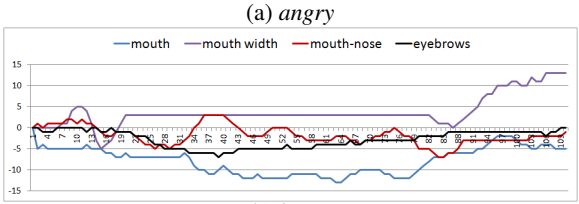
Figure 3: *Facial points detected in the depth maps of 3D frames ("happy" sequence): (a) Neutral frame; (b) Apex frame.*

Each feature f_k^t, measured for a particular subject, captures an absolute distance that depends from the face size as well as from the facial expression of the subject. As a consequence, these distances could be affected more by the identity of the subject than by his/her expression. In order to remove the unwanted dependency from the subject, each feature is normalized with respect to the distance between the inner eye points which is assumed to be an intrinsic feature of the subject and is invariant with respect to expression changes [BBDd11]. In addition, to remove the effect induced on expressions by the specific subject, each feature is managed in a differential form, by subtracting from its value that assumed in the initial frame of the sequence. Finally, the magnitude of features f_k^t is quantized into $2L+1$ discrete intervals. This is obtained by a stair-step function with a quantization step Δ:

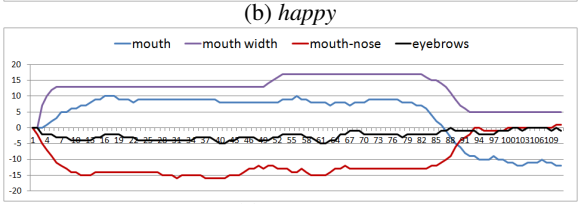
$$\tilde{f}_k^t = \begin{cases} L & if \quad f_k^t \geq L\Delta \\ i & if \quad f_k^t \in [i\Delta,(i+1)\Delta) \\ -i & if \quad f_k^t \in [-i\Delta,-(i+1)\Delta) \\ -L & if \quad f_k^t \leq -L\Delta. \end{cases}$$

Normalized and quantized distances $\tilde{f}_k^t$ are computed for each 3D frame of a dynamic facial sequence, thus constituting the *intra*-frame representation of the face. According to this, the facial expression recognition process relies on the analysis of the temporal behavior of these distances. Fig. 4 shows the temporal dynamic of distances f_k^t (i.e., before quantization) for 3D sequences of a same subject ("angry", "happiness" and "surprise" expression sequences are reported in (a), (b) and (c), respectively). For example, it can be observed that the "happy" expression is characterized by a convex shape of the *mouth width* curve and by a concave shape of the *mouth-nose* curve. Differently, the signatures for the "surprise" expression are given by the convex shape of the *mouth* and *mouth-nose* curves. For the "angry" expression, the concave shapes of the *mouth* and *eyebrows* curves are the most characterizing signatures. As a common behavior across different expressions, the features that most characterize each expression change from an initial reference value passing through an activation interval (*onset*), constituted by the sequence of frames in which the face changes from a neutral state to an expressive one, followed by an *apex* interval, where the expression reaches its maximum, with the sequence closed by a deactivation interval (*offset*), where the expressive face vanishes to a neutral one. This evidences the capability of the extracted features to capture the expression changes in the experimented 3D sequences that evolve through neutral appearance, low intensity, high intensity, and back to low intensity and neutral again (see Sect. 2).

(a) *angry*



(b) *happy*



(c) *surprise*

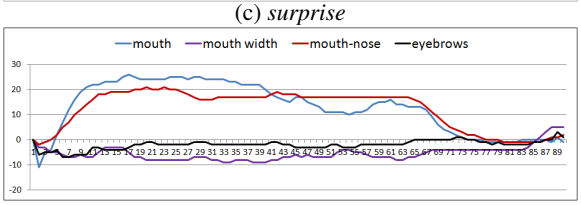


Figure 4: *Temporal dynamics of the distances f_k^t of a same subject for three sequences of the BU-4DFE: (a) Angry; (b) Happiness; (c) Surprise.*

In view of HMM training and classification, each feature $\tilde{f}_k^t$ is seen as an observation O_t of the hidden process represented by the specific facial expression to be modeled as detailed in the next Section.

4. Expression classification based on HMMs

Let $\lambda = \{A, B, \pi\}$ denote an HMM to be trained and N be the number of hidden states in the model. We indicate the states as $S = \{S_1, S_2, ..., S_N\}$, and the state at instant time t is q_t. The state transition probability distribution is indicated as $A = \{a_{ij}\}$, where $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$, with $1 \leq i, j \leq N$. In a discrete domain, each state of the model can emit a set of observation symbols, taken from a discrete alphabet corresponding to the physical output of the system being modeled. The individual symbols are indicated as $V = \{v_1, v_2, \ldots, v_M\}$, being M the number of distinct observation symbols. In our case, the set V includes all the possible instances that can be assumed by the discretized features $\hat{f}_k^t$ generated by the face representation of Sect. 3. Given an observation v_k, $B = \{b_j(k)\} = P(v_k \text{ at } t | q_t = S_j)$ is the observation probability distribution in state j, that is the probability that the observation k being produced from state j, independent of t. Finally, with $\pi = \{\pi_i\}$ is denoted the initial probability array, being $\pi_i = P(q_1 = S_i)$.

In our case, sequences of 3D frames constitute the temporal dynamics to be classified. Each prototypical expression is modeled by four HMMs, one for each feature $\hat{f}_k^t$ (a total of 24 HMMs is required, λ_k^{expr}, with $k = 0, 1, 2, 3$, and $expr \in \{an, di, fe, ha, sa, su\}$). 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. 5 exemplifies the structure of the HMMs in our framework.

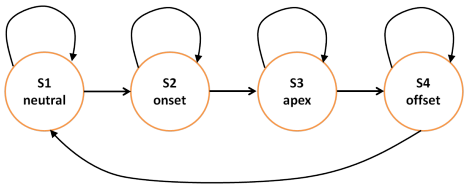


Figure 5: *Structure of the HMMs modeling a 3D facial sequence. The four states model, respectively, the neutral, onset, apex and offset frames of the sequence. As shown, from each state it is possible to remain in the state itself or move to the next one (left-right HMM).*

The training procedure of each HMM is summarized as follows:

- Observation sequences $O = \{O_1, O_2, \cdots, O_T\}$, are derived from the 3D expression sequences, where each O_t denotes an observation at time t expressed by the feature $\hat{f}_k^t$;
- The HMM λ is initialized with random values. The *Baum-Welch* algorithm [Rab89] is used to perform *unsupervised* learning from a set of training sequences, thus estimating the model parameter $\lambda = \{A, B, \pi\}$ when $P(O|\lambda)$ is maximized.

Given a 3D sequence to be classified, it is processed as in Sect. 3, so that each feature $\hat{f}_k^t$ corresponds to a *query* observation $O_k = \{O_k^1 \equiv \hat{f}_k^1, \ldots, O_k^T \equiv \hat{f}_k^T\}$. Then, the query observation O_k is presented to the six HMMs λ_k^{expr} that model the feature k for different expressions, and the *Viterbi* algorithm is used to determine the best *path* $\bar{Q}_k = \{\bar{q}_k^1, \ldots, \bar{q}_k^T\}$, which corresponds to the state sequence giving a maximum of likelihood to the observation sequence O_k. The likelihood along the best path, $p_k^{expr}(O_k, \bar{Q}_k | \lambda_k^{expr}) = \bar{p}_k^{expr}(O_k | \lambda_k^{expr})$, is considered as a good approximation of the true likelihood given by the more expensive *forward* procedure [Rab89], where all the possible paths are considered instead of the best one. This procedure is applied to the four features $\hat{f}_k^t$ of a sequence and the likelihoods of the corresponding HMMs are combined according to a weighted product rule. Finally, the sequence is classified as belonging to the class corresponding to the four HMMs whose combined log-likelihood along the best paths is the greatest one:

$$c = \arg\max_{expr} \prod_{k=0}^{3} \alpha_k \cdot \log \bar{p}_k^{expr},$$

where α_k is a parameter that weights the likelihood of the HMM classifying the individual feature $\hat{f}_k^t$, and c is the resulting class.

5. Experimental results

In the following, we present our preliminary facial expression recognition results obtained on the BU-4DFE database. Similarly to the approach in [SZPR11], our solution can effectively discriminate between *angry*, *happiness* and *surprise* expressions. So, with respect to the general formulation of the previous section, in the proposed experiments $expr$ takes value in the set $\{an, ha, su\}$. Data of 80 subjects have been considered in the experiments (in contrast to 60 subjects used in [SY08], [SZPR11] and [LTH11]), using the remaining 21 subjects for a preliminary tuning of the proposed algorithms. The 80 subjects are randomly partitioned into 10 sets, each containing 8 subjects, and 10-fold cross validation has been used for test, where at each round 9 of the 10 folds (72 subjects) are used for training while the remaining (8 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.

Results are reported in the confusion matrix of Tab. 1. Rows of the table are the *true* expressions to classify, whereas columns represent the results of the classification. It can be observed that the best classified expression is *angry* with a recognition accuracy of about 84%, whereas there is a greater confusion for the expressions of *happiness* and *surprise*. The average recognition accuracy is equal to 76.3%.

	Angry	*Happy*	*Surprise*
Angry	**83.75**	11.25	5
Happy	17.5	**75**	7.5
Surprise	13.75	16.25	**70**

Table 1: *Average confusion matrix (percentage values).*

We also note that the proposed feature extraction and classification algorithms can process 3D sequences at a rate of around 10 frames per second (on an Intel Centrino Duo at 2.2GHz with 2GB memory), thus permitting a real-time analysis.

5.1. Discussion and comparative evaluation

To the best of our knowledge, the only four works reporting results on expression recognition from dynamic sequences of 3D scans are those in [SY08], [SZPR11], [LTH11] and [FZSK11]. 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 [SY08] is not completely automatic and also presents an 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 pre-defined 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 [SZPR11] 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 ICP alignment of the 3D mesh of each frame with respect to a reference frame and 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: *anger*, *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 [LTH11], 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.

The most recent method in the literature is that proposed in [FZSK11]. The approach is fully automatic, and based on 4D registration and dense correspondences between subsequent frames of 3D facial sequences. In particular, two techniques are proposed and compared to identify 3D correspondences, namely, *spin-images* [JH99] computed around *Harris* corner points of 3D meshes, and *MeshHOG* descriptors computed in correspondence to *MeshDOG* detected points [ZBVH09]. Matching of these descriptors between frames, with RANSAC filtering, ensures points correspondence and permits rigid registration using *Procrustes* analysis. Then, the initial frame of the sequence is fitted to a 3D deformable face model (AFM) using ICP, and the AFM is deformed on the first frame minimizing an energy function whereas for the subsequent frames the deformation on the previous one is used to initialize the AFM. Finally, LBP on three orthogonal planes is used to describe the AFM sequence. An expression recognition rate of 74.63% on the six expressions of the BU-4DFE is reported (average on 507 sequences from 100 subjects). However, due to the very high computational cost required for feature extraction, registration and fitting, it is evident that the approach is only suited for off-line processing of the sequences.

With respect to these solutions, our approach has a lower accuracy (using more sequences), but is positively distinguished by the capability to run in real-time, without requiring manual intervention in both the training and testing phase (no manual annotation of facial points or supervised

learning are required). In addition, our solution has been tested on a larger subset of the BU-4DFE database.

6. Conclusions

In this paper, we have presented a fully automatic approach for facial expression recognition from 3D dynamic sequences (3D + time) of facial scans. The approach is targeted to have a low computational cost (to the best of our knowledge, it is the only one that can work in real time on 4D sequences) so that it can be used in applicative scenarios where time constraints are relevant. To this end, the modeling of the face content is obtained by automatically detecting in 3D a set of facial points and measuring distances between them. The temporal dynamics of the distances between facial points is then used as input to a set of HMMs capable to classify the dominant expression appearing in a temporal sequence.

As future work, we plan to extend our face representation approach in order to extract facial features capable to discriminate also between *disgust*, *fear* and *sad* expressions.

References

[BBDd11] BERRETTI S., BEN AMOR B., DAOUDI M., DEL BIMBO A.: 3d facial expression recognition using sift descriptors of automatically detected keypoints. *The Visual Computer 27*, 11 (Nov. 2011), 1021–1036. 5

[BdP*10] BERRETTI S., DEL BIMBO A., PALA P., BEN AMOR B., DAOUDI M.: Selected sift features for 3d facial expression recognition. In *20th Int. Conf. on Pattern Recognition* (Istanbul, Turkey, Aug. 2010), pp. 4125–4128. 2

[BM11] B.J. MATUSZEWSKI W. QUAN L.-K. S.: High-resolution comprehensive 3-d dynamic database for facial articulation analysis. In *Proc. IEEE International Conference on Computer Vision Workshops* (Barcelona, Spain, Nov. 2011), pp. 2128–2135. 1

[Ekm72] EKMAN P.: Universals and cultural differences in facial expressions of emotion. In *Proc. Nebraska Symposium on Motivation* (Lincoln, NE, 1972), vol. 19, pp. 207–283. 1

[FM87] FARKAS L. G., MUNRO I. R.: *Anthropometric Facial Proportions in Medicine*. Thomas Books, Springfield, IL, 1987. 4

[FZSK11] FANG T., ZHAO X., SHAH S., KAKADIARIS I.: 4d facial expression recognition. In *Proc. IEEE International Conference on Computer Vision Workshop* (Barcelona, Spain, Nov. 2011), pp. 1594–1601. 2, 7

[GMB10] GUPTA S., MARKEY M. K., BOVIK A. C.: Anthropometric 3D face recognition. *International Journal of Computer Vision 90*, 3 (Dec. 2010), 331–349. 3

[GWLT09] GONG B., WANG Y., LIU J., TANG X.: Automatic facial expression recognition on a single 3D face by exploring shape deformation. In *Proc. ACM Int. Conf. on Multimedia* (Beijing, China, Oct. 2009), pp. 569–572. 2

[JH99] JOHNSON A., HEBERT M.: Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence 21*, 5 (May 1999), 433–449. 7

[Kin10] KINECT: http://www.xbox.com. 1

[Low04] LOWE D.: Distinctive image features from scale-invariant key points. *International Journal of Computer Vision 60*, 2 (Nov. 2004), 91–110. 4

[LTH11] LE V., TANG H., HUANG T. S.: 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. 2, 6, 7

[MBD*11] MAALEJ A., BEN AMOR B., DAOUDI M., SRIVASTAVA A., BERRETTI S.: Shape analysis of local facial patches for 3d facial expression recognition. *Pattern Recognition 44*, 8 (Aug. 2011), 1581–1589. 2

[MMS08] MPIPERIS I., MALASSIOTIS S., STRINTZIS M. G.: Bilinear models for 3-D face and facial expression recognition. *IEEE Transactions on Information Forensics and Security 3*, 3 (Sept. 2008), 498–511. 2

[Rab89] RABINER L.: A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of IEEE 77*, 2 (Feb. 1989), 257–286. 6

[RKVW06] RAMANATHAN S., KASSIM A., VENKATESH Y. V., WAH W. S.: Human facial expression recognition using a 3D morphable model. In *Proc. IEEE Int. Conf. on Image Processing* (Atlanta, GA, Oct. 2006), pp. 661–664. 2

[SAD*08] SAVRAN A., ALYÜZ N., DIBEKLIOĞLU H., ÇELIKTUTAN O., GÖ B., SANKUR B., AKARUN L.: Bosphorus database for 3D face analysis. In *Proc. First COST 2101 Workshop on Biometrics and Identity Management* (May 2008). 1

[SD07] SOYEL H., DEMIREL H.: Facial expression recognition using 3D facial feature distances. In *Proc. Int. Conf. on Image Analysis and Recognition* (Aug. 2007), pp. 831–838. 2

[SY08] SUN Y., YIN L.: Facial expression recognition based on 3D dynamic range model sequences. In *Proc. Eur. Conf. on Computer Vision* (Marseille, France, Oct. 2008), pp. 58–71. 2, 6, 7

[SZPR11] SANDBACH G., ZAFEIRIOU S., PANTIC M., RUECKERT D.: 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. 2, 6, 7

[TH08] TANG H., HUANG T. S.: 3D facial expression recognition based on automatically selected features. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition* (Anchorage, AK, June 2008), pp. 1–8. 2

[WYWS06] WANG J., YIN L., WEI X., SUN Y.: 3D facial expression recognition based on primitive surface feature distribution. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition* (June 2006), vol. 2, pp. 1399–1406. 2

[YCS*08] YIN L., CHEN X., SUN Y., WORM T., REALE M.: 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. 1, 3

[YWS*06] YIN L., WEI X., SUN Y., WANG J., ROSATO M.: 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. 1

[ZBVH09] ZAHARESCU A., BOYER E., VARANASI K., HORAUD R.: Surface feature detection and description with applications to mesh matching. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition* (Miami Beach, FL, June 2009), pp. 373–380. 7

[ZP07] ZHAO G., PIETIKÄINEN M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 6 (June 2007), 915–928. 2