Eurographics Workshop on 3D Object Retrieval (2014)
B. Bustos, H. Tabia, J.-P. Vandeborre, and R. Veltkamp (Editors)

GeoTopo: Dynamic 3D Facial Expression Retrieval Using Topological and Geometric Information†

A. Danelakis [1] , T. Theoharis [1,2] and I. Pratikakis [3]

[1]Department of Informatics & Telecommunications, University of Athens, Greece
[2]Department of Computer & Information Science, Norwegian University of Science and Technology, Norway
[3]Department of Electrical & Computer Engineering, Democritus University of Thrace, GR-67100, Xanthi, Greece

Abstract

Recently, a lot of research has been dedicated to address the problem of facial expression recognition in dynamic sequences of $3D$ face scans. On the contrary, no research has been conducted on facial expression retrieval using dynamic $3D$ face scans. This paper illustrates the first results on the area of dynamic $3D$ facial expression retrieval. To this end, a novel descriptor is created, namely ***GeoTopo****, capturing the topological as well as the geometric information of the $3D$ face scans along time. Experiments have been implemented using the angry, happy and surprise expressions of the publicly available dataset $BU-4DFE$. The obtained retrieval results are very promising. Furthermore, a methodology which exploits the retrieval results, in order to achieve unsupervised dynamic $3D$ facial expression recognition, is presented. The aforementioned unsupervised methodology achieves classification accuracy comparable to the supervised dynamic $3D$ facial expression recognition state-of-the-art techniques.*

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 H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Retrieval models

1. Introduction

Facial expressions are generated by facial muscle movements, resulting in temporary deformation of the face. In recent years, automatic analysis of facial expressions has emerged as an active research area due to its various applications such as human-computer interaction, human behavior understanding, biometrics, emotion recognition, computer graphics, driver fatigue detection, and psychology. Ekman [EF78] was the first to systematically study human facial expressions. His study categorizes the prototypical facial expressions, apart from neutral expression, into six classes representing anger, disgust, fear, happiness, sadness and surprise. This categorization is consistent across different ethnicities and cultures. Furthermore, each of the six aforementioned expressions is mapped to specific movements of facial muscles, called Action Units (AUs). This led to the Facial Action Coding System ($FACS$), where facial changes are described in terms of AUs.

The recent availability of $4D$ data‡ has increased research interest in the field. The first dataset that consists of $4D$ facial data was $BU-4DFE$, presented by Yin *et al.* [YCS*08]. $BU-4DFE$ was created at the University of New York at Binghamton and was made available in 2006. It involves 101 subjects (58 females and 43 males) of various ethnicities. For each subject the six basic expressions were recorded. The $Hi4D-ADSIP$ dataset was presented by Matuszewski *et al.* in [MQS*12]. The dataset was created at University of Central Lancashire and is not available yet. It contains 80 subjects (48 females and 32 males) of various age and ethnic

† This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALES-3DOR (MIS 379516). Investing in knowledge Society through the European Social Fund.

‡ $4D$ will refer to $3D$ + time (dynamic $3D$); each element of such a sequence is a $3D$ frame.

DOI: 10.2312/3dor.20141043

origins. Each subject was recorded for seven basic expressions (anger, disgust, fear, happiness, sadness, surprise and pain). Finally, Yin *et al.* [ZYC*13] presented the *EAGER* dataset in 2013 to the research community. This dataset contains high-resolution spontaneous $3D$ dynamic facial expressions. It involves 41 subjects (23 females and 18 males) of various ethnicities. Each of the aforementioned datasets are accompanied by a number of facial landmarks marked on each $3D$ frame. Table 1 illustrates the publicly available $4D$ facial expression datasets.

A lot of research has been dedicated to address the problem of facial expression recognition in dynamic sequences of $3D$ face scans. On the contrary, to the best of our knowledge, no research on facial expression retrieval using dynamic $3D$ face scans appears in the bibliography. This paper illustrates the first results on the area of $4D$ facial expression retrieval. To this end, a novel descriptor is created, namely **GeoTopo**, capturing the topological, as well as, the geometric information of the $3D$ face scans along time. Experiments have been implemented using the angry, happy and surprise expressions of the publicly available dataset $BU-4DFE$. The obtained retrieval results are very promising. Furthermore, a methodology which exploits the retrieval results, in order to achieve unsupervised $4D$ facial expression recognition, is presented. The aforementioned methodology achieves classification accuracy comparable to the supervised $4D$ facial expression recognition state-of-the-art techniques.

The remainder of the paper is organized as follows. In Section 2, previous works on the field of $4D$ facial expression recognition are reviewed. In Section 3, the new **GeoTopo** descriptor is explicitely described and the proposed retrieval methodology is illustrated. In Section 4, the experimental results of the proposed methodology are presented and discussed. Finally, conclusions are drawn in Section 5.

2. Related Work

Due to the lack of previous work in $4D$ facial expression retrieval, the current section deals with recognition; however we concentrate on the descriptors and the $4D$ representation used, which are also related to the retrieval process. $4D$ video facial expression recognition methodologies will be reviewed and categorized based on the dynamic face analysis approach that they use. Dynamic face analysis enables robust detection of facial changes. Dynamic face analysis approaches can be divided into four categories: temporal tracking of facial landmarks, temporal tracking of facial critical points, mapping $3D$ facial scans onto a generic $3D$ face model and, finally, analyzing different facial surfaces in order to detect temporal facial changes.

2.1. Landmark Tracking-based Methods

Landmark tracking-based techniques aim to track areas around facial landmarks along $3D$ frames. Then, they detect temporal changes on geometry characteristics of the areas using appropriate features.

In [CVTV05], a $2D$ tracker was employed and the facial model's projection was warped by 22 tracked feature points. The depth of a vertex was recovered by minimizing the distance between the model and the range data. Lipschitz embedding embeds the normalized deformation of the model in a low dimensional generalized manifold. For classification, a probabilistic expression model was learned on the generalized manifold. In [RCY08], the composition of the descriptor and the classifier are the same as in [CVTV05] but in [RCY08] the $2D$ face texture is generated using a conformal mapping and model adaptation algorithm. The proposed coarse to-fine model adaptation approach between the planar representations was used and the correspondences are extrapolated back to the $3D$ meshes. A Linear Discriminant Analysis (LDA) classifier is implemented for the classification process. In [SCRY10], another version of [RCY08] is presented. Instead of a LDA classifier, a spatio-temporal Hidden Markov Model (HMM) is implemented. The HMM incorporates $3D$ surface feature characterization to learn the spatial and temporal information of faces. In [SRY08], an Active Appearance Model (AAM) was implemented in order for 83 key landmark vertices to be tracked through the $3D$ sequence. Radial basis functions are used to adapt the generic model to the range facial model. Each adapted vertex is assigned one of eight possible primitive surface labels, by exploiting its principal curvature. Thus, a range model is represented by a label map composed of all vertices' labels in the facial region. LDA is used to project the range model to an optimal feature space. For classification, a HMM classifier is used. The method presented in [SRY08] was taken a step further in [SY08], where radial basis functions are used, after positioning of the landmark vertices, in order to adapt the generic model to the range facial model. This method is more focused on facial expression recognition and less on facial AUs recognition. In [TM09] an Active Shape Model (ASM) is built in order for 81 $3D$ facial landmarks to be selected. The ASM is then fitted onto the data using the gradient information in the neighborhood of each landmark. The feature vectors combine geometric information of the landmarks and the statistics on the density of edges and curvature around the landmarks according to the $FACS$. In [TM10], an improved version of [TM09] is presented. This version is more focused on facial expression rather than facial action units recognition. It implements more classification rules achieving better classification accuracy than [TM09]. Finally, in [CSZY12], $3D$ landmark tracking is applied and the tracked landmarks are used for curvature-based feature extraction. For classification, a Support Vector Machine SVM classifier is exploited.

DATASET	YEAR	SIZE	CONTENT	LANDMARKS
$BU-4DFE$ [YCS*08]	2008	101 subjects	6 basic expressions	83 facial points
$Hi4D-ADSIP$ [MQS*12]	2012	80 subjects	7 basic expressions	84 facial points
$EAGER$ [ZYC*13]	2013	41 subjects	27 AUs	83 facial points

Table 1: *Publicly available* 3*D video facial expression datasets.*

2.2. Critical Point Tracking-based Methods

Critical points tracking-based techniques aim to track $3D$ model key points along $3D$ frames. Then, they detect temporal changes on spatial characteristics that are defined by these facial points and not by entire facial areas.

In [BDBP12a], automatic selection of points on the nose, eyes and mouth using z-buffers takes place. A face in a $3D$ frame is represented by computing and averaging distances between the detected facial points. These distances are then normalized, quantized and summed in a final descriptor. HMM is used for system training and classification. In [JLN*12] use critical points, providing a $3D$ shape for each frame, are initially estimated using Constrained Local Models (CLM) method. Then, the rigid transformation is removed from the $3D$ shape acquired and it is projected to $2D$. Procrustes normalization is applied on the $2D$ projections. For the classification task, the differences between the features of the actual shape and the features of the first (neutral) frame, were used for further normalization before SVM-based multi-class classification takes place.

2.3. 3D Facial Model-based Methods

Facial deformation-based techniques aim to generate descriptors based on the facial temporal deformations which occur due to facial expressions.

In [YWLB06], a tracking 3D model for estimating motion trajectories, which are used to construct a spatio-temporal descriptor called facial expression label map ($FELM$), is proposed. The tracking model is first aligned to the $3D$ face scan, and then deformed to fit the target scan by minimizing an energy function. The $FELM$ vector and the motion vector are concatenated to form the descriptor, which becomes the input to a LDA classifier. In [SZPR11], free form deformations are used in order to find a vector field reflecting facial motion. Next, $2D$ feature extraction takes place for every frame. All derived features are concatenated into one feature vector per frame in the image sequences, and these are used for classification. For classification, a HMM is used. In [SZPR12], a similar approach is adopted. This approach focuses on the facial regions which present the greatest amount of motion. The classification process in enriched by using GentleBoost (GB) classifiers in addition to HMM. In [FZSK11], a mesh matching procedure, based on facial vertex correspondence, is applied. Procrustes analysis is used to determine the correspondence transformation. To construct the final descriptor, the pixels of an image are labeled by thresholding each pixel's neighborhood with the center value. The results are translated into binary numbers, which codify local patterns of different types and are accumulated in a histogram over a predefined region. Temporal evolution is also considered. This histogram essentially becomes the descriptor of the region and the whole image can be described by a concatenation of such histograms. In [FZO*12], an enriched version of [FZSK11] is proposed. This version improves the face registration procedure. In [ZRY13], a new $4D$ spatio-temporal Nebula feature is proposed. Given a spatio-temporal volume, the data is voxelized and fit to a cubic polynomial. A label is assigned based on the principal curvature values, and the polar angles of the direction of least curvature are computed. The labels and angles for each feature are used to build a histogram for each region of the face. The concatenated histograms from each region construct the final feature vector. For the classification procedure the LDA classifier is implemented.

2.4. Facial Surface-based Methods

Facial surface-based techniques extract facial surfaces on different face depth levels. The final descriptor is generated by estimating the intersection along time between the face and each surface.

In [LTH11], facial level curves on the Z axis are created, at different heights h. Every facial point at height h belongs to the corresponding curve. Comparison between same level curves leads to a distance vector (descriptor) for each frame. The descriptors corresponding to individual frames are combined to create an augmented vector. Principal Component Analysis (PCA) and LDA are used to decrease the dimensionality of the descriptor and a HMM is employed for classification. In [DBAD*12], a new Deformation Vector Field (DVF) descriptor is proposed. The facial surfaces are represented by a set of parameterized radial curves emanating from the tip of the nose, which defines the novel descriptor. Then, a LDA-based transformation is used for dimensionality reduction. Finally, the Multiclass Random Forest (MRF) learning algorithm is exploited for the classification process.

3. Methodology

As discussed in Section 2, the large part of existing works on $4D$ facial expression analysis rely on facial landmarks/critical points, accurately identified on the face surface, in order to build the corresponding descriptors. The detection of these landmarks/critical points should be performed automatically, so that the resulting descriptor can also be automatically applicable, potentially in real-time.

The $3D$ model-based dynamic face analysis approaches have a major disadvantage. They cannot operate reliably when pose variation is presented along the dynamic $3D$ sequence of the expression. Because of this, the majority of the dynamic face analysis approaches are based on the detection of $3D$ landmarks/critical points along time frames. Facial expressions are closely linked to the positions of key-points of the face at given times. These approaches achieve acceptable classification accuracies.

Furthermore, the development of the *FACS* [EF78] gives a promising prospect for any future approaches. This system, which was introduced by psychologists to describe the various facial movements in terms of *AU*s (see Figure 1), has not yet received the attention it deserves in the field of $4D$ facial expression analysis.

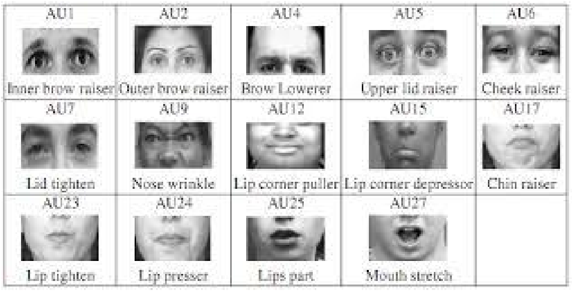

Figure 1: *The basic AUs as illustrated in Ekman's work.*

The aformetioned reasoning led to the creation of the **GeoTopo** descriptor. This spatio-temporal descriptor captures and combines facial geometric (based on curvatures) and topological (based on *FACS AU*s) information. It is a based on both landmark and critical point-tracking face analysis. In our work we will use the more general term "landmarks" to refer to both landmarks and critical points. To this end, eight facial landmarks, tracked on the $3D$ facial scans, are exploited (see Figure 2). More specifically, four landmarks for the eyes, two for the mouth, one for the nose and one for the chin are used. The focus of our work is on the descriptor creation rather than the tracking process. That is why, we have used the landmarks provided by $BU-4DFE$ dataset which were determined using the active appearance model technique [YCS*08]. The number of landmarks used here is less than the number that is usually utilized by the state-of-the-art techniques.

3.1. The GeoTopo Descriptor

The proposed descriptor captures geometric, as well as, topological information, which is achieved by the concatenation of two separate sub-descriptors, one expressing the facial geometry and one the facial topology.

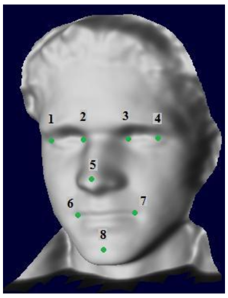

Figure 2: *8 facial landmarks used for the creation of **GeoTopo** descriptor.*

The geometric part of the **GeoTopo** descriptor is a simple $2D$ function (G), as illustrated in equation 1. Function G represents the maximum curvature of the j-th landmark (L_j) in the i-th $3D$ frame ($frame_i$).

$$G(i,j) = MaximumCurvature(frame_i, L_j) \qquad (1)$$

The topological sub-descriptor is also a $2D$ function (T), as illustrated in equation 2. Function T represents the value of the j-th feature, related to one or more AUs, in the i-th $3D$ frame. Ten features are selected in total. One of them is angular, four are areas and five express distances on the face. The calculations of the values of these ten features are performed using exclusively the $3D$ coordinates of the eight tracked landmarks (LMs) on each $3D$ time frame.

$$T(i,j) = \begin{cases} Angle_{i,j}(LMs) & : j = 1 \\ Area_{i,j}(LMs) & : j \in \{2,\dots,5\} \\ Distance_{i,j}(LMs) & : j \in \{6,\dots,10\} \end{cases} \qquad (2)$$

Each facial expression can be deconstructed into specific AUs, as illustrated in Table 2. There is a correspondence between each facial muscle and a number of AUs. The actual type of the AU is determined by the muscle temporal movement. Each of the ten selected features is directly related to one or more AUs of *FACS*, as illustrated in Table 3. $MEAN$ stands for the mean of two $3D$ points X, Y: $MEAN(X,Y) = \frac{X+Y}{2}$. The features have been selected in such a manner as to express the temporal motion of the AUs of the eyes, mouth and cheek. Moreover, according to the experimental results, these facial features are sufficient to distinguish the three expressions. In order to calculate the angle Ang, formed by three $3D$ points X, Y, Z, the following formula is used:

$$Ang = \arctan(|(D_1 \times D_2)| - (D_1 \cdot D_2))$$

where $D_1 = X - Y$, $D_2 = Y - Z$ and arctan, $|\,|$, $\times$ and $\cdot$ stand for the arctangent, 2^{nd} order norm, cross product and dot product respectively. For the calculation of the area formed by three $3D$ points, Heron's formula is used. Finally, for the calculation of facial distances, the euclidean distance is used. Figures 3, 4 and 5 illustrate the mapping of the selected ten features on a $3D$ face scan.

FACIAL EXPRESSION	ACTION UNITS
Angry	$AU4 + AU7 + AU23$
Disgust	$AU9 + AU14 + AU15$
Fear	$AU1 + AU5 + AU20 + AU25$
Happy	$AU6 + AU12$
Sad	$AU1 + AU15 + AU17$
Surprise	$AU1 + AU5 + AU26$

Table 2: *Facial expressions deconstruction into AUs.*

AU DESCRIPTION	FEATURE CODE	FEATURE TYPE	FEATURE VALUE
$AU1$: Inner Brow Raiser	#1	Angle	$\widehat{L_2, MEAN(L_2,L_3), L_5}$
$AU4$: Brow Lowerer	#1	Angle	$\widehat{L_2, MEAN(L_2,L_3), L_5}$
$AU5$: Lid Raiser	#2	Area	$\overbrace{L_1,L_2,L_5}^{AREA}$ or $\overbrace{L_3,L_4,L_5}^{AREA}$
$AU6$: Cheek Raiser	#3	Area	$\overbrace{L_1,L_5,L_6}^{AREA}$ or $\overbrace{L_4,L_5,L_7}^{AREA}$
$AU7$: Lid Tightener	#2	Area	$\overbrace{L_1,L_2,L_5}^{AREA}$ or $\overbrace{L_3,L_4,L_5}^{AREA}$
$AU9$: Wrinkler	#6	Distance	$\overline{MEAN(L_2,L_3),L_5}$
$AU12$: Lip Corner Puller	#7	Distance	$\overline{L_1,L_6}$ or $\overline{L_4,L_7}$
$AU14$: Dimpler	#4	Area	$\overbrace{L_6,L_7,L_5}^{AREA}$
$AU15$: Lip Corner Depressor	#7	Distance	$\overline{L_1,L_6}$ or $\overline{L_4,L_7}$
$AU17$: Chin Raiser	#5	Area	$\overbrace{L_6,L_7,L_8}^{AREA}$
$AU20$: Lip Strecher	#8	Distance	$\overline{L_6,L_7}$
$AU23$: Lip Tightener	#8	Distance	$\overline{L_6,L_7}$
$AU25$: Lips Part	#9	Distance	$\overline{L_5,L_8}$
$AU26$: Jaw Drop	#9	Distance	$\overline{L_5,L_8}$
Normalization Distance	#10	Distance	$\overline{L_1,L_8}$ or $\overline{L_4,L_8}$

Table 3: *Connecting AUs with mathematical features for **GeoTopo** descriptor.*

The concatenation of the aforementioned sub-descriptors, as illustrated in equations 1 and 2, produces the final **GeoTopo** descriptor.

3.2. Comparison between GeoTopo Descriptors

For the comparison between **GeoTopo** descriptors corresponding to different $4D$ data (query vs database descriptors), the Dynamic Time Warping (DTW) [SC07] algorithm was implemented. DTW is extremely efficient as a time-series similarity measure which minimizes the effects of shifting and distortion in time by allowing "elastic" transformation of time series in order to detect similar shapes with different phases. Given two time series $X = (x_1, x_2, \ldots, x_N)$ and $Y = (y_1, y_2, \ldots, y_M)$, N and M are positive integers, represented by the sequences of values DTW yields optimal solution in $O(M \cdot N)$ time. The closer to zero a returned DTW comparison value is, the more similar the two compared descriptors are, and thus, the more similar the two facial expressions. The retrieval results, using **GeoTopo** descriptor, are very encouraging and are presented in the following section.

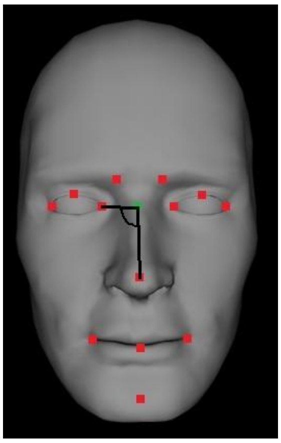

Figure 3: *Angle feature used for expressing $AU1$, $AU2$, $AU4$.*

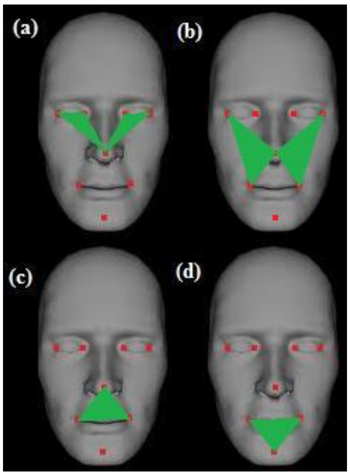

Figure 4: *Area features used for expressing (a) $AU5$ and $AU7$, (b) $AU6$, (c) $AU14$, (d) $AU17$.*

4. Experimental Results

The dataset we used to conduct our experimets is $BU-4DFE$. It was presented by Yin *et al.* [YCS*08] and was the first dataset consisting of faces recorded in $3D$ video.

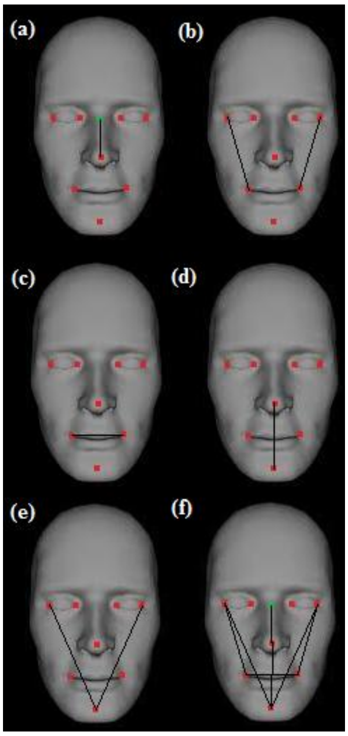

Figure 5: *Distance features used for expressing (a) $AU9$, (b) $AU12$ and $AU15$ (c) $AU23$ and $AU24$, (d) $AU27$, (e) Normalization distance, (f) overall AUs.*

It involves 101 subjects (58 females and 43 males) of various ethnicities. For each subject the six basic expressions (angry, disgust, fear, happy, sad and surprise) were recorded gradually from neutral face, outset, apex, offset and back to neutral, using the dynamic facial acquisition system *Di*3*D* (www.di3d.com) and producing roughly 60,600 3D face models (frames), with corresponding texture images. Each basic expression 3D video sequence lasts about four seconds. The temporal resolution of the 3D videos is 25 fps and each 3D model consists of approximately 35,000 vertices. Finally, each frame is associated with 83 facial landmark points. In Figure 6, examples of $BU-4DFE$ dataset are illustrated.

It should be noted that the facial data constituting the dataset are of good quality. However, inconsistencies are exhibited. Specifically, although in the database description [YCS*08], the authors state that each sequence contains an expression performed gradually from neutral appearance, low intensity, high intensity, and back to low intensity and neutral, it is not the case for some of the sequences (see Figure 7). Moreover, some videos contain corrupted meshes (see Figure 8) or they have obvious discontinuity. Finally, there are meshes that have spike shaped reconstruction artifacts around their borders. So, it is obvious that further improvement of the quality is a matter of significant importance. Berretti *et al.* [BDBP12b] presented a methodology in this direction, especially focusing on 3D static and dynamic facial data.

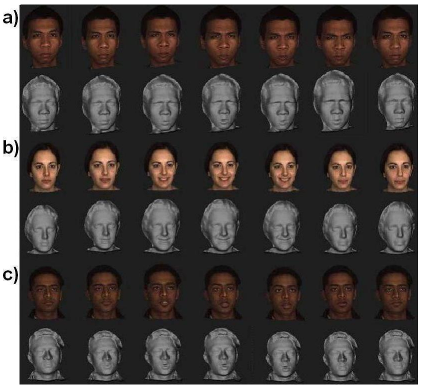

Figure 6: *Example of $BU-4DFE$ dataset including texture images and* 3*D models: (a) anger, (b) happiness, (c) surprise.*

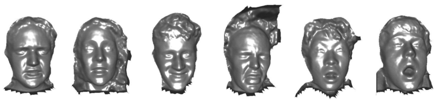

Figure 7: *Initial frames from $BU-4DFE$ dataset sequences in which the subjects do not start with a neutral expression.*

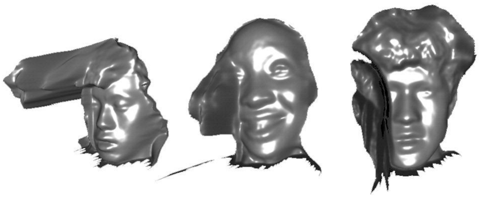

Figure 8: *Illustration of corrupted frames in the $BU-4DFE$ dataset.*

Experiments have been implemented using the angry, happy and surprise expressions of the publicly available dataset $BU-4DFE$. Only the dynamic 3D sequences were used and not the corresponding textures. It should be pointed out that, although there are dynamic 3D sequences containing serious artifacts (some subjects do not start with a neutral expression or express dual emotions and some sequences contain corrupted meshes or present obvious discontinuities), no manual corrective removals took place. Three expressions for all 101 subjects of the dataset were used. Thus, 303 dynamic 3D sequences, or over 30,300 3D frames were processed (each sequence consists of more than 100 3D frames). In all tests, the Leave-One-Subject-Out approach was employed.

Distance, angle, area and curvature values of the **GeoTopo** descriptor are weighed so that bigger weights correspond to landmarks around the mouth and eyes. The actual weights were experimentally determined and are given in Table 4. This table illustrates each feature inner weights (the weight of each angle, area, distance and landmark curvature) as well as the total weight of all angles, areas, distances and curvatures. Distance, angle and area values weigh more than curvature values, while distances outweigh all other values. In order to combine these values, L_1, L_2 and L_g fusions are used resulting in a new weighted mixed fusion.

WEIGHTS	FEATURE INNER WEIGHTS								FEATURE TOTAL WEIGHT
ANGLES	1								0.2
AREAS	0.1					0.3	0.3	0.3	0.25
DISTANCES	0.1				0.275	0.175	0.275	0.175	0.35
CURVATURES	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.2

Table 4: *Feature weights in **GeoTopo** descriptor.*

Several parameters had to be determined in order to conduct the experiments. Initially, descriptor normalization took place. Normalization sets the feature values of the **GeoTopo** descriptor in the interval $[0,1]$ and was implemented separately for angles, areas, distances and curvatures. Then a subtraction scheme was implemented; the descriptor values are not used as absolute values corresponding to the current time frame, but as differences of the current from the initial time frame. Next, the time window (T-window), which indicates the width of the neighboring (following and previous), $3D$ time frames that affect the current frame, should be defined. T-window value equal to 1 indicates that each $3D$ time frame is independent from other neighboring $3D$ time frames.

In Table 5 the retrieval evaluation metrics achieved by the **GeoTopo** descriptor, with respect to the T-window, are illustrated. In Figure 9 the precision-recall diagrams, with respect to the T-window, are presented. The best results are achieved for T-window equal to 1, but in general, the retrieval method is insensitive to T-window changes, as the results remain the same for T-window values higher than 4. The retrieval evaluation values are very promising, as they are all close to 1 and above 0.7.

T-WINDOW	NN	1st TIER	2nd TIER	DCG
1	0.88	0.74	0.9	0.89
$(\geq)$ 4	0.88	0.73	0.9	0.89

Table 5: *Retrieval evaluation for **GeoTopo** on $BU-4DFE$ (3 expressions).*

Besides retrieval, **GeoTopo** descriptor can be used in order to implement $4D$ facial expression recognition. This also allows our method to be compared against state-of-the-art methods whose performance is evaluated in terms of classification accuracy. Compared to the existing $4D$ facial expression approaches, the process illustrated here is completely unsupervised but remains comparable in terms of classification accuracy.

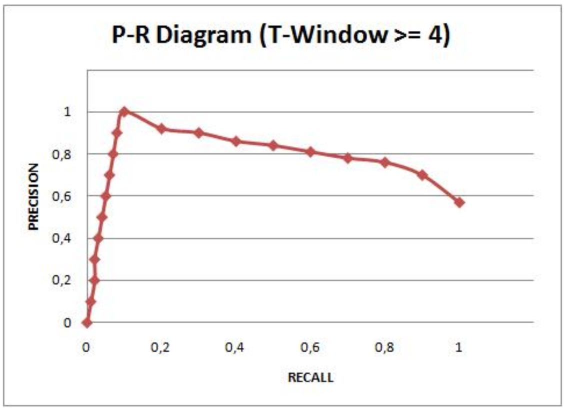

Figure 9: *Precision-Recall diagram for **GeoTopo** on $BU-4DFE$ (3 expressions).*

To achieve $4D$ facial expression recognition, by exploiting the $4D$ facial retrieval results of the **GeoTopo** descriptor, is straightforward. A k-NN classifier based on the retrieved results is used. In Table 6 the classification accuracies achieved by the **GeoTopo** descriptor, with respect to the variable k of the classifier, are outlined.

k	CLASSIFICATION ACCURACY (%)
3	96.67
5	93.33
10	93.33
15	96.67
20	96.67

Table 6: *Classification accuracies for **GeoTopo** on $BU-4DFE$ (3 expressions).*

Table 7 summarizes state-of-the-art methods on $4D$ facial expression recognition for 3 expressions from the $BU-4DFE$ dataset. It should be pointed out that Berretti *et al.* [BDBP12a] use a new automatic method for tracking their own landmarks instead of using the ones provided by $BU-4DFE$ dataset. The remaining two methods illustrated in Table 7 do not use critical points or any other landmarks to achieve expression recognition. In addition, Le *et al.* [LTH11] method (highlighted with *italic* on Table 7) use the sad instead of angry expression, for conducting their experiments. Finally, it is important to be mentioned that the classification accuracies shown at the table have been achieved after supervised recognition. Our method achieves unsupervised recognition. It can be concluded that the results of our unsupervised recognition outperform the supervised recognition results of state-of-the-art techniques.

METHOD	NUMBER OF EXPRESSIONS	CLASSIFICATION ACCURACY
Berretti *et al.* [BDBP12a]	3	76.30%
Sandbach *et al.* [SZPR11]	3	81.93%
Le et al. [LTH11]	*3*	*92.22%*
Proposed Method	**3**	**96.67%**

Table 7: *Overview of research work on dynamic* 3*D facial expression recognition for* $BU-4DFE$ *dataset.*

5. Conclusions

Dynamic $3D$ facial expression analysis constitutes a crucial open research field due to its applications in human-computer interaction, psychology, biometrics etc. In this paper, an approach for dynamic $3D$ facial expression retrieval is presented and the **GeoTopo** descriptor is proposed. **GeoTopo** captures the topological and the geometric information of $3D$ face scans along time. Experiments have been conducted on the angry, happy and surprise expressions of the publicly available dataset $BU-4DFE$. The obtained results are very promising and can be provided as ground truth for future retrieval techniques. Furthermore, a methodology which exploits the retrieval results, in order to achieve unsupervised dynamic $3D$ facial expression recognition, is presented. The aforementioned methodology achieves classification accuracy comparable to the supervised dynamic $3D$ facial expression recognition state-of-the-art techniques.

References

[BDBP12a] Berretti S., Del Bimbo A., Pala P.: Real-time expression recognition from dynamic sequences of 3D facial scans. In *EU Workshop on 3D Object Retrieval* (2012), pp. 85–92. 3, 7, 8

[BDBP12b] Berretti S., Del Bimbo A., Pala P.: Superfaces: A super-resolution model for 3D faces. In *Computer Vision – ECCV 2012. Workshops and Demonstrations*, vol. 7583. Springer Berlin Heidelberg, 2012, pp. 73–82. 6

[CSZY12] Canavan S. J., Sun Y., Zhang X., Yin L.: A dynamic curvature based approach for facial activity analysis in 3D space. In *CVPR Workshops* (2012), pp. 14–19. 2

[CVTV05] Chang Y., Vieira M. B., Turk M., Velho L.: Automatic 3D facial expression analysis in videos. In *IEEE Workshop AMFG '05* (2005), pp. 293–307. 2

[DBAD*12] Drira H., Ben Amor B., Daoudi M., Srivastava A., Berretti S.: 3D dynamic expression recognition based on a novel deformation vector field and random forest. In *ICPR '12* (2012), pp. 1104–1107. 3

[EF78] Ekman P., Friesen W.: *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press, Palo Alto, 1978. 1, 4

[FZO*12] Fang T., Zhao X., Ocegueda O., Shah S. K., Kakadiaris I. A.: 3D/4D facial expression analysis: An advanced annotated face model approach. *Image and Vision Computing 30*, 10 (2012), 738–749. 3

[FZSK11] Fang T., Zhao X., Shah S. K., Kakadiaris I. A.: 4D facial expression recognition. In *ICCV '11* (2011), pp. 1594–1601. 3

[JLN*12] Jeni L. A., Lőrincz A., Nagy T., Palotai Z., Sebők J., Szabó Z., Takács D.: 3D shape estimation in video sequences provides high precision evaluation of facial expressions. *Image and Vision Computing 30*, 10 (2012), 785 – 795. 3

[LTH11] Le V., Tang H., Huang T. S.: Expression recognition from 3D dynamic faces using robust spatio-temporal shape features. In *IEEE FG '11* (2011), pp. 414–421. 3, 7, 8

[MQS*12] Matuszewski B., Quan W., Shark L., McLoughlin A., Lightbody C., Emsley H., Watkins C.: Hi4D-ADSIP 3D dynamic facial articulation database. *Elsevier Image and Vision Computing 30*, 10 (2012), 713–727. 1, 3

[RCY08] Rosato M., Chen X., Yin L.: Automatic registration of vertex correspondences for 3D facial expression analysis. In *IEEE International Conference on Biometrics: Theory, Applications and Systems* (2008), pp. 1–7. 2

[SC07] Salvador S., Chan P.: Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal. 11*, 5 (2007), 561–580. 5

[SCRY10] Sun Y., Chen X., Rosato M. J., Yin L.: Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis. *IEEE Transactions on Systems, Man, and Cybernetics, Part A 40*, 3 (2010), 461–474. 2

[SRY08] Sun Y., Reale M., Yin L.: Recognizing partial facial action units based on 3D dynamic range data for facial expression recognition. In *FG '08* (2008), pp. 1–8. 2

[SY08] Sun Y., Yin L.: Facial expression recognition based on 3D dynamic range model sequences. In *Springer Proc. ECCV '08: Part II* (2008), pp. 58–71. 2

[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 FG '11* (2011), pp. 406–413. 3, 8

[SZPR12] Sandbach G., Zafeiriou S., Pantic M., Rueckert D.: Recognition of 3D facial expression dynamics. *Elsevier Image and Vision Computing 30*, 10 (2012), 762–773. 3

[TM09] Tsalakanidou F., Malassiotis S.: Robust facial action recognition from real-time 3D streams. In *CVPR '09* (2009), pp. 4–11. 2

[TM10] Tsalakanidou F., Malassiotis S.: Real-time 2D+3D facial action and expression recognition. *Elsevier Pattern Recognition 43*, 5 (2010), 1763–1775. 2

[YCS*08] Yin L., Chen X., Sun Y., Worm T., Reale M.: A high-resolution 3D dynamic facial expression database. In *IEEE Proc. FG '08* (2008), pp. 1–6. 1, 3, 4, 5, 6

[YWLB06] Yin L., Wei X., Longo P., Bhuvanesh A.: Analyzing facial expressions using intensity-variant 3D data for human computer interaction. In *Proc. ICPR '06* (2006), pp. 1248–1251. 3

[ZRY13] Zhang X., Reale M., Yin L.: Nebula feature: A space-time feature for posed and spontaneous 4D facial behavior analysis. In *IEEE FG '13* (2013). 3

[ZYC*13] Zhang X., Yin L., Cohn J. F., Canavan S., Reale M., Horowitz A., Liu P.: A high-resolution spontaneous 3D dynamic facial expression database. In *IEEE FG '13* (2013). 2, 3