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

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

1Department of Informatics & Telecommunications, University of Athens, Greece
2Department of Computer & Information Science, Norwegian University of Science and Technology, Norway
3Department 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.

© The Eurographics Association 2014.
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 explicitly 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.
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 is 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 [FZO12], 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 [DBAD12], 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.

<table>
<thead>
<tr>
<th>DATASET</th>
<th>YEAR</th>
<th>SIZE</th>
<th>CONTENT</th>
<th>LANDMARKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BU−ADFE</td>
<td>2008</td>
<td>101 subjects</td>
<td>6 basic expressions</td>
<td>83 facial points</td>
</tr>
<tr>
<td>HiAD−ADSIP</td>
<td>2012</td>
<td>80 subjects</td>
<td>7 basic expressions</td>
<td>84 facial points</td>
</tr>
<tr>
<td>EAGER</td>
<td>2013</td>
<td>41 subjects</td>
<td>27 AUs</td>
<td>83 facial points</td>
</tr>
</tbody>
</table>

Table 1: Publicly available 3D video facial expression datasets.
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 AUs (see Figure 1), has not yet received the attention it deserves in the field of 4D facial expression analysis.

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.

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 \((frame_i)\).

\[
G(i, j) = \text{MaximumCurvature}(frame_i, L_j) \quad (1)
\]

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.

\[
T(i, j) = \begin{cases} 
\text{Angle}_{i,j}(LMs) & : j = 1 \\
\text{Area}_{i,j}(LMs) & : j \in \{2, \ldots, 5\} \\
\text{Distance}_{i,j}(LMs) & : j \in \{6, \ldots, 10\}
\end{cases} \quad (2)
\]

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. \( \text{MEAN} \) stands for the mean of two 3D points \( X, Y \): \( \text{MEAN}(X, Y) = \frac{X + Y}{2} \). 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 \( \text{Ang} \), formed by three 3D points \( X, Y, Z \), the following formula is used:

\[
\text{Ang} = \arctan\left(\frac{|(D_1 \times D_2) - (D_1 \cdot D_2)|}{|D_1 - D_2|}ight)
\]

where \( D_1 = X - Y, D_2 = Y - Z \) and \( \arctan \), \( | \cdot | \), and \( \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \]}
for the arctangent, 2nd 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.

### Table 2: Facial expressions deconstruction into AUs.

<table>
<thead>
<tr>
<th>FACIAL EXPRESSION</th>
<th>ACTION UNITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>AU4 + AU7 + AU23</td>
</tr>
<tr>
<td>Disgust</td>
<td>AU9 + AU14 + AU15</td>
</tr>
<tr>
<td>Fear</td>
<td>AU1 + AU5 + AU20 + AU25</td>
</tr>
<tr>
<td>Happy</td>
<td>AU6 + AU12</td>
</tr>
<tr>
<td>Sad</td>
<td>AU1 + AU15 + AU17</td>
</tr>
<tr>
<td>Surprise</td>
<td>AU1 + AU5 + AU26</td>
</tr>
</tbody>
</table>

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

![Figure 3: Angle feature used for expressing AU1, AU2, AU4.](image3)

![Figure 4: Area features used for expressing (a) AU5 and AU7, (b) AU6, (c) AU14, (d) AU17.](image4)

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.

### 4. Experimental Results

The dataset we used to conduct our experiments is BU – ADFE. It was presented by Yin et al. [YCS08] and was the first dataset consisting of faces recorded in 3D video.
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 Di3D (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.

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.

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.

Figure 6: Example of BU – 4DFE dataset including texture images and 3D 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.
Distance, angle, area and curvature values of the GeoTopo descriptor are weighted 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 $L_\infty$ fusions are used resulting in a new weighted mixed fusion.

![Table 4: Feature weights in GeoTopo descriptor.](image)

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 it 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.

![Figure 9: Precision-Recall diagram for GeoTopo on BU – 4DFE (3 expressions).](image)

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.

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 retrieval 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.

![Table 6: Classification accuracies for GeoTopo on BU – 4DFE (3 expressions).](image)

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.
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