Facial Expression Retrieval Using 3-Dimensional Mesh Sequences

Danelakis E. Antonios*

National and Kapodistrian University of Athens

Department of Informatics and Telecommunications

adanelakis@di.uoa.gr

Abstract. Human emotions are often expressed by facial expressions and are generated by facial muscle movements. In recent years, 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. This dissertation introduces a new scheme for dynamic 3D facial expression retrieval. The new scheme employs novel descriptors which exploit facial mesh sequence information of automatically detected facial landmarks. A detailed evaluation of the new retrieval scheme is presented. Experiments have been conducted using the publicly available *BU−4DFE* and *BP4D−Spontaneous* datasets. The obtained results outperform the retrieval results of the state-of-the-art methodologies. Furthermore, the retrieval results are exploited in order to achieve *unsupervised* dynamic *3D* facial expression recognition. The aforementioned *unsupervised* procedure achieves better recognition accuracy compared to *supervised* dynamic *3D* facial expression recognition state-of-the-art techniques. Finally, we present a methodology for detecting primitive facial movements. The obtained results are mostly better than the state-of-the-art and more movements are detected.

1. Introduction

The process of extracting useful content information from large amounts of data, in an automated manner and based on an example or descriptive query, is called *content based information retrieval*. Common types of information that can benefit from such a retrieval process are: textual, visual, audio and video data and most recently, *3D* and *4D* (*3D* over time) data; the latter is also referred to as dynamic *3D* data or *3D* videos.

In recent years, through the creation of inexpensive *3D* scanners and the simplification of *3D* modelling software, a large volume of *3D* and *4D* data has been created. Some of the *4D* datasets that have recently been created involve human facial expressions. These datasets contain *3D* mesh sequences representing people of different ethnicities taking on a number of facial expressions. The creation of the aforementioned datasets gave rise to two new problems for the research community: The problem of *Facial Expression Recognition from 3D mesh sequences* and that of *Facial Expression Retrieval from 3D mesh sequences*. A lot of research has been dedicated to address the problem of *Facial Expression Recognition* in sequences of *3D* facial meshes. On the contrary, to the best of our knowledge, no research on *Facial Expression Retrieval* using *3D* facial mesh sequences appears in the bibliography. The present work thus addresses the latter problem.

Human emotions are often expressed by facial expressions instead of verbal communication. Facial expressions are generated by facial muscle movements, resulting in temporary deformation of the face. Ekman [1] 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 (*AU*s). This led to the Facial Action Coding System (*FACS*), where facial changes are described in terms of *AU*s.

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, engineering, human behavior understanding, biometrics, emotion recognition, computer graphics, driver fatigue detection and psychology.

1.1 Method Overview

This dissertation focuses on the problem of dynamic *3D* facial expression retrieval from large datasets. A lot of research has been dedicated to address the problem of facial expression

---

* Dissertation Advisor: Theoharis Theoharis, Professor

recognition in *4D* data. On the contrary, to the best of our knowledge, no research on facial expression retrieval in *4D* data appears in the bibliography.

In order to address this problem we develop a 3-step retrieval framework: (i) initially, eight *3D* facial landmarks are automatically detected on each *3D* facial mesh of the sequence. (ii) Next, the landmarks are used in order to create a descriptor for the dynamic *3D* facial expression sequence. (iii) Finally, distance functions are used in order for different descriptors (i.e. query descriptor vs dataset descriptor) to be compared and the retrieval list is produced. The pipeline of our scheme is illustrated in Figure 1.

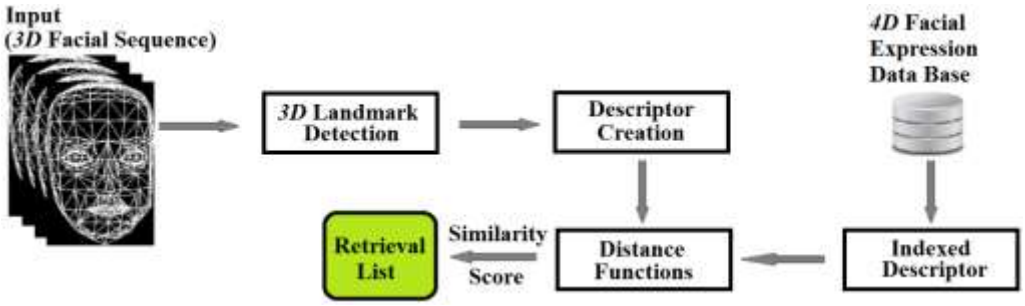

Figure 1: Pipeline of the proposed *4D* facial expression recognition scheme.

At first, eight *3D* facial landmarks are automatically detected on each *3D* facial mesh of the sequence. Each face *3D* mesh is, if not otherwise stated, defined by a set of points in the R^3 space (**vertices**) and a set of triangular **faces** defined in terms of the vertices.

The core of the problem is the computation of a feature set for each dynamic *3D* facial expression sequence. In this step, the structural and/or other special characteristics of the sequence are modelled and a descriptor that faithfully encodes the essence of the mesh sequence, in an efficient manner, is created. Feature selection is tightly connected to the corresponding application and can vary among different *4D* object retrieval systems.

Finally, each *3D* mesh sequence descriptor is used as a signature during the matching procedure. At this step, the signatures of the dynamic *3D* facial expressions, stored in the database, are compared to the corresponding signature of the query dynamic *3D* facial expression, using a specified metric, called *Distance Function*. The selected metric is dependent on both the selected features and the corresponding application. Finally, the response of the dynamic *3D* facial expression retrieval scheme is the set of dynamic *3D* facial expression(s) that correspond to the closest match (es) of the given user query.

1.2 Contributions

This thesis has made the following research contributions in the area of object retrieval: (1) Six new descriptors for the purpose of *Human Facial Expression Retrieval* from *3D* mesh sequences were proposed. For the creation of the descriptors we have used less landmarks than the state-of-the-art methods. (2) A novel mapping from facial features to primitive facial movements is proposed.

The descriptors developed and described in this dissertation are evaluated in terms of retrieval accuracy and demonstrated using both quantitative and qualitative measures via an extensive evaluation against state-of-the-art descriptors on standard datasets. This comparison illustrates the superiority of our descriptors compared to the state-of-the-art.

The overview of this thesis is as follows: In Section 2, the standard *4D* facial expression datasets are reviewed. In Section 3, state-of-the-art methods in the field of Human Facial Expression Recognition from *3D* Mesh Sequences are reviewed. In Section 4, the method for extracting specific *3D* facial landmarks from *3D* facial meshes is presented. In Section 5, the six descriptors, developed during this dissertation for the purpose of Human Facial Expression Retrieval from *3D* Mesh Sequences, are presented. In Section 6, distance functions for descriptor comparison purposes, are illustrated and compared. Section 7 presents the evaluation methodology and illustrates the extensive experimental results of the methods presented in this dissertation, against the state-of-the-art works on standard datasets. In Section 8, a supervised technique for detecting *AUs* is illustrated. Finally, in Section 9, conclusions are drawn.

2. 4D Facial Expression Datasets

The first dataset consisting of faces recorded in 3*D* video is *BU−4DFE*, presented by Yin *et al*. [2]. This dataset was made available in 2008. It involves 101 subjects (58 females and 43 males) of various ethnicities. For each subject the six basic expressions were recorded gradually from neutral face, outset, apex, offset and back to neutral, using the dynamic facial acquisition system *Di3D* (http://www.di3d.com) and producing roughly 60,600 3*D* facial meshes (frames), with corresponding texture images. Finally, each frame is associated with 83 facial landmark points.

Zhang *et al*. [3] presented the *BP4D-Spontaneous* dataset in 2013 to the research community. This dataset contains high-resolution spontaneous 3*D* dynamic facial expressions by encoding 27 *AUs* and their various combinations. It involves 41 subjects (23 females and 18 males) of various ethnicities. The subjects were 18-29 years of age. Each subject was recorded using the dynamic facial acquisition system *Di3D* (http://www.di3d.com). 328 3*D* sequences were created. Finally, each frame is associated with 83 facial landmark points. In Table 1, the basic characteristics of 3*D* video facial expression datasets are shown.

Table 1: Current publicly available datasets of *3D* facial expression mesh sequences.

Dataset	Year	Number of Subjects	Number of Expressions	Number of Landmarks
BU-4DFE	2008	101	6	83
BP4D-Spontaneous	2014	41	8	83

3. Related Work

Due to the lack of works in the area of *3D* video facial expression *retrieval* techniques, in this chapter we will review the *3D* video facial expression *recognition* state-of-the-art methodologies. We will focus on the descriptors which are the common necessities in both areas. The reader is also referred to the surveys presented in [4], [5], [6] and [7].

The typical operational pipeline employed by 3*D* video facial expression recognition methodologies is shown in Figure 2. *3D* video facial expression recognition methodologies take into account 3*D* facial data as 3*D* surfaces. Another common trait of these methodologies is the use of a variety of 3*D* dynamic face analysis techniques to detect and exploit the discrete and well studied facial muscle motions.

3D dynamic face analysis techniques can be divided into two major categories: Tracking–based and 3*D* facial model-based. Tracking-based techniques aim to track specific 3*D* facial model marks using appropriate tracking algorithms. On the other hand, 3*D* facial model-based techniques aim to exploit the facial deformations which take place due to a facial expression. These techniques often use alignment methods to achieve better results.

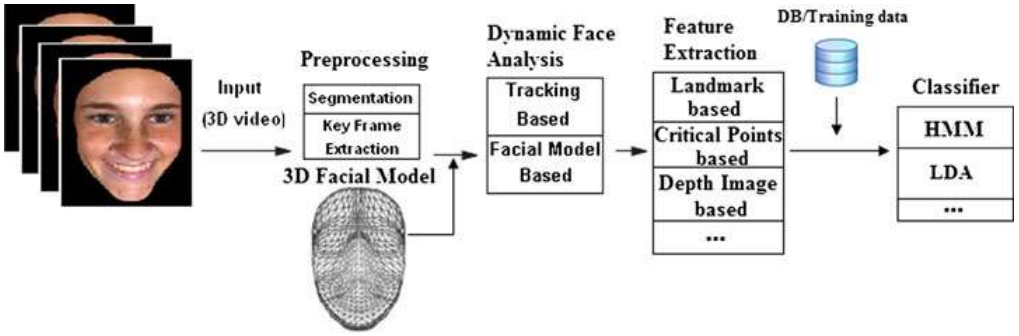

Figure 2: *3D* video facial expression recognition pipeline.

Tracking-based techniques can be further distinguished into two sub- categories: Landmark tracking-based and critical points tracking-based. In the first case, areas are tracked around specific facial landmarks along 3*D* frames and detect temporal changes on their geometry characteristics. In the latter case, techniques aim to track 3*D* model key points along time and detect temporal changes on spatial characteristics that are defined by these points.

Three dimensional facial model-based techniques can also be divided into two subcategories: Facial deformation-based, which aim to detect temporal deformations using a generic face model, and facial surface-based, which create facial surfaces on different face depth levels (i.e., different values on the *z*-axes). Then, estimate the intersection along time between the face and each surface, they extract the final descriptor. A summarization of the state-of-the-art methods is illustrated in Table 2. 'N/A' is used to indicate that the corresponding information is not available.

Table 2: Overview of research work on *3D* video facial expression recognition.

Method	*Dataset*	*Number of Expressions*	*3D Face Analysis*	*Features*	*Classifier*	*Automatic*	*Real-Time Suitability*	*Recognition Accuracy*
Chang *et al.* [8]	Proprietary	6	Landmark tracking	Generalized manifold + Texture	Bayes	NO	NO	N/A
Rosato *et al.* [9]	*BU-3DFE*	6	Landmark tracking	Generalized manifold + Texture	*LDA*	YES	NO	85.90%
Sun *et al.* [10]	*BU-4DFE*	6	Landmark tracking	Gradient + Curvature	*HMM*	YES	NO	94.37%
Tsalakanidou *et al.* [11]	Proprietary	4 (10 *AUs*)	Landmark tracking	Gradient + Curvature	*FACS*	YES	YES	84.00%
Tsalakanidou *et al.* [12]	Proprietary	4	Landmark tracking	Gradient + Curvature	*FACS*	YES	YES	85.00%
Sun *et al.* [13]	*BU-3DFE*	0 (8 *AUs*)	Landmark tracking	Curvature	*HMM*	NO	NO	87.10%
Sun *et al.* [14]	*BU-3DFE*	6	Landmark tracking	Curvature	*HMM*	NO	NO	90.44%
Canavan *et al.* [15]	*BU-4DFE*	6	Landmark tracking	Curvature	*SVM*	YES	NO	84.80%
Berretti *et al.* [16]	*BU-4DFE*	3	Critical point tracking	Average facial distances	*HMM*	YES	YES	76.30%
Jeni *et al.* [17]	*BU-4DFE*	6 (17 *AUs*)	Critical point tracking	Shape index	*SVM*	YES	NO	78.18%
Yin *et al.* [18]	*BU-3DFE*	6	Facial deformation	*FELM* + Motion vectors	*LDA*	NO	NO	80.20%
Sandbach *et al.* [19]	*BU-4DFE*	6	Facial deformation	Vector direction distribution	*GB + HMM*	YES	NO	64.46%
Sandbach *et al.* [20]	*BU-4DFE*	3	Facial deformation	Mean + *STD* of vector direction distribution	*HMM*	YES	NO	81.93%
Fang *et al.* [21]	*BU-4DFE*	6	Facial deformation	*LBP-TOP*	*SVM*	YES	NO	75.82%
Fang *et al.* [22]	*BU-4DFE*	6	Facial deformation	*LBP-TOP*	*SVM*	YES	NO	91.00%
Zhang *et al.* [23]	*BP4D-Spontaneous*	0 (27 *AUs*)	Facial deformation	Curvature + Polar angles	*SVM*	YES	NO	61.33%
Zhang *et al.* [23]	*BU-4DFE*	6	Facial deformation	Curvature + Polar angles	*SVM*	YES	NO	76.12%
Le *et al.* [24]	*BU-4DFE*	3	Facial surface	Chamfer distances	*HMM*	YES	NO	92.22%
Drira *et al.* [25]	*BU-4DFE*	6	Facial surface	*DVF*	*LDA + RMF*	YES	YES	93.21%

4. *3D* Facial Landmarks Detection

The first step of our proposed retrieval scheme is the detection of *3D* facial landmarks on the dynamic *3D* mesh sequence (see Figure 1). Eight facial landmarks, on the *3D* facial scans, are exploited. More specifically, four landmarks for the eyes, two for the mouth, one for the nose and one for the chin are used (see Figure 3). The number of landmarks used here is smaller than the number used by other state-of-the-art techniques.

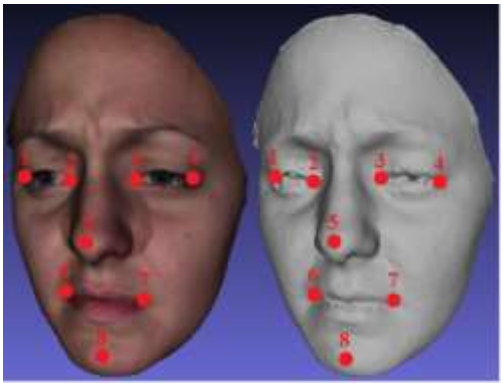

Figure 3: Eight facial landmarks used for the proposed retrieval scheme.

The landmarks are automatically detected using the state-of-the-art methodology previously developed by our team [26], making the proposed retrieval scheme self-contained.

The method presented in [26] performs automatic pose-invariant detection of landmarks on *3D* facial scans under large yaw variations, and is invariant to facial expressions. Three-dimensional information is exploited by fusing *3D* local shape descriptors to extract candidate landmark points. The shape descriptors include the shape index, a continuous map of principal curvature values of a *3D* object's surface and spin images, which are local descriptors of the object's *3D* point distribution.

Landmark detection takes part in two phases. In the training phase, a Facial Landmark Model (*FLM*) representing the landmark positions is created, shape index target values for each landmark are computed and spin image templates for each landmark are generated.

In the detection phase, the algorithm first detects candidate landmarks on the probe facial datasets, by exploiting the *3D* geometry-based information of shape index and spin images. The extracted candidate landmarks are then filtered out and labeled by matching them with the *FLM*. The facial landmark detection method is robust to rotations about the vertical facial axis up to 60 degrees and returns the detected pose; this information is used in our method in order to rotate facial instances that are not frontal.

5. Descriptors for *3D* Facial Mesh Sequences

The first two descriptors proposed in this chapter are spatial, which means that they are based only on spatial changes of the facial expressions across time. The remaining four are spatio-temporal, which means that they are based on both temporal and spatial changes of the facial expressions.

The motivation behind the proposed spatial, hybrid facial expression descriptors *GeoTopo* and *GeoTopo+* is the fact that some facial expressions, like happiness and surprise, are characterized by obvious changes in the mouth topology while others, like anger, fear and sadness, produce geometric but no significant topological changes.

The motivation behind the spatio-temporal descriptors is the expectation that the extra information offered by the temporal dimension can lead to a more accurate descriptor. In addition the descriptor can potentially be made more compact if it stores aggregations of attributes across the time dimension.

GeoTopo Descriptor (GEOmetric & TOPOlogic)

The proposed *GeoTopo* (*Geo*metric and *Topo*logical) 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 ($G_1(i,j)$), as illustrated in Equation (1). Function G_1 represents the maximum curvature of the *j*-th landmark (L_j) in the *i*-th *3D* mesh ($mesh_i$).

$$G_1\,(i,j) = MaxCurvature\,(mesh_i, L_j) \tag{1}$$

The topological sub-descriptor is also a *2D* function ($T(i, j)$), 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* mesh. 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 (*LM*s) in the *i*-th *3D* time mesh.

$$\boldsymbol{T}(\boldsymbol{i},\boldsymbol{j}) = \begin{cases} \boldsymbol{Angle}_{i,j}(\boldsymbol{LMs}) : j \in \{1\}, \\ \boldsymbol{Area}_{i,j}(\boldsymbol{LMs}) : j \in \{2,\ldots,5\}, \\ \boldsymbol{Distance}_{i,j}(\boldsymbol{LMs}) : j \in \{6,\ldots,10\} \end{cases} \tag{2}$$

Each facial expression can be deconstructed into specific *AUs*. There is a correspondence between each facial muscle and a number of *AUs*. The actual type of the *AU* is determined by the muscle's temporal movement. Figure 4 illustrates the mapping of the ten selected topological features on 3D facial mesh. The concatenation of the aforementioned sub-descriptors, as illustrated in Equations (1) and (2), produces the final *GeoTopo* descriptor: *GeoTopo* = (G_1 ++ *T*).

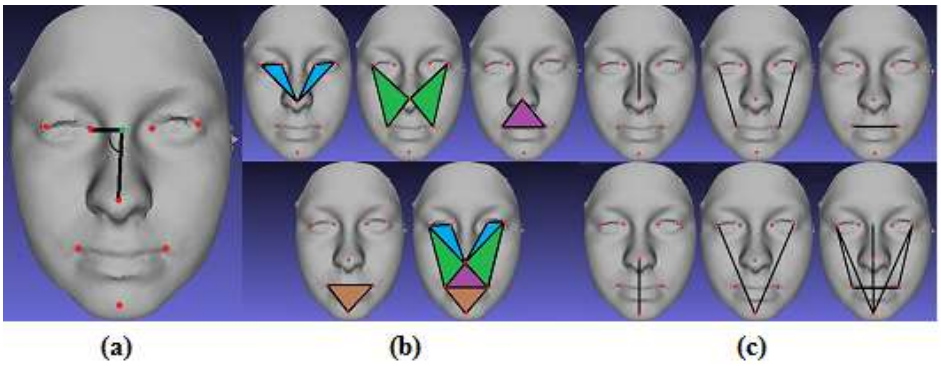

Figure 4: Topological features in use. (a) Angle, (b) Areas, (c) Distances.

GeoTopo+ Descriptor *(GEOmetric & TOPOlogic PLUS)*

GeoTopo+ is a spatial, hybrid descriptor which combines three sub-descriptors capturing topological, as well as, geometric information of the *3D* facial meshes. Two sub-descriptors are used for capturing the facial geometry, based on the heat kernel signature of the *3D* facial surface and the *3D* facial model vertices' normal vectors. The third sub-descriptor is used for capturing facial topology based on *FACS AUs*.

The first geometric sub-descriptor is a simple *2D* function ($G_2(i,j)$), as illustrated in Equation (3). Function G_2 represents the heat kernel signature (*HKS*) [27] of the *j*-th landmark (L_j) in the *i*-th *3D* mesh ($mesh_i$). *HKS* is based on the properties of the heat diffusion process on a shape. It is obtained by restricting the heat kernel in the temporal domain, thus obtaining a local effect.

$$G_2\,(i, j) = HKS\,(mesh_i, L_j) \tag{3}$$

The second geometric sub-descriptor of the *GeoTopo*+ descriptor is the *2D* function ($G_3(i,j)$) presented in Equation (4), which represents the normal vector of the *j*-th landmark (*Lj*) in the *i*-th *3D* mesh ($mesh_i$).

$$G_3\,(i, j) = NormalVector\,(mesh_i, L_j) \tag{4}$$

The topological sub-descriptor is the *2D* function $T(i,j)$ already presented in Equation (2). The concatenation of the aforementioned sub-descriptors, as illustrated in Equations (2), (3) and (4) produces the final *GeoTopo*+ descriptor: *GeoTopo*+ = (G_2 ++ G_3 ++*T*).

DCT-*GeoTopo* Descriptor *(Discrete Cosine Transformation – GeoTopo)*

DCT-GeoTopo uses only a subset of the topological sub-descriptor $T(i,j)$, illustrated in Equation (2). To produce the final descriptor, we apply the *Discrete Cosine Transformation* (*DCT*) on the temporal values of the features producing a transformed sequence for each feature. This transformation maps the features from the temporal to the frequency domain and thus the transformed features represent the spatio-temporal deformation of the initial features. Eight

features of the transformed sequences are selected to construct the final descriptor. Equation (5) represents the final descriptor; this is an *8D* vector irrespective of the number of meshes of the corresponding facial expression *3D* sequence.

$$\text{DCT-GeoTopo}=\begin{bmatrix} 2^{nd}\ DCT \text{ component for area with feature code } \#3, \\ 3^{rd}\ DCT \text{ component for area with feature code } \#3, \\ 3^{rd}\ DCT \text{ component for area with feature code } \#5, \\ 2^{nd}\ DCT \text{ component for distance with feature code } \#6, \\ 4^{\eta}\ DCT \text{ component for distance with feature code } \#6, \\ \text{M}ean\ DCT \text{ components value for distance with feature code } \#8, \\ 2^{nd}\ DCT \text{ component for additional distance of Figure 41}, \\ 2^{nd}\ DCT \text{ component for distance with feature code } \#10 \end{bmatrix} \quad (5)$$

WT-GeoTopo+* Descriptor *(Wavelet Transformation – GeoTopo+)

For the construction of the *WT-GeoTopo+*, the extracted landmarks are used in order to capture the geometric and topological information for each mesh the same way as in *GeoTopo+* descriptor. Then, we perform Wavelet Transformation [28], using *Gaussian Wavelets* at 64 different scales, on each temporal information sequence. We apply the *1D* Wavelet Transformation on all temporal sequences for every feature, as indicated in Equations (6) – (8). The term *WT*() indicates the Wavelet Transformation of the *1D* signal placed within the parenthesis, and the term * indicates the values of the landmarks over the whole set of *3D* meshes of the sequence.

$$W_{G1}(*, j) = WT(G_1(*, j)) \quad (6)$$
$$W_{G2}(*, j) = WT(G_2(*, j)) \quad (7)$$
$$W_T(*, j) = WT(T(*, j)) \quad (8)$$

For each transformed sequence we extract four feature aggregators: the *median*, the *mode* (the most frequently occurring element), the *norm* and the *root mean square* (*RMS*) of the sequence. Thus, we get a *4D* feature vector for each sequence as indicated by Equations (9) – (11). The number of *4D* feature vectors is constant (26 *4D* feature vectors), independent of the number of the frames of the sequence and constitutes the spatio-temporal information of the *3D* facial expression sequence.

$$ST_{G_1} = \{Mean(W_{G_1}(*, j)), Mode(W_{G_1}(*, j)), Norm(W_{G_1}(*, j)), RMS(W_{G_1}(*, j))\} \quad (9)$$
$$ST_{G_2} = \{Mean(W_{G_2}(*, j)), Mode(W_{G_2}(*, j)), Norm(W_{G_2}(*, j)), RMS(W_{G_2}(*, j))\} \quad (10)$$
$$ST_T = \{Mean(W_T(*, j)), Mode(W_T(*, j)), Norm(W_T(*, j)), RMS(W_T(*, j))\} \quad (11)$$

The concatenation of the above sub-descriptors produces the final descriptor: *WT-GeoTopo+* = *(ST_{G1} ++ ST_{G2} ++ ST_T)*.

CVD* Descriptor *(Coordinate Vector Descriptor)

CVD descriptor uses only six of the extracted landmarks (1st, 4th, 5th, 6th, 7th and 8th of Figure 3) and captures their positional information for each frame and, therefore, for the entire dynamic *3D* mesh sequence. The positional information, actually corresponds to the coordinate values of each one of the six facial landmarks.

As a pre-processing step, for each facial mesh, we perform translation so that the nose tip (3rd landmark) coincides with the center of the coordinate system. Thus, even though the initial data are registered, we create even better consistency between the *3D* meshes of the sequence. Next, the Z-coordinates of the 1st, 4th, 6th, 7th and 8th landmark and the Y-coordinates of the 6th, 7th and 8th landmark are kept. The above procedure produces, for each *3D* frame, a feature vector of length eight. Normalization sets the feature values of each vector in the interval [0, 1] and has been performed for each vector item separately. Finally, the set of all the normalized feature vectors of a dynamic *3D* facial expression sequence are averaged, resulting in a single *8D* feature vector.

The function for calculating the *8D* coordinate vector for each *3D* frame *f*, FC_f , is given in Equation (12) and the function for calculating the averaged *8D* coordinate vector, *CVD* , of an entire *3D* sequence is given in Equation (13). *N* is the number of *3D* frames in the sequence.

By capturing specific landmarks' successive coordinate values, even the slightest facial motions are detected. Furthermore, in the proposed descriptors presented earlier in this dissertation, the motion of each landmark is attached to one (in case distance feature is used)

or two (in case angle or area feature is used) other landmarks. On the contrary, the *CVD* coordinate vector captures each landmark's behavior independently of all the other landmarks. The experimental results show that positional features are more descriptive than relational features. However, the results can be further improved using the Wavelet Transformation of the following section.

$$FC_f = \begin{Bmatrix} Z \text{ Coordinate of the } 1^{st} \text{ landmark.} \\ Z \text{ Coordinate of the } 4^{th} \text{ landmark.} \\ Z \text{ Coordinate of the } 6^{th} \text{ landmark.} \\ Z \text{ Coordinate of the } 7^{th} \text{ landmark.} \\ Z \text{ Coordinate of the } 8^{th} \text{ landmark.} \\ Y \text{ Coordinate of the } 6^{th} \text{ landmark.} \\ Y \text{ Coordinate of the } 7^{th} \text{ landmark.} \\ Y \text{ Coordinate of the } 8^{th} \text{ landmark.} \end{Bmatrix} \quad (12)$$

$$CVD = \frac{\sum_{f=1}^{f=N} FC_f}{N} \quad (13)$$

WT-CVD$_b$* Descriptor *(Wavelet Transformation – CVD)

The positional information of the *CVD* descriptor can be used as a basis for the creation of a hybrid spatio-temporal descriptor. To this effect, we perform Wavelet Transformation, using *Gaussian Wavelets* at 64 different scales, on the coordinate information. We apply the *1D* Wavelet Transformation on the *8D* coordinate vector, as indicated in Equation (13). The term *WT*()$_b$ indicates the Wavelet Transformation of the *1D* signal placed within the parenthesis for scale *b*, and the term * indicates that all the components of the vector are used for the calculation.

$$WT - CVD_b = WT(CVD(*))_b, b = 1,...,64 \quad (14)$$

WT-CVD$_b$ is a spatio-temporal descriptor of constant length. Unlike *WT-GeoTopo+* descriptor, we haven't extracted any feature aggregators for the 64 wavelet transformed sequences. If we would have used aggregators, the information for each axis would be confused.

We prefer *Wavelet Transformation* as it has better resolution than *Fourier* and *Cosine Transformation* [29]. This means that each coefficient of the transformation, which expresses both frequency and time domain information, is created in such a way as to capture as much as possible and as precise as possible frequency-time information.

6. Distance Functions

A *Distance Function* is a mathematical expression which is applied on two given time series and produces a scalar metric as output. This output is a non-negative integer number and is called *similarity score*. The similarity score is a mathematical expression of how similar the two input time series are. If the similarity score equals 0, then the two input time series are exactly the same and thus, we have maximum similarity. In general, the bigger the similarity score is, the less similar the two input time series are.

The selection of a distance function is not trivial. It is dependent on both the features that have been selected for the descriptor and the envisaged application. Apart from the initial intuition, the selection of distance function involves extensive experimentation.

The only distance function appropriate for comparing two time series of different length is the Dynamic Time Warping (*DTW*) [30]. That is why, it is suitable for comparing *GeoTopo* and *GeoTopo+* descriptors, since their length differs and is dependent on the number of the *3D* meshes of the mesh sequence corresponding to the descriptor. *DTW* minimizes the effects of shifting and distortion in time by allowing "elastic" transformation of a time series in order to detect similar shapes with different phases. Unfortunately, *DTW* is time consuming compared to other distance functions.

Experimental results show that, in the case of the *WT-GeoTopo+* and *DCT-GeoTopo* spatio-temporal descriptors the most suitable distance functions are *Square of Euclidean Distance* and *Kullback-Leibler Divergence* respectively. Still, there are cases where *DTW* produces significantly better results even for sequences of the same length; such cases are the *CVD* and *WT-CVD$_b$* descriptors.

7. Experimental Results

The experimental evaluation is based on the Precision-Recall curves (or *P-R* Diagram) and five quantitative measures: Nearest Neighbor (*NN*), First Tier (*FT*), Second Tier (*ST*), *E*-measure (*E*-m) and Discounted Cumulative Gain (*DCG*) [31] for the classes of each corresponding dataset. The datasets used for conducting the experiments are the ones presented in Section 2.

Retrieval Evaluation

Several parameters had to be determined in order to conduct the experiments. Initially, descriptor normalization took place. Normalization sets the feature values in the interval [0, 1]. Then a subtraction scheme was implemented; the descriptor values are not used as absolute values corresponding to the current time mesh, but as differences of the current from the initial time mesh.

Table 3 illustrates the retrieval evaluation of the proposed descriptors, compared to the state-of-the-art descriptors for *BU-4DFE* dataset. Table 4 illustrates the corresponding evaluation for *BP4D-Spontaneous* dataset. The proposed descriptors outperform the state-of-the-art in both cases.

Table 3: Comparison of descriptors for *BU-4DFE* dataset.

Descriptor	NN	FT	ST	DCG
WT-CVD$_b$ **[32]**	**0.82**	**0.73**	**0.95**	**0.92**
CVD [32]	0.82	0.72	0.95	0.92
WT-GeoTopo+ [33]	0.81	0.65	0.76	0.92
DCT-GeoTopo [34]	0.75	0.61	0.66	0.86
GeoTopo+ [35]	0.73	0.56	0.77	0.91
GeoTopo [36]	0.71	0.55	0.73	0.89
Berretti *et al.* [16]	0.60	0.50	0.70	0.88
Distribution Vectors	0.52	0.41	0.59	0.82
Curvature	0.47	0.40	0.60	0.82
LBP-TOP	0.43	0.37	0.54	0.80
FELM	0.42	0.38	0.56	0.80
Gradient	0.40	0.36	0.54	0.79
Shape Index	0.35	0.37	0.54	0.79

Table 4: Comparison of descriptors for *BP4D-Spontaneous* dataset.

Descriptor	NN	FT	ST	DCG
WT-CVD$_b$ **[32]**	**0.76**	**0.62**	**0.72**	**0.84**
CVD [32]	0.53	0.60	0.71	0.81
WT-GeoTopo+ [33]	0.75	0.61	0.69	0.83
DCT-GeoTopo [34]	0.70	0.58	0.59	0.78
GeoTopo+ [35]	0.67	0.55	0.72	0.83
GeoTopo [36]	0.61	0.52	0.69	0.82
Berretti [16]	0.59	0.49	0.69	0.81
Distribution Vectors	0.50	0.41	0.57	0.76
Curvature	0.39	0.34	0.47	0.71
LBP-TOP	0.39	0.34	0.47	0.71
FELM	0.63	0.35	0.48	0.77

Gradient	0.38	0.33	0.46	0.71
Shape Index	0.30	0.32	0.47	0.70

Recognition Evaluation

To achieve recognition, a majority voting scheme is implemented among the *k*-top retrieval results. The query expression is classified as belonging to the outvoting class. Experimental results showed that optimal recognition accuracy is achieved for $k = 11$.

Table 5 illustrates the recognition evaluation of the proposed descriptors, compared to the other descriptors, proposed in this thesis, as well as the state-of-the-art descriptors for *BU-4DFE* dataset. Obviously, three descriptors proposed here, outperform the state of the art descriptors by far. Three of our proposed descriptors were tested, for the first time in terms of retrieval, *in BP4D-Spontaneous* dataset and the results are illustrated in Table 6.

Table 5: Comparison of recognition accuracy for *BU-4DFE* dataset.

Method	Nr Facial Expressions	Recognition Accuracy
WT-CVD [32]	**6**	**100.0%**
CVD [32]	6	100.0%
WT-GeoTopo+ [33]	6	96.04%
Sun [10]	6	94.37%
Drira [25]	6	93.21%
Fang [22]	6	91.00%
DCT-GeoTopo [34]	6	90.83%
GeoTopo+ [35]	6	90.00%
Canavan [15]	6	84.80%
GeoTopo [36]	6	84.18%
Berretti [16]	6	79.40%
Jeni [17]	6	78.18%
Zhang [23]	6	76.12%
Fang [6]	6	75.82%
Sandbach [19]	6	64.60%

Table 6: Comparison of recognition accuracy for *BP4D-Spontaneous* dataset.

Method	Nr Facial Expressions	Recognition Accuracy
WT-CVD$_b$ [32]	**8**	**100. 0%**
CVD [32]	8	100.0%
GeoTopo+ [32]	8	88.56%

8. Action Units Detection

The detection of facial Action Units lies in the core of facial analysis and, while related to retrieval, has far broader applications such as biometrics, interaction, behavioral analysis and animation control.

The close connection between the *AU*s and the topological features of the *GeoTopo+* descriptor makes *T* appropriate for the detection of active *AUs* across a dynamic *3D* facial mesh sequence.

We train a classifier on the topological features of Equation (2) for the case of an *AU* activation/no activation. Then, we can decide whether an *AU* is activated/not activated across a facial mesh sequence. The training and the testing sets are defined using the 5-fold cross validation method. The results, compared to the state-of-the-art are illustrated in Table 7. We achieve mostly better detection and we are capable of detecting twelve more *AU*s.

Table 7: Detection of activated *AU*s (%) for *BP4D-Spontaneous* dataset.

*AU*s	Proposed Method	Method [22]	AUs	Proposed Method	Method [22]
1	65.6	58.4	15	61.2	69.0
2	51.9	64.8	16	59.1	-
4	63.3	63.1	17	75.3	65.6
5	57.9	-	18	93.2	-
6	67.4	68.8	19	85.1	-
7	77.4	58.9	20	64.6	-
9	58.1	-	22	90.7	-
10	66.9	66.4	23	61.8	61.4
11	92.0	-	24	59.4	67.6
12	64.8	59.1	27	81.4	-
13	97.8	-	28	59.6	-
14	64.4	59.1	30	96.1	-

9. Conclusions

The present dissertation proposes a robust scheme for facial expression retrieval from sequences of *3D* facial meshes. Our scheme consists of three steps: (i) Detection of landmarks for each *3D* facial mesh of the sequence, (ii) Creation of the descriptor for the sequence and (iii) Comparison of different descriptors (i.e. query descriptor vs dataset descriptors).

The descriptors developed and described in this dissertation are evaluated in terms of retrieval accuracy and demonstrated using both quantitative and qualitative measures via an extensive evaluation against state-of-the-art descriptors on well-known, publicly available datasets. This comparison illustrates the superiority of our descriptors compared to the state-of-the-art.

Furthermore, a technique which exploits the retrieval results, in order to achieve unsupervised dynamic *3D* facial expression recognition is presented. The proposed unsupervised technique exhibits improved performance against supervised state-of-the-art techniques. Finally, the features of the topological part of the *GeoTopo+* descriptor are used for supervised *AU* activation detection, from dynamic *3D* facial mesh sequences. The detection performance of the proposed technique improves on the state-of-the-art for most *AU*s while it can detect twelve more *AU*s than the state-of-the-art.

References

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

[2] L. Yin, X. Wei, Y. Sun, J. Wang and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *IEEE Proceedings on FGR '06*, 2006, pp. 211–216.

[3] X. Zhang, L. Yin, J. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu and J. Girard, "BP4D-Spontaneous: A high resolution spontaneous 3D dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692-706, 2014.

[4] A. Danelakis, T. Theoharis and I. Pratikakis, "3D mesh video retrieval: A survey," in *Proceedings on 3DTV-CON '12*, 2012, pp. 1–4.

[5] A. Danelakis, T. Theoharis and I. Pratikakis, "A Survey on Facial Expression Recognition in 3D Video Sequences," *Multimedia Tools and Applications*, vol. 74, no. 15, pp. 5577-5615, 2014.

[6] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah and I. A. Kakadiaris, "3D facial expression recognition: A perspective on promises and challenges," in *IEEE Proceedings on FG '11*, 2011, pp. 603–610.

[7] G. Sandbach, S. Zafeiriou, M. Pantic and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image and Vision Computing*, vol. 30, no. 10, pp. 683–697, 2012.

[8] Y. Chang, M. B. Vieira, M. Turk and L. Velho, "Automatic 3D facial expression analysis in videos," in *IEEE Workshop AMFG '05*, 2005, pp. 293–307.

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

[10] Y. Sun, X. Chen, M. J. Rosato and L. Yin, "Tracking vertex flow and model adaptation for threedimensional spatiotemporal face analysis," *IEEE Transactions on Systems Man and Cybernetics Part A*, vol. 40, no. 3, pp. 461–474, 2010.
[11] F. Tsalakanidou and S. Malassiotis, "Robust facial action recognition from real-time 3D streams," in *CVPR '09*, 2009, pp. 4–11.
[12] F. Tsalakanidou and S. Malassiotis, "Real-time 2D+3D facial action and expression recognition," *Pattern Recognition*, vol. 43, no. 5, pp.1763–1775, 2010.
[13] Y. Sun, M. Reale and L. Yin, "Recognizing partial facial action units based on 3D dynamic range data for facial expression recognition," in *IEEE FG '08*, 2008, pp 1–8.
[14] Y. Sun and L. Yin, "Facial expression recognition based on 3D dynamic range model sequences," in *Proceedings on ECCV '08*, 2008, pp. 58–71.
[15] S. J. Canavan, Y. Sun, X. Zhang and L. Yin, "A dynamic curvature based approach for facial activity analysis in 3D space," in *CVPR '12 Workshops*, 2012, pp. 14–19.
[16] S. Berretti, A. D. Bimbo and P. Pala, "Automatic facial expression recognition in real-time from dynamic sequences of 3D face scans," *The Visual Computer*, vol. 29, no. 12, pp. 1333-1350, 2013.
[17] L. A. Jeni, A. Lörincz, T. Nagy, Z. Palotai, J. Sebök, Z. Szabó and D. Takács, "3D shape estimation in video sequences provides high precision evaluation of facial expressions," *Image and Vision Computing*, vol. 30, no. 10, pp. 785-795, 2012.
[18] L. Yin, X. Wei, P. Longo and A. Bhuvanesh, "Analyzing facial expressions using intensity-variant 3D data for human computer interaction," in *Proceedings on ICPR '06*, 2006, pp. 1248–1251.
[19] G. Sandbach, S. Zafeiriou, M. Pantic and D. Rueckert, "Recognition of 3D facial expression dynamics," *Image and Vision Computing*, vol. 30, no. 10, pp. 762–773, 2012.
[20] G. Sandbach, S. Zafeiriou, M. Pantic and D. Rueckert, "A dynamic approach to the recognition of 3D facial expressions and their temporal models," in *IEEE FG '11*, 2011, pp. 406–413.
[21] T. Fang, X. Zhao, S. K. Shah and I. A. Kakadiaris, "4D facial expression recognition," in *ICCV '11*, 2011, pp. 1594-1601.
[22] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah and I. A. Kakadiaris, "3D/4D facial expression analysis: An advanced annotated face model approach," *Image and Vision Computing*, vol. 30, no. 10, pp.738–749, 2012.
[23] X. Zhang, M. Reale and L. Yin, "Nebula feature: a space-time feature for posed and spontaneous 4D facial behavior analysis," in *IEEE FG '13*, 2013, pp. 1-8.
[24] V. Le, H. Tang and T. S. Huang, "Expression recognition from 3D dynamic faces using robust spatiotemporal shape features," in *IEEE FG '11*, 2011, pp. 414–421.
[25] H. Drira, B. B. Amor, M. Daoudi, A. Srivastava and S. Berretti, "3D dynamic expression recognition based on a novel deformation vector field and random forest," in *ICPR '12*, 2012, pp. 1104–1107.
[26] P. Perakis, G. Passalis, T. Theoharis and I. A. Kakadiaris, "3D facial landmark detection under large yaw and expression variations," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1552–1564, 2013.
[27] J. Sun, M. Ovsjanikov and L. Guibas, "A concise and provably informative multi-scale signature based on heat diffusion," in *SGP '09 Eurographics Association*, 2009, pp. 1383-1392.
[28] I. Daubechies, *Ten lectures on wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.
[29] R. Q. Quiroga, O. W. Sakowitz, E. Basar and M. Schürmann, "Wavelet transform in the analysis of the frequency composition of evoked potentials," *Brain Research Protocols*, vol. 8, no. 1, pp. 16-24, 2001.
[30] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," Intelligent Data Analysis, vol. 11, no. 5, pp. 561–580, 2007.
[31] P. Shilane, P. Min, M. M. Kazhdan, and T. A. Funkhouser, ''The princeton shape benchmark,'' in *IEEE Computer Society*, 2004, pp. 167–178.
[32] A. Danelakis, T. Theoharis and I. Pratikakis, "A Spatio-temporal Wavelet-based Descriptor for Dynamic 3D Facial Expression Retrieval and Recognition," *Pattern Recognition*, submitted, 2015.
[33] A. Danelakis, T. Theoharis and I. Pratikakis, "A Robust Spatio-Temporal Scheme for Dynamic 3D Facial Expression Retrieval," *The Visual Computer*, pp. 1-13, 2015.
[34] A. Danelakis, T. Theoharis and I. Pratikakis, "A Spatio-Temporal Descriptor for Dynamic 3D Facial Expression Retrieval and Recognition," in *3DOR '15 Eurographics Association*, 2015, pp. 63- 70.
[35] A. Danelakis, T. Theoharis, I. Pratikakis and P. Perakis, "An Effective Methodology for Dynamic 3D Facial Expression Retrieval," *Pattern Recognition*, 2015.
[36] A. Danelakis, T. Theoharis and I. Pratikakis, "Geotopo: Dynamic 3D facial expression retrieval using topological and geometric information," in *3DOR '14 Eurographics Association*, 2014, pp. 1-8.