Noname manuscript No.
(will be inserted by the editor)

Action unit detection in $3D$ facial videos with application in facial expression retrieval and recognition

Antonios Danelakis · Theoharis Theoharis · Ioannis Pratikakis

Received: date / Accepted: date

Abstract This work introduces a new scheme for action unit detection in $3D$ facial videos. Sets of features that define action unit activation in a robust manner are proposed. These features are computed based on eight detected facial landmarks on each facial mesh that involve angles, areas and distances. Support vector machine classifiers are then trained using the features of the descriptor in order to perform action unit detection. The proposed AU detection scheme is used in a dynamic $3D$ facial expression retrieval and recognition pipeline, highlighting the most important AUs, in terms of providing facial expression information, and at the same time, resulting in better performance than the state-of-the-art methodologies.

Keywords Dynamic $3D$ mesh sequence · Action unit detection · Facial expression retrieval · Facial expression recognition.

A. Danelakis
Department of Computer & Information Science, Norwegian University of Science and Technology, 7034, Trondheim, Norway
Tel.: +47-73591447
E-mail: antonios.danelakis@ntnu.no

T. Theoharis
Department of Computer & Information Science, Norwegian University of Science and Technology, 7034, Trondheim, Norway
Tel.: +47-73591447
E-mail: theotheo@ntnu.no

I. Pratikakis
Department of Electrical & Computer Engineering, Democritus University of Thrace, 67100, Xanthi, Greece
Tel.: +30-25410-79586
E-mail: ipratika@ee.duth.gr

1 Introduction

Human emotions are often expressed by facial expressions instead of verbal communication. Facial expressions are generated by facial muscle movements, resulting in temporary deformations of the face. The detection of facial AUs lies at the core of facial analysis and has emerged as an active research area in recent years due to its broad applications in human-computer interaction, biometrics, facial expression recognition, computer graphics and psychology.

Ekman [19] was the first to systematically study human facial expressions. His study categorizes the prototypical facial expressions, apart from the 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 onto specific primitive movements of facial muscles, called *Action Units* (AUs) as indicated in Table 1. There are 44 AUs in total, but approximately 20 of them are common (see Figure 1). This led to the *Facial Action Coding System* ($FACS$), where facial changes are described in terms of AUs.

Table 1 Facial expression deconstruction into AUs.

FACIAL EXPRESSION	ACTION UNITS
Anger	{ $AU4$, $AU5$, $AU7$, $AU23$ }
Disgust	{ $AU9$, $AU15$, $AU16$ }
Fear	{ $AU1$, $AU4$, $AU5$, $AU7$, $AU20$, $AU27$ }
Happiness	{ $AU6$, $AU12$ }
Sadness	{ $AU1$, $AU15$, $AU17$ }
Surprise	{ $AU1$, $AU5$, $AU26$, $AU27$ }

There are a lot of works in the literature when it comes facial AU detection, or relative areas, using $2D$ images [76,24,8,61,46,91,2,25,15,89] or $2D$ video [34,17,82,3,44,77,31,98,38,32,1,86,75,65,30,42,41,10,62,52,66,27, 83,63,64,84,36,35,37,23,28,102,106]. In addition, fusion of multiple modalities, including video and audio data [74] and removal of speaking effects [88] are reported for further accuracy improvement in real world applications. Two very concise surveys on AU detection techniques based on $2D$ images and $2D$ videos can be found in [45,22]. Most of the works are tested on private data sets. The results of their experimental testing would unlikely hold if pose or lighting variations existed. When using $2D$ images or $2D$ image sequences the facial data are prone to illumination changes and pose variations that affect the perceived geometry and appearance of facial features. In addition, subtle skin deformations that characterize AU activations are difficult to be captured by a $2D$ camera.

In recent years, the proliferation of inexpensive $3D$ scanners and the simplification of $3D$ modeling software has resulted in a large volume of $3D$ and $4D$ data ($3D$ mesh sequences over time or $3D$ videos). Some of the $4D$ data

Fig. 1 The basic *AU*s as illustrated in Ekman's work [19].

sets that have recently been created involve human facial data. These data sets contain 3*D* videos representing people of different ethnicities taking on a number of facial expressions which are encoded in different *AU*s. These data sets can be used for retrieval and recognition purposes.

In order for the aforementioned 2*D* modality problems to be handled, 3*D* facial data can be recruited for detecting *AU*s [53, 54, 4, 60, 95, 103, 104, 58, 57, 56, 73]. With 3*D* data, lighting and head pose variations are no longer issues of concern. The works presented in [60, 59] highlights the advantages of the 3*D* over the 2*D* modality when it comes to *AU* detection.

Temporal information on $3D$ data can potentially further improve the accuracy of AU detection, but AU detection using $4D$ facial data is still a virgin area and the subject of this paper. There are only five papers, found in the literature, that deal with $4D$ facial data. In [79], AU detection is achieved using a set of mathematical rules, combined with facial anatomy heuristics on extracted facial areas. However, the usage of such heuristics restricts the number of AUs that can be handled. In addition, the authors use a proprietary data set for their experiments. In [50], AU detection is achieved using a curvature-based feature based on 83 extracted facial landmarks. However, when using a plain geometry-based feature, such as curvature, the information provided by the facial topology is not exploited. Method [14] presents a descriptor capturing the topological and geometric information of a $3D$ facial mesh sequence. Although the aforementioned descriptor was designed for facial expression retrieval purposes, its topological part is completely AU-based, which makes it possible to use it for AU detection. The fact, however, that it was originally designed for facial expression recognition restricts its flexibility in terms of AU detection. It must be pointed out that $GT+$ only uses 10 AU-based features, while the proposed method use almost twice as many. In [81], Haar features are extracted from the dynamic $3D$ facial data. Random Forests are then used to perform the final AU estimation. One drawback of this work is the necessity to accurately estimate a large number of facial landmarks that can lead to system failures. Finally, in [69], the authors apply the Active Appearance Model (AAM), on a selected set of AUs, in order to track facial features across the $3D$ model sequences. The more AUs the more time consuming the aforementioned technique gets. A Hidden Markov Model (HMM) is employed to recognize the partial AUs. The last four aforementioned methods use the publicly available data set $BP4D - Spontaneous$, for their experiments.

The majority of $4D$ facial AU detection techniques, discussed in the present Section, implement facial analysis based on facial landmark points' temporal tracking. They do so in order to take advantage of the strong connection between facial deformations and positions of facial key-points at given times.

The main problem to be solved by the techniques dealing with $4D$ data is the creation of descriptors, taking into account both spatial and temporal information, which can accurately distinguish activation of different AUs. These descriptors are computational models that are used as digital signatures of different AUs. The novelty and contribution of this paper consists of:

- A set of novel features that reflect the dynamics of facial AUs and expressions is proposed.
- The detection of a larger number of AUs than state-of-the-art techniques of the same modality ($3D$ facial sequences) but also most techniques of other modalities (only the $2D$ work presented in [46] detects more AUs). Thus, the proposed technique can serve a broader spectrum of applications.

- The usage of fewer landmarks than the state-of-the-art leads to a speed-up when it comes to descriptor construction. Thus, the proposed technique can be integrated in real-time applications.
- Higher accuracy in *AU* detection, compared to state-of-the-art techniques of the same modality, in most experiments.
- Application in dynamic $3D$ facial expression recognition and retrieval, outperforming the state-of-the-art.
- Determination of the subset of *AU*s that are the most essential information contributors, in terms of facial expressions.

The rest of the paper is organized as follows: Section 2 describes the proposed *AU* detection scheme using $3D$ facial videos. Section 3 presents the evaluation methodology and extensive experimental results of the *AU* detection scheme against state-of-the-art works on standard data sets. In Section 4, future challenges are discussed and in Section 5 conclusions are drawn.

2 AU detection scheme

The pipeline of the supervised *AU* detection scheme proposed here is illustrated in Figure 2. The input to the procedure is a $3D$ facial mesh sequence data set. These sequences are divided into the training and the testing subsets, based on a 5-fold cross validation procedure [16]. The training process is applied on the training subset of the data and the testing process on the testing subset of the data.

The training process consists of four steps:

1. Extract eight facial landmarks for each $3D$ mesh of a facial mesh sequence belonging to the training set.
2. Normalize for translation by ensuring that the nose tip is at the center of the coordinate system.
3. Create the descriptor for each mesh sequence by extracting the proposed features.
4. Train classifiers based on the descriptor features.

The testing process consists of the following steps:

1. The first three steps of the training process are repeated.
2. Decide on *AU* activations using classifiers.

If an *AU* is labeled as activated, then it is recorded on the *AU* activation list of the testing sequence, which is the output of the pipeline. Next, each step will be analyzed.

2.1 $3D$ landmark detection

According to [87], the most essential facial areas for the recognition of emotions/expressions are the eyes and the mouth, which is in agreement with

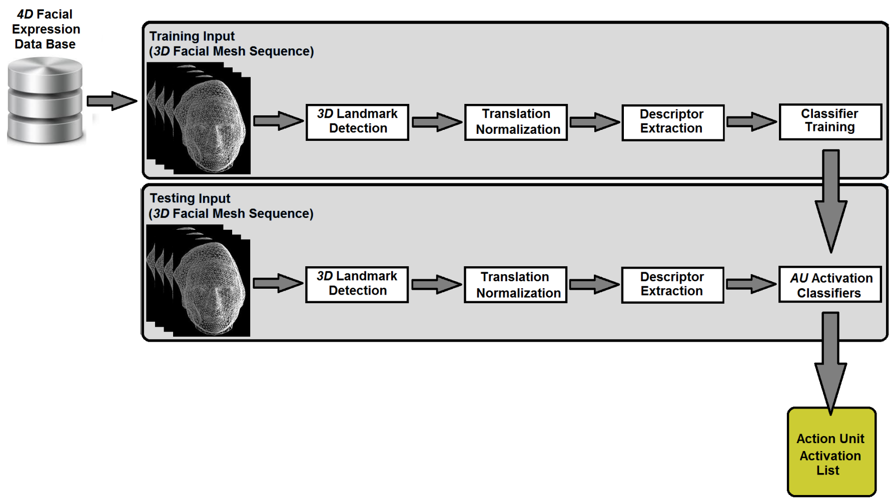

Fig. 2 Pipeline of the proposed *AU* detection scheme.

common intuition. We thus concentrate on the robust extraction of a minimal number of landmarks that define these facial areas and extract eight facial landmarks on each $3D$ facial mesh as illustrated in Figure 3.

We should note here that the number of landmarks used by our method is significantly smaller than the number used by other state-of-the-art techniques. Any $3D$ facial landmark detection technique from the bibliography can be used, as long as it accurately provides the aforementioned eight landmarks. The main concern of the present work is not landmark detection, but the construction of a robust facial movement descriptor. For experimental purposes, we have automatically detected the landmarks using the state-of-art methodology previously developed by our team [47], making the proposed scheme self-contained. This facial landmark detection method is robust to noise, rotations about the vertical facial axis of up to 60 degrees and returns the detected pose; this information is used by our method in order to rotate facial instances that are not frontal.

We are making the $3D$ facial landmarks publicly available for the data sets used throughout this study ($BP4D-Spontaneous$, dynamic $3D$ $FACS$), so that other researchers can experiment on the same basis. The landmarks can be found at `https://vc.ee.duth.gr/Face4D/BP4D-S/`.

2.2 Translation normalization using the nose tip

After the extraction of the $3D$ landmarks on each mesh, we perform translation normalization so that the nose tip (5^{th} landmark) coincides with the origin of the coordinate system (see Figure 4). Thus, we create better correspondence between the $3D$ meshes of each sequence.

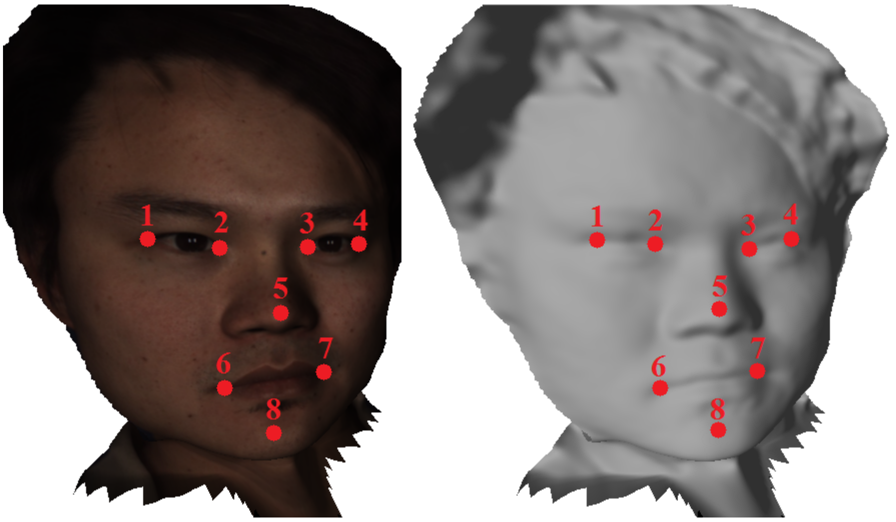

Fig. 3 Eight facial landmarks used in the proposed AU detection scheme. Here they are marked on the (a) $2D$ texture and (b) $3D$ facial mesh.

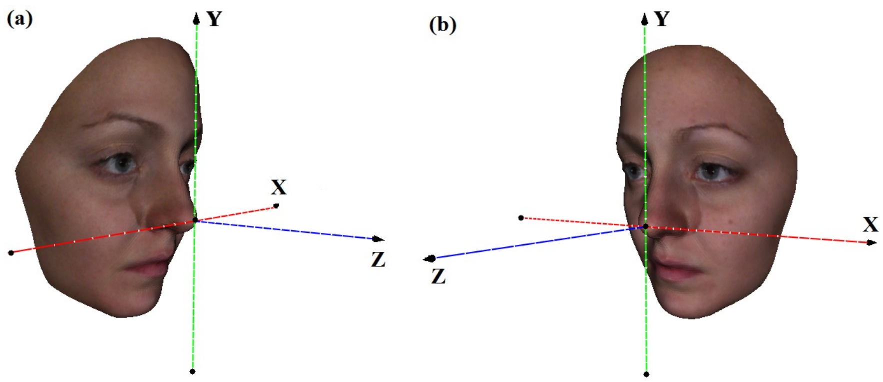

Fig. 4 Nose tip transition to the center of the coordinate system.

2.3 Descriptor extraction

The descriptor implemented within the proposed dynamic $3D$ AU detection scheme captures the topological information of a given $3D$ facial mesh sequence. The extracted eight critical facial landmarks and the temporal facial movements are exploited, as indicated by $FACS$ theory, in order for appropriate descriptor features to be selected. Each feature of the descriptor is directly mapped onto one or more AUs of $FACS$.

The descriptor is a function $T(i, j)$, see Equation 1, which represents the value of the j-th feature (related to one or more AUs) in the i-th $3D$ facial mesh of the sequence. The descriptor thus represents the dynamic behavior of AUs along a $3D$ facial expression sequence. Out of the 17 features, two are angles, eight are based on facial areas and seven are distances on the face. The calculations of the values of these features are performed exclusively using the $3D$ coordinates of the eight landmarks on each $3D$ facial mesh. The set of landmarks will be denoted as LMs, see Equations 2, 3.

$$T(i,j) = \begin{cases} Angle_{i,j}(\{LMs\}) & : j \in \{1,2\} \\ Area_{i,j}(\{LMs\}) & : j \in \{3,\dots,9\} \\ Distance_{i,j}(\{LMs\}) & : j \in \{10,\dots,16\} \end{cases} \tag{1}$$

where

$$LMs = \{P_1, P_2, \dots, P_8\} \tag{2}$$

and

$$P_i = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} \tag{3}$$

The features of the proposed descriptor are based on the eight extracted landmarks, and have been selected in such a manner as to express the temporal behavior of the AUs of the eyes, mouth, chin and jaw, as those are the most important and frequently used AUs of the human face [19]. The motivation behind the feature selection follows. Each of the 16 selected features can be thought of as a facial muscle which has been directly related to one or more AUs of $FACS$, as illustrated in Table 2. The features have been chosen in order to maximize the discrimination of their value ranges between activated and the corresponding non-activated AUs. In case more than one AU are mapped onto the same feature, only one of them can be active at a time, due to human facial anatomy; different AU activations will correspond to different value ranges of the shared feature. In other words, the activated facial AU is defined by the dynamic behavior of the corresponding features values. These features produce discrete dynamic behaviors which indicate the currently activated AU. According to the experimental results, these facial features are sufficient to detect the activation of 24 AUs.

In Table 2, function $MEAN(P_1, P_2)$ stands for the mean of two $3D$ points P_1 and P_2: $MEAN(P_1, P_2) = \frac{P_1+P_2}{2}$. To calculate the angle $Angle(P_1, P_2, P_3)$ formed by three $3D$ points P_1, P_2 and P_3 on P_2, the following formula is used:

$$Angle(P_1, P_2, P_3) = \arctan(|(D_1 \times D_2)| - (D_1 \cdot D_2)) \tag{4}$$

where $D_1 = P_1 - P_2$, $D_2 = P_2 - P_3$. For the calculation of the area of a triangle formed by three $3D$ points P_1, P_2 and P_3, returned by the function $Area(\Delta(P_1, P_2, P_3))$, Heron's formula is used. Finally, $Distance(P_1, P_2)$ denotes the euclidean distance between two $3D$ points P_1 and P_2. LM_i denotes the i-th extracted landmark as per Figure 3. Figures 5, 6 and 7 illustrate the mapping of the selected 16 features on a $3D$ facial mesh. Figure 8 illustrates the selected 16 features on the corresponding activated AUs. It should be pointed out that, in order to achieve scale invariance, distance features are normalized with respect to the distance formed between the outer part of the eye and the jaw (see Figure 7 image (e)), which is expected to be the biggest valued distance. Towards the same direction, area features are normalized with respect to the area formed by the outer part of both eyes and the jaw (see Figure 6 image (h)).

The key advantages of the proposed features compared to the state-of-the-art can be summed up as follows: 1) fewer landmarks are used, thus achieving a computational speed up of the descriptor, which results into a potential real-time procedure, 2) Geometrical information, which is usually found in the state-of-the-art works, is complemented by topological information. This combination is advantageous as some AUs (i.e. the ones involved in happiness expression) cause obvious topological but negligible geometrical changes of the face while other AUs (i.e. the ones involved in sadness expression) act the opposite way.

Table 2 Connecting AUs with features of the proposed descriptor.

AU DESCRIPTION	FEATURE NUMBER	FEATURE TYPE	FEATURE VALUE
$AU1$: Inner Brow Raiser $AU4$: Brow Lowerer	1	Angle	$Angle(LM_2, MEAN(LM_2, LM_3), LM_5)$
$AU2$: Outer Brow Raiser	2	Angle	$Angle(LM_1, MEAN(LM_2, LM_3), LM_5)$
$AU5$: Lid Raiser $AU7$: Lid Tightener	3	Area	$\frac{Area(\Delta(LM_1,LM_2,LM_5))+Area(\Delta(LM_3,LM_4,LM_5))}{2}$
$AU6$: Cheek Raiser	4	Area	$\frac{Area(\Delta(LM_1,LM_5,LM_6))+Area(\Delta(LM_4,LM_5,LM_7))}{2}$
$AU9$: Nose Wrinkle	10	Distance	$Distance(MEAN(LM_2, LM_3), LM_5)$
$AU10$: Upper Lip Raiser $AU19$: Tongue Show $AU24$: Lip Pressor	5	Area	$\frac{Area(\Delta(LM_5,LM_6,LM_7))+Area(\Delta(LM_6,LM_7,LM_8))}{2}$
$AU11$: Nasolabial Deepener	6	Area	$\frac{Area(\Delta(LM_2,LM_5,LM_6))}{Area(\Delta(LM_3,LM_5,LM_7))}$
$AU12$: Lip Corner Puller $AU15$: Lip Corner Depressor	11	Distance	$\frac{Distance(LM_1,LM_6)+Distance(LM_4,LM_7)}{2}$
$AU13$: Cheek Puffer	7	Area	$\frac{Area(\Delta(LM_1,LM_2,LM_6))+Area(\Delta(LM_3,LM_4,LM_7))}{2}$
$AU14$: Dimpler	8	Area	$Area(\Delta(LM_5, LM_6, LM_7))$
$AU16$: Lower Lip Depressor $AU17$: Chin Raiser	9	Area	$Area(\Delta(LM_6, LM_7, LM_8))$
$AU18$: Lip Puckerer $AU22$: Lip Funneler	12	Distance	$\frac{Distance(LM_5,LM_8)}{Distance(LM_6,LM_7)}$
$AU20$: Lip Strecher $AU23$: Lip Tightener	13	Distance	$Distance(LM_6, LM_7)$
$AU27$: Mouth Strech	14	Distance	$\frac{Distance(LM_1,LM_8)+Distance(LM_4,LM_8)}{2}$
$AU28$: Lip Suck	15	Distance	$\frac{Distance(LM_6,LM_8)+Distance(LM_7,LM_8)}{2}$
$AU30$: Jaw Sideways	16	Distance	$\frac{Distance(LM_5,LM_6)}{Distance(LM_5,LM_7)}$

2.4 AU detection as classification

The last step deals with the training/use of the classifiers. Since there are no huge training $4D$ facial data sets available, justifying the choice of deep learning techniques, we have opted for standard *Support Vector Machine* (SVM)

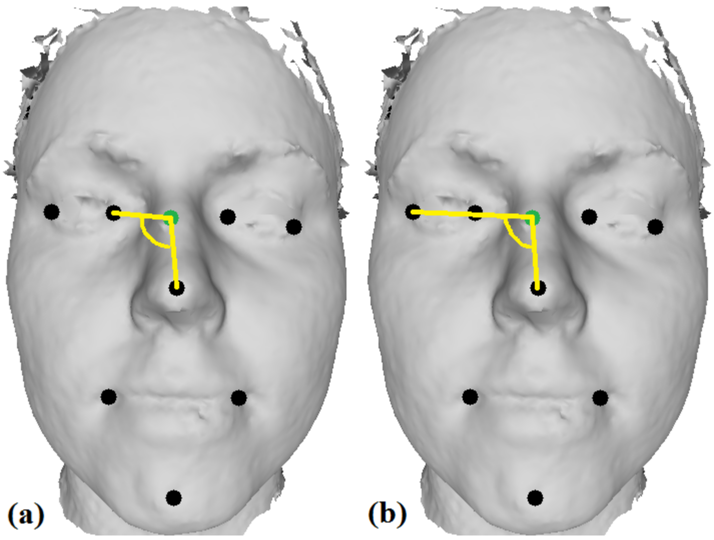

Fig. 5 Angular features used to model (a) $AU1$, $AU4$ and (b) $AU2$.

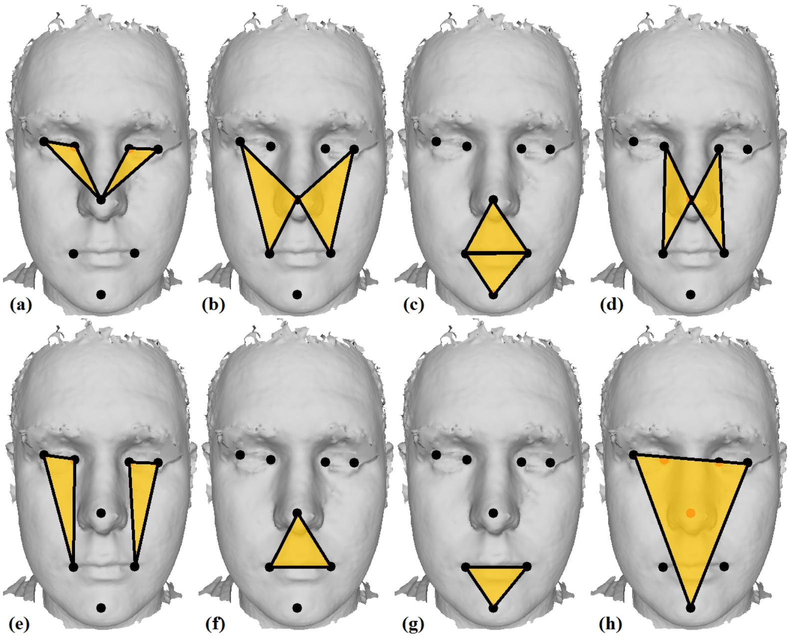

Fig. 6 Area features used to model (a) $AU5$, $AU7$, (b) $AU6$, (c) $AU10$, $AU19$, $AU24$, (d) $AU11$, (e) $AU13$, (f) $AU14$ and (g) $AU16$, $AU17$, (h) Normalization area.

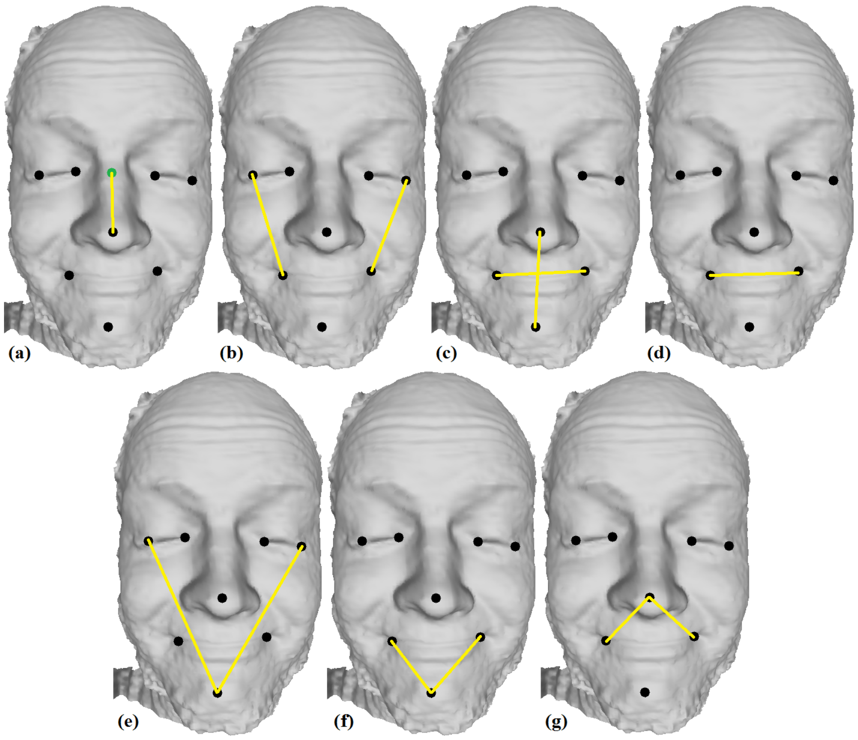

Fig. 7 Distance features used to model (a) $AU9$, (b) $AU12$, $AU15$ (c) $AU18$, $AU22$, (d) $AU20$, $AU23$, (e) $AU27$ (f) $AU28$ and (g) $AU30$.

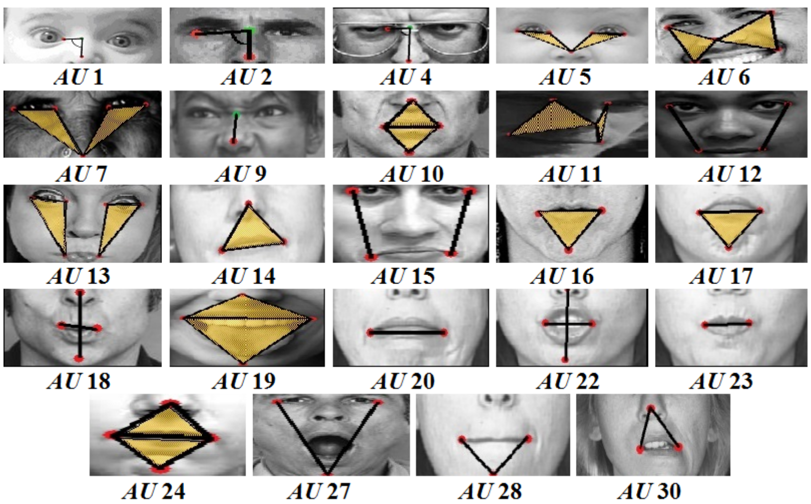

Fig. 8 Selected features on the corresponding 24 activated action units.

classifiers. We use 24 SVM classifiers, each corresponding to one AU. The *Radial Basis Function* (RBF) is used as kernel function, where $||x-x'||$ is the squared Euclidean Distance between two feature vectors x, x'. The cost parameter C and the scaling factor σ of the $SVM$are free parameters and their values were determined by implementing a "grid-search" on C and σ using cross-validation (for our case $\sigma = 0.2$ and $C = 0.1$). The SVM which classifies the i-th AU as activated or not, is trained by using the feature vectors of the descriptors of the training sequences corresponding to the i-th AU.

In the testing process, the feature vector of the descriptor of a testing sequence, which corresponds to the i-th AU, is given as input to the trained SVM corresponding to the same AU. The SVM, then, decides if the i-th AU is activated within the testing sequence or not. If the AU is classified as activated, then it is recorded on the AU activation list of the testing sequence, which is the output of the proposed scheme.

3 Experimental results

This section illustrates the experimental results of the proposed methodology. Initially, the data sets used for conducting experiments are discussed. Then, the results of the AU detection scheme, which is the main task of the proposed procedure, are illustrated. The achieved results are compared against state-of-the-art techniques using the same modality. Moreover, as additional applications, a new descriptor is presented, based on our AU detection scheme. The aforementioned descriptor is used for $4D$ facial expression retrieval and recognition. The performance on these tasks is also illustrated and compared against the state-of-the-art.

3.1 Data sets

Experiments have been conducted on the standard $4D$ facial data set $BP4D-Spontaneous$ [101], which is publicly available for a small fee and encodes AUs on $4D$ facial data, as well as the dynamic $3D$ $FACS$ [11] data set. These are the only available data sets containing $4D$ facial data accompanied with AU encoding. This is why all the AU detection techniques working on $3D$ video modality use one of the aforementioned data sets. One additional often used data set is $BU-4DFE$ [93] which only codes facial expressions, and is less challenging than $BP4D-Spontaneous$. On the contrary, $BP4D-Spontaneous$ codes both facial expressions and AUs and that is why it is selected for experimental purposes over the $BU-4DFE$ data set.

$BP4D-Spontaneous$ [101] involves 41 subjects (23 females and 18 males) of various ethnicities. For each subject, eight expressions (anger, disgust, embarrassment, fear, happiness, pain, sadness and surprise) were recorded and consist of the following phases: neutral face, outset, apex, offset and back to neutral face. The dynamic facial acquisition system $Di3D$ (www.di3d.com) was used and produced roughly 328 $3D$ videos, lasting about 25 seconds each. The corresponding texture images are provided for each $3D$ mesh of a $3D$ mesh sequence. The temporal resolution of the $3D$ videos is 25 fps and each $3D$ mesh consists of approximately 35,000 vertices. Finally, each $3D$ mesh is associated with 83 facial landmark points which are provided with the data set. The $BP4D-Spontaneous$ data set includes spontaneous facial behavior for all the provided facial expressions, making the experiments harder, and annotates the corresponding set of AU activations. In Figure 9, examples of $BP4D-Spontaneous$ data are illustrated.

In general the facial data constituting the data set are of good quality. However, inconsistencies are exhibited as some videos contain occluded meshes (see Figure 10). It should be pointed out that, despite the $3D$ data artifacts, no corrective actions took place so that our results can be comparable to those of future methods that may report results on this data set. In addition, the existence of inconsistencies highlights the true strength of the proposed descriptor.

All eight expressions for all 41 subjects of the data set were employed. Only the dynamic $3D$ mesh sequences were used (not the corresponding textures). Thus, 328 dynamic $3D$ sequences were processed. The $BP4D-Spontaneous$ data set contains approximately 360.500 $3D$ meshes in total. In order to reduce the processing time, temporal subsampling was performed; this was possible without sacrificing accuracy as we saw by inspection that the sequences of the data set were very dense with little variation between successive frames. The ratio of the subsampling was 1:6, reducing the number of $3D$ meshes which were finally processed to 60.100.

Finally, as far as the $BP4D-Spontaneous$ data set is concerned, we use the proposed technique in two variations. The first includes all 24 AUs presented in Table 2 and is destined for AU detection purposes; however we have also included it in the retrieval and recognition experiments for comparative purposes. The second variation includes only the 13 AUs which are directly connected to the six prototypical facial expressions according to $FACS$, and are presented in Table 1 ($AU26$ is excluded, as the $BP4D-Spontaneous$ data set does not provide corresponding ground truth). This variation is destined for facial expression retrieval and recognition purposes. The motivation for the second variation is to filter only the meaningful information of the AUs directly connected to the relevant facial expressions, as irrelevant AUs may be translated into noise during the training process of the SVM classifiers.

The second data set used for experiments is dynamic $3D$ $FACS$ [11]. This data set involves 10 subjects (6 females and 4 males) of average age 23.6 years. Each subject activates from 19 to 97 different AUs both individually and in combinations. The peak expression frame of each sequence has been manually $FACS$ coded by certified $FACS$ experts. The data set consists of 519 $3D$ videos. Each sequence lasts from 5 to 10 seconds depending on the complexity of the recorded AU. The temporal resolution of the $3D$ videos is 60 fps and each $3D$ mesh consists of approximately 30,000 vertices. However, the creators of the data set provide, approximately, 90 $3D$ frames per sequence.

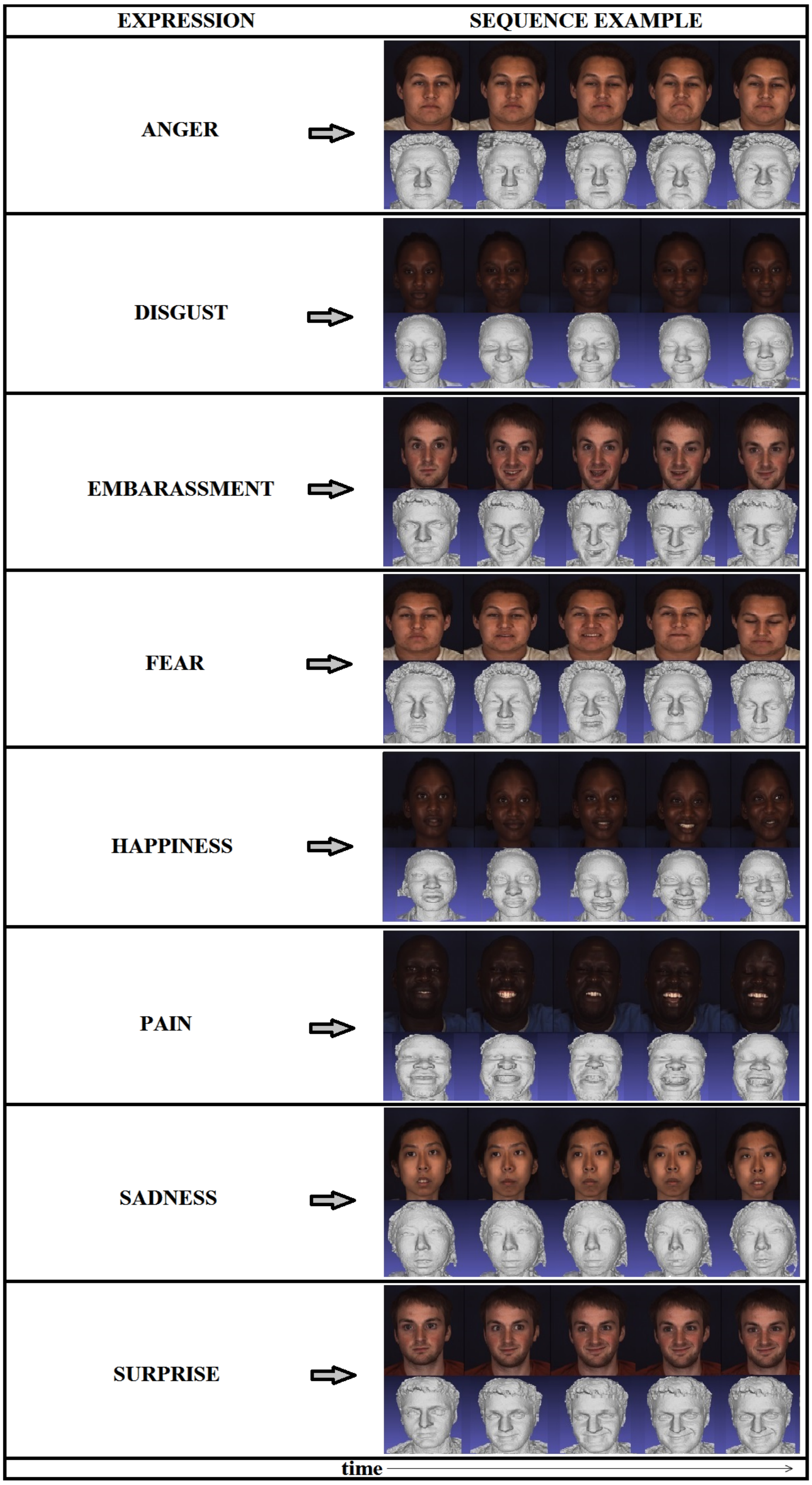

Fig. 9 Representative examples of expressions in $BP4D - Spontaneous$.

Fig. 10 Illustration of occluded mesh in the $BP4D - Spontaneous$ data set.

Corresponding texture images are provided for each $3D$ mesh of a $3D$ mesh sequence but were not used. Finally, no facial landmarks are provided for the $3D$ frames. In Figure 11, examples of the dynamic $3D$ $FACS$ data are illustrated.

The sequences of the dynamic $3D$ $FACS$ data set are very dense. In fact, the temporal resolution of this data set is more than twice than that of the $BU4D - Spontaneous$ data set. This means that the variation between successive frames in dynamic $3D$ $FACS$ data is quite low. For this reason we have performed, once again, sequence subsampling on the temporal domain. The ratio of the subsampling was 1:4. The eight landmarks for each $3D$ frame of the dynamic $3D$ $FACS$ data set were manually extracted. This highlights the ability of our method to be smoothly combined with any facial landmark detection technique as long as the aforementioned technique detects the eight landmarks specified within subsection 2.1.

The dynamic $3D$ $FACS$ data set encodes only AUs and not facial expressions. Consequently, we have used the aforementioned data set only for AU detection and not for facial expression retrieval/recognition purposes. We detect the subset of AUs which contains the intersection of the set of all the 24 AUs that can be detected by our method, and the set of all the AUs provided that are jointly encoded for all the subjects of the data set (not the same AUs are encoded for each subject of the dynamic $3D$ $FACS$ data set).

3.2 AU detection results

The AU detection methods presented in [50,81,69] are the only methods of the state-of-the-art which are tested on the publicly available data set $BP4D - Spontaneous$. That is why we can only be reliably compared with the aforementioned methods.

Table 3 illustrates the detection results of both variations of our proposed technique and the state-of-the-art on $BP4D - Spontaneous$. 'N/A' stands for not available information. Bold indicates the best performance among the methods compared. Notice that the variants of the proposed method generally outperform the state-of-the-art. There are only 7 exceptions ($AU1$, $AU2$, $AU4$, $AU5$, $AU15$, $AU24$, $AU27$) where we are outperformed by [50,81,69]. An explanation for this is that the methods of [50,69] use 83 landmarks, while we use only 8. Thus, [50,69] use more primary information while we use much less. The evaluation presented in Table 3, was calculated based on valid AU detection within the entire $3D$ facial mesh sequence.

Table 4 illustrates that we detect more AUs than state-of-the-art techniques. The proposed 24 AU detection method detects seventeen more AUs ($AU6$, $AU7$, $AU9$, $AU10$, $AU11$, $AU12$, $AU13$, $AU14$, $AU16$, $AU17$, $AU18$, $AU19$, $AU22$, $AU23$, $AU24$, $AU28$ and $AU30$) than the method of [69], twelve more AUs ($AU5$, $AU9$, $AU11$, $AU13$, $AU16$, $AU18$, $AU19$, $AU20$, $AU22$, $AU27$, $AU28$ and $AU30$) than the method of [50], thirteen more AUs ($AU5$, $AU9$, $AU11$, $AU13$, $AU16$, $AU18$, $AU19$, $AU20$, $AU22$, $AU24$, $AU27$, $AU28$ and $AU30$) than the method of [81], and ten more ($AU2$, $AU10$, $AU11$, $AU13$, $AU18$, $AU19$, $AU22$, $AU24$, $AU28$ and $AU30$) than [14]. It achieves an average per sequence detection rate of 80.45% for these 24 AUs, compared to 63.51% for the 12 AUs of [50] and 77.85% for the 14 AUs of [14]. In most cases, the proposed 13 AU detection method performs slightly better than the proposed 24 AU detection method. This is because, in the first case, there are fewer classes and the SVM classification problem is simpler.

The proposed AU detection scheme functions on a per mesh basis. Thus, we are capable of evaluating our scheme in terms of valid AU detections per $3D$ mesh of the sequence. Similar results can be drawn using the method of [14]. Table 5 illustrates the detection results on the $BP4D - Spontaneous$ data set for the per frame case. 'N/A' stands for not available information. The average detection rate of the proposed 24 AUs detection method is 66.23% compared to 65.50% for [14]. The proposed 13 AUs detection method also performs slightly better than the proposed 24 AUs detection method in the per mesh case for the reason explained above. Notice that per sequence detection is better than per mesh detection. This is because, in the first case we take into consideration temporal information in addition to the spatial information.

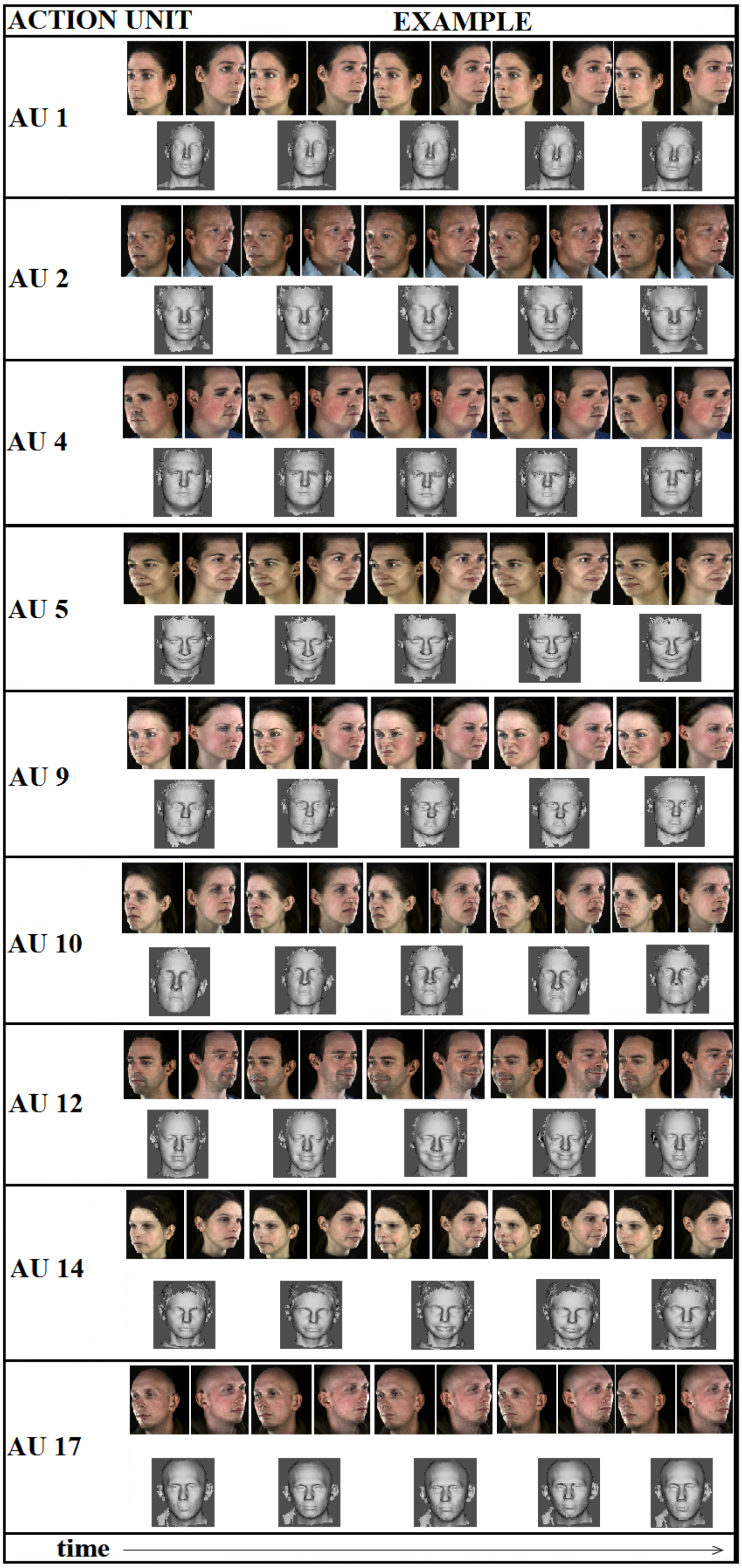

Fig. 11 Representative examples of AUs' encoding in the dynamic $3D$ $FACS$ data set. Top rows: $2D$ images from 2 viewpoints, bottom rows: $3D$ meshes.

Table 3 *AU* detection results compared to the state-of-the-art on *BP4D − Spontaneous* data set, per sequence.

*AU*s	1	2	4	5	6	7	9	10	11	12	13	14
Proposed Method (24 *AU*s)	62.9%	55.3%	63.8%	76.4%	84.6%	90.4%	72.6%	**86.9%**	**91.6%**	**80.5%**	**99.3%**	**86.9%**
Proposed Method (13 *AU*s)	65.8%	N/A	64.1%	75.8%	**85.0%**	**90.8%**	**72.7%**	N/A	N/A	80.0%	N/A	N/A
Sun *et al.* [69]	**93.0%**	**91.0%**	**89.0%**	**82.0%**	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Reale *et al.* [50]	58.4%	64.8%	63.1%	N/A	68.8%	58.9%	N/A	66.4%	N/A	59.1%	N/A	59.1%
Tulyakov *et al.* [81]	60.0%	50.0%	56.0%	N/A	79.0%	70.0%	N/A	77.0%	N/A	**81.0%**	N/A	66.0%
GeoTopo+ **[14]**	62.5%	N/A	63.3%	76.1%	84.2%	90.1%	72.5%	N/A	N/A	80.2%	N/A	86.7%

*AU*s	15	16	17	18	19	20	22	23	24	27	28	30
Proposed Method (24 *AU*s)	69.2%	73.5%	88.5%	**93.0%**	**85.2%**	**77.9%**	**90.3%**	77.4%	65.3%	89.0%	**74.2%**	**96.1%**
Proposed Method (13 *AU*s)	69.4%	**74.2%**	**88.7%**	N/A	N/A	77.8%	N/A	**77.7%**	N/A	89.0%	N/A	N/A
Sun *et al.* [69]	**94.0%**	N/A	N/A	N/A	N/A	67.0%	N/A	N/A	N/A	**99.0%**	N/A	N/A
Reale *et al.* [50]	69.0%	N/A	65.6%	N/A	N/A	N/A	N/A	61.4%	**67.6%**	N/A	N/A	N/A
Tulyakov *et al.* [81]	56.0%	N/A	63.0%	N/A	N/A	N/A	N/A	N/A	60.0%	N/A	N/A	N/A
GeoTopo+ **[14]**	68.9%	73.3%	88.4%	N/A	N/A	77.6%	N/A	77.3%	N/A	88.8%	N/A	N/A

Table 4 *AU*s detected by the proposed and state-of-the-art techniques.

METHOD	**DETECTED *AU*s**	**NUMBER OF DETECTED *AU*s**
Sun *et al.* [69]	*AU*1, *AU*2, *AU*4, *AU*5, *AU*15, *AU*20, *AU*27	7
Tsalakanidou *et al.* [79]	*AU*1, *AU*2, *AU*4, *AU*5, *AU*7, *AU*9, *AU*12, *AU*15, *AU*25, *AU*26, *AU*27	11
Tulyakov *et al.* [81]	*AU*1, *AU*2, *AU*4, *AU*6, *AU*7, *AU*10, *AU*12, *AU*14, *AU*15, *AU*17, *AU*23	11
Reale *et al.* [50]	*AU*1, *AU*2, *AU*4, *AU*6, *AU*7, *AU*10, *AU*12, *AU*14, *AU*15, *AU*17, *AU*23, *AU*24	12
Proposed Method (13 *AU*s)	*AU*1, *AU*4, *AU*5, *AU*6, *AU*7, *AU*9, *AU*12, *AU*15, *AU*16, *AU*17, *AU*20, *AU*23, *AU*27	13
GeoTopo+ [14]	*AU*1, *AU*4, *AU*5, *AU*6, *AU*7, *AU*9, *AU*12, *AU*14, *AU*15, *AU*16, *AU*17, *AU*20, *AU*23, *AU*27	14
Proposed Method (24 *AU*s)	*AU*1, *AU*2, *AU*4, *AU*5, *AU*6, *AU*7, *AU*9, *AU*10, *AU*11, *AU*12, *AU*13, *AU*14, *AU*15, *AU*16, *AU*17, *AU*18, *AU*19, *AU*20, *AU*22, *AU*23, *AU*24, *AU*27, *AU*28, *AU*30	24

Table 5 *AU* detection results on *BP4D − Spontaneous* data set, per $3D$ mesh.

*AU*s	1	2	4	5	6	7	9	10	11	12	13	14
Proposed Method (24 *AU*s)	80.0%	**63.1%**	74.9%	**74.9%**	**54.0%**	**48.6%**	69.6%	**51.6%**	**76.1%**	**50.3%**	**63.0%**	**57.3%**
Proposed Method (13 *AU*s)	**80.2%**	N/A	**75.1%**	74.2%	53.8%	46.5%	**73.2%**	N/A	N/A	50.1%	N/A	N/A
GeoTopo+ **[14]**	79.6%	N/A	74.2%	74.5%	53.5%	47.9%	69.3%	N/A	N/A	49.8%	N/A	57.1%

*AU*s	15	16	17	18	19	20	22	23	24	27	28	30
Proposed Method (24 *AU*s)	66.4%	**84.8%**	**57.6%**	**70.5%**	**63.7%**	69.2%	**71.6%**	63.9%	**63.7%**	72.0%	**69.4%**	**73.4%**
Proposed Method (13 *AU*s)	**69.0%**	84.7%	**57.6%**	N/A	N/A	**69.7%**	N/A	**69.8%**	N/A	**72.1%**	N/A	N/A
GeoTopo+ **[14]**	66.1%	84.1%	57.2%	N/A	N/A	68.8%	N/A	63.4%	N/A	71.6%	N/A	N/A

The automatically detected landmarks, provided by the algorithm of [47] during the first step of the pipeline, have an average estimation error of 2.53 mm compared to the ground truth landmarks, provided by the data set. Table 6 illustrates the performance of the proposed technique when using the ground truth versus the automatically detected landmarks. It is clear that the ground truth landmarks make only a marginal contribution in AU detection, indicating that the proposed features are robust AU detectors.

Table 6 AU detection results on $BP4D-Spontaneous$ data set, per $3D$ mesh, when using automated and ground truth landmarks.

AUs	1	2	4	5	6	7	9	10	11	12	13	14
Automated Landmarks (24 AUs)	62.9%	55.3%	63.8%	76.4%	84.6%	90.4%	72.6%	86.9%	91.6%	80.5%	99.3%	86.9%
Ground Truth Landmarks	63.0%	56.4%	66.1%	77.5%	86.8%	91.5%	74.8%	88.0%	92.7%	82.7%	99.5%	88.1%
AUs	15	16	17	18	19	20	22	23	24	27	28	30
Automated Landmarks (24 AUs)	69.2%	73.5%	88.5%	93.0%	85.2%	77.9%	90.3%	77.4%	65.3%	89.0%	74.2%	96.1%
Ground Truth Landmarks	71.4%	74.6%	89.6%	96.3%	87.4%	79.2%	91.4%	78.5%	68.6%	91.2%	76.4%	97.3%

AU detection experiments were also conducted using the dynamic $3D$ $FACS$ data set. The results are illustrated in Table 7 for the per sequence case. A drawback of this data set is that not the same AUs are encoded for each subject. Thus, we detect the AUs which belong to the intersection of the set of 24 AUs that can be detected by our method, and the set of AUs that are jointly encoded for all the subjects of the data set. There are nine such AUs ($AU1$, $AU2$, $AU4$, $AU5$, $AU9$, $AU10$, $AU12$, $AU14$ and $AU17$). Finally, the encoding provided is only on a per sequence basis. Thus, experiments of AU detection on a per frame basis could not be performed due to the lack of ground truth.

Table 7 AU detection results on the dynamic $3D$ $FACS$ data set per sequence.

AUs	1	2	4	5	9	10	12	14	17
Proposed Method	72.3%	89.9%	88.5%	87.7%	87.4%	89.5%	87.6%	89.1%	90.0%

3.3 Facial expression retrieval results

Based on the proposed AU detection scheme using $4D$ data, we illustrate a spatio-temporal binary descriptor which can be used to perform $4D$ facial expression retrieval. Each dynamic $3D$ facial mesh of the $BP4D-Spontaneous$ data set represents a single facial expression. In all tests, Leave-One-Out cross validation was employed. Initially, for each $3D$ facial mesh sequence, used as a query, we apply the AU detection scheme. We thus construct a list indicating which of the 24 (or 13) tested AUs are activated within the sequence. This list is the basis for a spatio-temporal descriptor as indicated in Equation 5.

The descriptor of the query sequence is then compared against the descriptors of the remaining sequences of the data set using the *Hamming Distance* [9]. An ascending sort of the sequences based on the similarity scores gives us the retrieval list for the query.

$$B_i = \left\{ \begin{array}{l} 1 : \text{if the } i\text{-th } AU \text{ is activated within the sequence} \\ 0 : \text{if the } i\text{-th } AU \text{ is not activated within the sequence} \end{array} \right\} \quad (5)$$

The retrieval results are compared in terms of retrieval evaluation metrics: Nearest Neighbor (NN), first/second tier, Discounted Cumulative Gain (DCG) and precision-recall diagram (PR diagram). Table 8 illustrates the comparison of the proposed retrieval methodologies against the state-of-the-art $4D$ facial expression retrieval techniques on the $BP4D - Spontaneous$ data set. Both variations of the proposed technique outperform the state-of-the-art. The *Proposed Methods with ground truth AUs* are based on the proposed methodology, except that the binary descriptor B_i is constructed based on the ground truth for the 24 (or 13) AUs instead of the proposed AU detection scheme, thus effectively assuming 100% detection rate for all 24 (or 13) AUs. Therefore, the performance (in terms of retrieval and recognition) of the ground truth-based methods are upper bounds for any AU-based technique. Notice that the ground truth-based method using 13 AUs achieves a higher upper bound than the one using 24 AUs. This is because these specific 13 AUs are directly connected to facial expressions [19], and thus, they are more suitable for describing them. The remaining 11 AUs act as noise during the training process of the SVM classifiers.

In Figure 12 the PR diagrams of the proposed retrieval methodologies are presented in relation to the best state-of-the-art retrieval methodologies.

Table 8 Comparison of state-of-the-art and the proposed methodologies for the retrieval process on $BP4D - Spontaneous$ data set.

METHOD	NN	1st TIER	2nd TIER	DCG
Proposed Method with ground truth AUs (13 AUs)	**0.83**	**0.71**	**0.90**	**0.91**
Proposed Method (13 AUs)	**0.72**	**0.66**	**0.79**	**0.86**
Proposed Method with ground truth AUs (24 AUs)	**0.69**	**0.57**	**0.74**	**0.86**
Proposed Method (24 AUs)	**0.68**	**0.55**	**0.72**	**0.84**
$GeoTopo+$ [14]	0.67	0.55	0.72	0.83
$FELM$ [94]	0.63	0.35	0.48	0.77
Danelakis *et al.* [13] (extended to 8 expressions)	0.61	0.52	0.69	0.82
Berretti *et al.* [5]	0.59	0.49	0.69	0.81
Distribution Vectors [55]	0.50	0.41	0.57	0.76
Curvature [70,71,68,78,80,6,100]	0.39	0.34	0.47	0.71
$LBP - TOP$ [21,20]	0.39	0.34	0.47	0.71
Gradient [68,78,80]	0.38	0.33	0.46	0.71
Shape Index [29]	0.30	0.32	0.47	0.70

3.4 Facial expression recognition results

The AU-based descriptor B_i, illustrated in Equation 5, can be used to perform $4D$ facial expression recognition, in order to test the performance of the

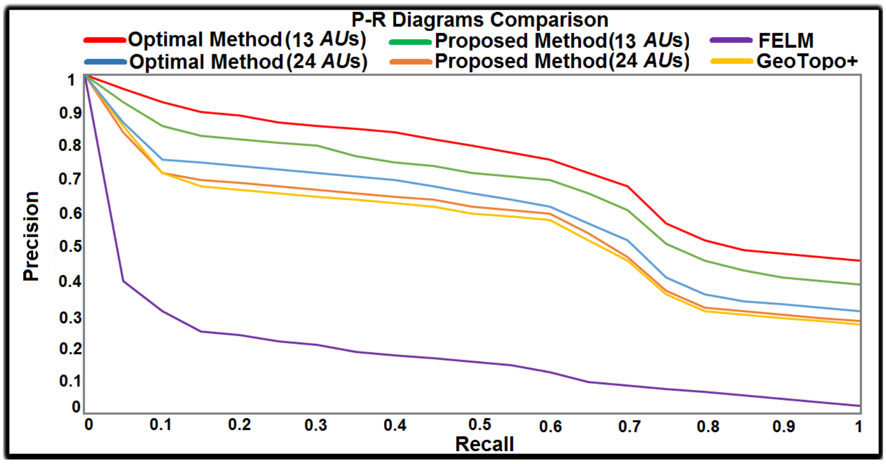

Fig. 12 PR diagram of the proposed methods and the top retrieval methodologies for $BP4D-Spontaneous$ data set.

descriptor against state-of-the-art methods whose performance is evaluated in terms of classification accuracy. To this effect, a supervised procedure is used. The sequences of the $BP4D-Spontaneous$ data set are divided into training and testing subsets based on 5-fold cross validation. The training subset is used to train an SVM classifier on the features of the binary descriptor B_i. The RBF is used again as kernel function. Then, given a sequence of the testing set, the trained SVM classifier is recruited to classify it into one of the eight possible classes/expressions.

Although there are plenty of methods using $2D$ images, $2D$ videos and $3D$ images for facial expression recognition [26,12,90,96,92,74,43,97,99,72,67,39, 40,49,56,18,48], we are focused on the $3D$ video modality techniques [7,94, 51,79,85,71,33,55,14,105]. Only the technique presented in [14] is tested on the $BP4D-Spontaneous$ data set and, thus, it can be reliable compared to the proposed method. Table 9 presents the results of $4D$ facial expression recognition for all the expressions of the $BP4D-Spontaneous$ data set. The proposed AU-based $4D$ recognition methodology variations exhibit the best classification results. The recognition performance of the 13 AU detection method is better than the 24 AU detection method, which, once again, proves the superiority of the first subset of AUs when it comes to facial expression information. The ground truth-based methods achieve significantly better results; this is because the binary descriptors of these methods achieve better training for the SVM recognition classifier. It should be pointed out that the $4D$ recognition methodology presented is [14] is unsupervised, while our proposed procedures are supervised.

Table 9 Overview of research work on dynamic $3D$ facial expression recognition on the $BP4D - Spontaneous$ data set.

METHOD	NUMBER OF EXPRESSIONS	CLASSIFIER TRAINING	CLASSIFICATION ACCURACY
Proposed Method with ground truth AUs (13 AUs)	**8**	**YES**	**98.90%**
Proposed Method with ground truth AUs (24 AUs)	**8**	**YES**	**97.60%**
Proposed Method (13 AUs)	**8**	**YES**	**92.23%**
Proposed Method (24 AUs)	**8**	**YES**	**90.60%**
GeoTopo+ [14]	8	NO	88.56%

3.5 Time efficiency

The experiments required a respectable amount of time in order to be executed. On a machine with an Intel Core i7 CPU at 3.5GHz and 16GB RAM memory, approximately 5 seconds per query sequence are needed. The time required for landmark extraction is not included as the landmark extraction method can vary. The proposed method is potentially real-time as the part following the landmark extraction *is* real-time. The landmark extraction, used here, requires approximately 3 minutes per mesh and is not real-time but, according to the authors of the paper [47], it has a real-time potential.

4 Future challenges

The $BP4D{-}Spontaneous$ and the dynamic $3D$ $FACS$ data sets, used throughout this work, are of good quality. This fact allows our descriptors to be tested within an almost ideal environment. Our proposed method depends on the landmark extraction method that is used. The implemented extraction algorithm [47], has been shown to be robust to noise, especially that coming from facial expressions and their axis variations. A challenge would be to test the proposed procedure to noisy real data. Such facial data are not yet available when it comes to $4D$ modality.

Recently, deep learning methods are being applied when it is not easy to design good features and large labeled training data sets are available. Such large data sets are not currently available in the case of $4D$ facial expressions. So, no reliable deep learning can be integrated, while human intuition has resulted in a good set of features. The construction of artificial or real big $4D$ data sets containing face data constitute a major future challenge. If such data sets become available, then deep learning procedures could also be recruited on the field.

5 Conclusions

AU detection lies at the heart of facial analysis. Applications, such as facial expression retrieval and recognition, can be based on AU detection. The present work proposes a robust scheme for AU detection based on 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 a descriptor of topological information, and (iii) Training classifiers using the features of the aforementioned descriptor.

Our method employed significantly fewer landmarks than state-of-the-art techniques. We proposed a set of features, based on the landmarks, which extract topological information from a $3D$ facial sequence, resulting in a descriptor for such data. These features are used to train classifiers which perform AU detection.

The detection performance of the proposed scheme improves on the state-of-the-art for most AUs while it can detect significantly more AUs than previous techniques that use $4D$ data. Among the multitude of applications of AU detection, we illustrate facial expression retrieval and recognition. To this end, the optimal combination of AUs for the description of facial expression is highlighted. The retrieval performance is comparable to the state-of-the-art while the recognition performance outperforms it.

References

1. Ashraf, A.B., Lucey, S., Cohn, J.F., Chen, T., Ambadar, Z., Prkachin, K.M., Solomon, P.E.: The painful face pain expression recognition using active appearance models. Image and Vision Computing **27**(12), 1788 – 1796 (2009)
2. Baltruaitis, T., Mahmoud, M., Robinson, P.: Cross-dataset learning and person-specific normalisation for automatic action unit detection. In: Automatic Face and Gesture Recognition (FG), pp. 1–6 (2015)
3. Bartlett, M., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Fully Automatic Facial Action Recognition in Spontaneous Behavior. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR), pp. 223 – 230 (2006)
4. Bayramoglu, N., Zhao, G., Pietikäinen, M.: CS-3DLBP and geometry based person independent 3d facial action unit detection. In: International Conference on Biometrics, pp. 1–6 (2013)
5. Berretti, S., Del Bimbo, A., Pala, P.: Automatic Facial Expression Recognition in Real-time from Dynamic Sequences of 3D Face Scans. Vis. Comput. **29**(12), 1333–1350 (2013)
6. Canavan, S.J., Sun, Y., Zhang, X., Yin, L.: A dynamic curvature based approach for facial activity analysis in 3D space. In: CVPR Workshops, pp. 14–19 (2012)
7. Chang, Y., Vieira, M.B., Turk, M., Velho, L.: Automatic 3D facial expression analysis in videos. In: IEEE Workshop AMFG '05, pp. 293–307 (2005)
8. Chen, J., Liu, X., Tu, P., Aragones, A.: Learning person-specific models for facial expression and action unit recognition. Pattern Recogn. Lett. **34**(15), 1964–1970 (2013)
9. Choi, S.S., Cha, S.H.: A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics pp. 43 – 48 (2010)
10. Chu, W.S., De La Torre, F., Cohn, J.F.: Selective transfer machine for personalized facial action unit detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3515–3522 (2013)
11. Cosker, D., Krumhuber, E., Hilton, A.: A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling. In: Proc. ICCV '11, pp. 2296–2303 (2011)
12. Dahmane, M., Meunier, J.: Prototype-based modeling for facial expression analysis. IEEE Transactions on Multimedia **16**(6), 1574–1584 (2014)
13. Danelakis, A., Theoharis, T., Pratikakis, I.: GeoTopo: Dynamic 3D Facial Expression Retrieval Using Topological and Geometric Information. In: Proc. 3D Object Retrieval 2014, pp. 1–8 (2014)

14. Danelakis, A., Theoharis, T., Pratikakis, I., Perakis, P.: An effective methodology for dynamic 3D facial expression retrieval. Pattern Recognition **52**, 174 – 185 (2016)
15. Dapogny, A., Bailly, K., Dubuisson, S.: Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. International Journal of Computer Vision pp. 1–17 (2016)
16. Devijver, P.A., Kittler, J.: Pattern recognition: A statistical approach. Prentice Hall (1982)
17. Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J.: Classifying Facial Actions. IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(10), 974 – 989 (1999)
18. Drira, H., Ben Amor, B., Daoudi, M., Berretti, S.: A Dense Deformation Field for Facial Expression Analysis in Dynamic Sequences of 3D Scans, pp. 148–159. Springer International Publishing (2013)
19. Ekman, P., Friesen, W.: Facial action coding system: A technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto (1978)
20. 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), 738–749 (2012)
21. Fang, T., Zhao, X., Shah, S.K., Kakadiaris, I.A.: 4D facial expression recognition. In: ICCV '11, pp. 1594–1601 (2011)
22. Fasel, B., Luettin, J.: Automatic facial expression analysis: A survey. Pattern Recognition **36**(1), 259–275 (1999)
23. Gao, G., Xu, M., Shen, J., Ma, H., Yan, S.: Cast2face: Assigning character names onto faces in movie with actor-character correspondence. IEEE Transactions on Circuits and Systems for Video Technology **26**(12), 2299–2312 (2016)
24. Gehrig, T., Ekenel, H.: A common framework for real-time emotion recognition and facial action unit detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1 – 6 (2011)
25. Gudi, A., Tasli, H.E., den Uyl, T.M., Maroulis, A.: Deep learning based facs action unit occurrence and intensity estimation. In: Automatic Face and Gesture Recognition (FG), pp. 1–5 (2015)
26. Huang, Y., Li, Y., Fan, N.: Robust symbolic dual-view facial expression recognition with skin wrinkles: Local versus global approach. IEEE Transactions on Multimedia **12**(6), 536–543 (2010)
27. Jaiswal, S., Valstar, M.: Deep learning the dynamic appearance and shape of facial action units. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–8 (2016)
28. Jeni, L.A., Girard, J.M., Cohn, J.F., Torre, F.D.L.: Continuous au intensity estimation using localized, sparse facial feature space. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–7 (2013)
29. Jeni, L.A., Lórincz, A., Nagy, T., Palotai, Z., Sebók, J., Szabó, Z., Takács, D.: 3D shape estimation in video sequences provides high precision evaluation of facial expressions. Image and Vision Computing **30**(10), 785 – 795 (2012)
30. Jiang, B., Valstar, M.F., Pantic, M.: Action unit detection using sparse appearance descriptors in space-time video volumes. In: Automatic Face Gesture Recognition and Workshops, pp. 314–321 (2011)
31. Khademi, M., Morency, L.P.: Relative facial action unit detection. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1090 – 1095 (2014)
32. Kotsia, I., Zafeiriou, S., Pitas, I.: Texture and shape information fusion for facial expression and facial action unit recognition. Pattern Recognition **41**(3), 833 – 851 (2008)
33. Le, V., Tang, H., Huang, T.S.: Expression recognition from 3D dynamic faces using robust spatio-temporal shape features. In: IEEE FG '11, pp. 414–421 (2011)
34. Lien, J.J.J., Kanade, T., Cohn, J., Li, C.C.: Subtly Different Facial Expression Recognition and Expression Intensity Estimation. In: IEEE Conference on Computer Vison and Pattern Recogntion, pp. 853 – 859 (1998)

35. Liu, L., Cheng, L., Liu, Y., Jia, Y., Rosenblum, D.S.: Recognizing complex activities by a probabilistic interval-based model. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp. 1266–1272 (2016)
36. Liu, Y., Nie, L., Han, L., Zhang, L., Rosenblum, D.S.: Action2activity: Recognizing complex activities from sensor data. Computer Vision and Pattern Recognition **abs/1611.01872** (2016). URL `http://arxiv.org/abs/1611.01872`
37. Liu, Y., Nie, L., Liu, L., Rosenblum, D.S.: From action to activity: Sensor-based activity recognition. Neurocomputing **181**(Supplement C), 108 – 115 (2016)
38. Lucey, P., Cohn, J.F., Matthews, I., Lucey, S., Sridharan, S., Howlett, J., Prkachin, K.M.: Automatically detecting pain in video through facial action units. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **41**(3), 664–674 (2011)
39. Maalej, A., Amor, B.B., Daoudi, M., Srivastava, A., Berretti, S.: Local 3d shape analysis for facial expression recognition. In: 2010 20th International Conference on Pattern Recognition, pp. 4129–4132 (2010)
40. Maalej, A., Amor, B.B., Daoudi, M., Srivastava, A., Berretti, S.: Shape analysis of local facial patches for 3d facial expression recognition. Pattern Recognition **44**(8), 1581–1589 (2011)
41. van der Maaten, L., Hendriks, E.: Action unit classification using active appearance models and conditional random fields. Cognitive Processing **13**(2), 507–518 (2012)
42. Mahoor, M.H., Zhou, M., Veon, K.L., Mavadati, S.M., Cohn, J.F.: Facial action unit recognition with sparse representation. In: Automatic Face Gesture Recognition and Workshops, pp. 336–342 (2011)
43. Mao, Q., Rao, Q., Yu, Y., Dong, M.: Hierarchical bayesian theme models for multipose facial expression recognition. IEEE Transactions on Multimedia **19**(4), 861–873 (2017)
44. Pantic, M., Patras, I.: Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics **36**(2), 433 – 449 (2006)
45. Pantic, M., Rothkrantz, L.J.M.: Automatic Analysis of Facial Expressions: The State of the Art. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12), 1424 – 1445 (2000)
46. Pantic, M., Rothkrantz, L.J.M.: Facial action recognition for facial expression analysis from static face images. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **34**(3), 1449–1461 (2004)
47. Perakis, P., Theoharis, T., Kakadiaris, I.A.: Feature fusion for facial landmark detection. Pattern Recognition **47**(9), 2783 – 2793 (2014)
48. Pinto, S.C.D., Mena-Chalco, J.P., Lopes, F.M., Velho, L., Cesar, R.: 3d facial expression analysis by using 2d and 3d wavelet transforms. In: Image Processing (ICIP), 2011 18th IEEE International Conference on, pp. 1281–1284. IEEE (2011)
49. Ramanathan, S., Kassim, A., Venkatesh, Y.V., Wah, W.S.: Human facial expression recognition using a 3d morphable model. In: 2006 International Conference on Image Processing, pp. 661–664 (2006)
50. Reale, M., Zhang, X., Yin, L.: Nebula feature: A space-time feature for posed and spontaneous 4D facial behavior analysis. In: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1 – 8 (2013)
51. 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, pp. 1–7 (2008)
52. Ruiz, A., Van de Weijer, J., Binefa, X.: From emotions to action units with hidden and semi-hidden-task learning. In: The IEEE International Conference on Computer Vision (ICCV), pp. 3703–3711 (2015)
53. Sandbach, G., Zafeiriou, S., Pantic, M.: Binary pattern analysis for 3D facial action unit detection. In: Proceedings of the British Machine Vision Conference, BMVC 2012, Surrey, UK, pp. 119.1 – 119.12 (2012)
54. Sandbach, G., Zafeiriou, S., Pantic, M.: Local normal binary patterns for 3D facial action unit detection. In: 19th IEEE International Conference on Image Processing (ICIP), pp. 1813 – 1816 (2012)
55. Sandbach, G., Zafeiriou, S., Pantic, M., Rueckert, D.: Recognition of 3D facial expression dynamics. Elsevier Image and Vision Computing **30**(10), 762–773 (2012)

56. Savran, A., Sankur, B.: Automatic detection of facial actions from 3d data. In: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pp. 1993–2000 (2009)
57. Savran, A., Sankur, B.: Detecting action units on 3d faces. In: 2010 IEEE 18th Signal Processing and Communications Applications Conference, pp. 300–303 (2010)
58. Savran, A., Sankur, B.: Detecting 3d facial action units via registration. In: 2011 IEEE 19th Signal Processing and Communications Applications Conference (SIU), pp. 371–374 (2011)
59. Savran, A., Sankur, B., Bilge, M.T.: Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units. Pattern recognition **45**(2), 767–782 (2012)
60. Savran, A., Sankur, B., Taha Bilge, M.: Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units. Pattern Recognition **45**(2), 767–782 (2012)
61. Senechal, T., Bailly, K., Prevost, L.: Impact of action unit detection in automatic emotion recognition. Pattern Analysis and Applications **17**(1), 51–67 (2014)
62. Senechal, T., McDuff, D., el Kaliouby, R.: Facial action unit detection using active learning and an efficient non-linear kernel approximation. In: The IEEE International Conference on Computer Vision (ICCV) Workshops, pp. 10–18 (2015)
63. Senechal, T., Rapp, V., Salam, H., Seguier, R., Bailly, K., Prevost, L.: Combining aam coefficients with lgbp histograms in the multi-kernel svm framework to detect facial action units. In: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pp. 860–865. IEEE (2011)
64. Senechal, T., Rapp, V., Salam, H., Seguier, R., Bailly, K., Prevost, L.: Facial action recognition combining heterogeneous features via multikernel learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **42**(4), 993–1005 (2012)
65. Simon, T., Nguyen, M.H., Torre, F.D.L., Cohn, J.F.: Action unit detection with segment-based svms. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 2737–2744 (2010)
66. Song, Y., McDuff, D., Vasisht, D., Kapoor, A.: Exploiting sparsity and co-occurrence structure for action unit recognition. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, pp. 1–8 (2015)
67. Soyel, H., Demirel, H.: Optimal feature selection for 3d facial expression recognition using coarse-to-fine classification. Turkish Journal of Electrical Engineering & Computer Sciences **18**(6), 1031–1040 (2010)
68. 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), 461–474 (2010)
69. Sun, Y., Reale, M., Yin, L.: Recognizing partial facial action units based on 3d dynamic range data for facial expression recognition. In: Automatic Face Gesture Recognition, pp. 1–8 (2008)
70. Sun, Y., Reale, M., Yin, L.: Recognizing partial facial action units based on 3D dynamic range data for facial expression recognition. In: FG '08, pp. 1–8 (2008)
71. Sun, Y., Yin, L.: Facial expression recognition based on 3D dynamic range model sequences. In: Springer Proc. ECCV '08: Part II, pp. 58–71 (2008)
72. Tang, H., Huang, T.S.: 3d facial expression recognition based on properties of line segments connecting facial feature points. In: 2008 8th IEEE International Conference on Automatic Face Gesture Recognition, pp. 1–6 (2008)
73. Tao, D., Song, M., Li, X., Shen, J., Sun, J., Wu, X., Faloutsos, C., Maybank, S.J.: Bayesian tensor approach for 3-d face modeling. IEEE Transactions on Circuits and Systems for Video Technology **18**(10), 1397–1410 (2008)
74. Tawari, A., Trivedi, M.M.: Face expression recognition by cross modal data association. IEEE Transactions on Multimedia **15**(7), 1543–1552 (2013)
75. Tian, Y.I., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(2), 97–115 (2001)
76. Tian, Y.L., Kanade, T., Cohn, J.F.: Recognizing Action Units for Facial Expression Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(2), 97 – 115 (2001)

77. Tong, Y., Liao, W., Ji, Q.: Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(10), 1683 – 1699 (2007)
78. Tsalakanidou, F., Malassiotis, S.: Robust facial action recognition from real-time 3D streams. In: CVPR '09, pp. 4–11 (2009)
79. Tsalakanidou, F., Malassiotis, S.: Real-time 2D+3D facial action and expression recognition. Pattern Recognition **43**(5), 1763 – 1775 (2010)
80. Tsalakanidou, F., Malassiotis, S.: Real-time 2D+3D facial action and expression recognition. Elsevier Pattern Recognition **43**(5), 1763–1775 (2010)
81. Tulyakov, S., Vieriu, R.L., Sangineto, E., Sebe, N.: Facecept3d: Real time 3d face tracking and analysis. In: The IEEE International Conference on Computer Vision (ICCV) Workshops, pp. 29–33 (2015)
82. Valstar, M., Pantic, M.: Fully Automatic Facial Action Unit Detection and Temporal Analysis. In: Computer Vision and Pattern Recognition Workshop (CVPRW), pp. 149 – 149 (2006)
83. Valstar, M.F., Almaev, T., Girard, J.M., McKeown, G., Mehu, M., Yin, L., Pantic, M., Cohn, J.F.: Fera 2015 - second facial expression recognition and analysis challenge. In: Automatic Face and Gesture Recognition (FG), pp. 1–8 (2015)
84. Valstar, M.F., Sánchez-Lozano, E., Cohn, J.F., Jeni, L.A., Girard, J.M., Zhang, Z., Yin, L., Pantic, M.: Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. arXiv preprint arXiv:1702.04174 (2017)
85. Venkatesh, Y.V., Kassim, A.K., Murthy, O.V.R.: Resampling approach to facial expression recognition using 3d meshes. In: 2010 20th International Conference on Pattern Recognition, pp. 3772–3775 (2010)
86. Walecki, R., Rudovic, O., Pavlovic, V., Pantic, M.: Variable-state latent conditional random fields for facial expression recognition and action unit detection. In: Automatic Face and Gesture Recognition, vol. 1, pp. 1–8 (2015)
87. Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., Kissler, J.: Mapping the emotional face. how individual face parts contribute to successful emotion recognition. PLOS ONE **12**(5), 1–15 (2017). URL `https://doi.org/10.1371/journal.pone.0177239`
88. Wu, C.H., Wei, W.L., Lin, J.C., Lee, W.Y.: Speaking effect removal on emotion recognition from facial expressions based on eigenface conversion. IEEE Transactions on Multimedia **15**(8), 1732–1744 (2013)
89. Xie, L., Shen, J., Han, J., Zhu, L., Shao, L.: Dynamic multi-view hashing for online image retrieval. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, pp. 3133–3139 (2017)
90. Yan, J., Zheng, W., Xu, Q., Lu, G., Li, H., Wang, B.: Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech. IEEE Transactions on Multimedia **18**(7), 1319–1329 (2016)
91. Yce, A., Gao, H., Thiran, J.P.: Discriminant multi-label manifold embedding for facial action unit detection. In: Automatic Face and Gesture Recognition (FG), pp. 1–6 (2015)
92. Yeasin, M., Bullot, B., Sharma, R.: Recognition of facial expressions and measurement of levels of interest from video. IEEE Transactions on Multimedia **8**(3), 500–508 (2006)
93. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3D dynamic facial expression database. In: IEEE Proc. FG '08, pp. 1–6 (2008)
94. Yin, L., Wei, X., Longo, P., Bhuvanesh, A.: Analyzing facial expressions using intensity-variant 3D data for human computer interaction. In: Proc. ICPR '06, pp. 1248–1251 (2006)
95. Yudin, E., Wetzler, A., Sela, M., Kimmel, R.: Improving 3d facial action unit detection with intrinsic normalization. In: Proceedings of the 1st International Workshop on DIFFerential Geometry in Computer Vision for Analysis of Shapes, Images and Trajectories (DIFF-CV 2015), 5, pp. 1–10 (2015)
96. Zafeiriou, S., Pitas, I.: Discriminant graph structures for facial expression recognition. IEEE Transactions on Multimedia **10**(8), 1528–1540 (2008)
97. Zen, G., Porzi, L., Sangineto, E., Ricci, E., Sebe, N.: Learning personalized models for facial expression analysis and gesture recognition. IEEE Transactions on Multimedia **18**(4), 775–788 (2016)

98. Zeng, J., Chu, W.S., l. Torre, F.D., Cohn, J.F., Xiong, Z.: Confidence preserving machine for facial action unit detection. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 3622–3630 (2015)
99. Zhang, T., Zheng, W., Cui, Z., Zong, Y., Yan, J., Yan, K.: A deep neural network-driven feature learning method for multi-view facial expression recognition. IEEE Transactions on Multimedia **18**(12), 2528–2536 (2016)
100. 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)
101. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.: BP4D Spontaneous: A high-resolution spontaneous 3D dynamic facial expression database. Image and Vision Computing **32**(10), 692–706 (2014)
102. Zhao, K., Chu, W.S., De la Torre, F., Cohn, J.F., Zhang, H.: Joint patch and multi-label learning for facial action unit detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
103. Zhao, X., Dellandréa, E., Chen, L., Samaras, D.: Au recognition on 3d faces based on an extended statistical facial feature model. In: Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on, pp. 1–6. IEEE (2010)
104. Zhao, X., DellandréA, E., Zou, J., Chen, L.: A unified probabilistic framework for automatic 3d facial expression analysis based on a bayesian belief inference and statistical feature models. Image and Vision Computing **31**(3), 231–245 (2013)
105. Zhen, Q., Huang, D., Wang, Y., Chen, L.: Muscular movement model-based automatic 3d/4d facial expression recognition. IEEE Transactions on Multimedia **18**(7), 1438–1450 (2016)
106. Zhu, Y., la Torre, F.D., Cohn, J.F., Zhang, Y.J.: Dynamic cascades with bidirectional bootstrapping for spontaneous facial action unit detection. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–8 (2009)