Eurographics Workshop on 3D Object Retrieval (2010)
I. Pratikakis, M. Spagnuolo, T. Theoharis, and R. Veltkamp (Editors)

Person Independent 3D Facial Expression Recognition by a Selected Ensemble of SIFT Descriptors

Stefano Berretti[†1], Boulbaba Ben Amor[2], Mohamed Daoudi[2] and Alberto Del Bimbo[1]

[1]Dipartimento di Sistemi e Informatica, University of Firenze, Firenze, Italy
[2]Institut TELECOM, TELECOM Lille 1, LIFL (UMR 8022), France.

Abstract

Facial expression recognition has been addressed mainly working on 2D images or videos. In this paper, the problem of person-independent facial expression recognition is addressed on 3D shapes. To this end, an original approach is proposed that relies on selecting the minimal-redundancy maximal-relevance features derived from a pool of SIFT feature descriptors computed in correspondence with facial landmarks of depth images. Training a Support Vector Machine for every basic facial expression to be recognized, and combining them to form a multi-class classifier, an average recognition rate of 77.5% on the BU-3DFE database has been obtained. Comparison with competitors approaches using a common experimental setting on the BU-3DFE database, shows that our solution is able to obtain state of the art results.

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

1. Introduction

Applications in several different areas, such as computer graphics and human-machine interaction, require methods capable to automatically recognize facial expressions. The first studies on this subject date back to the late 70s with the pioneering work of Ekman [Ekm72]. In these studies, it is evidenced that the *basic facial expressions* can be categorized into six classes, representing *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*, plus the *neutral* expression. The *Facial Action Coding System* was developed by Ekman and Friesen [EF77] to code the facial expressions through the movement of face points as described by the *action units*. This work inspired many researchers to analyze facial expressions in 2D by tracking facial features and measuring the amount of facial movements in images and videos. Almost all of the methods developed in 2D use distributions of facial features as inputs to classification systems, and the outcome is one of the facial expression classes. These approaches mainly differ in the facial features selected and the classifier used to distinguish among the different facial expressions.

Recently, thanks to the increasing availability of effective devices capable to acquire high resolution 3D data, there has been a progressive shift from 2D to 3D approaches, in order to perform face recognition and facial expression recognition. The main motivation for this is that solutions based on 3D face scans feature less sensitivity to lighting conditions and pose variations. Whereas many solutions have appeared to perform 3D face recognition [KPT*07], [MBO08], [SSDK09], still few works have taken advantage of the 3D facial geometric information to perform facial expression recognition. Few years ago, the first solutions trying to automatically perform facial expression recognition based on 3D face scans have been proposed using very small databases and a limited set of facial expressions [RKVW06]. Recently, the availability of new facial expression databases, like those constructed at the *Binghamton University* (BU-3DFE) [YWS*06], and at the *Boğaziçi University* (Bosphorus database) [SAD*08], has pushed the research on this

† The authors would like to thank the region Nord-Pas de Calais, France, for a visiting Professorship to Stefano Berretti under the program Ambient Intelligence. This research was also supported partially by the project FAR3D ANR-07-SESU-004.

DOI: 10.2312/3DOR/3DOR10/047-054

topic. In particular, the BU-3DFE has become the de-facto standard database for comparing facial expression recognition algorithms. This is due to the fact that, differently from other 3D face datasets, the BU-3DFE provides a precise categorization of facial scans according to the six facial basic expressions plus the neutral one, also providing different levels of the expression intensities. Most of the works on 3D facial expression recognition can be categorized as based on: *generic facial model* or *feature classification*. In the first category, a general face model is trained with some prior knowledge, such as feature points, shape and texture variations or local geometry labels. A dense correspondence between faces is usually required to build the generic model. For example, in [RKVW06] a correspondence is established between faces with expression and their neutral pair by minimizing an energy function. A *Morphable Expression Model* (MEM) is constructed by applying *Principal Component Analysis* (PCA) to different expressions, so that new expressions can be projected into points in a low-dimensional space constructed by the eigen-expressions obtained by MEM. Expression classification is performed by comparing *Euclidean* distance among projected points in the eigen-expression space, and a recognition rate of over 97% is reported on a small and private dataset (just 25 subjects with 4 expressions per subject are included). An approach inspired by the advances of *ant colony* (ACO) and *particle swarm intelligence* (PSO) is proposed in [MMPS08]. In this work, first anatomical correspondence between faces is established using a generic 3D deformable model and 83 facial landmarks of the BU-3DFE. Then, surface points are used as a basis for classification, according to a set of classification rules which are discovered by an ACO/PSO based rule discovery algorithm. The performance of the algorithm evaluated on the BU-3DFE scored a total recognition rate of 92.3%. In [MMS08], face recognition and facial expression recognition are performed jointly by decoupling identity and expression components with a bilinear model. An elastically deformable model algorithm that establishes correspondence among a set of faces is proposed. Construction of the model relies on manually identified landmarks which are used to establish points correspondence in the training stage. Fitting these models to unknown faces enables face recognition invariant to facial expressions and facial expression recognition with unknown identity. A quantitative evaluation of the technique is conducted on the publicly available BU-3DFE face database with an overall 90.5% facial expression recognition. In [GWLT09], the shape of an expressional 3D face is approximated as the sum of a basic facial shape component, representing the basic face structure and neutral-style shape, and an expressional shape component that contains shape changes caused by facial expressions. The two components are separated by first learning a reference face for each input non-neutral 3D face then, based on the reference face and the original expressional face, a facial expression descriptor is constructed which accounts for the depth changes of rectangular regions around eyes and mouth. Average recognition rates of 71.63% and 76.22% have been reported on the BU-3DFE, not using and using a reference neutral scan for each subject, respectively.

Approaches in the second category, extract features from the 3D scan and classify them into different expressions. In [WYWS06], a geometry feature based facial expression descriptor is proposed, and the BU-3DFE database is used for the first time. The face is subdivided into 7 regions using manually annotated landmarks, and primitive surface features are classified into basic categories including *ridge*, *ravine*, *peak*, *saddle*, etc. using surface curvatures and their principal directions. They reported the highest average recognition rate of 83.6% using the primitive facial surface features and a LDA classifier. The facial expressions of *happiness* and *surprise* were reported to be the best well identified with accuracies of 95% and 90.8%, respectively. Comparison with the results obtained using the *Gabor-wavelet* and the *Topographic Context* 2D appearance feature based methods, showed that the 3D solution outperformed the 2D methods. 3D facial expression recognition on the BU-3DFE database has been also performed in [SD07]. Among the 83 facial landmarks labeling the 3D faces of the BU-3DFE, only six distance measures maximizing the differences of facial expressions are selected. These six distance values are used to form a distance vector for the representation of facial expressions as defined by the MPEG-4 *Facial Definition Parameter Set* [PF05]. The results obtained from a neural network classifier using the 3D distance vectors reaches up to 98.3% in the recognition of *surprise* facial expression, whereas the average recognition performance is 91.3%. In [TH08], a set of candidate features composed of normalized *Euclidean* distances between 83 facial landmarks of the BU-3DFE are first extracted. Then, a feature selection method based on maximizing the average relative entropy of marginalized class-conditional feature distributions is used to retain just the most informative distances. Using a regularized multi-class *AdaBoost* classification algorithm, a 95.1% average recognition rate for the six basic facial expressions is obtained on a subset of BU-3DFE. The neutral facial expression is not classified rather, as a preprocessing step, its features serve as fiducial measures that are subtracted from the features of the six basic facial expressions of the corresponding subject. The approach proposed in [VKM09] uses a modified PCA to classify facial expressions using only the shape information at a finite set of fiducial points which are extracted from the 3D neutral and expressive faces of the BU-3DFE database. The approach uses 2D texture images of the face to mark interest regions around the eyebrows, eyes, nose and mouth, and extracts facial contours in those regions with the help of an active contour algorithm. Then, these contours are uniformly sampled and the sampled points are mapped onto the 3D dataset in order to generate a shape and color descriptor of the interest-regions. An average recognition rate of 81.67% is reported.

From the analysis above, it emerges that the large part

Figure 1: *BU-3DFE: the six basic facial expressions of a sample face at the highest level of intensity.*

of existing works on 3D facial expression recognition rely on the presence of landmarks accurately identified on the face surface. Methods based on *generic facial model*, use landmarks to establish correspondences between faces in the construction of a deformable template face. Usually, these approaches are also computationally demanding due to the deformation process. Solutions based on *feature classification*, in many cases compute distances between landmarks and evaluate how these distances change between the expressional and neutral scans. The fact that several landmarks are not automatically detectable and the precision required for their positioning demand for manual annotation during enrollment of train or test scans. Furthermore, several solutions require a neutral scans for every subject in order to evaluate the differences that every expression generates with respect to the neutral reference model. This limits the applicability of many approaches.

A few recent works have shown that local descriptors computed around salient keypoints can be usefully applied to describe 3D objects. In [MBO08], a 3D keypoint detector and descriptor inspired to the *Scale Invariant Feature Transform* (SIFT) [Low04], has been designed and used to perform 3D face recognition through a hybrid 2D+3D approach that also uses the SIFT detector and descriptor to index 2D texture face images. In [MZ09], SIFT are used to detect and represent salient points in multiple 2D range images derived from 3D face models for the purpose of 3D face recognition. A similar idea is used in [OF09] to perform 3D object retrieval by visual similarity, but in this case points of a sampling grid are used, and SIFT descriptors are computed for them. Finally, SIFT descriptors have been also used in [ZTLH09] to perform 2D expression recognition from non-frontal face images.

Grounding on these studies, in this work we propose to use local descriptors to perform 3D facial expression recognition. Differently from existing approaches, we exploit the local characteristics of the face around a limited set of points to perform 3D recognition of facial expressions. The proposed approach is based on computing SIFT descriptors around a set of facial landmarks and using them as feature vector to represent the face. Before to perform classification, a feature selection approach is used to identify a subset of features with *minimal-redundancy and maximal-relevance* among the large set of features extracted with SIFT. The set of selected features is finally used to feed a set of classifiers based on *Support Vector Machines* (SVM). As emerges from the experimental evaluation, the proposed approach is capable to achieve state of the art results on the BU-3DFE, without using neutral scans, and just relying on few landmarks. In the current implementation, some landmarks are given manually and some others are obtained automatically. In perspective, all of them can be automatically identified.

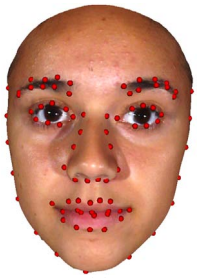
The rest of the paper is organized as follows: In Sect.2, the salient characteristics of the BU-3DFE are described in order to provide some of the choices that guide our solution. In Sect.3, the main characteristics of the SIFT descriptor are summarized, and its adaptation to our case is presented. The feature selection approach used to reduce the set of SIFT features, and the SVM based classification of the selected features are addressed in Sect.4. Experiments carried out with the proposed approach, with results and comparative evaluation are reported in Sect.5. Finally, discussion and conclusions are given in Sect.6.

2. The BU-3DFE database

The BU-3DFE database was recently constructed at the *Binghamton University*. It was designed to provide 3D facial scans of a large population of different subjects each showing a set of prototypical emotional states at various levels of intensities. There are a total of 100 subjects in the database, divided between female (56 subjects) and male (44 subjects). The subjects are well distributed across different ethnic groups or racial ancestries, including *White*, *Black*, *East-Asian*, *Middle-East Asian*, *Hispanic-Latino*, and others. During the acquisition, each subject was asked to perform the neutral facial expression as well as the six basic facial expressions defined by Ekman, namely *anger* (AN), *disgust* (DI), *fear* (FE), *happiness* (HA), *sadness* (SA), and *surprise* (SU). Each facial expression has four levels of intensities, namely *low*, *middle*, *high* and *highest*, except the neutral facial expression that has only one intensity level. Thus, there are 25 3D facial expression scans for each subject, resulting in 2500 3D facial expression scans in the database. As an example, Fig.1 shows the six basic facial expressions of a sample 3D face at the highest level of intensity.

Each of the 3D facial expression scan is also associated with a raw 3D face mesh, a cropped 3D face mesh, a pair of

texture images with two-angle of view (about $+45$ and -45^o away from the face frontal normal), a frontal-view texture image, a set of 83 landmark points, and a facial pose vector. These data give a complete 3D description of a face under a specific facial expression. In this paper, we only use the cropped 3D face mesh model and the 83 landmark points marked on it as shown in Fig.2(a). It can be observed that the landmarks are mainly distributed around the most distinguishing traits of the face, that is, *eyes*, *eyebrows*, *nose* and *mouth*, as summarized in Fig.2(b). A more detailed description of the BU-3DFE database can be found in [YWS*06].



(a)

	#points	*indices*
right eye	8	1 - 8
left eye	8	9 - 16
right eyebrow	10	17 - 26
left eyebrow	10	27 - 36
nose	12	37 - 48
mouth	20	49 - 68
face boundary	15	69 - 83

(b)

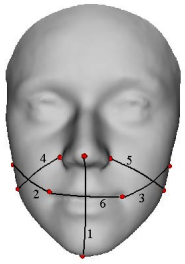
Figure 2: *BU-3DFE: (a) The 83 facial landmarks evidenced on a neutral expression 3D face scan; (b) The number of landmarks and their indices grouped by different parts of the face.*

3. SIFT feature representation of range facial images

In order to capture salient features that characterize different facial expressions in 3D, we followed the idea to use local descriptors around landmarks of the face. The SIFT feature extraction algorithm has been used for this purpose. According to its original formulation, SIFT includes a *keypoint detector* and a *feature extractor*. A detailed explanation of the SIFT keypoint detection and feature extraction is given by Lowe [Low04]. In the following, we summarize the main idea of SIFT and its adaptation to our context.

SIFT have been defined for 2D gray-scale images and cannot be directly applied to 3D face scans. However, the 3D information of scanned faces can be captured through *range images* that use the gray-scale of every pixel to represent the depth of a scan. According to this, SIFT keypoint detector could be applied to the range images in order to extract image keypoints. Actually, by definition SIFT keypoints are mainly located at corner points of an image. As a consequence, SIFT keypoints detected on range images can be too few or no keypoints can be detected in regions of the face that are relevant to discriminate between facial expressions (like the cheek). Due to this, facial landmarks located in important morphological regions of the face and identified using semi-automatic solutions (i.e., manually and automatically detected landmarks are considered) are used as keypoints instead. For these landmarks, the SIFT feature extractor is run so as to obtain a *SIFT descriptor*. Briefly, a SIFT descriptor for a small image patch, for example of size 4×4, is computed from the gradient vector histograms of the pixels in the patch. There are 8 possible gradient directions, and therefore the total size of the SIFT descriptor is $4 \times 4 \times 8 = 128$ elements. This descriptor is normalized to enhance invariance to changes in illumination (not relevant in the case of range images), and transformed in other ways to ensure invariance to scale and rotation as well. These properties make the SIFT descriptor capable to provide compact and powerful local representation of the range image and, as a consequence, of the face surface.

In the case of BU-3DFE, we performed some steps to derive the facial landmarks and to transform 3D face scans into range images. The 83 landmarks provided with every 3D face scan (see Sect.2) have been used as keypoints. However, many of these landmarks are not automatically detectable, both in 2D or 3D, and do not include regions of the face whose local information can be useful for expression recognition (like the cheek). So, we automatically located 85 additional landmarks on the face to be used as keypoints. These are selected by uniformly sampling the lines connecting on the face surface some fiducial points, and have the advantage to be automatically located just starting from the start and end points of the lines. The lines to which the additional landmarks belong to are shown in Fig.3(a), whereas their indices are reported in the table of Fig.3(b).



(a)

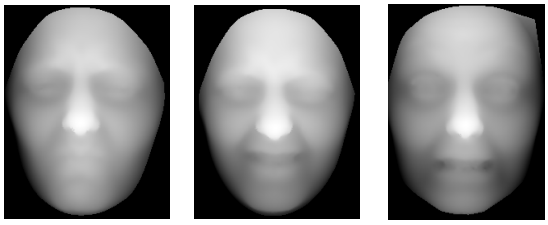
line	*#points*	*indices*
1	12	84 - 95
2	16	96 - 111
3	16	112 - 127
4	16	128 - 143
5	16	144 - 159
6	9	160 - 168

(b)

Figure 3: *(a) The six lines along which the 85 additional keypoints are located; (b) The number of landmarks and their indices grouped according to the surface line they belong to.*

Once the overall set of landmarks has been identified, face scans of the BU-3DFE database are transformed to range images considering a frontal view of the scan. Before to extract the range images, some preprocessing was also applied to the face scans in the dataset. In particular, a sphere of radius 130*mm* centered on the nose tip was used to crop the 3D face (the nose tip has been detected using the approach in [MBO07]). Then, spikes in the 3D face were removed using median filtering in the z-coordinate. Holes were filled using cubic interpolation, and 3D scans were re-sampled on an uniform square grid at 0.7*mm* resolution. Finally, we re-

mapped the landmarks provided on the 3D scans to the corresponding range images. As an example, Fig.4 shows the range images derived from the 3D face scans of a same subject under three different facial expressions.



(a) (b) (c)

Figure 4: *Range images derived from 3D scans of a same subject, for the expressions (highest level of intensity): (a) anger; (b) disgust; (c) fear.*

Finally, SIFT descriptors have been extracted using the following setting (the publicly available implementation of SIFT in [VF08] has been used in our experimentation):

- For each range image, the $83+85$ landmarks have been used as keypoints where to compute SIFT descriptors. If the keypoints are occluded by the face, we simply compute the SIFT descriptor in the corresponding positions of the image;
- SIFT descriptors are computed at scale equal to 3, whereas the preferred SIFT orientation angle is computed;
- The orientation histograms of 4×4 sample regions of each keypoint are used to calculate the SIFT descriptor. By computing the 128-dimensional SIFT descriptor at the $83+85$ sparse keypoints, a 21504-dimensional feature vector is obtained to represent each range image.

To reduce the dimensionality and improve the significativeness of the features, only the features with *maximal-relevance* and *minimal-redundancy* have been selected using the feature selection analysis reported in Sect.4.

4. Feature selection

Feature selection is mainly motivated by the *dimensionality curse*, which states that in presence of a limited number of training samples, each one represented as a feature vector in R^n, the mean accuracy does not always increase with vector dimension (n). Rather, the classification accuracy increases until a certain dimension of the feature vector, and then decreases. In other words, the higher the dimensionality of the feature space, the higher the number of training samples required to achieve the same classification accuracy. Therefore, the challenge is to identify m out of the n features which will yield similar, if not better, accuracies as compared to the case in which all the n features are used in a classification task.

In the proposed analysis, feature selection is performed using the *minimal-redundancy maximal-relevance* (mRMR) model [PLD05]. For a given classification task, the aim of mRMR is to select a subset of features by taking into account the ability of features to identify the classification label, as well as the redundancy among the features. These concepts are defined in terms of the *mutual information* between features.

Given two discrete random variables x and y, taking values in $\{s_i\}_{i=1}^N$, their joint probability $P(x,y)$ and the respective marginal probabilities $P(x)$ and $P(y)$, the mutual information between x and y is defined as the difference between the *Shannon*'s entropy of x, and the conditional entropy of x given y, that is: $I(x,y)=H(x)-H(x|y)$, where the entropy is used as measure of the uncertainty of a random variable. In practice, this expression states that if from the uncertainty of x is subtracted the uncertainty of x once y is known, the information (in bits) that the variable y provides about x is obtained. According to this, mutual information provides a measure of the dependency of variables, and can also be computed as:

$$I(x,y)=\sum_{i=1}^{N}\sum_{j=1}^{N}P(s_i,s_j)\log\frac{P(s_i,s_j)}{P(s_i)P(s_j)}. \qquad (1)$$

The work in [PLD05] proposes to jointly maximize the dependency between a feature variable x_i and the classification variable l and minimize the dependency between pairs of feature variables x_i, x_j. Thus, the task of feature selection is posed as selecting from the complete set of n features S_n, a subset S_m of $m<n$ features that maximizes:

$$\frac{1}{m}\sum_{x_i\in S_m}I(x_i,l)-\frac{1}{\binom{m}{2}}\sum_{x_i,x_j\in S_m}I(x_i,x_j). \qquad (2)$$

This expression takes into account the relevance of features with the class label while penalizing redundancy among them. Since the search space of subsets of m elements in R^m is too big to be explored in practice, S_m is determined incrementally by means of a forward search algorithm. Having a subset S_{m-1} of $m-1$ features, the feature $x_i\in\{S_n-S_{m-1}\}$ that determines a subset $\{x_i,S_{m-1}\}$ maximizing Eq.(2) is added. It can be shown that this nested subset strategy is equivalent to iteratively optimize the following condition:

$$\max_{x_i\in S_n-S_{m-1}}\left(I(x_i,l)-\frac{1}{m-1}\sum_{x_j\in S_{m-1}}I(x_j,x_i)\right). \qquad (3)$$

Experiments in [PLD05] show that for subsets of more than 20 features, the S_m obtained with this method achieves more accurate classification performances than the subset obtained by maximizing the $I(S_m,l)$ value (that is, the mutual information between the whole subset of variables and the classification label l), while the required computation cost is significantly lower.

4.1. SVM classification of relevant SIFT features

In our approach, the mRMR algorithm is applied to the set of 21504-dimensional feature vectors representing the faces $v_f = (f_1, \ldots, f_{21504})$. Each vector is constructed by concatenating the 128-dimensional SIFT descriptors for the face landmarks, orderly from 1 to 83+85. A data discretization is applied to the vectors as preprocessing step. This is obtained by computing the mean value μ_k and the standard deviation σ_k for every feature f_k. Then, discretized values $\hat{f}_k$ are obtained according to the following rule:

$$\hat{f}_k = \begin{cases} 2 & if \ f_k < \mu_k - \alpha \cdot \sigma_k \\ 3 & if \ \mu_k - \alpha \cdot \sigma_k \leq f_k \leq \mu_k + \alpha \cdot \sigma_k \\ 4 & if \ f_k > \mu_k + \alpha \cdot \sigma_k, \end{cases} \quad (4)$$

being α a parameter that regulates the width of the discretization interval (it is equal to 0.2 in our experiments).

The overall set of discretized feature vectors is used to feed the mRMR algorithm so as to determine the features which are more relevant in discriminating between different facial expressions of 3D face scans of different subjects. The facial expression recognition problem is a multi-classification task that, in our approach, is faced as a combination of separated instances of *one*-vs-*all* classification subproblems. For each subproblem, face scans showing one expression are assumed as targets (positive examples), whereas all the other scans with any different expression are considered as negative examples. Repeatedly, the target expression is changed among the six basic expressions provided by the BU-3DFE, so that the sets of positive and negative examples change. Due to this, mRMR feature selection is performed independently for every classification subproblem. In general, this results into different features providing the minimal-redundancy and maximal-relevance for the purpose of discriminating across different facial expressions. Then, just the most relevant features identified for every expression are retained from the original feature vectors in order to train the classifiers. This results into vectors $v_{\hat{f}}^{expr} = (\hat{f}_{p_1}, \ldots, \hat{f}_{p_{Nexpr}})$, where $p_1, \ldots, p_{Nexpr}$ are the indices of the features components selected in the original vector, and the subscript the label of a particular expression.

The selected features are then used to perform facial expression recognition using a maxima rule between six *one*-vs-*all* SVM classifiers, each with a radial basis function kernel of standard deviation equal to one (the publicly available SVMLight implementation of SVMs has been used: `http://svmlight.joachims.org/`).

5. Results

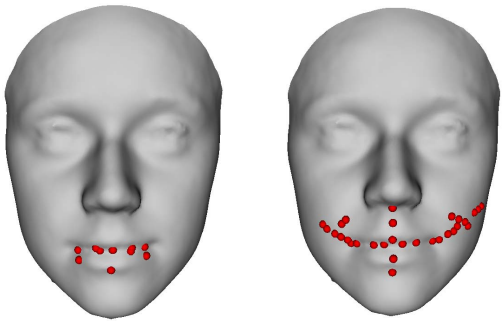
Experiments on the BU-3DFE database have been conducted using similar setup as in [GWLT09]. In particular, we performed a series of experiments in each of which 60 randomly selected subjects are used with the two highest-intensities scans for each of the six basic facial expressions (i.e., each experiment includes 720 scans). The random selection of the subjects approximately guarantees that, in each experiment, the person and gender independency are preserved, and a good distribution of the subjects across the various ethnic groups. In each experiment, six *one*-vs-*all* SVM classifiers, one for each expression, are trained and tested using the feature vectors $v_{\hat{f}}^{expr}$ and *10-fold cross validation*. According to this, the 60 subjects are split into ten subsets, each containing 6 subjects. Of the 10 subsets, one subset is retained to test the model, and the remaining 9 subsets are used as training data, that is, the training set contained 54 subjects (648 scans), and the test set contained 6 subjects (72 scans). The ratio between positive and negative examples in the train and test subsets is equal to the ratio existing in the original dataset. Using 10-fold cross validation, training is repeated 10 times, with each of the 10 subsets used exactly once as the test data. Finally, the results from the ten steps are averaged to produce a single estimation of the performance of the classifier for the experiment. In this way, all observations are used for both training and test, and each observation is used for test exactly once. However, as pointed out in [GWLT09], since average recognition accuracies can vary from experiment to experiment, in order to permit a fair generalization and obtain stable expression recognition accuracies, we run 100 independent experiments and averaged the results (1000 train and test sessions in total).

According to the feature selection analysis of Sect.4, just the most relevant SIFT features identified with mRMR are used to perform 3D facial expression recognition. Table 1 summarizes, for the six basic expressions, the outcomes of mRMR by using the pair (l, r), where: l is the landmark index according to the numbering reported in Fig.2(b), and Fig.3(b); and r represents the relevance (given in percentage) of the feature selected for the l landmark. The relevance value is obtained as the mutual information value returned by the mRMR algorithm, normalized to the mutual information of the most important feature. Actually, it should be noted that mRMR can select and use for classification, none, one or more than one of the 128 features of the SIFT descriptor at a particular landmark. From the table, it can be observed that landmarks from which the most relevant SIFT features are extracted vary across expressions. In addition, the six values reported in each column show as the relevance decreases, so that just a few features among the 21504-dimensional SIFT feature vector are sufficient to perform expression recognition. In particular, in the recognition experiments we found optimal results using 12, 12, 16, 8, 14, and 12 features, respectively, for the *anger*, *disgust*, *fear*, *happy*, *sad*, and *surprise* expressions. It is relevant to note that, among all expressions, SIFT features of 40 landmarks are used for recognition, and these landmarks do not include those provided by BU-3DFE around eyes, eyebrows and nose (see Fig.5). Furthermore, just 10 of the BU-3DFE landmarks belong to this set (they are all located on the mouth). Since the additional landmarks that we introduced can be detected au-

rank	*Anger* **l,r**	*Disgust* **l,r**	*Fear* **l,r**	*Happiness* **l,r**	*Sad* **l,r**	*Surprise* **l,r**
1	64, 100.0%	102, 100.0%	125, 100.0%	153, 100.0%	160, 100.0%	90, 100.0%
2	68, 88.5%	64, 96.2%	96, 94.4%	153, 99.1%	64, 97.3%	99, 96.8%
3	68, 88.0%	101, 94.6%	60, 94.4%	134, 98.9%	160, 96.9%	89, 96.1%
4	166, 87.6%	165, 93.5%	168, 91.3%	135, 98.5%	161, 96.8%	62, 95.5%
5	64, 84.3%	153, 90.6%	113, 91.1%	116, 95.6%	168, 96.0%	64, 95.3%
6	62, 83.4%	90, 87.6%	165, 90.2%	115, 95.2%	61, 95.4%	91, 94.8%

Table 1: *BU-3DFE: Most relevant features for every expression according to mRMR. Each cell reports the pair (l,r), where l is the landmark number according to the numbering of Fig.2(b) and Fig.3(b), and r is the corresponding relevance (in percentage). Values in each column are ordered by decreasing relevance scores.*

tomatically, this opens the way to a completely automatic expression recognition approach. In addition, these results seem to indicate that, with our approach, expression recognition can be performed mainly considering the mouth and cheek regions.



(a) (b)

Figure 5: *The landmarks for which at least one feature of the SIFT descriptor is selected by mRMR for expression classification: (a) The 10 landmarks selected from the 83 manually located in the BU-3DFE; (b) The 30 landmarks selected from the additional 85 automatically located.*

	An	*Di*	*Fe*	*Ha*	*Sa*	*Su*
An	**81.7%**	0.9%	3.3%	4.2%	8.1%	1.7%
Di	3.3%	**73.6%**	2.6%	7.8%	0.0%	12.6%
Fe	2.6%	14.5%	**63.6%**	9.2%	0.8%	9.2%
Ha	0.9%	4.5%	6.9%	**86.9%**	0.8%	0.0%
Sa	30.1%	0.0%	0.0%	3.4%	**64.6%**	1.8%
Su	1.8%	1.7%	1.7%	0.0%	0.0%	**94.8%**

Table 2: *BU-3DFE: Average confusion matrix.*

Using these mRMR features, the results of 3D facial expression classification are reported in Tab.2, considering the average *confusion matrix* as performance measure. Rows of the table are the true expressions to classify, whereas columns represent the results of the classification. It can be observed that some expressions (like *happiness* and *surprise*) are recognized with very high accuracies, whereas it results more difficult to identify *sadness* (high confusion with *anger*) and *fear* (which is confused mainly with *disgust*). The overall recognition rate is equal to 77.54%. Looking to Fig.1, the difficulty to discriminate between *sadness* and *anger* expressions can be motivated by the fact that these two expressions are very similar and can be discriminated mainly looking to the eyes region. However, eye regions typically produce noisy data when acquired with 3D scanners, and SIFT features are not stable for the corresponding landmarks. As a consequence, the SIFT features of the eyes are not selected by mRMR, and just those around the mouth are not sufficiently discriminating.

Finally, in Tab.3 the results of our approach are compared against those reported in [GWLT09] on a same experimental setting. In this setting, the average recognition accuracies are computed by performing 100 independent runs, each including 10-fold cross validation on the two highest intensities scans of 60 randomly selected subjects (720 scans per independent experiment). For the approaches in [SD07], [WYWS06] and [TH08] results are replicated from those reported in [GWLT09]. Some differences between the approaches listed in the table should be noted: Soyel [SD07] uses distances between manually identified landmarks (11 in total); Tang [TH08] uses both distances between manual landmarks (83 in total), and neutral scans to normalize distances; Wang [WYWS06] uses manual landmarks (64 in total) to segment face regions; Gong [GWLT09] obtains its best results subtracting neutral scans from depth region masks of the eyes and mouth. In comparison, our approach does not use neutral scans, but just relies, after mRMR, on 10 manually detected landmarks selected from those provided by the BU-3DFE, plus 30 landmarks selected among the additional that are located automatically. In particular, it can be observed that our approach outperforms other solutions, with larger differences with respect to works that do not use neutral scans.

6. Conclusions

In this paper, we investigate the problem of person independent facial expression recognition from 3D facial scans. We propose an original automatic feature selection approach

	This work	*Gong*	*Wang*	*Soyel*	*Tang*
AVG	**77.54%**	76.22%	61.79%	67.52%	74.51%

Table 3: *BU-3DFE: Comparison of this work with respect to the works of Gong et al. (Gong) [GWLT09], Wang et al. (Wang) [WYWS06], Soyel et al. (Soyel) [SD07], and Tang et al. (Tang) [TH08]. The average (AVG) expression recognition rates computed on all the six expressions, and all the independent experiments are reported.*

based on minimizing the redundancy between features, maximizing, at the same time, their relevance in terms of mutual information, and apply it to a complete pool of SIFT descriptors computed on a set of facial landmarks given on 3D face scans. Using a multi-class SVM classification, and a large set of experiments an average facial expression recognition rate of 77.54% is obtained for the six basic facial expressions, on the publicly available BU-3DFE database.

The experimental evidence resulting from this work, suggests that a limited set of landmarks are sufficient for the computation of SIFT descriptors with good classification capability for different facial expressions. According to this, as future work we will investigate the automatic detection of a reduced set of landmarks so as to define a completely automatic approach. Further experiments will be performed to evaluate the robustness of the approach when applied to the BU-3DFE facial expression scans with the lower and medium levels of expression intensities.

References

[EF77] EKMAN P., FRIESEN W. V.: *Manual for the the Facial Action Coding System*. Consulting Psychologist Press, Palo Alto, CA, 1977. 1

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

[GWLT09] GONG B., WANG Y., LIU J., TANG X.: Automatic facial expression recognition on a single 3d face by exploring shape deformation. In *Proc. ACM Int. Conf. on Multimedia* (Beijing, China, Oct. 2009), pp. 569–572. 2, 6, 7, 8

[KPT*07] KAKADIARIS I. A., PASSALIS G., TODERICI G., MURTUZA N., LU Y., KARAMPATZIAKIS N., THEOHARIS T.: Three-dimensional face recognition in the presence of facial expressions: An annotated deformable approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 4 (Apr. 2007), 640–649. 1

[Low04] LOWE D.: Distinctive image features from scale-invariant key points. *Int. Journal of Computer Vision 60*, 2 (Nov. 2004), 91–110. 3, 4

[MBO07] MIAN A. S., BENNAMOUN M., OWENS R.: An efficient multimodal 2d-3d hybrid approach to automatic face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 11 (Nov. 2007), 1927–1943. 4

[MBO08] MIAN A. S., BENNAMOUN M., OWENS R.: Keypoint detection and local feature matching for textured 3d face recognition. *Int. Journal of Computer Vision 79*, 1 (Aug. 2008), 1–12. 1, 3

[MMPS08] MPIPERIS I., MALASSIOTIS S., PETRIDIS V., STRINTZIS M. G.: 3d facial expression recognition using swarm intelligence. In *Proc. IEEE Int. Conf. on Acoustic, Speech, and Signal Processing* (Mar. 2008), pp. 2133–2136. 2

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

[MZ09] MAYO M., ZHANG E.: 3d face recognition using multiview key point matching. In *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance* (Genoa, Italy, Sept. 2009), pp. 290–295. 3

[OF09] OHBUCHI R., FURUYA T.: Scale-weighted dense bag of visual features for 3d model retrieval from a partial view 3d model. In *Proc. Workshop on Search in 3D and Video* (Kyoto, Japan, Sept. 2009). 3

[PF05] PANDZIC I., FORCHHEIMER R.: *MPEG-4 Facial Animation: the Standard, Implementation and Applications*. Wiley, 2005. 2

[PLD05] PENG H., LONG F., DING C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 8 (Aug. 2005), 1226–1238. 5

[RKVW06] RAMANATHAN S., KASSIM A., VENKATESH Y. V., WAH W. S.: Human facial expression recognition using a 3d morphable model. In *Proc. IEEE Int. Conf. on Image Processing* (Atlanta, GA, Oct. 2006), pp. 661–664. 1, 2

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

[SD07] SOYEL H., DEMIREL H.: Facial expression recognition using 3d facial feature distances. In *Proc. Int. Conf. on Image Analysis and Recognition* (Aug. 2007), pp. 831–838. 2, 7, 8

[SSDK09] SAMIR C., SRIVASTAVA A., DAOUDI M., KLASSEN E.: An intrinsic framework for analysis of facial surfaces. *Int. Journal of Computer Vision 82*, 1 (Apr. 2009), 80–95. 1

[TH08] TANG H., HUANG T. S.: 3d facial expression recognition based on automatically selected features. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition* (Anchorage, AK, June 2008), pp. 1–8. 2, 7, 8

[VF08] VEDALDI A., FULKERSON B.: VLFeat: An open and portable library of computer vision algorithms, 2008. `http://www.vlfeat.org/`. 5

[VKM09] VENKATESH Y. V., KASSIM A. A., MURTHY O. V. R.: A novel approach to classification of facial expressions from 3d-mesh datasets using modified pca. *Pattern Recognition Letters 30*, 12 (Sept. 2009), 1128–1137. 2

[WYWS06] WANG J., YIN L., WEI X., SUN Y.: 3d facial expression recognition based on primitive surface feature distribution. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition* (June 2006), vol. 2, pp. 1399–1406. 2, 7, 8

[YWS*06] YIN L., WEI X., SUN Y., WANG J., ROSATO M.: A 3d facial expression database for facial behavior research. In *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition* (Southampton, UK, Apr. 2006), pp. 211–216. 1, 4

[ZTLH09] ZHENG W., TANG H., LIN Z., HUANG T. S.: A novel approach to expression recognition from non-frontal face images. In *Proc. IEEE Int. Conf. on Computer Vision* (Kyoto, Japan, Sept. 2009), pp. 1901–1908. 3