A Set of Selected SIFT Features for 3D Facial Expression Recognition

Stefano Berretti, Alberto del Bimbo, Pietro Pala, Boulbaba Ben Amor, Daoudi Mohamed

2010 International Conference on Pattern Recognition

A Set of Selected SIFT Features for 3D Facial Expression Recognition

Stefano Berretti, Alberto Del Bimbo, Pietro Pala
Dipartimento di Sistemi e Informatica
Università di Firenze
Firenze, Italy
{*berretti,delbimbo,pala*}*@dsi.unifi.it*

Boulbaba Ben Amor, Mohamed Daoudi
Institut TELECOM/TELECOM Lille 1
LIFL, Université de Lille 1
Lille, France
{*boulbaba.benamor,mohamed.daoudi*}*@telecom-lille1.fr*

***Abstract*—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 computes SIFT descriptors on a set of facial landmarks of depth images, and then selects the subset of most relevant features. Using SVM classification of the selected features, an average recognition rate of 77.5% on the BU-3DFE database has been obtained. Comparative evaluation on a common experimental setup, shows that our solution is able to obtain state of the art results.**

***Keywords*-3D facial expression recognition; feature selection; svm classification;**

I. Introduction

Methods capable to automatically recognize facial expressions are required in several different areas, such as *computer graphics* and *human-machine interaction*. Early work on this subject analyzed facial expressions in 2D images and videos by tracking facial features and measuring the amount of facial movements. These approaches were inspired by the pioneering work of Ekman [1], that proposed a categorization of *basic facial expressions* into six classes representing *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*. Recently, the increasing availability of effective devices capable to acquire high resolution 3D data, has determined a progressive shift from 2D to 3D approaches. This is mainly motivated by the fact that deformations of facial morphology due to expression changes in 3D are expected to be more easily detectable than in 2D. Moreover, the availability of new 3D facial expression databases, like that constructed at the *Binghamton University* (BU-3DFED) [11], has further pushed the research on 3D facial expression recognition. In general, previous work 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 constructed using dense correspondence and some prior knowledge on a set of training faces. For example, in [6] an elastically deformable model algorithm that establishes correspondence among a set of faces is proposed. Fitting the elastically deformable model to unknown faces enables face recognition invariant to facial expressions and facial expression recognition with unknown identity. In [2], the shape of an expressional 3D face is approximated as the sum of a basic facial shape component, and an expressional shape component. The two components are separated by learning a reference face for each input non-neutral 3D face, then a facial expression descriptor is constructed which accounts for the depth changes of rectangular regions around eyes and mouth. Approaches in the second category, extract features from 3D facial scans and classify them into different expressions. In [10], the face is subdivided into regions using manually annotated landmarks, and surface curvatures and their principal directions in the regions are categorized and used to recognize different facial expressions. Comparison with results obtained on the BU-3DFED using the *Gabor-wavelet* and the *Topographic Context* 2D appearance feature based methods, showed that the 3D solution outperformed the 2D methods. In [8], six *Euclidean* distances between some facial landmarks labeling the 3D faces of the BU-3DFED, have been selected and used to form a distance vector and train a neural network classifier. In [9], a set of normalized *Euclidean* distances between 83 facial landmarks of the BU-3DFED are extracted. Then, maximizing the average relative entropy of marginalized class-conditional feature distributions, just the most informative distances are retained and classified using a regularized multi-class *AdaBoost* algorithm.

A few recent works have shown that salient keypoints and local descriptors can be effectively used to describe 3D objects. In [5], a 3D keypoint detector and descriptor inspired to the *Scale Invariant Feature Transform* (SIFT) [3], has been designed and used to perform 3D face recognition. In [4], 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. In 2D, SIFT descriptors have been also used to perform 2D expression recognition from non-frontal face images [12].

In this work we propose to use local face descriptors to perform 3D facial expression recognition. Differently from existing approaches, we exploit the local characteristics of the face by computing SIFT descriptors around a small set of facial landmarks identified on range images, and using them as feature vector to represent the face. Then, a feature selection approach is used to identify the subset of most features features among the set of SIFT features. The selected features are finally used to feed a set of *Support*

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.1002

Vector Machines (SVM) classifiers. Experimentation on the BU-3DFED, shows that the proposed approach ia capable to achieve state of the art results, without using neutral scans and just relying on few landmarks that, in perspective, can be automatically identified.

The rest of the paper is organized as follows: In Sect.II, the salient features of the BU-3DFED are summarized. In Sect.III, the SIFT descriptor is briefly presented with its adaptation to our case. 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.IV. Results obtained with the proposed approach and a comparative evaluation are reported in Sect.V. Finally, discussion and conclusions are given in Sect.VI.

II. The BU-3D facial expression database

The BU-3DFED has been recently 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 (56 female and 44 male), well distributed across different ethnic groups. During the acquisition, each subject was asked to perform the neutral facial expression as well as the six basic facial expressions, i.e., *anger* (AN), *disgust* (DI), *fear* (FE), *happiness* (HA), *sadness* (SA), and *surprise* (SU). Each facial expression has four levels of intensities (*low*, *middle*, *high* and *highest*), except the neutral facial expression that has only one intensity level (2500 3D facial expression scans in total). Each 3D face scan is also associated with a cropped 3D face mesh, and a set of 83 landmarks manually located on the *eyebrows*, *eyes*, *nose*, *mouth* and *face contour*. As an example, Fig.1 shows the six basic facial expressions of a sample 3D face at the highest level of intensity.

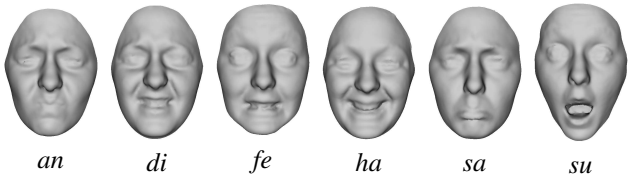
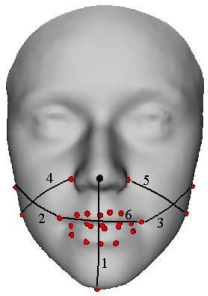


Figure 1. Basic facial expressions of a sample face (highest level of intensity).

III. SIFT descriptors of range facial images

In order to capture salient features characterizing different facial expressions in 3D, we followed the idea to use local descriptors around landmarks of the face. To this end, we used the *SIFT feature extraction* algorithm, adapting it to our particular case. In the original formulation [3], SIFT has 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, facial landmarks located in important morphological regions of the face are used as keypoints (so by-passing the SIFT *keypoint detector*), and the SIFT feature extractor is run on these keypoints so as to obtain a SIFT *descriptor*. Briefly, a SIFT descriptor of 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, and transformed to ensure invariance to scale and rotation as well. These properties make the SIFT descriptor capable to provide compact and powerful local representations of the range image and, consequently, of the face surface.



(a)

component	*#points*
nose (BU-3DFED)	2
mouth (BU-3DFED)	20
face contour (BU-3DFED)	5
line 1	12
line 2	16
line 3	16
line 4	16
line 5	16
line 6	9

(b)

Figure 2. (a) The subset of the BU-3DFED landmarks, and the lines along which the 85 additional landmarks are located; (b) Number of landmarks for every face component they belong to.

In the case of BU-3DFED, we performed some steps to derive the facial landmarks and to transform 3D face scans into range images. The BU-3DFED, provides 83 landmarks with every 3D face, but in a preliminary experimentation we found that SIFT descriptors extracted on many of these are not significant to discriminate between different facial expressions. So, we considered a subset including 20 landmarks on the mouth, 2 on the nose, and 5 on the face contour (27 in total). Further, we identified 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 subset of BU-3DFED landmarks, and the lines to which the additional landmarks belong to are shown in Fig.2(a), whereas their number and grouping are reported in the table of Fig.2(b).

Once the overall set of landmarks has been identified, face scans of the BU-3DFED 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 $130mm$ centered on the nose tip (automatically detected and included in the landmarks set) was used to crop the 3D face.

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 $0.7mm$ resolution. Finally, we re-mapped the landmarks identified on the 3D scans to the corresponding range images.

SIFT descriptors have been extracted using the following setting: *(i)* For each range image, the 27+85 landmarks have been used as keypoints where to compute SIFT descriptors; *(ii)* SIFT descriptors are computed at scale equal to 3, whereas the preferred SIFT orientation angle is computed; *(iii)* 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 27+85 sparse keypoints, a 14336-dimensional feature vector is obtained to represent each range image.

To reduce the dimensionality and improve the significativeness of the description, just the most relevant features have been selected and classified.

IV. Selection and classification of relevant SIFT features

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. Therefore, the challenge is to identify m out of the n features which 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 [7]. 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 "*l*", as well as the redundancy among the features, according to the following equation:

$$\max_{x_i \in S_n - S_{m-1}} \Big(I(x_i, l) - \frac{1}{m-1} \sum_{x_j \in S_{m-1}} I(x_j, x_i) \Big). \quad (1)$$

In this equation, having a subset S_{m-1} of $m-1$ features variables x_i, the feature $x_i \in \{S_n - S_{m-1}\}$ that determines a subset $\{x_i, S_{m-1}\}$ maximizing the relevance of features with the class label whereas penalizing redundancy among them, is added. In the same equation, $I(.,.)$ measures the *mutual information* between two discrete random variables (i.e., the difference between the *Shannon*'s entropy of the first variable, and the conditional entropy of the first variable given the second one).

In our approach, the mRMR algorithm is applied to the set of 14336-dimensional feature vectors representing the faces $v_f = (f_1, \ldots, f_{14336})$. Each vector is constructed by concatenating the 128-dimensional SIFT descriptors for the face landmarks, orderly from 1 to 27+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:

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

where the α parameter (set equal to 0.2 in our experiments) regulates the width of the discretization interval.

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-3DFED, so that the sets of positive and negative examples change. Due to this, mRMR feature selection is performed independently for every classification subproblem. 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. In particular, in the expression recognition experiments we found optimal results using 12, 12, 16, 8, 14, and 12 features out of the 14336, respectively, for the *anger*, *disgust*, *fear*, *happy*, *sad*, and *surprise* expressions. 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.

V. Experimental results

Experiments on the BU-3DFED have been performed using a similar setup as in [2]. In particular, since average recognition accuracies can vary from experiment to experiment, in order to permit a fair generalization and obtain stable expression recognition accuracies, we have run 100 independent experiments and averaged the results. In every experiment, 60 randomly selected subjects are considered, each with the two scans of highest-intensities expressions for each of the six basic facial expressions (i.e., 720 scans per experiment). The random selection of the subjects approximately guarantees that the person and gender independency are preserved, and a good distribution of the subjects across

the various ethnic groups is achieved. In each experiment, six *one*-vs-*all* SVM classifiers, one for each expression, are trained and tested using the selected features as determined in Sect.IV, and *10-fold cross validation* (1000 train and test sessions in total).

Table I
AVERAGE CONFUSION MATRIX.

	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%**

Using the mRMR features, the results of 3D facial expression classification are reported in Tab.I, considering the average *confusion matrix* as performance measure. 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%.

Finally, in Tab.II the results of our approach are compared against those reported in [2]. In fact, in Gong et al. (Gong) [2], the performance of the approaches in Wang et al. (Wang) [10], Soyel et al. (Soyel) [8], and Tang et al. (Tang) [9], are obtained on a same experimental setting. The set up used in this work is more challenging in that, differently from [2] where 60 subjects were selected, in our case the 60 selected subjects varied randomly from experiment to experiment. 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, like [8] and [10].

Table II
AVERAGE (AVG) EXPRESSION RECOGNITION RATES FOR OUR APPROACH AND THE WORKS IN [2], [10], [8], [9].

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

VI. CONCLUSIONS

In this paper, we investigate the problem of person independent facial expression recognition from 3D facial scans. We propose an original feature selection approach based on minimizing the redundancy between features, maximizing, at the same time, their relevance in terms of mutual information, and apply it to 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 on the publicly available BU-3DFED, an average facial expression recognition rate at the state of the art is obtained.

ACKNOWLEDGMENTS

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.

REFERENCES

[1] P. Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska Symposium on Motivation*, pages 207–283, Lincoln, NE, 1972.

[2] B. Gong, Y. Wang, J. Liu, and X. Tang. Automatic facial expression recognition on a single 3d face by exploring shape deformation. In *Int. Conf. on Multimedia*, pages 569–572, Beijing, China, Oct. 2009.

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

[4] M. Mayo and E. Zhang. 3d face recognition using multiview key point matching. In *Int. Conf. on Advanced Video and Signal Based Surveillance*, pages 290–295, Genoa, Italy, Sept. 2009.

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

[6] I. Mpiperis, S. Malassiotis, and M. G. Strintzis. Bilinear models for 3-d face and facial expression recognition. *IEEE Transactions on Information Forensics and Security*, 3(3):498–511, Sept. 2008.

[7] H. Peng, F. Long, and C. Ding. 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):1226–1238, Aug. 2005.

[8] H. Soyel and H. Demirel. Facial expression recognition using 3d facial feature distances. In *Int. Conf. on Image Analysis and Recognition*, pages 831–838, Aug. 2007.

[9] H. Tang and T. S. Huang. 3d facial expression recognition based on automatically selected features. In *Int. Conf. on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, AK, June 2008.

[10] J. Wang, L. Yin, X. Wei, and Y. Sun. 3d facial expression recognition based on primitive surface feature distribution. In *Int. Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 1399–1406, June 2006.

[11] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato. A 3d facial expression database for facial behavior research. In *Int. Conf. on Automatic Face and Gesture Recognition*, pages 211–216, Southampton, UK, Apr. 2006.

[12] W. Zheng, H. Tang, Z. Lin, and T. S. Huang. A novel approach to expression recognition from non-frontal face images. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 1901–1908, Kyoto, Japan, Sept. 2009.