3D FACIAL EXPRESSION RECOGNITION USING ZERNIKE MOMENTS ON DEPTH IMAGES

Nicholas Vretos, Nikos Nikolaidis and Ioannis Pitas

Informatics and Telematics Institute
Centre for Research and Technology Hellas, Greece
and
Department of Informatics, Aristotle University of Thessaloniki
Thessaloniki 54124, Greece Tel,Fax: +30-2310996304
e-mail: nikolaid,pitas@aiia.csd.auth.gr

ABSTRACT

In this paper we propose a new method for 3D facial expression recognition. We make use of the Zernike moments, which are calculated in the depth image of a 3D facial point cloud. Combining, the Zernike moments along with the 3D point clouds and the depth images, we succeed in tackling problems arising in facial expression recognition due to affine transformations of the data, such as translation, rotation and scaling which, in other approaches are considered very harmful in the overall accuracy of a facial expression recognition algorithm. Support vector machines are used in order to classify the previously extracted features. Results are drawn in two publicly available databases for 3D facial expression recognition.

1. INTRODUCTION

Recent advances in image and video processing are often oriented towards human centered analysis of the image content. Face is perhaps the most important element of the human body since it is the basic means through which humans recognize other humans. Moreover the face and its expressions are the main channels humans use to communicate their feelings. Many algorithms have been built in order to detect [1], track [2] and recognize [3] human faces. In [4], Ekman has established, under an anthropological investigation, six main facial expressions which are used in order to communicate emotions between humans. Those are until now considered as the most inter racially interpretable facial expressions. Most of the efforts in facial expression recognition are targeted at recognizing those six facial expressions or a subset of them. These expressions, which stem from the early years of homo sapiens, where evolved due to the need of communication between humans before spoken language development and are used until now but in a different context and more or less subconsciously. Those six facial expressions are related to the emotions of anger, surprise, happiness, disgust, fear and sadness [4]. Often in facial expression recognition, a seventh class is considered which models the neutral face.

Facial expression recognition in video sequences and still images is a very important research topic with applications in human-centered interfaces, ambient intelligence, behavior analysis etc. One consideration that has to be taken into account when designing facial expression recognition algorithms is the fact that a facial expression is a dynamic process that evolves over time and includes three stages [5]: an *onset* (attack), an *apex* (sustain) and an *offset* (relaxation). Many facial expression recognition algorithms operate on the video frame (or still image) that corresponds to the expression *apex*. Based on the approach used, facial expression recognition algorithms can be classified in two main categories: image-based (or image feature-based) and model-based ones. Each approach has its own merits. For instance, image-based algorithms are faster, as no complex image preprocessing is usually involved. On the other hand, model-based approaches employ a 2D or 3D face model, whose fitting on the facial image [6] implies significant computational cost. Despite being computationally demanding, model-based approaches are popular, because they can capture essential geometrical information for facial expression recognition [6]. An overview of the state of the art in facial expression recognition can be found in [7].

3D object description based on 3D point clouds has become very popular during the last years due to the technological advances in the field of 3D point clouds acquisition and manipulation e.g. by using new more accurate and less expensive 3D scanners and digitizers. In human-computer interfaces, 3D point clouds have already been used in facial expressions recognition and face recognition tasks with satisfactory results so far [8],[9]. The main issue that has to be tackled by these methods, is the 3D point cloud matching/registration and although algorithms exist, which provide some solution to the problem they usually are computationally expensive and practically not robust for noisy point clouds [10].

In this paper we propose a new method for 3D facial expression recognition which makes use of facial depth images, constructed from the facial 3D point cloud, and involves the Zernike image moments in order to create a feature vector which is then passed to a support vector machine (SVM) trained to recognize the 6 basic facial expressions. With this approach, we eliminate the need of matching the points of two facial 3D point clouds and we provide robustness against rotation, translation and scaling of the facial 3D point clouds. The paper is organized as follows: in section 2, we discuss the method of converting a 3D point cloud into a depth image. In section 3, the Zernike based feature vector is analyzed and its useful prop-

The research leading to these results has received funding from the European Community's Seventh Framework Programe (FP7/2007-2013) under grant agreement nº 211471 (i3DPost).

erties of robustness are discussed. In section 4, results on 2 publicly available databases are presented. Finally, conclusions are drawn in section 5

2. CONSTRUCTION OF DEPTH IMAGES FROM 3D FACIAL POINT CLOUDS

Depth images, also known as, range images, xyz maps, surface profiles and 2.5D images, are representations of 3D scenes by means of 2D images, where intensity at each pixel encodes the relative depth from a reference plane [11]. During the last decade depth images gained attention in many areas of 3D image processing, mainly due to the ease of extraction from standard commercial scanners [11] and the underlying explicit spacial organization provided from the 2D image support (i.e. the image x-y plane).

In human specific tasks like 3D face recognition and 3D facial expressions recognition, depth images have not been exploited thoroughly yet. Although some works exists in 3D face recognition [12], in 3D facial expression recognition, to the best of the authors knowledge, no work exists that makes use of depth images to represent the facial information.

Before proceeding to the depth image calculation we perform a Principal Components Analysis (PCA) based registration step to align the 3D facial points clouds corresponding to different faces. The method we use is the one in [13]. Therein, it has been shown that due to the geometry of the facial 3D points cloud the maximum variance axes are consistent between different face models. Thus, alignment of the 3D point clouds can be achieved by identifying and align these axes. Although in practice, mainly due to outliers, this assumption may not always, hold the resulting errors are only slight rotations of the maximum variance axes and do not alter the results of facial expression recognition significantly. Moreover, as will be detailed later on, such errors are tackled by using the Zernike moments, which are robust towards rotation.

The reference xy-plane, from which depths/intensities are calculated, is the plane defined by the minimum z value of the 3D point cloud and the plane defined by the two major axes of the PCA. This way different 3D facial point clouds have, if not equal, similar reference xy-planes. It has to be emphasized that small differences in the reference z point (i.e. between different depth images) do not impose a problem due to an adaptive histogram equalization, which is performed subsequently as will be detailed shortly.

Depth images, derived from 3D point clouds, can be modeled as a map from $\mathbb{N}^2$ to $[0, 1]$:

$$f(\phi(x), \phi(y)) = z, \tag{1}$$

where x, y, z are the Cartesian coordinates of the 3D points cloud, with z normalized in $[0, 1]$, and $\phi()$ a function from $\mathbb{R}$ to N, where $N \times N$ the final image resolution. Function $\phi()$ is the quantization function used to map real spacial coordinates to the image support. In our case $\phi()$ is a cubic interpolation function. In Figure 1, an example is shown for a 3D facial point cloud.

In order to provide better 3D facial expression recognition rates, and solve the problem of different z values between depth images due to different reference planes, we process the depth images by performing adaptive histogram equalization (AHE). In Figure 2, a facial depth image that has been equalized is shown.

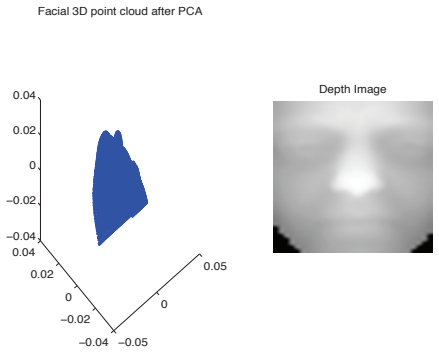

Fig. 1. Transformation of a facial 3D points cloud to facial depth image.

Fig. 2. Preprocessing of a facial depth image with AHE.

3. ZERNIKE MOMENTS ON DEPTH IMAGES

In [14], Zernike has introduced a set of complex polynomials, which form an orthogonal set over the interior of the unit circle. Based on these polynomials, in [15], the Zernike moments were proposed in order to tackle several problems arising from the use of raw moments in image processing such as redundancy of the moments, as well as, difficulty in the recovery of the image from these moments, due to high computational burden. Another issue related to Zernike moments, other than the orthogonality, is that they are robust to rotation due to their circular nature.

The Zernike polynomials are defined in the unit circle as:

$$V_{nm}(x, y) = V_{nm}(\rho, \theta) = R_{nm}(\rho) \exp(jm\theta), \tag{2}$$

where $n \in \mathbb{N}$, $m \in \mathbb{Z}$ subject to constraints $n - |m|$ even and $|m| \leq n$ and ρ, θ the polar coordinates corresponding to x, y Cartesian ones. The radial polynomial $R_{nm}(\rho)$ is defiened as:

$$R_{nm}(\rho) = \sum_{s=0}^{n-\frac{|m|}{2}} (-1)^s \cdot \frac{(n-s)!}{s! \left(\frac{n+|m|}{2} - s\right)! \cdot \left(\frac{n-|m|}{2} - s\right)!} \rho^{n-2s}. \tag{3}$$

Table 1. First 5 orders of Zernike Moments.

Order	Moments	N^o of moments
0	A_{00}	1
1	A_{11}	1
2	A_{20}, A_{22}	2
3	A_{31}, A_{33}	2
4	A_{40}, A_{42}, A_{44}	3
5	$A_{51}, A_{53}, A55$	3

Finally, the Zernike moments of order n and repetition m of an image described by a function $g(x, y)$ can be defined as follows:

$$A_{mn} = \frac{n+1}{\pi} \sum_x \sum_y g(x, y) V^*(\rho, \theta), \; x^2 + y^2 \leq 1. \quad (4)$$

For a more detailed insight of the Zernike moments the interested reader may refer to [14]. In our case we use the first 20 Zernike moments of all repetitions, which results in a total of 121 moments that will be the final length of the feature vector. In Table 1, Zernike moments up to 5^{th} order are shown.

Before calculating the Zernike moments, we extend our original depth image into an image of larger dimensions, in such a way that, most of the information of the depth image can be included in the Zernike limiting circle. To do so we calculate the square, whose inscribed circle is at the same time circumscribed in the square defined from the largest dimension of the original facial depth image. In more detail, let consider that the initial image dimensions are $N_1 \times N_2$ with $N_1 > N_2$. The enlarged image dimensions are $N'_1 \times N'_1$, where N'_1 is calculated as $N'_1 = N_1\sqrt{2}$. The pixels of the border added around the original facial depth image in order to reach the new dimensions are set to 0. Figure 3, illustrates this process.

Once the feature vector is calculated for all facial depth images, we use an SVM multiclass classifier [6], trained for the 6 facial expressions mentioned before. SVMs were chosen due to their good performance in various practical pattern recognition applications [16]-[19], and their solid theoretical foundations. SVMs minimize an objective function under certain constraints in the training phase, so as to find the support vectors, and subsequently use them to assign labels to the test set. Many SVM variants exist. These include both linear and non linear forms, with different kernels being used in the latter.

4. EXPERIMENTAL RESULTS

We have performed experiments on 2 different 3D Facial Expressions Databases. The BU3DFE [20], and the Bosphorous [21]. The BU3DFE consist of 100 individuals (44 male and 56 female) with 6 emotions each and with 4 levels of intensity at each emotion. This results in a total of 2400 different point clouds. In our case we used only the high intensity facial expression level, thus, we have 600 different facial 3D point clouds. On the other hand, the Bosphorus database, consists of 105 individuals which, do not perform all six facial expressions. Only 65 out of 105 have a full facial expression set. This results, in our case in a total of 390 (i.e. 6×65) different 3D facial point clouds.

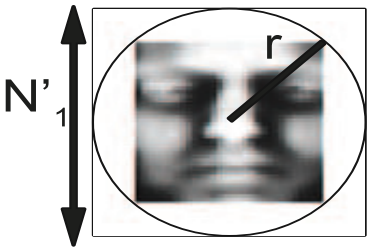

Fig. 3. Extension process to include all available information to the Zernike circle.

Table 2. Confusion matrix for 6 facial expression on the BU3DFE database.

	Ang	Dis	Fea	Hap	Sad	Sup
Ang	61.0	5.0	6.0	1.0	19.0	1.0
Dis	10.0	79.0	12.0	4.0	4.0	6.0
Fea	3.0	6.0	63.0	8.0	7.0	6.0
Hap	1.0	2.0	9.0	86.0	2.0	0
Sad	25.0	4.0	5.0	0	68.0	6.0
Sup	0	4.0	5.0	1.0	0	81.0

In both cases we followed the same approach (described in section 2) for the creation of facial depth images, although, in [21], predefined facial depth images are provided. We have run several experiments in order to calibrate the SVM parameters through a grid search for different kernels (i.e. linear, rbf and polynomial). Results are presented in Table 2 and Table 3. These results are the outcome of a person-based 5-fold cross validation test. That is, we exclude 20% of the individuals at both databases and we train with the remaining 80% of the individuals. This procedure ensures that when recognizing the facial expressions of an individual none of its image are included in the training set. Doing a greedy 5-fold cross validation where expressions from the same individual may be included, at the same time, in the training and the test sets, can results in erroneous interpretation of the results.

The achieved overall accuracy in 3D facial expression recognition was 73% for BU3DFE and 60.5% for Bosphorus, respectively. The best accuracy with respect to a particular facial expression is 86% for happiness in the BU3DFE database, while 92.3% is achieved for the same expression in Bosphorus database. Despite the fact that other methods that operate on 3D point clouds have been shown to achieve better facial expression recognition rates in the same databases, the superiority of these methods is rather artificial as they are based on the assumption that manually selected points and/or landmarks will be available.

Table 3. Confusion matrix for 6 facial expression on the Bosphorus database.

	Ang	Dis	Fea	Hap	Sad	Sup
Ang	70.8	13.8	9.2	0	23.1	1.5
Dis	12.3	58.5	6.2	1.5	18.5	3.1
Fea	3.1	7.7	43.1	1.5	4.6	46.2
Hap	0	6.2	1.5	92.3	3.1	0
Sad	13.8	10.8	6.2	4.6	50.8	1.5
Sup	0	3.1	33.8	0	0	47.7

5. CONCLUSIONS

In this paper a new method for 3D facial expression recognition has been proposed. The use of facial depth images combined with Zernike moments is exploited and results are drawn for 2 public 3D facial expressions databases. The calculated features are robust against affine transformations of the facial 3D points clouds, which makes the proposed method efficient. 3D facial expressions recognition rates 73%, as well as, the ease of extracting depth images from facial 3D points clouds, provides evidence that the method can be used efficiently for real life applications.

At this point, we have to emphasize the fact that our approach can be considered as a 3D facial expressions recognition approach that is applicable to real situations. Most attempts so far use a predefined set of manually selected points or landmarks, which makes these methods difficult to use in practice. In our approach, due to the use of depth images there is no need for such a landmark selection step. Everything is handled in the facial depth image level, where the use of Zernike moments assures the rotation invariance, while the mapping to the image xy-plane that create images of the same size provides scale invariance to the framework.

6. REFERENCES

[1] E. Hjelmas and B.K. Low, "Face detection: A survey," *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236–274, 2001.

[2] G.R. Bradski et al., "Computer vision face tracking for use in a perceptual user interface," *Intel Technology Journal*, vol. 2, no. 2, pp. 12–21, 1998.

[3] W. Zhao, R. Chellappa, PJ Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *Acm Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[4] P. Ekman, "Facial expression and emotion," *Personality: Critical Concepts in Psychology*, vol. 48, pp. 384–92, 1998.

[5] M. Pantic and LJM Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[6] I. Kotsia and I. Pitas, "Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, 2007.

[7] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on Pattern Analysis and Machine Intelligence*, pp. 39–58, 2008.

[8] M.O. Irfanoglu, B. Gokberk, and L. Akarun, "3D shape-based face recognition using automatically registered facial surfaces," in *Proceedings of International Conference on Pattern Recognition*, 2004, vol. 4, pp. 183–186.

[9] B. Gökberk, M.O. İrfanoğlu, and L. Akarun, "3D shape-based face representation and feature extraction for face recognition," *Image and Vision Computing*, vol. 24, no. 8, pp. 857–869, 2006.

[10] G. Vosselman, S. Dijkman, et al., "3D building model reconstruction from point clouds and ground plans," *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, vol. 34, no. 3/W4, pp. 37–44, 2001.

[11] R.J. Campbell and P.J. Flynn, "A survey of free-form object representation and recognition techniques," *Computer Vision and Image Understanding*, vol. 81, no. 2, pp. 166–210, 2001.

[12] A. Scheenstra, A. Ruifrok, and R. Veltkamp, "A survey of 3D face recognition methods," in *Audio-and Video-Based Biometric Person Authentication*. Springer, 2005, pp. 891–899.

[13] N. Vretos, N. Nikolaidis, and I. Pitas, "A model-based facial expression recognition algorithm using Principal Components Analysis," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2010, pp. 3301–3304.

[14] F. Zernike, "Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode," *Physica*, vol. 1, no. 7-12, pp. 689–704, 1934.

[15] A. Khotanzad and Y.H. Hong, "Invariant image recognition by Zernike moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 489–497, 1990.

[16] A. Tefas, C. Kotropoulos, and I. Pitas, "Using support vector machines to enhance the performance of elasticgraph matching for frontal face authentication," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 735–746, 2001.

[17] H. Drucker, D. Wu, and V.N. Vapnik, "Support Vector Machines for Spam Categorization," *IEEE IEEE Transactions on Neural Networks*, vol. 10, no. 5, 1999.

[18] A. Ganapathiraju, J. Hamaker, and J. Picone, "Support Vector Machines for Speech Recognition," *Fifth International Conference on Spoken Language Processing*, 1998.

[19] M. Pontil and A. Verri, "Support vector machines for 3 D object recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 6, pp. 637–646, 1998.

[20] L. Yin, X. Wei, Y. Sun, J. Wang, and M.J. Rosato, "A 3D facial expression database for facial behavior research," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*. IEEE, 2006, pp. 211–216.

[21] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," *Biometrics and Identity Management*, pp. 47–56, 2008.