Methodology and Performance Analysis of 3-D Facial Expression Recognition Using Statistical Shape Representation

Wei Quan, Bogdan J. Matuszewski, Lik-Kwan Shark

ADSIP Research Centre, University of Central Lancashire
{WQuan, BMatuszewski1, LShark}@uclan.ac.uk

Charlie Frowd

School of Psychology, University of Central Lancashire
CFrowd@uclan.ac.uk

Abstract

This paper presents the methodology and performance of a statistical shape representation for automatic facial expression analysis in 3-D. The core of the method uses the statistical shape modelling technique with the deformable model-based surface matching process, which is capable of simulation and interpretation of 3-D human facial expressions. Using the proposed method, a 3-D face is represented by a low-dimensional shape space vector conveying information about face shape. Since the method relies only on the 3-D shape, it is inherently invariant to changes in the background, illumination, and viewing angle, which are the difficulties often suffered in 2-D facial expression analysis. Using 3-D static facial data from the BU-3DFE database as well as the 3-D dynamic facial expression database recently built by the authors in the ADSIP Research Centre, the paper also reports on the performance of the proposed facial expression representation. Furthermore, to demonstrate the effectiveness of the proposed facial expression representation, a comparison is made with human performance by involving a number of human participants to validate the facial expressions in the ADSIP database.

Keywords: *3-D facial expression analysis, statistical shape modelling, BU-3DFE database, ADSIP database*

1. Introduction

Over the last two decades, the advance in imaging technology and ever increasing computing power have led to the rapidly growing interest to achieve automatic analysis of facial expressions. One reason for such interest is the wide spectrum of possible applications in diverse areas, such as human computer interaction [1], pain assessment [2], clinical psychology [3], and identification of concentration level (detecting drivers' tiredness) [4]. From the security perspective, automatic analysis of facial expressions has been investigated in the context of lie detection during interrogation [5], and in aiding face recognition.

Automatic analysis of facial expressions is a difficult task due to its inherent subjective nature. Broadly speaking, there are three major challenges which need to be addressed in the research of facial expression analysis. The first challenge is to build representative and accurate database which would cover a range of typical facial expressions at different intensity levels made by different face types in terms of subject age, gender and ethnic origin, with verified accuracy of expression representation. Such databases have been partially

constructed for specific types of facial expressions including 2-D, 3-D or dynamic databases [6], [7], [8].

The second challenge is to extract representative features for expression interpretation. This may include use of original data (pixels or vertices), or more structured data. e.g. shape, motion vectors, colour and spatial configurations of the face with its components. Usually it is expected that the extracted features should have reduced dimensionality compared to the original input facial data [9]. The most widely used representation methods include the topographic context [10], primitive surface feature distribution [11], statistical shape model [12], active appearance model [13], localized geometric model [14], etc.

The third challenge is to find a suitable classification algorithm to categorize the facial expressions by using the extracted facial features. Some frequently used classification methods include linear discriminant analysis, support vector machines, or Hidden Markov Models (HMM) [15].

This paper focuses on a statistical shape representation of facial expressions proposed by the authors [16], [17], which is based on 3-D face matching using the statistical shape model. The so-called shape space vector (SSV) of the statistical shape model which controls the change of shape is postulated as a significant feature for analysis of facial expressions. This feature models the high dimensional shape variations observed in the training data set using projection on a low dimensional shape space. In order to extract the SSV feature, a statistical shape model needs to be constructed first with the available dense point correspondences, and this is followed by the model fitting process, which iteratively matches the shapes of the built model to the new input faces. The proposed method analyzes facial articulations completely in the 3-D space, since 3-D data is more accurate in the representation of human faces, and improves the recognition performance when compared to 2-D [18]. In this work, two publicly available facial expression databases are used to demonstrate the performance of the proposed facial expression representation, they are BU-3DFE database [7] and ADSIP database [8]. In order to establish a ground truth criterion, facial expressions in the ADSIP database are validated by a number of human participants, thereby enabling a comparison to be made between human performance and automatic recognition performance.

The remainder of this paper is organized as follows. Section 2 introduces the methodology of the statistical shape model. Section 3 discusses the databases used for performance evaluation, and the database validation procedure. Results of facial expression analysis using the SSV-based feature are presented in Section 4. Concluding remarks and possible future work are given in Section 5.

2. Statistical Shape Model

2.1. Construction of Statistical Shape Model

Developed based on the point distribution model (PDM), the statistical shape model (SSM) was first proposed by Cootes et al. [19]. The basic principle of SSM is to exploit the redundancy in the structural regularity of the given shapes, thereby enabling a shape to be described using fewer parameters, i.e., reducing the dimensionality of the shapes. In order to build a correct SSM, the dense point correspondences between 3-D faces in the training set need to be determined. In this work, this is achieved in three steps. The first step is to identify the corresponding anatomical landmarks on the reference face and other training faces. The second step is to warp the reference face to different training faces using thin-plate spline (TPS) transformation that is calculated based on the selected landmarks as control points. The third step is to estimate the point correspondence between warped reference face and different

training faces based on the closest distance metric. Using the correspondences estimated, standard principal component analysis (PCA) is applied to calculate the mean and eigen-vectors for all of the training data. The detail of the implementation of this stage can be found in [16], [20].

2.2. Model Matching

In order to extract the SSV for each individual input 3-D face, the built shape model has to match the face shape. This is done by the two processing steps, namely, initial model alignment and model refinement. The initial model alignment is achieved by the modified iterative closest point (ICP) method with similarity transformation [21]. The whole process starts from the alignment of the input face and model by using their centroid points, then determining the dense point correspondences between them according to the closest distance metric. Using the estimated correspondences, the similarity transformation which aligns the input face to the model is calculated. Subsequently the algorithm re-determines the correspondences based on the transformed input face and the model. The process iterates until its associated cost function converges to its local minimum.

Having the model and input face globally matched by the initial model alignment, the model refinement continues the model matching process to yield a better match by deforming the model. This iterative process is a combination of the ICP method and the least-squares projection onto the shape space, estimating in turn the pose and shape parameters between the two faces. The pose parameters include translation vector, rotation matrix and scaling factor, whereas the shape parameters are defined by the SSV. The ICP method is used firstly for finding the dense point correspondences and computing the pose parameters, and then followed by the least-squares projection, which estimates the SSV. In order to improve the accuracy of the shape representation using the model matching method described above, the use of adaptive size of the SSV was proposed in [17], which prevents the model from deforming into inappropriate shapes during the model refinement process. Some examples produced by the model matching process are shown in Figure 1, where the test input faces were selected from the ADSIP database and the shape model was built using 450 faces from the BU-3DFE database.

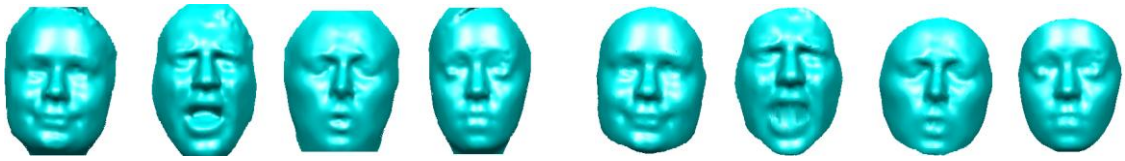

Figure 1. Examples of the Model Matching; (a) Input faces selected from ADSIP database, (b) matched shape model.

3. 3-D Facial Expression Databases

In order to evaluate the proposed method for automatic analysis of 3-D facial expressions, a comprehensive 3-D facial database is required. Two of the recently developed 3-D facial expression databases are chosen for the performance evaluation of the proposed method, and they are BU-3DFE and ADSIP databases.

3.1. BU-3DFE Database

The BU-3DFE database aims to aid the general understanding of facial behaviour and 3-D structure of facial expression on a detail level [12]. It was developed at the Binghamton University, New York, and consists of a set of static 3-D facial expression scans with both texture and shape information. The database includes 100 subjects, with age ranging from 18 to 70 years old, with a variety of ethnic origins including White, Black, East-Asian, Middle-East Asian, Indian and Hispanic Latino. Each subject performed seven expressions, which include a neutral expression plus six universal expressions: happiness, disgust, fear, angry, surprise, and sadness. The details of this database can be found in [7]. Some of the facial examples which have been used for the evaluation of the proposed method are shown in Figure 2.

Figure 2. Examples of 3-D face scan from the BU-3DFE database.

3.2. ADSIP Database

The intention of the ADSIP database is to initially build a pioneering facial expression database with a relatively small number of human participants in arguably the most realistic format, 3-D dynamic [8]. As mentioned before, recognition based on 3-D static data potentially offers a better performance for detection of facial expressions than 2-D, however, it still does not convey all the information necessary for robust and accurate expression identification. For example, there are robust facial motion cues which could facilitate the perception of subtle facial expressions [22]. Using sophisticated video alignment and editing techniques, recent work has also started to unpack the importance of the various facial areas involved in emotional expression [23]. It is clear that while some emotions can be recognized very well statically from the face alone, for example happiness, others require extra facial information. For instance, both the expressions of agreement and disagreement require so-called 'rigid head motion' (RHM) – an up-down or left-right head shake, respectively. Others, such as fear, 'don't know' (clueless) and 'don't understand' (confusion) normally involve RHM to some extent, and the perception of the expression is improved when head movements are observed [24].

The ADSIP database was developed in the ADSIP Research Centre at the University of Central Lancashire. In its first release it contains 10 subjects (8 female and 2 males), with age ranging from 19 to 23 years. Each subject was asked to perform seven expressions at three levels of intensity labelled as: mild, normal and extreme. The order in which the expressions were carried out was fixed and followed the pattern: anger, disgust, fear, happiness, sadness, surprise and pain. The 'mild' expression was recorded first in each case, followed by 'normal' and then 'extreme'. All the facial expressions start from the neutral expression. A set of 21 sequences (7 expressions on 3 levels) was collected from each subject, which results in a total of 210 sequences. The frame rate was set to 24 frames per second as normal video recording speed, and the duration of each sequence is 3 seconds. Furthermore, each face scan was accompanied with video recording for the purpose of database validation by human

observers. Participants in face scans were all graduates from the Performance Art Department at the University of Central Lancashire. It has been established that the use of actors, or trainee actors, is a valid approach, as it enables fairly accurate facial expressions to be captured [23]. Figure 3 shows some snap-shots of two dynamic 3-D facial expressions recorded in the ADSIP database.

Figure 3. Examples of 3-D dynamic facial expression recorded in the ADSIP database (female and male).

3.3. Validation of ADSIP Database

In order to investigate the difference in the understanding of facial expressions between the human observers and computerized facial expression analysis algorithms, as well as to obtain the "calibrated" ground truth data for the assessment of the computer based facial analysis algorithms, the ADSIP database was validated by the human observers. The invited participants were 10 staff and students from the University of Central Lancashire. They comprised of 6 females and 4 males with an average age of 34.8 years. A bespoke computer program was designed to present the recorded video clips and collect the confidence ratings given by the observer. Participants were shown one video clip at a time and were asked to enter their confidence ratings against seven expressions' categories: anger, disgust, fear, happiness, pain, sadness and surprise, with the confidence ratings selected from the range of 0 to 100% for each category. To reflect a possible observers' confusion about expressions for a given video clip, ratings could be distributed over the various expression categories as long as scores added up to 100%. Examples of some snapshots from the video clips are shown in Figure 4.

Confidence scores were extracted for the expression in each video clip. These data are presented in Table 1. It can be seen that happiness expressions were given near perfect confidence scores, and anger, pain and fear were the worst rated at around 50%. Also, the 'normal' intensity level was somewhat better rated than 'mild', and 'extreme' was also somewhat better than 'normal'. Table 2 shows the confidence confusion matrix for the seven expressions. It can be seen that the observers were again very confident about recognizing the happiness expression whereas the fear expression was often confused with the surprise expression.

Table 1. Mean confidence scores for seven expressions

Intensity	Anger (%)	Disgust (%)	Fear (%)	Happiness (%)	Sadness (%)	Surprise (%)	Pain (%)	Mean (%)
Mild	47.5	51.5	43.3	90.3	72.9	57.4	42.6	57.9
Normal	56.6	78.3	41.5	94.3	75.6	62.0	51.4	65.7
Extreme	61.4	80.7	48.4	96.0	74.0	75.7	56.2	70.3
Mean(%)	55.2	70.2	44.4	93.5	74.2	65.0	50.0	64.6

Table 2. Confidence confusion matrix for human observers

Input/Output	Anger (%)	Disgust (%)	Fear (%)	Happiness (%)	Sadness (%)	Surprise (%)	Pain (%)
Anger	**55.39**	26.03	5.19	0.00	5.13	5.31	2.94
Disgust	7.70	**68.86**	5.22	0.00	8.47	4.59	5.16
Fear	3.80	9.02	**46.90**	0.00	7.13	23.90	9.26
Happiness	0.27	0.98	0.71	**92.95**	1.15	2.35	1.59
Sadness	4.07	5.87	3.63	0.71	**74.15**	3.22	8.33
Surprise	0.60	7.54	21.84	1.04	2.46	**64.64**	1.88
Pain	4.94	9.45	9.46	2.30	18.96	3.85	**51.04**

4. Results of Facial Expression Recognition

Using the introduced 3-D facial expression databases, the performance of the proposed method was evaluated by applying three well known (off-the-shelf) classification algorithms with the proposed facial features namely, linear discriminate analysis (LDA), quadratic discriminant classifier (QDC) and nearest neighbour classifier (NNC). The details of these classification algorithms are beyond the scope of this paper but can be found in most of the textbooks on pattern recognition [25].

Figure 4. Expressions from one of the actors at three intensity levels (01='mild', 02='normal', 03='extreme'). Images show single frames taken from the camcorder at roughly the peak of expression.

4.1. Evaluation using BU-3DFE Database

The first performance evaluation was carried out based on the BU-3DFE database. A statistical shape model was built using 450 randomly selected training faces from the BU-3DFE database. These training faces covered a wide range of ages, ethnic origins and expressions. 900 faces were used for the testing. Each of the testing face was represented by a SSV which was extracted through the model matching process as described previously. In the validation, all the testing faces were divided into 6 subsets with each subset containing 150 randomly selected faces. During the evaluation, one of the subset was selected as the test subset while the remaining subsets were used as the training set. Such experiment was repeated six times with a different subset selected as the test subset each time. The LDA classifier achieved the highest recognition rate of 81.89%. It is a reasonable result when compared with the first pioneer work on facial expression recognition using the same database which was carried out by Wang et al. [10]. They claimed that about 83% correct

recognition was achieved in classifying the same six universal expressions using the LDA approach. Table 3(a) shows the confusion matrix of the LDA classifier. It can be seen that the anger, happiness, sadness and surprise expressions are all classified with above 80% accuracy, whereas the fear expression is only classified correctly for around 73%. This is similar to the validation results for the ADSIP database reported earlier, which showed the human observers often mix up the fear expression with other expressions.

4.2. Evaluation with ADSIP Database

The performance evaluation using the ADSIP database is an extension of that based on the BU-3DFE database. It is used to check stability of the proposed methodology with a statistical shape model built from one database and tested using a completely different database. Furthermore, it is also used to assess the performance of the proposed algorithm against human observers. The statistical shape model built, as explained before, based on 450 faces randomly selected from the BU-3DFE database, was used for the performance evaluation. One hundred static faces were randomly chosen as the testing faces from the 3-D dynamic facial sequences in the ADSIP database, each approximately at the maximum of the expression. The selected test set contained 10 subjects with six different facial expressions as well as various intensities. With each testing face represented by its SSV, three standard classifier, LDA, QDC, and NNC, were applied for the facial expression classification. The LDA classifier again achieved the highest recognition rate of 72.84%. Although this result is worse than the result obtained on the BU-3DFE faces, it is around 2.5% higher than the mean recognition rate achieved by the human observers for those 'extreme' expressions in the same ADSIP database shown in Table 1, and it is much better than the human observers' mean recognition rate for the other two intensity levels of expressions. The confusion matrix of the LDA is shown in Table 3(b). It can be seen that the happiness is the most recognizable expression and followed by the surprise expression. Fear and sadness are the expressions which are often confused with others.

Table 3. Confusion matrix of the LDA classifier; (a) BU-3DFE database, (b) ADSIP database

Input/Output	Anger (%)	Disgust (%)	Fear (%)	Happiness (%)	Sadness (%)	Surprise (%)
Anger	**82.64**	3.48	4.17	3.47	4.86	1.39
Disgust	7.64	**78.47**	3.48	5.56	2.08	2.78
Fear	4.17	3.47	**72.59**	12.50	5.56	1.39
Happiness	2.78	5.56	8.33	**83.33**	0.00	0.00
Sadness	4.17	3.47	11.11	0.00	**81.25**	0.00
Surprise	0.00	0.00	4.17	2.78	0.00	**93.06**

(a)

Input/Output	Anger (%)	Disgust (%)	Fear (%)	Happiness (%)	Sadness (%)	Surprise (%)
Anger	**73.08**	11.54	0.00	1.92	13.46	0.00
Disgust	11.54	**71.15**	9.62	1.92	1.92	3.85
Fear	1.92	11.54	**55.77**	17.31	13.46	0.00
Happiness	2.08	0.00	10.83	**87.08**	0.00	0.00
Sadness	22.92	0.00	10.42	0.00	**66.67**	0.00
Surprise	0.00	6.25	8.33	0.00	2.08	**83.33**

(b)

5. Conclusions and Future Work

This paper presents the methodology and performance of a statistical shape representation for automatic analysis of facial expressions in 3-D. The method is based on the statistical shape modelling technique combined with deformable surface matching process. In order to examine the performance of the proposed method on the facial expression analysis task, two recent developed 3-D facial expression databases have been used. The evaluation result on the BU-3DFE database shows that the proposed method is capable of simulation and interpretation of 3-D human facial expressions. The test preformed with the ADSIP database indicates that the method can be used with data collected using different acquisition protocols which could be quite different from the protocol used to acquire the training data. The

obtained results are compared with human observers' results showing similar recognition rates.

Currently the ADSIP database is being expanded by increasing the number of subjects and aims to reach the final goal of 100 subjects. Revision of the expression captured also needs to be considered. It would be advantageous to be more specific about certain expressions. For example, surprise clearly comes in two forms, pleasant and unpleasant. It would be useful to differentiate these (and potentially capture both types). Considerations will also be given to increasing the number of different expressions. Potential candidates include agreement and disagreement, with potential applications for the film and animation industry, and confusion and clueless, for psychological research. So far all the analysis work using the proposed method is based on the 3-D static. The proposed method and testing is now being extended to 3-D dynamic for the ADSIP database, thereby enabling analysis of the facial expressions using not only the spatial knowledge but also the temporal information of subtle movements on the face. The analysis result will be published in near future.

Acknowledgement

The work presented in this paper has been supported by the TeRaFS projects (EPSRC Project No. EP/H024913/1).

References

[1] A. Pentland, "Looking at People: Sensing for Ubiquitous and Wearable Computing," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 107-119, 2000.

[2] S. Brahnam, C. Chuang, F. Y. Shih, and M. R. Slack, "Machine Recognition and Representation of Neonatal Facial Displays of Acute Pain," Artificial Intelligent in Medicine, vol.36, no.3, pp. 211-222, 2006.

[3] S. D. Pollak and P. Sinha, "Effects of Early Experience on Children's Recognition of Facial Display of Emotion," Developmental Psychology, vol. 38, no. 5, pp. 784-791, 2002.

[4] E. Vural, M. Cetin, G. Littlewort, M. Bartlett, and J. Movellan, "Drowsy Driver Detection Through Facial Movement Analysis," Lecture Notes in Computer Science, no. 4796, pp. 6-18.

[5] H. Wang and N. Ahuja, "Facial Expression Decomposition," Ninth IEEE International Conference on Computer Vision, 2003.

[6] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive Database for Facial Expression Analysis," 4th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46-53, 2000.

[7] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3D Facial Expression Database for Facial Behaviour Research," 7th International Conference on Automatic Face and Gesture Recognition (FG2006), pp. 211-216, 2006.

[8] C. D. Frowd, B. J. Matuszewski, L.-K. Shark, and W. Quan, "Towards a Comprehensive 3D Dynamic Facial Expression Database," WSEAS International Conference on Multimedia, Internet and Video Technology, 2009.

[9] H. Park and J. Park, "Analysis and Recognition of Facial Expression Based on Point-Wise Motion Energy," Lecture Notes in Computer Sciences, vol. 3212, no. 2004, pp. 700-708, 2004.

[10] J. Wang and L. Yin, "Static Topographic Modelling for Facial Expression Recognition and Analysis," Computer Vision and Image Understanding, pp. 19-34, 2007.

[11] J. Wang, L. Yin, X. Wei, and Y. Sun, "3D Facial Expression Recognition Based on Primitive Surface Feature Distribution," IEEE International Conference on Computer Vision and Pattern Recognition, pp. 17-22, 2006.

[12] W. Quan, B. J. Matuszewski, and L.-K. Shark, "3-D Facial Expression Representation Using Statistical Shape Models," BMVA Symposium on 3D Video, Analysis, Display and Applications, 2008.

[13] T. Hong, Y.-B. Lee, Y.-G. Kim, and H. Kim, "Facial Expression Recognition Using Active Appearance Model," Lecture Notes in Computer Sciences, vol. 3972, no. 2006, pp. 69-76, 2006.

[14] A. Saxena, A. Anand and A. Mukerjee, "Robust Facial Expression Recognition Using Spatially Localized Geometric Model," International Conference on Systemic, Cybernetics and Informatics, pp. 124-129, 2004.

[15] Y. Sun and L. Yin, “Facial Expression Recognition Based on 3D Dynamic Range Model Sequences,” 10th European Conference on Computer Vision (ECCV08), 2008.

[16] W. Quan, B. J. Matuszewski, L.-K. Shark and D. Ait-Boudaoud, “3-D Facial Expression Biometrics Using Statistical Shape Models”, Special Issue on Recent Advances in Biometrics Systems: A Signal Processing Perspective, EURASIP Journal on Advances in Signal Processing, vol. 2009, pp. 1-17, 2009.

[17] W. Quan, B. J. Matuszewski, and L.-K. Shark, “Improved 3-D Facial Representation through Statistical Shape Model”, IEEE International Conference on Image Processing, 2010.

[18] M. Pantic & L. J. M. Rothkrantz, "Expert System for Automatic Analysis of Facial Expressions," Image and Vision Computing, vol. 18, no. 2000, pp. 881-905, 2000.

[19] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, “Active Shape Models – Their Training and Application,” Computer Vision and Image Understanding, vol. 61, no. 1, pp. 38-59, 1995.

[20] W. Quan, B. J. Matuszewski, and L.-K. Shark, "3-D Facial Expression Representation Using B-spline Statistical Shape Model," Vision, Video and Graphic Workshop, British Machine Vision Conference, 2007.

[21] P. J. Besl and N. D. McKay, "A Method for Registration of 3-D Shapes," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 1, pp. 33-80, 2000

[22] Z. Ambadar and J. F. Cohn, “Deciphering the Enigmatic Face: The Importance of Facial Dynamics in Interpreting Subtle Facial Expression,” Psychological Science, vol. 16, pp. 403-410, 2005.

[23] M. Nusseck, D. W. Cunningham, C. Wallraven, and H. H. Bulthoff, “The Contribution of Different Facial Regions to the Recognition of Conversational Expressions,” Journal of Vision, vol. 8, pp. 1-23, 2008.

[24] C. Busso, M. Grimm, and S. Narayanan, “Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis,” IEEE Transaction on Audio, Speech, and Language Processing, vol. 15, pp. 1075-1086.

[25] R. O. Duda, P. E. Hart, and D. G. Stork, “Pattern Classification,” John Wiley & Sons Inc, 2001.