SPONTANEOUS AND NON-SPONTANEOUS 3D FACIAL EXPRESSION RECOGNITION USING A STATISTICAL MODEL WITH GLOBAL AND LOCAL CONSTRAINTS

Diego Fabiano and Shaun Canavan

University of South Florida

ABSTRACT

In this paper, we propose a novel method for 3D facial expression recognition based on a statistical shape model with global and local constraints. We show that the combination of the global shape of the face, along with local shape index-based information can be used to recognize a range of expressions. These expressions include happiness, sadness, surprise, embarrassment, fear, nervousness, anger, disgust, and pain. We give insights into which features are important for facial expression recognition through statistical analysis. We also show that our proposed method outperforms the current state-of-the-art methods on spontaneous and non-spontaneous facial data.

Index Terms— Expression, classification, statistical, model, spontaneous, non-spontaneous

1. INTRODUCTION

It has been argued that recognizing emotions is one of the most important aspects of human intelligence [22]. Facial expressions are one way in which emotions are elicited, and the research into automatically recognizing these emotions has been successful in the past decade. This is due in part to the release of larger and more challenging facial databases [14], [15], [19], [21], [25], [28], [29], [34], [36], [37], [38], [39], [40]. Combined, these databases contain multiple terabytes of multimodal data, allowing for rapid growth of new methods to accurately recognize a wide range of challenging facial expressions. These datasets are integral to the success of expression recognition research, as it has been shown that 2D data struggles to provide a realistic resource for consistent and reliable expression recognition [26], [29]. The challenges encountered with 2D data collection, such as pose variation, lighting, and color, can be overcome with the use of 3D datasets; which makes them a natural fit for our study.

Expression recognition is largely an unsolved problem, with applications in video games, education, entertainment, intelligent transportation systems, pain recognition systems in the health industry, and behavior analysis. With a focus on these types of applications, the BU-4DFE [36] face database has been successfully used in recent years for facial expression analysis. Sun et al. [30] developed an approach establishing correspondence between 3D models over time. Using these correspondences, they applied Spatio-Temporal Hidden Markov Models that represent facial information by looking at both intra-frame and inter-frame changes for facial expression recognition; they achieved a classification accuracy of 83.7%. Similar work was done by Yin et al. [36] with a person-independent experiment, on 60 subjects, where a two-dimensional Hidden Markov Model was used to learn the temporal relations of the facial regions, to classify expressions, achieving an accuracy of 90.44%.

Drira et al. [13] proposed a new Deformation Vector field, based on Riemannian facial shape analysis, which describes local deformation of the face over time. Using their proposed approach, along with a random forest [5], they achieved 93% accuracy while classifying expressions. Fang et al. [17] used a Support Vector Machine with a Radial Basis Function kernel to classify 3D facial expressions. They achieved 84.1% when using the geometrical coordinates as the feature vector and 91% when using the normal of the coordinates. Danelakis et al. [11] proposed the GeoTopo+ descriptor which exploits the landmarks of the face, to create three sub-descriptors that capture the topological and geometric information of the face. Using these descriptors, they performed unsupervised facial expression recognition on the BU-4DFE and BP4D [38], [39] databases, achieving accuracies of 90% and 88.56% respectively.

Motivated by these works, we propose a method for facial expression recognition on spontaneous and non-spontaneous 3D facial data. A summary of the main contributions of the paper are detailed below.

(1) We propose a novel method for facial expression recognition using a statistical shape model with global and local constraints.
(2) We compare and analyze the proposed method on spontaneous and non-spontaneous 3D facial data.
(3) We show that the proposed method outperforms the current state of the art on two public databases.

2. STATISTICAL SHAPE MODEL WITH GLOBAL AND LOCAL CONSTRAINTS

Certain regions of the face inherently have important information for recognizing emotions, such as the mouth and eyes. To model these key features, facial landmarks and local curvatures of the surrounding neighbors of each landmark are used. A statistical model of shape from facial landmarks is created; which includes the eyes, nose, mouth, eyebrows, and contour of the face. To construct this statistical model, training data is first annotated with L landmarks, which allows for modeling of both global and local constraints of the face. An overview of model construction and landmark detection are given in the following sub-sections. For a detailed analysis of the method, the reader may reference the work on shape index-based statistical shape models [7], [8].

2.1. Global constraints

The global constraints of the face shape are constructed from landmark features that represent key features (e.g. eyes, mouth, and nose). Training data is manually annotated with L landmarks around these features. A $n \times n$ patch is constructed, using the corresponding (u, v) coordinates around each landmark on the training data. Given a training set of size M SI-SSM models, each with L patches of size $n \times n$, a parameterized model $S_G = (x_1, y_1, z_1, \ldots, x_N, y_N, z_N)$ that defines the global face shape is constructed. Principal component analysis (PCA) is then applied to the training data to learn the modes of variation, allowing us to approximate new global face shapes.

2.2. Local constraints

The local constraints of the face shape are constructed from shape index values [12] that represent a local $n \times n$ patch around each key facial landmark from the global constraints. Given a training set of M models, each with L patches, a local parametrized model $S_L = (SI_1, \ldots, SI_N)$ that defines the local face shape is constructed. PCA is applied to this model in the same manner as the global constraints, allowing us to approximate new local face shapes.

2.3. Combined feature model

Given the parameterized models S_G and S_L, a feature model that combines both global and local constraints, is constructed. The combination of these models allows us to move the local patches on the face while still maintaining the global constraint of the overall face shape through the feature vector $S_{GL} = (x_1, y_1, z_1, \ldots, x_N, y, z_N, SI_1, \ldots, SI_N)$. By using shape index values to constrain and move the combined model, it does not suffer from problems of global lighting variation [30], as they are invariant to this. The combined feature model is used for facial landmark detection; which is detailed in the next sub-section.

2.4. Landmark detection

To detect landmarks on 3D geometric face meshes, the correlation between the combined global and local feature vectors, as well as the input mesh, is computed. To do this, a sufficient starting point for the model is found on the mesh, and the weight parameters of the global shape are learned by uniformly varying the weight vector; this generates new instances of the model. Iterative closest point [4] is then used to minimize the distance between the combined model and the mesh. The instance of the model that results in the lowest matching score is used as the initialization. Once initialized, a local patch-based correlation score is computed between the model and the input mesh through a cross-correlation template matching scheme [20]. This score is computed for all shape index-based patches, which are summed for a final correlation score (*CS*). This is used as a baseline to perform a patch-based correlation search to move the local patches into a position that better represents the true face shape.

After initialization, a local patch-based search is performed by creating a new patch, consisting of shape index values of the same size around each of the cells of the original $n \times n$ patches. A new correlation score is computed for each of the new patches. The landmark feature is then moved to the patch that results in the best correlation score. The new correlation scores for each patch are then summed to give a new *CS*. This continues until the model converges with the mesh. See figure 1 for sample detected landmarks on the BU-4DFE [36] and BP4D [38], [39] databases.

3. EXPERIMENTS AND EVALUATION

Using the 3D facial landmarks detected from the method detailed in section 2, we perform facial expression recognition on two facial expression databases. We then analyze spontaneous and non-spontaneous data and compare our results to the current state of the art. An overview of the data used, experimental design, results, and analysis are detailed in the following sub-sections.

3.1. Facial expression databases

To conduct our experiments, we used two state-of-the-art 3D facial databases. The first database is the BU-4DFE [36]; that consists of 101 subjects displaying 6 prototypic facial expressions (anger, happiness, fear, disgust, sadness, and surprise). There are 58 females and 43 male subjects, with a variety of ethnic and racial ancestries with an age range of 18-45. The data was collected in a controlled environment in which the subjects were instructed to perform specific emotions; therefore, we designate the BU-4DFE as a non-spontaneous database for the rest of the paper. The expressions were captured at 25 frames per second, where each expression sequence is approximately 100 frames, giving a total of over 60,000 frames of data. Each 3D model has a resolution of approximately 35,000 vertices (figure 1).

The second database is the BP4D [38], [39]. This was used in the Facial Expression Recognition and Analysis Challenge 2015 [31] and 2017 [32]. It was developed to promote the exploration of spatiotemporal features in subtle facial expressions. There are 23 females and 18 male subjects displaying 8 expressions each (happy, sad, surprise, fear, nervous, pain, anger, disgust). The data was collected in an environment where subjects naturally interacted with an interviewer; we attribute the BP4D as a spontaneous database. The subjects are between 18-29 years of age; 11 Asian, 4 Hispanic, 6 African-American, and 20 Euro-American ethnicities are represented.

3.2. Experimental design

Based on the model detailed in section 2, we detected 83 facial landmarks on the spontaneous and non-spontaneous datasets. From these landmarks, we constructed a 249-dimension vector to represent each face and expression (83 3D landmarks (x, y, z)). By using these 83 landmarks, we can reduce the original model vector of approximately 105,000 dimensions (obtained from 35,000 3D vertices of the original model) to a 249-dimension vector and still maintain a high degree of accuracy for expression recognition. We detected these 83 facial landmarks on 60,402 models in the non-spontaneous dataset and 367,474 models in the spontaneous dataset. We then used these facial landmarks to train a separate random forest (RF) [5], which is a collection of regression trees [6], where the mean output of all trees is taken as the final output, for each dataset. We used a 10-fold cross-validation scheme where the data is randomly split into 10 subsets, where nine are used for training and the other is used for testing. This is done a total of 10 times, where each subset is used to test. The average error of all iterations is taken to help reduce overfitting. We show that using this experimental design results in accurate recognition of spontaneous and non-spontaneous facial expressions.

3.3. Experimental Results and Analysis

By training our models with the landmark feature vectors, we achieved a max classification rate of 99.9934% (60,398 out of 60,402) in the non-spontaneous dataset [36] and 99.6974% (366,362 out of 367,474) in the spontaneous dataset [38], [39]. See tables 1 and 2 for the confusion matrices on non-spontaneous and spontaneous environments.

Table 1. Confusion matrix, for facial expression recognition on non-spontaneous data, showing number of classified instances. Expression key: E1-Happiness; E2-Sadness; E3-Surprise; E4-Fear; E5-Anger; E6-Disgust.

	E1	**E2**	**E3**	**E4**	**E5**	**E6**
E1	**9973**	0	0	0	0	0
E2	0	**10142**	0	0	0	0
E3	0	1	**9947**	0	0	0
E4	0	0	1	**10043**	0	0
E5	0	0	0	0	**10122**	2
E6	0	0	0	0	0	**10171**

The proposed method was able to recognize most of the non-spontaneous expressions with 100% accuracy (4 misclassified models from 3 different expressions). From the incorrect classifications, surprise was misclassified as sadness once, fear was misclassified as surprise once, and anger was misclassified as disgust twice. The proposed method is also able to recognize spontaneous expressions with a high degree of accuracy, misclassifying a total of 1,112 facial models out of 367,474. From the incorrectly classified spontaneous facial data, embarrassment was incorrectly classified as the seven other expressions the most, at 38% of the time (425 out of 1112). This can be attributed to embarrassment being a complex, self-conscious emotion, where multiple behaviors occur over the emotion [33].

Table 2. Confusion matrix, for facial expression recognition on spontaneous data, showing number of classified instances. Expression key: E1-Happiness; E2-Sadness; E3-Surprise; E4-Embarrassment; E5-Fear; E6-Pain; E7-Anger; E8-Disgust.

	E1	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**
E1	**47544**	0	0	63	9	4	16	3
E2	4	**65506**	1	4	0	0	13	1
E3	24	14	**12482**	91	23	16	56	6
E4	73	5	9	**60245**	43	8	77	7
E5	22	2	5	91	**52013**	12	46	6
E6	3	0	0	39	13	**44518**	53	4
E7	22	2	4	58	24	3	**68890**	5
E8	12	2	8	79	11	6	12	**15164**

To test the utility of our landmark features, we also ran the same classification scheme with a support vector machine (SVM). Due to the size of the spontaneous data and the time complexity of SVMs, we ran it on the non-spontaneous data, achieving a classification accuracy of 99.62%.

The accuracy of these experimental results can be attributed to the proposed method’s ability to model the large variations that are present in the data. In the non-spontaneous data, the mouth region contains a large portion of the emotion variation. In the spontaneous data, there is variation in the mouth region. However, it is not the only significant region of variation; the regions around the eyes also detail large variations. To further study this, we evaluate the top-ranked features across spontaneous and non-spontaneous data and compare them to gain understanding about which features are statistically important for each data type. We rank all 249 features based on the information gained from classifying facial expressions, which is defined as the reduction in entropy of the expression class after the feature is observed. The top 10% ranked features (25) of each model were then selected for further analysis. This can be seen in figure 1 and more details are given in table 3.

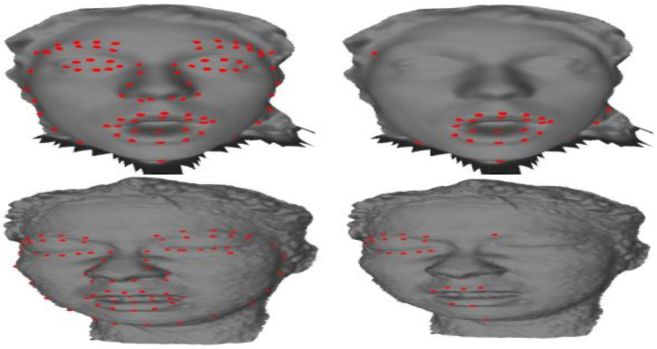

Figure 1. Surprise expression from non-spontaneous data (top), and spontaneous data (bottom). Left side shows 83 landmarks, right side shows top 10% of ranked features.

Table 3. Number of features in the corresponding facial regions of top ranked landmark features.

	Non-spontaneous	Spontaneous
Mouth	19	6
Contour of face	6	3
Eyes	N/A	8
Eyebrows	N/A	8

As seen in table 3, spontaneous and non-spontaneous data display variation in different facial regions when emotions are expressed. This difference in variation is due, partially, to the nature of the datasets we are comparing (spontaneous vs. non-spontaneous). In the non-spontaneous data, the dominance of the mouth and contour of the face over the other landmarks can be attributed to the lack of natural movement on the experiment in which the data was collected. On the other hand, the dominant features for the spontaneous setting are not only the mouth and the contour of the face, but the eyebrows, and more importantly the eyes; this agrees with physiological literature that describes the eyes as a fundamental feature for expression recognition [27]. The natural reaction from a subject to a stimulus, such as a loud noise or a hand being held into a bucket of ice water, leads to an increase in corporal movement compared to the non-spontaneous environment; this justifies an even distribution across the importance of the features in the spontaneous data. The proposed method can accurately model these complex variations in emotion across spontaneous and non-spontaneous data, with a high degree of accuracy.

To further analyze using a statistical shape model with global and local constraints on non-spontaneous and spontaneous data, we looked at the standard deviation of the variation in emotion for the top-ranked landmarks. The non-spontaneous database presented a minimum standard deviation of 9.6445 and a maximum 10.7695. The persistent results for the standard deviation across the landmarks show that they are an effective method for facial expression recognition due to the consistent change across the landmarks. In other words, landmarks change in their position in different ways, but the amount they move is stable. In the spontaneous setting, a minimum standard deviation of 13.7695 and maximum of 15.555 was presented. The results for the spontaneous dataset validate what was observed in a non-spontaneous setting, even though there is more change across the data (higher standard deviations) the variance across the standard deviations is still low. Again, the facial landmarks vary in a consistent manner. This consistent variation in the facial landmarks across spontaneous and non-spontaneous data further details the power of the proposed method for expression recognition.

We also compared our results to the current state of the art in 3D facial expression recognition. Table 4 shows comparisons on non-spontaneous data and table 5 shows comparisons on spontaneous data. It is important to note that the majority of works on this spontaneous data detail action unit recognition [1], [31], [38], [39], not expression recognition as proposed in this study. As can be seen in tables 4 and 5, our proposed method outperforms the state of the art on spontaneous and non-spontaneous facial data.

Table 4. Comparison of proposed method to state of the art on non-spontaneous data for facial expression recognition.

Method	# of expressions	Accuracy
Proposed Method (RF)	**6**	**99.99%**
Proposed Method (SVM)	6	99.62%
Drira et al. [13]	6	93.21%
Abbasnejad et al. [1]	6	91.22%
Fang et al. [17]	6	91.00%
Yin et al. [36]	6	90.44%
Danelakis et al. [11]	6	90.00%
Canavan et al. [9]	6	84.80%
Sun et al. [30]	6	83.70%
Beretti et al. [3]	6	79.40%
Jeni et al. [18]	6	78.18%
Reale et al. [23]	6	76.12%
Yang et al. [35]	6	75.90%
Fang et al. [16]	6	75.82%

Table 5. Comparison of proposed method to state of the art on BP4D database for facial expression recognition.

Method	# of expressions	Accuracy
Proposed Method (RF)	**8**	**99.69%**
Danelakis et al. [11]	8	88.56%

As shown in tables 4 and 5, the proposed method outperforms current state of the art on both types of data with an increase of 11.13% accuracy on spontaneous data and an increase of at least 6.78% accuracy on non-spontaneous data.

4. DISCUSSION

Across our experiments, results, and analysis, we have presented a novel method for facial expression recognition; which outperforms the current state of the art in both spontaneous and non-spontaneous settings. We have shown that the different types of data (spontaneous and non-spontaneous) display variation in different facial regions, showing an important difference in the features that are relevant for expression recognition. The non-spontaneous data showed most variation in expression near the mouth, while the spontaneous data showed variation across multiple facial regions, including the mouth and eyes. We have demonstrated that the proposed method is powerful for expression recognition due to the consistent nature of facial landmark variation. The proposed method outperforms the current state of the art on two public databases [36], [38], [39].

We are interested in further studying the complexity of expressions (e.g. embarrassment), as well as testing on larger and more challenging datasets [40]. To facilitate this, we will further develop our proposed method by incorporating multi-modal data, such as physiological, thermal, and action units, to gain insight into which modalities contribute the most to facial expression recognition. We will also compare our hand-crafted features to a deep feature representation.

5. REFERENCES

[1] I. Abbadnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey, "Using synthetic data to improve facial expression analysis with 3D convolutional networks," CVPR 2017.

[2] T. Baltrusaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," *Face and Gesture,* 2015.

[3] S. Berretti, A. Bimbo, and P. Pala, "Automatic facial expression recognition in real-time from dynamic sequences of 3D face scans," *Vis. Comput.,* 29(12):1333-1350, 2013.

[4] P. Besl, and N. McKay, "A method of registration of 3D shapes," IEEE Trans. on PAMI 14, pp. 239-256, 1922.

[5] L. Breiman, "Random forests," *Machine Learning,* 45(1):5-32, 2001.

[6] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and Regression Trees," Wadsworth and Brooks, Monterey, CA, 1984.

[7] S. Canavan and L., Yin, "Feature detection and tracking on geometric mesh data using a combined global and local shape model for face analysis," BTAS, 2015.

[8] S. Canavan, L. Yin, et al., "Landmark localization on 3D/4D range data using a shape index-based statistical shape model with global and local constraints," *Computer Vision and Image Understanding,* 139, 136-148, 2015.

[9] S. Canavan, Y. Sun, X. Zhang, and L. Yin, "A dynamic curvature based approach for facial activity analysis is 3D space," *Computer Vision and Pattern Recognition Workshops,* 2012.

[10] F. Chang, T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "ExpNet: Landmark-free, deep, 3D facial expressions," arXiv: preprint arXiv: 1802.00542, 2018.

[11] A. Danelakis, T. Theoharis, I. Pratikakis, and P. Perakis, "An effective methodlogy for dynamic 3D facial expression retrieval," *Pattern Recognition,* 52, pp. 174-185, 2016.

[12] J. Koenderink and A. van Doorn, "Surface shape and curvature scales," *Image and Vision Computing,* 10(8): 557-564, 1992.

[13] H. Drira, B. Amor, M. Daoudi, A. Srivastava, S. Berretti, "3D dynamic expression recognition based on a novel deformation vector field and random forest," *International Conference on Pattern Recognition,* 2012.

[14] D. Cosker, E. Krumhuber, and A. Hilton, "A facs valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling," *ICCV,* 2011.

[15] G. Fanelli, J. Gall, et al., "A 3-d audio-visual corpus of affective communication," *IEEE Trans. on Multimedia,* 2010.

[16] T. Fang, X. Zhao, S. Shah, and I. A. Kakadiaris, "4D facial expression recognition," *International Conference on Computer Vision,* 2011.

[17] T. Fang, et al., "3D/4D facial expresion analysis: an advanced annotated face model approach," *Image and Vision Computing,* 30(10):738-749, 2010.

[18] L. Jeni, et al., "3D shape estimation in video sequences provides high precision evaluation of facial expression," *Image and Vision Computing,* 30(10):785-795, 2012.

[19] S. Koelstra, C. Muhl, et al., "Deap: A database for emotion analysis; using physiological signal," *IEEE Transactions on Affective Computing,* 3(1):18-31, 2012.

[20] J. Lewis, "Fast Template Matching," Vision Interface, 1995.

[21] G. McKeown, M. Valstar, et al. "The semaine database: An annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing,* 3(1):5-17, 2012.

[22] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. on PAMI,* 23(10):1175-1191, 2001.

[23] M. Reale, X. Zhang, and L. Yin, "Nebula feature: a space-time feature for posed and spontaneous 4D facial behavior analysis," *Face and Gesture,* 2013.

[24] G. Sandbach, et al., "Recognition of 3D facial expression dynamics," *Image Vision Computing,* 30(10):762-773. 2012.

[25] A. Savran, H. Alyuz, et al., "Bosphorus database for 3d face analysis," *Biometrics and Identity Management,* 47-56, 2008.

[26] A. Savran, S. Bulent, and M. T. Bilge, "Comparative evaluation of 3D vs. 2D modality for automatic detection of facial action units," *Pattern Recognition,* 45(2): 767-782, 2012.

[27] M.Schurgin, J. Nelson, S. Lida, H. Ohira, J. Y. Chiao, and S. Franconeri, "Eye movements during emotion recognition in faces," *Journal of Vision,* 14(13), 2014.

[28] M. Soleymani, J. Lichtenauer, et al., "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. on Affective Computing,* 3(1):42-55, 2012.

[29] G. Stratou, A. Ghosh, et al., "Exploring the effect of illumination on automatic expression recognition using the ict-3drfe database," *Image and Vision Computing,* 30(10), 2012.

[30] Y. Sun, X. Chen, et al., "Tracking vertex flow and model adaptation for 3D spatio-temporal face analysis," IEEE Trans. on SMC-A, 40(3)::461-474, 2010.

[31] M. Valstar, J. Girard, et al., "FERA '15 2nd facial expression recognition and analysis challenge," *Face and Gesture,* 2015.

[32] M. Valstar, E. S.-Lozano, et al., "FERA 2017 – Addressing head pose in the third facial expression recognition and analysis challenge," arXiv: 1702.04174, 2017.

[33] K. Weir, "A complex emotion," *Monitor on Psychology,* 43(10): 62, 2012.

[34] S. Wang, Z. Liu, et al., "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Trans. on Multimedia,* 12(7), 2010.

[35] H. Yang, and L. Yin, "CNN based 3d facial expression recognition using masking and landmark features," International Confrence on Intelligent Interaction, 2017.

[36] L. Yin, X. Chen, et al., "A high-resolution 3d dynamic facial expression database," *Face and Gesture,* 2008.

[37] L. Yin, X. Wei, et al., "A 3d facial expression database for facial behavior research," *Face and Gesture,* 2006.

[38] X. Zhang, L. Yin, et al., "BP4D-Spontaneous: A high resolution 3D dynamic facial expression database," *Image and Vision Computing,* 32(10):692-706, 2014.

[39] X. Zhang, L. Yin, et al., "A high resolution spontaneous 3D dynamic facial expression database," *Face and Gesture,* 2013.

[40] Z. Zhang, J. Girard, et al., "Multimodal spontaneous emotion corpus for human behavior analysis," *Computer Vision and Pattern Recognition,* 2016.