PROCEEDINGS OF SCIENCE

Training on Statistical Feature Models of Action Units for 3D Facial Expression Recognition

Zhenjiang Dong
Shanghai Jiaotong University
Shanghai, 200240, China
E-mail: dong.zhenjiang@zte.com.cn

Xia Jia
ZTE Corporation
Nanjing, 210012, China
E-mail: jia.xia@zte.com.cn

Wanshun Gao[1]
Xi'an Jiaotong University
Xi'an, 710049, China
E-mail: feiwuzefang@163.com

Kai Wang
Xi'an Jiaotong University
Xi'an, 710049, China
E-mail: wk19890107@stu.xjtu.edu.cn

Most of the existing studies on 3D Facial Expression Recognition (FER) are messaged-based approaches, which only detect the already known six universal expressions. In this paper, we describe the group of global and local features used to comprehensively characterize facial activities. These features are further used to train Statistical Feature Models (SFMs) associated with each Action Unit (AU). The occurrence probability of a specific AU on an input textured 3D face model is then computed. The results demonstrate that the evidence of AUs is of importance for applying AU space to evaluate expressions.

ISCC2017
16-17 December 2017
Guangzhou, China

[1]This work is supported by the National Nature Science Foundation of China (Grant No. 61303121, Grant No. 71731009, Grant No. 71732006, Grant No. 71742005, Grant No. 91546119), the Ministry of Education & China Mobile Joint Research Fund Program (MCM20160302), Science and Technology Planning Project of Guangdong Province (Grant No. 2014B040404043) and ZTE Industry-Academia-Research Cooperation Funds.

PoS(ISCC 2017)021

1.Introduction

The analysis of facial expressions has various purposes and applications, involving human computer interface, face recognition, psychological studies, tiredness detection, and face animation, etc. There are two main streams of facial expression recognition in the literature, which are judgment-based approaches and sign-based approaches. The first category directly associates specific facial patterns with a predefined number of discrete classes. The most commonly used ones are the six basic expressions [1]. The second category abstracts and codes facial muscle activities by facial action Action units Units (AUs) in FACS [2] and then claims that the combination of detected AU can be interpreted into a variety of expression states by high-level decision making, using Emotional FACS (EMFACS) rules [3], for instance [3].

Textured 3D faces, which capture both facial texture and geometry information, have gained increased increasing interests for Facial Expression Recognition (FER). Meanwhile, most of the existing studies on 3D FER are messaged-based approaches which only detect the six universal expressions [4 - 9]. On the other hand, work on AU detection for FER is mostly based on 2D texture images or 2D videos [10]. In [11], 14 AUs were detected and exploited on image sequences. Dynamic texture was used to recognize 18 AUs on image sequences in [12]. Furthermore, most work on AU-based FER simply performs AUs detection [13] and there are very few studies really conducting interpretation of detected AUs for explicit FER through high-level decision making (e.g., EMFACS rules).

In this paper, we propose a statistical AU space is proposed to which can be act as an alternative for FER through the interpretation of AU combinations. In this statistical AU space, the origin represents the neutral state while axes represent AUs which that are describable by a statistical learning. A facial expression can thus be represented by a point in the statistical AU space. The coordinate of an expression is large in an AU axis when the underlying statistical models predict a high occurrence probability of the corresponding AU in the expression display. This coordinate would be small in case that it's predicted action is low. To estimate the location of an displayed expression in the statistical AU space, we propose to make use of similarity scores as computed in [14], by using statistical feature models.

Section 2 is a brief introduction on the frameworkIn section 2 we briefly introduce the framework. The use of this statistical AU space for explicit FER is then described in Section 3. The experimental results are presented in Section 4 and our conclusion is drawn in Section 5.

2. Building Statistical AU Space

The occurrence probabilities of AUs displayed on a face are computed as the sum of matching scores between both global and local features extracted from a textured 3D facial model and the corresponding statistical feature models associated with AUs that we have been previously proposed [14]. To begin with, we describe there is the description of the set of global and local features used to comprehensively characterize facial activities. These features are further used to train Statistical Feature Models (SFMs) associated with each AU. The occurrence probability of a specific AU on an input textured 3D face model is then computed.

2.1 Extraction of Global and Local Features

AUs or expressions cause face deformations. These deformations impact three facial representations, which are facial texture, facial morphology and facial geometry. Specifically, facial texture describe spots, wrinkles, furrow and so forth that we can be observed during a facial expression. Facial morphology contains a group of landmarks depicting different facial organs. Facial geometry aims to describe exact facial surface shape, e.g. such as cheek swelling, mouth opening, delivered by 3D faces. Different facial expressions differently impact these three facial properties. For instance, without moving the eye corners, the texture in the eye region is changed significantly by AU7 and AU43. The texture and local geometry in the mouth region are changed mostly by AU24, with less effects on landmarks. Therefore, as textured 3D face data has these three face representations, it is essential to extract different features simultaneously. There are in total In total, 15 features of AUs are extracted from these three facial modalities. They are the offsets of landmarks *D* which are away from a mean neutral face, intra distances of these landmarks *L* for the morphology modality, the texture patches around each landmark *G* and the local binary patterns at five scales *LBPt* 1−5 for the texture modality, the local range patches in the vicinity of each landmark, *Z*, the shape index for each local range patch *SI* and *LBPs* computed on the range patches also at 5 scales *LBPr* 1−5 for the geometry modality. The details on the extraction of these features can be found in referred to [14].

2.2 Statistical Feature Models

Given a the feature as extracted previously, a SFM is trained using PCA for each AU to model the corresponding AU occurrence evidence. Based on a training set of 3D face models displaying a particular AU i_x , all the 15 features are extracted. For each feature F_l, PCA is applied to preserve the 95% major of the variation modes of the features extracted from the 3D face models displaying the AU i_x .

$$F_l^{i_x} = \bar{F}_l^{i_x} + P_l^{i_x} b_l^{i_x} \tag{2.1}$$

where $\bar{F}_l^{i_x}$ is the mean feature, $P_l^{i_x}$ is the set of eigenvectors calculated from PCA, and $b_l^{i_x}$ is a set of parameters which obey Gaussian distributed with the mean of zero and the standard deviation of $\sigma_l^{i_x}$ where *j* refers to each parameter of $b_l^{i_x}$. Given a new face *k*, a feature instance $\hat{F}_{lk}^{i_x}$ from the learnt model can be obtained by evaluating first the best parameter $b_l^{i_x}$ using feature F_{lk} extracted from *k* :

$$b_l^{i_x} = P_l^{i_x T}\left(F_{lk} - \bar{F}_l^{i_x}\right) \tag{2.2}$$

We set a boundary ($\pm 0.5\sigma_l^{i_x}$) is set for the corresponding parameter in $b_l^{i_x}$ to form $\hat{b}_l^{i_x}$ in order to restrain the face deformation and thus to build up their separability. Then To follow that, the feature $\hat{F}_{lk}^{i_x}$ instance is computed by inputting $\hat{b}_l^{i_x}$ in Eq. 1.

2.3 Computation of AU Matching Scores

The occurrence probability of a specific AU given by an observed feature from a 3D face model can then be computed as the matching score between the observed feature and its approximation, by using the corresponding SFM for the given AU. Specifically, the matching

score $Q_{lk}^{i_x}$ between the feature F_{lk} and its instance $\hat{F}_{lk}^{i_x}$ is computed as the normalized correlation response as defined in Eq. 3.

$$Q(F_{lk}^{i_x}) = \langle \frac{F_{lk}}{\|F_{lk}\|}, \frac{\hat{F}_{lk}^{i_x}}{\|\hat{F}_{lk}^{i_x}\|} \rangle \tag{2.3}$$

The occurrence evidence of a specific AU given by a textured 3D model is then simply a weighted sum of these matching scores for all the 15 features as previously extracted from the 3D face model. The matching score computation computing process uses SFM. Specifically, the final matching score of an input face model *k* that displaying displays an AU i_x is then a weighted sum of the matching scores produced by all features.

$$Sc_k^{i_x} = \sum_{l=1}^{N_c} A_l Q_{lk}^{i_x} \tag{2.4}$$

Wwhere A_l is a set of weights (all set to 1). The higher matching score achieved by an AU i_x on in the input face model *k* indicates that the input face model has a higher occurrence probability of the AU i_x .

3. Using the Statistical AU Space Forinterpreting for Facial Expression Interpretation

The statistical AU space is constructed with AU bases. The origin of the space is neutral; , along and the positive direction of each AU axis is occurrence probability associated with the AU. As there are 44 AUs in FACS, the full AU space should have 44 dimensions. An expression displayed on a static data is thus represented by a point in this AU space. An expression from an image sequence is thus represented by a continuous line, describing the occurrence probability at each instantinstantly.

Now, given a the learning dataset of textured 3D face models displaying a combination of AUs, we can compute the occurrence evidence for each AU on an input 3D face model can be computed, through using the use of matching scores. The input 3D face model displaying a facial expression can thus be represented by a point within this AU space, with the neutral state as its origin. For example, the expression of surprise can be a combination of AU2 (Outer Brow Raiser), AU26 (Jaw Drop) and AU1 (Inner Brow Raiser), so that relatively higher scores are expected on the axes of these AUs and lower scores on axes of other AUs. Meanwhile, it is quite hard to have get enough training data for all these 44 AUs. In this paper, we conducted a preliminary evaluation is conducted for applying statistical AU space to 3D FER and we carried out the experiments using the Bosphorus dataset [15] are carried out.

Specifically, we train a set of SFMs corresponding to different AUs for each feature are trained. For a face, we extract 15 features are extracted as described in section 2.1 and compute 15 sets of scores using SFM are computed as described mentioned in Section 2.3. These score sets can be either summed together and and then fed into one SVM classifier[16], or fed to 15 SVM classifiers respectively. When testing on a given face with expression, a SVM classifier outputs seven probabilities in relation corresponding to the neutral and other six universal expressions. For the method using one classifier, expression is recognized by choosing the highest value in the probability set. For the method using 15 classifiers, 15 sets of probabilities from all classifiers are firstly summed to obtain the probability set for recognition.

4. Experimental Results

As there is no accessible 3D face data displaying all 44 AUs, face facial scans displaying 16 AUs from the Bosphorus database[15] are used for training on SFMs. These AUs are listed in Table 1. 3D face data describing the six universal and neutral expressions are involved in score computation. Their scores are further used for training and testing for on AU space- based facial expressions interpretation.

Action Unit	Description	Action Unit	Description
2	Outer Brow Raiser	4	Brow Lowerer
7	Lid Tightener	9	Nose Wrinkler
10	Upper Lip Raiser	12	Lip Corner Puller
14	Dimpler	17	Chin Raiser
18	Lip Puckerer	22	Lip Funneler
24	Lip Presser	26	Jaw Drop
27	Mouth Stretch	28	Lip Suck
34	Puff	43	Eyes Closure

Table 1: Description of Action Units as defined in FACS. The first and third columns list the AUs detected in the experiments, the second and fourth columns give describe the corresponding AUs.

The statistical AU space is thus constructed with only these 16 AU axes. A ten-fold person-independent cross-validation is done Specifically, we carried out a ten-fold person-independent cross-validation for the computation of the statistical AU space coordinates for each facial scan. In each round, facial scans displaying AU from 45 subjects are used to train SFMs and then are computedconfigured with the coordinates of all the facial scans displaying expression from the remaining 5 subjects in the statistical AU space, i.e. the matching scores on the 16 AU axes. Once after computed the computation of the coordinates of all face scans from the 50 subjects in this statistical 16 AU space is completed, we used a leave-one-subject-out methods for FER is adopted, where which is, in each round, scores from 49 subjects are used selected to train SVMs and the left subject are used for testing.

Input \ Output	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutral
Anger	76.0%	11.0%	0.0%	0.0%	11.0%	0.0%	2.0%
Disgust	4.0%	82.0%	4.0%	2.0%	8.0%	0.0%	0.0%
Fear	0.0%	2.0%	62.0%	0.0%	2.0%	30.0%	4.0%
Happiness	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
Sadness	10.0%	4.0%	0.0%	0.0%	74.0%	0.0 %	12.0 %
Surprise	0.0%	0.0%	8.0%	0.0%	0.0%	92.0%	0.0%
Neutral	2.0%	0.0%	0.0%	0.0%	0.0%	18.0%	80.0%

Table 2: Confusion Matrix of the Expression Recognition based on Statistical AU Space.

The overall average recognition rate for of the two methods are 79.3% and 80.9% respectively. Due to the space limitation, we show the confusion matrix is only shown on with the second method in Table 2, where 15 SVM classifiers are adopted. The performance can certainly be improved, only if instead of a statistical AU space with only 16 axes, one constructs an an AU space having with higher dimensions by including most of the AUs which are highly relevant to the six universal expressions. Indeed, some important AUs required for the

interpretation of the six universal expressions are missing within this 16 AU space, including in particular AU1 (Inner Brow Raiser), AU15 (Lip Corner Depressor) and AU25 (Lips Part). Another source approach of for performance improvement is a the more accurate computation of occurrence evidences of AUs from 3D facial scans. Indeed, as a matter of fact, we can see as shown from in Table 2, the recognition rates for of surprise and disgust are rather good, while the recognition rates for those of anger and fear can be much improved are in need of enhancement. When connected to making connection with our former work on the recognition of AUs using SFMs [14], it can be observed we can observe from Table 3 that the detection of AUs relevant to disgust and surprise is quite accurate, while the detection of AUs relevant to anger and fear is less accurate, which implies the inaccuracy of that the occurrence evidence of these AUs as their coordinates in the statistical AU space are not accurate. This suggests indicate that a reliable inference of occurrence evidence of AUs is of importance signifiance for applying AU space to evaluate expressiones.

Input \ Expression	**Related AUs**
Anger	AU7*** (78.3%); AU17*** (80.0%); AU22* (90.0%); AU26* (91.7%)
Disgust	AU9** (81.7%); AU10** (95%); AU26* (91.7%)
Fear	AU4*** (75.0%); AU26* (91.7%); AU27* (91.7%)
Happiness	AU12** (85.0%)
Sadness	AU4** (75.0%); AU17** (80.0%)
Surprise	AU2*** (90%); AU26**(91.0%); AU27** (91.7%)

Table 3：Relevance between Expressions and AUs

The AUs in the column 'Related AUs' are those detected in our former work [14], an example of high level decision rules for interpreting AU into emotions; number of stars indicates the importance of an AU for the relevant expression; the percentage in brackets are positive rates of AU in our former work [14].

4. Conclusion

We have proposed a statistical AU space, which can be used taken as an alternative to FER, according to sign-based approaches. Such a statistical AU space partially releases the strong copes with the constraint that all involved AUs in a facial expression should shall be correctly precisely detected when being applied toying the existing high-level decision making, e.g., EMFACS rules, for FER. Meanwhile, this preliminary study on statistical AU space-based approach leaves much space for further improvement on FER. Firstly, one needs there is a necessity to provide reliable occurrence evidence of AUs on facial scans, as FER is then carried out on the coordinates of the statistical AU space. Secondly, the choice of axes of the statistical AU space is also importantimportant, as they should include the most important significant AUs required for the facial expression interpretation.

References

[1] P. Ekman and W. V. Friesen, "*Constants across cultures in the face and emotion,*" journal of Personality and Social Psychology, vol. 17, no. 2, pp. 124–129, 1971.

[2] P. Ekman and W. V. Friesen, "*Facial action coding system: A technique for the measurement of facial movement,*" Consulting Psychologists, 1978.

[3] http://face-and-emotion.com/dataface/general/homepage.jsp.

[4] S. Berretti, A. Del Bimbo, P. Pala and B. Ben Amor, and M. Daoudi, "*A set of selected SIFT features for 3D facial expression recognition*," ICPR, 2010.

[5] H. Tang and T. S. Huang, "*3D facial expression recognition based on automatically selected features*," CVPR workshop, pp. 1–8, 2008.

[6] H. Soyel and H. Demirel, "*3D facial expression recognition with geometrically localized facial features*," Symposium on Com. Sci. and Info. Tech., pp. 1–4, 2008.

[7] J. Wang, L. Yin, X. Wei, and Y. Sun, "*3D facial expression recognition based on primitive surface feature distribution*," CVPR, pp. 1399–1406, 2006.

[8] I. Mpiperis, S. Malassiotis, and M.G. Strintzis, "*Bilinear models for 3D face and facial expression recognition*," IEEE Trans. on Info. Fore. and Secu., vol. 3, no. 3, pp. 498–511, 2008.

[9] S. Ramanathan, A. Kassim, Y.V. Venkatesh, and W. S. Wah, "*Human facial expression recognition using a 3D morphable model*," ICIP, pp. 661–664, 2006.

[10] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, "*A survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions*," IEEE Trans. on PAMI,vol.31, No.1, pp. 39–58, 2009.

[11] Y. Tong, W. Liao, and Q. Ji, "*Facial action unit recognition by exploiting their dynamic and semantic relationships,*" IEEE Trans. of PAMI, vol. 29, no. 10, pp. 1683– 1699, 2007.

[12] S. Koelstra, M. Pantic, and I. Patras, "*A dynamic texture-based approach to recognition of facial actions and their temporal models*," IEEE Trans. of PAMI,, vol. 32, no. 11, pp. 1940 –1954, 2010.

[13] A. Savran and B. Sankur, "*Automatic detection of facial actions from 3D data*," ICCV09: Workshop on Human Computer Interaction, 2009.

[14] X. Zhao, E. Dellandrea, L. Chen, and D. Samaras:, "*Au´ recognition on 3D faces based on an extended statistical facial feature model*," Inter. Conf. on BTAS, pp. 1 –6, Sep. 2010.

[15] A. Savran, N. Alyuz, H. Dibeklioglu, O. Celiktutan, B. Gokberk, B. Sankur, and L. Akarun, "*Bosphorus database for 3D face analysis*," The First COST 2101 Workshop on Bio. and Ident. Manag., 2008.

[16] C. Chang and C. Lin, "*LIBSVM: a library for support vector machines*," Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.