Evaluation of Texture and Geometry For Dimensional Facial Expression Recognition

Ligang Zhang[1], Dian Tjondronegoro[1], Vinod Chandran[2]
[1]Faculty of Science and Technology; [2]Faculty of Built Environment and Engineering
Queensland University of Technology, Brisbane, QLD 4000, Australia
ligzhang@gmail.com, dian@qut.edu.au, v.chandran@qut.edu.au

Abstract—**Facial expression recognition (FER) algorithms mainly focus on classification into a small discrete set of emotions or representation of emotions using facial action units (AUs). Dimensional representation of emotions as continuous values in an arousal-valence space is relatively less investigated. It is not fully known whether fusion of geometric and texture features will result in better dimensional representation of spontaneous emotions. Moreover, the performance of many previously proposed approaches to dimensional representation has not been evaluated thoroughly on publicly available databases. To address these limitations, this paper presents an evaluation framework for dimensional representation of spontaneous facial expressions using texture and geometric features. SIFT, Gabor and LBP features are extracted around facial fiducial points and fused with FAP distance features. The CFS algorithm is adopted for discriminative texture feature selection. Experimental results evaluated on the publicly accessible NVIE database demonstrate that fusion of texture and geometry does not lead to a much better performance than using texture alone, but does result in a significant performance improvement over geometry alone. LBP features perform the best when fused with geometric features. Distributions of arousal and valence for different emotions obtained via the feature extraction process are compared with those obtained from subjective ground truth values assigned by viewers. Predicted valence is found to have a more similar distribution to ground truth than arousal in terms of covariance or Bhattacharya distance, but it shows a greater distance between the means.**

Keywords- facial expression recognition, dimensional space, continuous value, SIFT, FAP

I. INTRODUCTION

Facial expression is an important channel for humans to perceive attitudes, express opinions and convey reactions in human-human interaction. Accordingly, facial expression recognition (FER) becomes increasingly significant in many areas, such as human-computer interaction, patient monitoring and driver condition assessment. A suitable reliable representation of facial expressions will improve the feasibility for use in real applications.

Facial expressions can be generally represented in three ways: discrete emotions recognized universally across different cultures (e.g. happiness and surprise), facial action units (AUs) defined in the facial action coding system (FACS), and dimensional spaces. Compared with discrete emotions and AUs, dimensional spaces [1] have the advantage of using values along continuous axes to represent a *wide* range of emotions. It can provide unique insights into the relationship between emotions and emotional intensity. Representing emotions in a dimensional space is suitable for applications such as image retrieval and clustering according to emotional content. The most popular dimensional space is arousal-valence (AV) as shown in Fig. 1, where the arousal axis denotes the level of activation, while the valence axis stands for the degree of pleasantness.

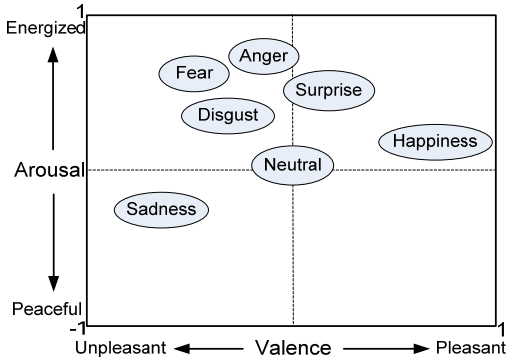


Figure 1. Arousal-valence dimensional space.

The majority of present work [2] on FER focus on discrete emotions and AUs, and relatively little attention has been paid to dimensional emotion recognition. Most approaches [3] reported on FER in dimensional spaces attempt to quantize the dimensions into an arbitrary number of levels, such as the four quadrants [4], or negative and positive emotions [5] using multi-modal information (e.g. facial images combined with shoulder and audio cues). Essentially these approaches belong to the discrete emotion classification set. Some recent studies [6] attempt to predict human affect in a continuous dimensional space using multimodal fusion of verbal and nonverbal behavioral events such as audio features and head events rather than facial expressions. Studies [7] have demonstrated that facial expressions contribute for 55% to the effect of the spoken message. Efforts [8] have also been made to map the facial expressions into a dimensional space using manifold learning techniques. The manifold spaces in these approaches, however, are not linked with dimensional values.

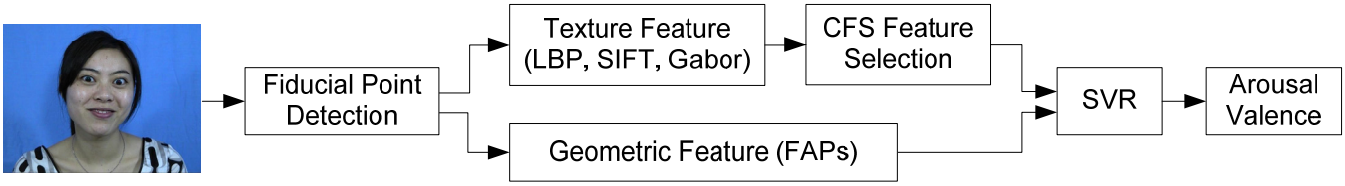


Figure 2. Framework of the evaluation system.

A few previous publications in this area have reported on recognizing facial expressions using continuous dimensional values. Michael *et al.* [9] estimated three space attributes, valence, activation and dominance, using Gabor features and a neuro-fuzzy classifier. Yangzhou *et al.* [10] built a linear mapping to represent expressional face images in the arousal-valence (AV) space. Yeasin *et al.* [11] mapped facial expressions into levels of interest based on a three-dimensional space and the intensity of optical flow. Nicolaou *et al.* [12] predicted facial expressions into an AV space learning the non-linear dependence between input from 20 facial points and the desired output over a pre-defined temporal window. All these methods have either only used texture or geometric features. As texture and geometry convey complementary and important information about facial expressions [13], it is worthwhile investigating whether a fusion can improve the performance of dimensional emotion recognition. Texture features, as used in these approaches, are designed for frontal faces and sensitive to large face movements. This makes them difficult to apply to real systems and it is desirable to evaluate performance on spontaneous images obtained from human conversations and annotated independently. All works except for [12] have been evaluated using self-built ground truths of dimensional values that are not assessable publicly. This paper will address these issues.

A framework is adopted to evaluate the performance of fusing different texture features with facial animation parameter (FAP) based geometric features for representing facial expressions in a continuous arousal-valence dimensional space. Three most widely used texture descriptors, LBP, SIFT and Gabor, are extracted around 53 fiducial points derived automatically from a well-trained active shape model (ASM). A subset of the most discriminative texture features of each type is selected using the correlation-based feature selection (CFS) algorithm. The geometric features consist of 43 distances between fiducial points defined based on FAPs. FAP based distances have been demonstrated to be a sparse, compact, yet information-rich representation of the facial shape [14]. Each type of text feature set is evaluated by itself and fused with geometric features. Regression of arousal and valence on two recently introduced public spontaneous databases, NVIE and Semaine, is achieved using support vector regression (SVR).

The rest of the paper is organized as follows. Section II describes details of the framework of the evaluation system. Section III presents the experimental results. Conclusions are drawn in Section IV.

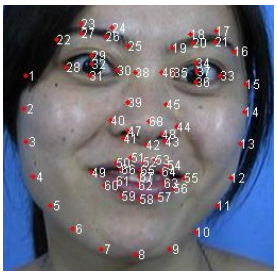


Figure 3. 68 fiducial points for training an ASM.

II. EVALUATION FRAMEWORK

Fig. 2 shows the framework of the evaluation system. For an input image, the face is located using the Viola-Jones detector [15] and 68 fiducial facial points are detected using a well-trained ASM. LBP, SIFT and Gabor texture features are extracted around each of 53 interior points, and the vectors from all points are concatenated into a final vector for each type of feature. A subset of the most discriminative texture features is selected using the CFS algorithm. The geometric feature vector is composed of 43 distance features defined based on an ASM and FAPs. Feature vectors are put into a SVR with a radial basis function kernel for regressing arousal and valence dimensions of emotions. Performances of each of the three textures individually, FAP geometry feature on its own, and feature level fusion of each texture feature with the geometric feature are then evaluated on the NVIE and Semaine databases.

A. Face and Fiducial Point Detection

For an input image, the face is first detected using the widely used Viola-Jones detector, and 68 facial fiducial points are then detected using an ASM [16]. ASM is known for its robustness in fitting and tracking fiducial points in human faces. To train the ASM, we collected 100 images from the internet with different emotions and different poses ranged from -20 to 20 degrees. The 68 fiducial points as shown in Fig. 3 are manually annotated with x and y locations. The trained ASM is anticipated to work well on faces with normal face movements. It has been observed that the points in the face boundary (index from 1 to 15 in Fig. 3) are not always accurately detected by the ASM due to face shape changes between subjects and face movements. Moreover, the regions around these points contain background information and do not provide reliable texture features. Therefore, only 53 interior points (index from 16 to 68 in Fig. 3) are used to extract texture features.

B. Texture Feature Extraction

To achieve a degree of tolerance to face movements and pose changes, the texture features are extracted in a patch around each of 53 interior points, and the features of all points are then combined into a feature vector. This method of extracting texture features around fiducial points have been successfully used in building robust features in FER [17] algorithms. Three texture features, including LBP, Gabor and SIFT, are used and compared here due to their excellent performance in FER [2].

(1) Local binary patterns (LBP) [18] label each pixel in an image by applying the center value as thresholds to neighborhood pixels and considering the result as a binary number, then collecting up the occurrence of different binary patterns, yielding a histogram as the texture descriptor of the image. LBP has the advantage of tolerance against illumination changes and computational simplicity. In this paper, we collect uniform patterns from a 14×18 patch centered at each point, resulting in a histogram with 2,597 bins for all points.

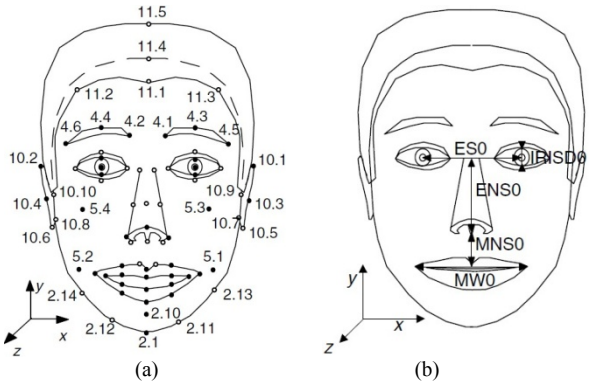
(2) Gabor features are extracted by performing multi-orientation and multi-scale filtering on an image. Following the common setting of Gabor parameters, this paper uses five scales $\lambda_\theta = 4 \times 2^{\sqrt{m-1}}$ $\mathrm{m} = (1,...,5)$ and eight orientations $\theta_n = \pi(n-1)/8$ $\mathrm{n} = (1,\ldots,8)$ Gabor filters. Therefore, we obtain 40 Gabor magnitude coefficients for each point and a final feature vector with 2,120 elements.

(3) Scale-invariant feature transform (SIFT) [19] is a distinctive invariant feature set that is suitable for describing local textures. It is known to be invariant to image scale and rotation, and also can provide robust matching across a substantial range of affine distortions, changes in 3D viewpoint, noise and illumination. In this paper, the SIFT descriptor is computed from the gradient vector histograms of the pixels in a 4×4 patch around each point. Given 8 possible gradient orientations, each descriptor contains 128 elements and the final feature vector contains 6,784 elements.

C. Geometric Feature Extraction

Geometric Facial animation parameters (FAPs) [20] are defined in the ISO MPEG-4 standard (part 2, visual) to allow the animation of synthetic face models. They are based on the study of minimal perceptible actions and are closely related to muscle actions. FAPs contain 68 parameters that are either high level parameters describing visemes and expressions, or low level parameters describing displacements of the single points of the face as shown in Fig. 4a. Therefore, FAPs can provide a concise representation of the evolution of facial expression, and can represent a complete set of basic facial actions, including head motion, tongue, eye and mouth control. Furthermore, FAPs also can handle arbitrary faces through the use of FAP units (FAPUs), which are defined as the fractions of distances between key points as shown in Fig. 4b.

Geometric features include 43 distances between 53 interior points detected by ASM. As listed in Table I, these distances are calculated based on FAPs. Because the ASM produces several points on the eyebrow that are around the middle, there are several features for FAP No. 33 (marked FAP 33*).



(a) (b)

Figure 4. (a) A subset of feature points defined in the MPEG-4 FAPs standard and (b) FAP units defined based on ratios of distances between the marked key features.

TABLE I. DISTANCES BETWEEN FACIAL POINTS DEFINED BY FAPS

FAP No.	Distance	FAP No.	Distance	FAP NO.	Distance
3	Dy(52,58)	29	Dy(29,31)	38	Dx(35,19)
4	Dy(65,42)	30	Dy(34,36)	51	Dy(52,42)
5	Dy(62,42)	31	Dy(30,25)	52	Dy(58,42)
6	Dx(49,42)	32	Dy(35,19)	55	Dy(50,42)
7	Dx(55,42)	33*	Dy(32,23)	56	Dy(54,42)
8	Dy(66,42)	33*	Dy(32,24)	57	Dy(60,42)
9	Dy(64,42)	33*	Dy(32,26)	58	Dy(56,42)
10	Dy(61,42)	33*	Dy(32,27)	61*	Dx(30,40)
11	Dy(63,42)	34*	Dy(37,17)	61*	Dx(30,39)
12	Dy(49,42)	34*	Dy(37,18)	62*	Dx(35,44)
13	Dy(55,42)	34*	Dy(37,20)	62*	Dx(35,45)
19	Dy(29,32)	34*	Dy(37,21)	63	Dy(35,68)
20	Dy(34,37)	35	Dy(28,22)	64	Dx(35,68)
21	Dy(31,32)	36	Dy(33,16)	-	-
22	Dy(36,37)	37	Dx(30,25)	-	-

Note: Dx(M,N) and Dy(M,N) indicate the distances between two points indexed M and N in the horizontal and vertical directions respectively. The indices M and N of the points are based on the 53 interior points in Fig. 3.

Similarly, there are multiple distances for FAP No. 34, 61 and 62. The distances defined based on FAPs have been demonstrated as a sparse, compact, yet information-rich representation of the facial shape [14]. Compared with the commonly used facial movement vectors obtained in multi-frames, distance features have the merits of being robust to translations and rotations of the facial geometry, and do not require compensation for face movements. Therefore, they are suitable for working on real-world images in the proposed approach. To provide invariance to different faces, all distances are normalized based on FAPUs.

D. Discriminative Texture Feature Selection

Feature selection aims to select a subset of the most discriminative features from the texture feature vector. It has been shown that discriminative LBP bins selected by Adaboost achieve better performance than using all bins [21]. However, Adaboost cannot be directly used for feature selection in the regression problem here. Instead, we use the correlation-based feature selection (CFS) for this task and CFS has also been successfully applied for feature selection in predicting dimensional emotions [6].

CFS [22] is a simple, fast correlation based filter algorithm suitable for both classification and regression problems. It is designed based on the principle that good feature subsets are highly correlated with the ground truth class labels, yet un-correlated with other feature subsets. It evaluates the merit of a feature subset and only selects those with the highest scores. The core of CFS can be expressed as:

$$Q_s = k\overline{r_{cf}} \Big/ \sqrt{k + k(k-1)\overline{r_{ff}}} \quad (1)$$

where Q_s is the quality or merit of a feature subset S containing k features, $\overline{r_{cf}}$ the average feature-class correlation, and $\overline{r_{ff}}$ the average feature-feature correlation. To save searching time, the first best search is used. Starting with an empty feature set, the first best search generates all possible single feature expansions and selects the subset with the highest evaluation. The search stops when the number of selected features reaches a preset limit (300 in this case).

III. EXPERIMENTS

A. Databases

The natural visible and infrared facial expression (NVIE) database [23] is a newly developed comprehensive platform for both spontaneous and posed facial expression analysis. The spontaneous part consists of image sequences from onset to apex, collected from 105, 111, 112 subjects under front, left and right illumination respectively. The spontaneous expressions are induced by showing subjects film clips deliberately collected from the internet, resulting in images with face movements and different sizes of faces. All the visible apex images are labeled by five students with arousal and valence values ranging from -1 to 1. In this paper, only the images with final evaluated annotations under front and right illumination are used. After removing those failed during face and fiducial point detection, we get a total of 1,027 images. Fig. 5 shows sample images and their arousal and valence values.

Figure 5. Samples of NVIE images with Arousal and Valence values.

The Semaine corpus [24] is also a new database that contains videos with emotionalized conversations. Subjects are video recorded while holding a conversation with an operator who plays four different roles to evoke emotional reactions. The video is recorded at 49.979 frames per second at a spatial resolution of 780×580 pixels. All the videos are annotated by up to 4 raters with five affective dimensions (arousal, valence, power, expectation and overall intensity) as continuous values between -1 and 1. The available dataset consists of 100 conversational and 50 non-conversational recordings of approximately 5 minutes each, from 20 participants aging from 22 to 60. In this paper, only low-quality conversational videos are used, and 54 videos are selected by excluding those with start and end sessions. We then select 50 frames from each video so that these frames are evenly distributed over the video. After face and fiducial point detection, we obtain 2,474 frames for the experiment. The average annotation values of arousal and valence from all raters are used as the final annotation for each frame.

B. Experimental Set-Up

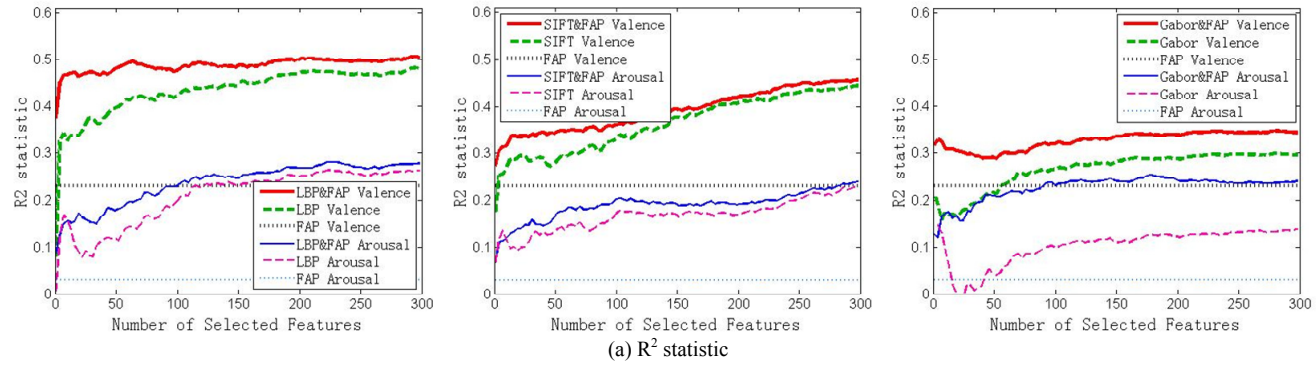
10 random subject-independent cross-validations are conducted to evaluate the performance in regressing arousal and valence dimensions. To be specific, we first divide all images into different sets according to the subject identity. Then we randomly select 10% for the testing set and the other 90% for the training set, and repeat the process 10 times to generate average performance. The performance is evaluated using four measurements: The R^2 statistic, Pearson correlation coefficient (CC), mean linear error (MLE), and Bhattacharyya distance (DB).

(1) The R^2 statistic measures the proportion of the variation of the observations around the mean that is explained by the fitted regression model. A value of 1 means the prediction results and ground truth are perfectly fitted, while a negative value means the data does not help the prediction. (2) Pearson correlation coefficient (CC) measures the strength of a linear relationship between two variables. It is defined as the covariance of the two variables divided by the product of their standard deviations. The absolute value of correlation coefficient is less than or equals to 1 where 1 corresponds to that two variables are perfectly correlated, while 0 implies no relationship. The larger the coefficient, the stronger is the association between two variables. (3) Mean linear error (MLE) measures the average of the absolute error between the predicted results and the ground true of the quantity being estimated. (4) Bhattacharyya distance (DB) measures the similarity of two probability distributions, by taking both the mean and covariance of the data into account. A distance close to 0 means that two distributions are similar, a larger value indicates a bigger difference.

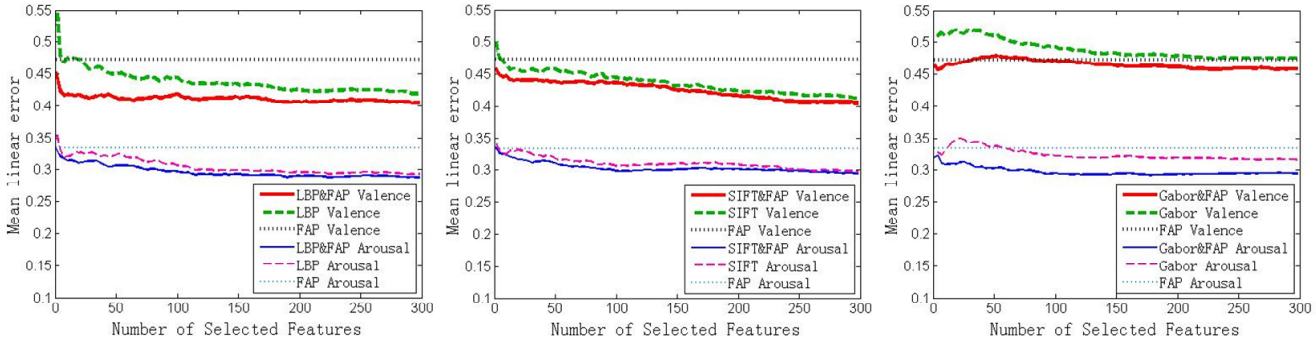
C. Performance Evaluation on NVIE database

Fig. 6 demonstrates the R^2 statistic and mean linear error (MLE) of the SVR generated arousal and valence values using different number of texture features plus 43 FAP features. Pearson correlation coefficients (CC) and Bhattacharyya distances (BD) obtained are similar to the R^2 statistic and MLE, respectively, and they are not shown here due to space limitation. Three types of feature combinations are used: texture feature alone, geometry (FAP) feature alone, and their fusion.

For all three texture features, fusion with geometry (FAP) features leads to only small performance improvements over using texture alone, but a significantly better performance than using FAP alone. Take the R^2 statistic of LBP features as an example, the result obtained using LBP+FAP is 0-8.6% higher than using LBP alone and 9.6-25.2% higher than using FAP alone when regressing arousal. When regressing valence, LBP+FAP achieves 2.4-14.0% higher R^2 statistic values than using LBP alone and 22.1-27.5% higher values than using FAP alone respectively. Similar results in terms of classification accuracy improvement were observed in our previous work on classifying basic discrete emotions and discriminating posed versus spontaneous emotions using fusion of texture and



(a) R^2 statistic



(b) Mean linear error (MLE)

Figure 6. Regression results of arousal and valence obtained using three texture features and FAP features on the NVIE database. As can be seen, for both arousal and valence, fusion of texture and FAP has a much better performance over FAP alone, but only a small performance improvement over texture alone. LBP+FAP achieve better overall performance than SIFT+FAP and Gabor+FAP for all measurements.

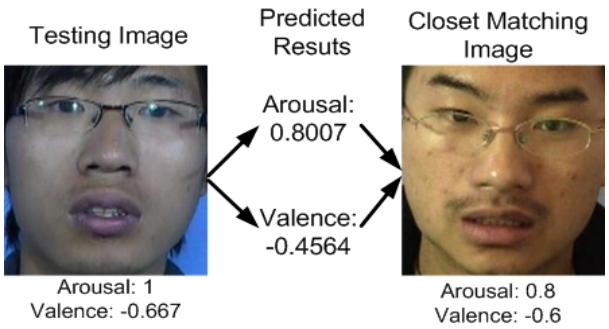


Figure 7. An example showing the values of Arousal and Valence predicted by the SVR for a testing image using LBP+FAP and the closest matching train set image for these values. Ground truths of arousal and valence are listed below each image.

TABLE II. REGRESSION RESULTS OBTAINED USING 100 TEXTURE FEATURES PLUS 43 FAP FEATURES ON THE NVIE DATABASE. THE BOLD FIGURES ARE THE BEST RESULTS AMONG ALL FEATURES.

	Arousal				Valence			
	R^2	CC	MLE	BD	R^2	CC	MLE	BD
LBP+ FAP	**0.230**	**0.498**	0.297	**0.053**	**0.475**	**0.690**	**0.415**	0.028
LBP	0.199	0.470	0.307	0.056	0.421	0.649	0.444	0.038
SIFT +FAP	0.202	0.487	0.299	**0.053**	0.361	0.611	0.436	**0.027**
SIFT	0.171	0.455	0.307	0.060	0.330	0.585	0.445	0.037
Gab.+ FAP	**0.230**	0.497	**0.296**	0.061	0.319	0.570	0.472	0.061
Gab.	0.098	0.367	0.322	0.090	0.266	0.519	0.493	0.090
FAP	0.029	0.345	0.333	0.101	0.230	0.564	0.472	0.053

geometry on the NVIE database. Geometrical FAP features only marginally improved performance in these cases. As the number of texture features increases, performance differences between texture plus FAP and texture become smaller.

LBP+FAP show the best overall performance for both arousal and valence. LBP+FAP regression on valence gives the best R^2 statistic and CC, and nearly the best MLE and BD values jointly with SIFT+FAP. Gabor+FAP show the worst performance of all the fused combinations. LBP+FAP and Gabor+FAP perform similarly when regressed to arousal and are marginally better than SIFT+FAP with respect to the R^2 statistic, MLE, and CC, while LBP+FAP and SIFT+FAP outperform Gabor+FAP in terms of BD. The highest overall performance of using LBP is probably due to its tolerance to illumination variations, shifting of key points from inaccurate ASM detection, and image scale changes [25]. Note that the facial images used here are directly derived from the Viola-Jones face detector with any pro-processing, such as illumination normalization and face alignment. Fig. 7 shows a testing image and its closet match in the train set to the arousal and valence values predicted by the SVR using LBP+FAP features.

There is a performance difference between arousal and valence. Cluster plots of the predicted and ground truth values were compared and it was observed that the mean values were shifted more for valence than arousal. This may explain the larger MLE values in Table II despite having higher correlation (R^2 and CC). Bhattacharya distances are nevertheless lower for valence in general. Better correlation appears to co-occur with

TABLE III. PERFORMANCE COMPARISON WITH PERVIOUS WORK

Met-hod		Arousal		Valence	
	Modality	CC	MLE	CC	MLE
Our	Facial expression	0.498	0.297	0.690	0.415
[9]	Facial expression	0.53	0.30	0.45	0.31
[6]	Head motion	0.204	0.208	0.037	0.258
	Event+audio+video	0.699	0.153	0.165	0.245

TABLE IV. REGRESSION RESULTS OF VIDEO FRAMES ON SEMAINE DATABASE

Met-hod	Arousal				Valence			
	R^2	CC	MLE	BD	R^2	CC	MLE	BD
LBP+ FAP	-0.307	-0.070	0.217	0.119	**0.002**	**0.241**	**0.206**	0.152
LBP	-0.355	-0.087	0.225	**0.103**	-0.005	0.221	**0.206**	0.187
FAP	**-0.282**	**0.003**	**0.213**	**0.103**	-0.161	0.093	0.225	**0.117**

higher error. It is worth noting that a result in this evaluation is contrary to the result presented in audio analysis [26] where arousal gets a higher R^2 than valence (0.583 versus 0.281). This result confirms psychological evidence [3] as well as the result in [12] indicating that visual cues (as opposed to audio) are more indicative of valence than arousal.

Table II gives the numerical values of the regression results of R^2, CC, MLE and BD obtained for arousal and valence using 100 LBP features and 43 FAP features. The best results in each column are highlighted in bold. LBP+FAP attains the best overall performance for both arousal and valence with R^2 statistic of 0.230 and 0.475, and CC of 0.498 and 0.690 for arousal and valence, respectively. The results confirm the benefits of fusing texture and geometry to improve the performance of dimensional emotional regression.

D. Performance Comparison with Previous Work

We also compared results obtained using LBP+FAP with these reported in previous work as show in Table III. Note that the works [9] and [6] are based on images selected from the VAM Corpus and videos segmented from the Semaine database. In addition, the results in [9] are obtained based on facial expressions, while those in [6] are obtained using audio, video, and event features, individually and in combination.

Table III shows that LBP+FAP has comparable performance to previous work, evaluated in terms of CC and MLE. It outperforms the work [9] which uses the same modality with 0.24 higher CC for valence and a 0.003 lower MLE for arousal, but it has a 0.032 lower CC for arousal and a 0.105 higher MLE for valence. Compared with the results in [6], LBP+FAP demonstrates better CC, but poorer MLE. The higher MLE using LBP+FAP is, to some extent, due to the fact that we do not restrict the predicted values of arousal and valence into a range of [-1, 1], while the previous work sets such a restriction. It also can be seen the last row in table III that fusion of multiple-modalities helps to improve the regression to dimensional representations of emotion.

E. Performance Tests on Semaine Video Frames

An interesting question is whether arbitrary still frames extracted from video segments with known emotion labels will result in similar performance using the LBP and FAP fusion method. To answer this question, we run an experiment on 2,474 frames from the Semaine database. Table IV shows the regression results obtained from this data using 100 LBP features and 43 FAP features. From the table, we can see that (a) the correlation between predicted values and the ground truth is poor for arousal, using texture features, geometric features and their fusion, (b) for valence, the correlation is poor for geometric features but not so bad with texture and fusion leads to a marginal improvement. Arbitrary video frames can contain faces with expressions not necessarily consistent with the emotion label of the entire video segment. Annotations in the Semaine database may rely on audio and head movement information within the video segment, which are absent in the arbitrarily selected still frames. There can also be larger head pose variations. Nevertheless, geometric features play a more important role for arousal in indicating the level of activation, whereas texture features are more important for valence in representing the degree of pleasantness in video. Further, regressing emotions using only facial expressions may be inadequate unless the expressive still images are appropriately selected.

IV. CONCLUSIONS

This paper evaluates the performance of recognizing spontaneous facial expressions in a continuous arousal-valence dimensional space using texture (LBP, Gabor, SIFT) and geometric (FAP) features. Experimental evaluations in terms of four measurements (R^2, CC, MLE, and BD) on the NVIE database demonstrate that fusion of texture and FAP features leads to only small performance improvements over texture alone, but a significant improvement over FAP alone, for both arousal and valence. Valence and Arousal behave differently and higher correlation appears to be accompanied by greater mean error values after regression. Dimensional emotion regression does not work well for arbitrarily selected still frames from annotated video segments but there still exists a fair correlation of regressed valence with ground truth values using texture features and this is improved by fusion with geometric features.

ACKNOWLEDGMENT

The research in this paper use the USTC-NVIE database collected under the sponsor of the 863 project of China, and the Semaine Database collected for the Semaine project (www.semaine-db.eu).

REFERENCES

[1] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology,* vol. 39, pp. 1161-1178, 1980.

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 31, pp. 39-58, 2009.

[3] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int'l Journal of Synthetic Emotion,* vol. 1, pp. 68-99, 2009.

[4] G. Caridakis, K. Karpouzis, M. Wallace, L. Kessous, and N. Amir, "Multimodal user's affective state analysis in naturalistic interaction," *Journal on Multimodal User Interfaces,* vol. 3, pp. 49-66, 2010.

[5] M. A. Nicolaou, H. Gunes, and M. Pantic, "Audio-Visual Classification and Fusion of Spontaneous Affective Data in Likelihood Space," in

Pattern Recognition (ICPR), 20th International Conference on, 2010, pp. 3695-3699.

[6] F. Eyben, M. Wollmer, M. F. Valstar, H. Gunes, B. Schuller, and M. Pantic, "String-based audiovisual fusion of behavioural events for the assessment of dimensional affect," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 322-329.

[7] A. Mehrabian, "Communication without words," *Psychology Today,* vol. 2 (9), pp. 52-55, 1968.

[8] Y. Chang, C. Hu, R. Feris, and M. Turk, "Manifold based analysis of facial expression," *Image and Vision Computing,* vol. 24, pp. 605-614, 2006.

[9] M. Grimm, D. G. Dastidar, and K. Kroschel, "Recognizing emotions in spontaneous facial expressions," in *Proceedings: International Conference on Intelligent Systems and Computing (ISYC)*, 2006.

[10] D. Yangzhou, B. Wenyuan, W. Tao, Z. Yimin, and A. Haizhou, "Distributing expressional faces in 2-D emotional space," in *Proceedings of the 6th ACM international conference on Image and video retrieval* Amsterdam, 2007, pp. 395-400.

[11] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *Multimedia, IEEE Transactions on,* vol. 8, pp. 500-508, 2006.

[12] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative RVM regression for dimensional and continuous emotion prediction," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 16-23.

[13] S. Mingli, T. Dacheng, L. Zicheng, L. Xuelong, and Z. Mengchu, "Image Ratio Features for Facial Expression Recognition Application," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* vol. 40, pp. 779-788, 2010.

[14] T. Hao and T. S. Huang, "3D facial expression recognition based on automatically selected features," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, 2008, pp. 1-8.

[15] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision,* vol. 57, pp. 137-154, 2004.

[16] 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, pp. 38-59, 1995.

[17] S. Berretti, A. D. Bimbo, P. Pala, B. B. Amor, and M. Daoudi, "A Set of Selected SIFT Features for 3D Facial Expression Recognition," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010, pp. 4125-4128.

[18] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition,* vol. 29, pp. 51-59, 1996.

[19] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision,* vol. 60, pp. 91-110, 2004.

[20] I. S. Pandzic and R. Forchheimer, *MPEG-4 facial animation: the standard, implementation and applications*: Wiley, 2002.

[21] C. Shan and T. Gritti, "Learning discriminative lbp-histogram bins for facial expression recognition," in *Proc. British Machine Vision Conference*, 2008.

[22] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Doctoral Dissertation, The University of Waikato, Department of Computer Science*, 1999.

[23] W. Shangfei, L. Zhilei, L. Siliang, L. Yanpeng, W. Guobing, P. Peng, C. Fei, and W. Xufa, "A Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Inference," *Multimedia, IEEE Transactions on,* vol. 12, pp. 682-691, 2010.

[24] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *Multimedia and Expo (ICME), IEEE International Conference on*, 2010, pp. 1079-1084.

[25] T. Gritti, C. Shan, V. Jeanne, and R. Braspenning, "Local features based facial expression recognition with face registration errors," in *Automatic Face & Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, 2008, pp. 1-8.

[26] Y. Yi-Hsuan, L. Yu-Ching, S. Ya-Fan, and H. H. Chen, "A Regression Approach to Music Emotion Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 16, pp. 448-457, 2008.