Image and Vision Computing 32 (2014) 1067–1079

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Representation of facial expression categories in continuous arousal–valence space: Feature and correlation☆

Ligang Zhang [a,b,*], Dian Tjondronegoro [b], Vinod Chandran [b]

[a] Faculty of Computer Science and Engineering, Xi'an University of Technology, 5 South Jinhua Road, Xi'an 710048, China
[b] Science and Engineering Faculty, Queensland University of Technology, 2 George St, Brisbane 4000, Australia

ARTICLE INFO

Article history:
Received 24 February 2014
Received in revised form 19 August 2014
Accepted 12 September 2014
Available online 6 October 2014

Keywords:
Facial expression recognition
Dimensional space
Continuous axis
Correlation
Categorized emotion

ABSTRACT

Representation of facial expressions using continuous dimensions has shown to be inherently more expressive and psychologically meaningful than using categorized emotions, and thus has gained increasing attention over recent years. Many sub-problems have arisen in this new field that remain only partially understood. A comparison of the regression performance of different texture and geometric features and the investigation of the correlations between continuous dimensional axes and basic categorized emotions are two of these. This paper presents empirical studies addressing these problems, and it reports results from an evaluation of different methods for detecting spontaneous facial expressions within the arousal–valence (AV) dimensional space. The evaluation compares the performance of texture features (SIFT, Gabor, LBP) against geometric features (FAP-based distances), and the fusion of the two. It also compares the prediction of arousal and valence, obtained using the best fusion method, to the corresponding ground truths. Spatial distribution, shift, similarity, and correlation are considered for the six basic categorized emotions (i.e. anger, disgust, fear, happiness, sadness, surprise). Using the NVIE database, results show that the fusion of LBP and FAP features performs the best. The results from the NVIE and FEEDTUM databases reveal novel findings about the correlations of arousal and valence dimensions to each of six basic emotion categories.

1. Introduction

Facial expression is an important means of perceiving attitude, expressing opinion and conveying reactions during human-to-human interactions and is also an indicator of fatigue or boredom. Accordingly, automated facial expression recognition (FER) is useful in areas such as human–computer interaction, psychological monitoring, and driver condition assessment. For FER facial expressions must be represented in some feature spaces of numerical or symbolic values. Facial expressions can be represented in three ways: categorized emotions (e.g. happiness and surprise), facial action units defined in the facial action coding system (FACS), and continuous dimensional spaces (e.g. arousal and valence). A dimensional space representation can provide unique insights into the intensity of emotions and the relationship between different emotions, and thus contain more psychologically meaningful information. Correlations between dimensional values and emotion categories can guide the recognition of a categorized emotion using continuous dimensional axes and can be useful in applications such as video content indexing [1].

Most of the current FER approaches use categorized emotions and facial action units. The use of dimensional representation is less common. Most FER algorithms in dimensional space quantize the dimensions into intervals such as high and low, and only few have used continuous values along the dimensional axes. There have been recent benchmarking studies, such as the 2012 and 2013 Audio/Visual Emotion Challenges (AVEC) [2,3]. Most of the approaches compared have used either texture or geometric features, not both. It is established that fusion leads to better performance for categorized emotions as it does with most classification problems. Whether a fusion of texture and geometry features can lead to better performance in the continuous arousal and valence dimensions has not been experimentally investigated. Further, nearly all the current knowledge of the correlation between continuous emotion dimensions and categorized emotions is directly adopted from psychological, cognitive, or neuroscience studies [4,5]. From these studies, it can be reasonably assumed that negative valence with negative arousal corresponds to sadness or boredom, but this is more of an abstract and relatively ambiguous correspondence and no explicit mapping between the two descriptions has been established [6]. No other work has been found that computes and estimates the correlations of arousal and valence dimensions to categorized emotions using publicly available databases. A key unanswered

☆ This paper has been recommended for acceptance by Stefanos Zafeiriou.

* Corresponding author at: Faculty of Computer Science and Engineering, Xi'an University of Technology, 5 South Jinhua Road, Xi'an 710048, China. Tel.: +61 7 3138 5074.

E-mail addresses: lg.zhang@xaut.edu.cn (L. Zhang), dian@qut.edu.au (D. Tjondronegoro), v.chandran@qut.edu.au (V. Chandran).

http://dx.doi.org/10.1016/j.imavis.2014.09.005
0262-8856/© 2014 Elsevier B.V. All rights reserved.

research question in this context is: Does arousal exhibit higher correlation with a categorized emotion (e.g. happiness) than valence? This paper will address such questions and issues.

A framework is proposed to evaluate the performance of different texture features fused with geometric distance features for representing facial expressions in a continuous arousal–valence space. The texture features include discriminative subsets of three most widely used texture descriptors: local binary patterns — LBP, scale-invariant feature transform — SIFT and Gabor filter outputs that have shown state-of-the-art FER performance [7]. The geometric features are 43 distances between fiducial points defined based on facial animation parameters (FAPs). Each type of texture feature set is evaluated by itself and evaluated fused with geometric features. Arousal and valence are regressed from the features using support vector regression (SVR). The best-performing combination of LBP and FAP features is adopted for further investigations. The predicted arousal and valence values using this combination are then compared with the corresponding ground truths, considering aspects such as spatial distribution, shift, similarity, and correlation for each of the six basic categorized emotions. Two databases, NVIE and FEEDTUM, are used to estimate the correlations between emotion dimensions and categorized emotions. The results are benchmarked with previous findings in psychological studies.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the evaluation framework. Section 4 presents experimental results. Conclusions are drawn in Section 5.

2. Related work

2.1. Facial expression representation

Facial expressions are represented using (a) emotion categories, (b) action units and (c) dimensional space.

The emotion category theory classifies an expression into one of predefined categories, such as six basic emotions — anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA), and surprise (SU), which are universal across different cultures and human ethnicities [8], and non-basic emotions such as interest, agreement or disagreement, and pain.

Facial action units (AUs) are defined in the FACS system developed by Ekman and Friesen [9]. FACS defines 44 different AUs and each AU may correspond to different facial muscle movements that could generate a certain facial action. The benefit of AUs is that thousands of expressions and subtle facial signals (e.g. frown and wink) can be expressed by the combination of relatively few AUs.

Dimensional theory describes emotions using continuous axes in an N-dimensional space, in which each emotion is represented as a point or a region. This theory is based on the assumption that emotion is best described in terms of latent dimensions rather than discrete categories [4]. It was first investigated in the field of psychology, and then developed into several equivalent two and three dimensional (2D and 3D) representations. Examples of such affective dimensions are power, valence, activation or arousal, and expectancy.

Fig. 1 shows the most popular arousal–valence (AV) dimensional space [4] and the distribution of the six basic categorized emotions plus neutral in this space based on psychological studies. Each basic emotion represents a bipolar entity being a part of the same emotional continuum. The arousal axis denotes the level of activation, while the valence axis stands for the degree of pleasantness.

Dimensional spaces [4] have the advantage of being able to represent a wide range of emotions, especially those spontaneous non-prototypical ones in real-life data such as boredom and interest. Studies [10] have demonstrated that in real situations, pure expressions of prototypical emotions are less frequently elicited and blends of emotional displays are often shown by humans. A dimensional space can provide insight into the emotional intensity, as well as the similarity and contrast between different categorized emotions. Although emotion dimensions can be inherently more expressive in comparison to categorized emotions, no explicit mapping between the two descriptions has been established [6]. The correlation between dimensions and categorized emotions is more of an abstract and relatively ambiguous correspondence. From Fig. 1, only a rough location in the AV space can be obtained for each basic categorized emotion, but not the precise values or partitioning of the AV dimensional axes.

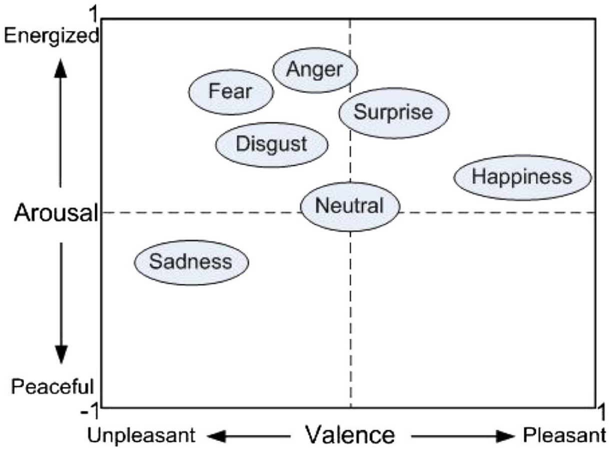


Fig. 1. Arousal–valence (AV) dimensional space. Revised from [4].

2.2. Dimensional facial expression recognition

Recent surveys on dimensional emotion recognition using single or multiple modalities (e.g. audio, gesture, facial expression) can be found in [11–13]. Eyben et al. [14], for example, attempted to predict human emotions 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. In this paper, attention is restricted to studies that use facial expression information.

Most approaches [11,16] to emotion recognition in dimensional spaces quantize the dimensions into a number of intervals, such as the four quadrants [17], or negative and positive emotions [18] using multi-modal information such as facial images combined with shoulder and audio cues. The 2011 Audio/Visual Emotion Challenge [19] focused on a binary classification (i.e. high and low) of each emotion dimension from arousal, expectation, power, and valence. Typical submissions include Wollmer et al. [20] that adopted long short-term memory modeling of optical flow based visual features and a set of audio features, and Ramirez et al. [21] that used latent-dynamic conditional random fields modeling of fused audio and high-level visual features such as smile intensity and gaze direction. Essentially these approaches belong to the discrete emotion classification set. Only a few studies investigated the use of continuous dimensional values. Nicolaou et al. [15] fused facial expressions, shoulder gestures, and audio cues for continuously predicting arousal and valence dimensions on the SAL database. The extracted points and audio features are used individually or in combinations via different levels of fusion, and further input into a neural network (i.e. BLSTM-NN) or support vector regression (SVR). Recent studies in the 2012 and 2013 Audio/Visual Emotion Challenge [2,3] focused on the fusion of visual facial features with audio features for modeling continuous emotion dimensions. For instance, Baltrusaitis [22] compared the use of visual features (geometric shape and LBP-TOP texture) and audio features for modeling five continuous emotion dimensions using Continuous Conditional Random Fields. The fusion of audio and visual cues has also been used for continuous estimation of arousal and valence dimensions [23] and for depression recognition [24].

Some publications have reported on representing facial expressions using continuous dimensional values. Grimm et al. [25] estimated three space attributes, valence, activation and dominance, using Gabor features and a neuro-fuzzy classifier. Yangzhou et al. [26] built a linear mapping to represent expressional face images in the arousal–valence (AV) space. Yeasin et al. [27] mapped facial expressions into levels of interest based on a three-dimensional space and the intensity of optical flow. Nicolaou et al. [28] 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. Other approaches map facial expressions into a dimensional space using manifold learning techniques, such as Lipschitz embedding [29], locality preserving projections [30] and locally linear embedding [31]. The manifold spaces in these approaches, however, are not linked with dimensional values. Fusion of geometric features with different types of texture features was not investigated for dimensional FER in these studies.

As texture and geometry convey complementary and important information about facial expressions [32], it is worthwhile investigating whether a fusion of the two can improve the performance of dimensional emotion recognition. It is desirable to evaluate performance on spontaneous data obtained from human conversations and annotated independently.

No other work has experimentally evaluated and validated the correlation of each emotion dimension to categorized emotions based on machine vision FER systems. Nicolaou et al. [6] is the only study to the best of our knowledge that investigated the correlation between emotion dimensions and categorized emotion using audio and facial features. Each categorized emotion (i.e. anger, happiness, sadness, contempt, or amusement) was predicted using a set of emotion dimensions consisting of valence, arousal, power, expectation and intensity, but the correlation of each dimension to categorized emotion was not addressed.

This paper is aimed at addressing the gaps in knowledge mentioned above. It presents a framework to represent facial expressions in a continuous dimensional arousal–valence (AV) space using a fusion of texture and geometric features. Texture features are extracted from fiducial facial points to achieve high robustness. Performance is compared based on the public NVIE database. It investigates the performance of alternative feature sets in this framework and selects the best for further investigations. Predication results of arousal and valence are then compared with the corresponding ground truths considering the four aspects of spatial distribution, shift, similarity, and correlation. This is done for each of the six basic categorized emotions on the NVIE and FEEDTUM databases.

3. Evaluation framework

Fig. 2 shows the framework of the evaluation system. Each input image belongs to one of six basic emotion categories (happiness in the example shown in Fig. 2). In each image the face is located using the Viola–Jones detector [33] and 68 fiducial facial points are detected using an active shape model (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 correlation-based feature selection (CFS) algorithm. The geometric feature vector is composed of 43 distance features defined based on an ASM and FAPs. Feature vectors are fed into an SVR with a radial basis function (RBF) 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 database. The regressed values of arousal and valence are then compared with the ground truths for each of six basic categorized emotions on the same NVIE database and across the FEEDTUM database.

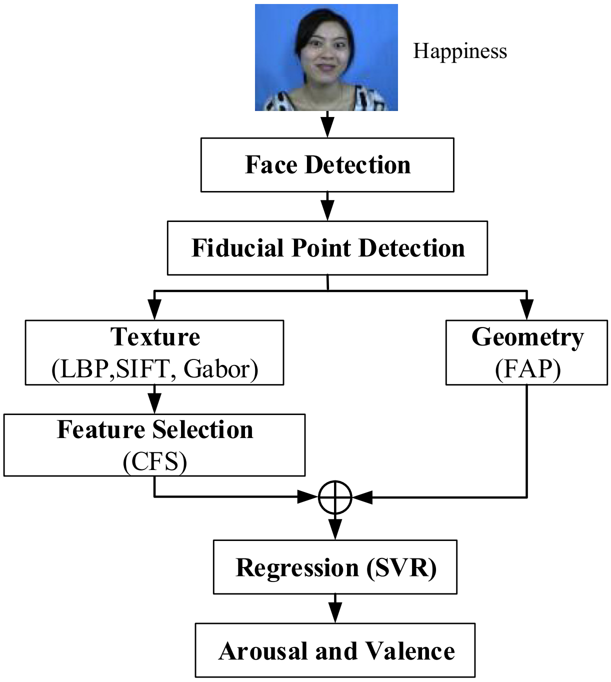


Fig. 2. Framework of the evaluation system.

3.1. Face and fiducial point detection

The face is first detected using the widely used Viola–Jones detector, 68 facial fiducial points are then detected using an ASM [34]. The ASM is known for its robustness in fitting and tracking fiducial points in human faces. To train the ASM, 100 images were collected from the Internet, with different emotions and different poses ranging from −20° to 20° around the Y axis (yaw). 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 subjected to normal facial 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 changes in the shape of the face between subjects and due to 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 and geometric features.

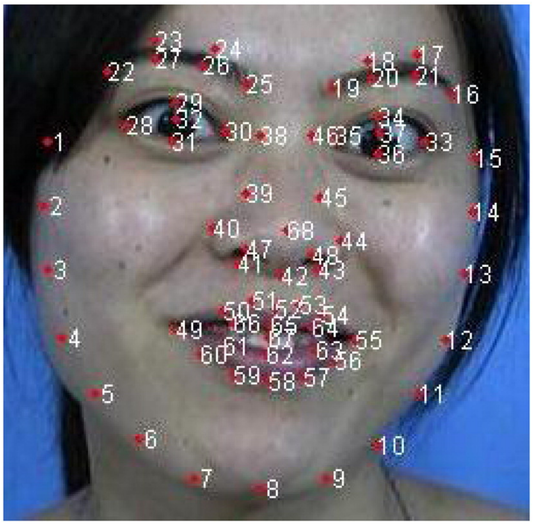


Fig. 3. 68 fiducial points for training an ASM.

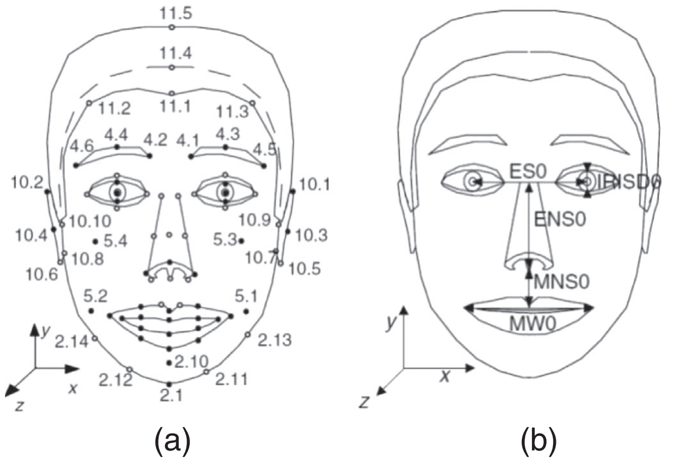


Fig. 4. (a) A subset of feature points defined in the MPEG-4 FAPs standard and (b) FAP units defined based on ratios of distances between the marked key features.

3.2. Texture feature extraction

To achieve a degree of tolerance to facial movements and pose changes, the texture features are extracted from patches around each of the 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 has been successfully used in building robust features in FER [35] algorithms. Three texture features, LBP, Gabor and SIFT, are used and compared here. They are known to have very good performance in FER [7].

1) Local binary patterns (LBP) [36] reduce each pixel in an image to a binary pattern by considering a neighborhood of pixels and applying the center value as a threshold to yield a binary number at each neighbor. A histogram depicting the frequency of occurrence of different binary patterns yields the texture descriptor of the image. LBP has the advantage of tolerance against illumination changes and computational simplicity. In this paper, uniform patterns $LBP^{u2}_{8,2}$ with 59 labels [37] from a 14 × 18 patch centered at each point are collected, resulting in a histogram with 2,597 bins for all points.
2) Gabor features [38] 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 $\lambda_\theta = 4 \times 2^{\sqrt{m-1}}\ m = (1, \ldots, 5)$ and eight orientations $\theta_n = \pi(n-1)/8\ n = (1, \ldots, 8)$ Gabor filters. Therefore, 40 Gabor magnitude coefficients are computed for each of 53 points, leading to a final feature vector with 2,120 elements.
3) Scale-invariant feature transform (SIFT) [39] 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.

3.3. Geometric feature extraction

Geometric facial animation parameters (FAPs) [40] 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.

Table 1
Distances between facial points defined by FAPs.

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

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.

The geometric features used include 43 of the distances between 53 interior points detected by the ASM. As listed in Table 1, 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*). Similarly, there are multiple distances for FAPs Nos. 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 [41]. Compared to the commonly used facial movement vectors obtained in multi-frames, distance features have the merits of being robust to translations and in-plane rotations of the facial geometry, and do not require compensation for facial movements. Therefore, they are suited for working on real-world images in the proposed system. To provide invariance to different faces, all distances are normalized based on FAPUs.

3.4. 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 [42] achieve better performance than using all bins [43]. However, Adaboost cannot be directly used for feature selection in the regression problem here. Instead, the correlation-based feature selection (CFS) is used for this task and CFS has also been successfully applied previously for feature selection in predicting dimensional emotions [14].

CFS [44] 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:

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

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.

3.5. Emotion regression

Support vector machine (SVM) is a supervised learning algorithm that is widely used for analyzing data and recognizing patterns. SVMs can also be applied to regression problems by the introduction of an alternative loss function [45]. Then it is called support vector regression (SVR) and the goal is to optimize the generalization bounds for regression by a loss function, which is used to weight the actual error of the point with respect to the distance from the correct prediction. SVR has shown great potential for use in regression tasks with high dimensional input data, since its optimization does not depend on the dimensionality of the input space. In this paper, a multiple-class SVR with a RBF kernel is adopted for the regression from features to continuous arousal and valence dimensions.

4. Experiments

This section compares regression results. The three types of texture features, the geometric features and their fusion (denoted as "LBP + FAP", for instance) are investigated using the NVIE database. Performance is also compared with previously published results. The arousal and valence values for each of the six basic categorized emotions, obtained using the best combination, the LBP and FAP fusion method, are then compared with the corresponding ground truths considering the aspects of spatial distribution, shift, similarity, and correlation over two databases — NVIE and FEEDTUM.

4.1. Databases

1) The natural visible and infrared facial expression (NVIE) database [46] 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 changes in the size of faces. All the visible apex images are rated by five students, with arousal and valence values ranging from −1 to 1, and also the probabilities of six basic categorized emotions ranging from 0 to 2. This makes it possible to study the correlations of arousal and valence dimensions to the basic categorized emotions. In this paper, the average annotation value of arousal or valence from all raters is used as the final annotation for each image. All images are also classified into one of six basic categorized emotions according to their emotional probabilities. Fig. 5 shows sample images located in the AV space according to the assigned arousal and valence values (faces are manually cropped).

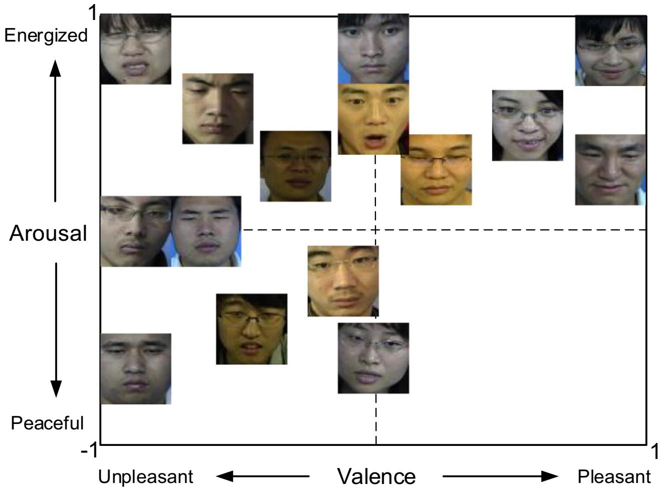


Fig. 5. Distribution of NVIE image samples in the arousal-valence space.

2) The facial expressions and emotions from the Technical University Munich (FEEDTUM) database [47] were collected to investigate the effects of different facial expressions. It contains 399 video sequences from 18 subjects. Each subject performed each of the six basic emotions and the neutral emotion three times, and each sequence starts and finishes with a neutral state. The project attempted to awaken real emotions by playing video clips or still images after a short introduction phase instead of telling the person to play a role. It lets the observed people react as naturally as possible and allows head movements in all directions.
3) The SEMAINE corpus [48] 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 min each, from 20 participants aging from 22 to 60.

For the experiments here, only the front and right illuminated NVIE images that have final annotations are used. Five FEEDTUM images with different emotional intensities are selected from each sequence. For the SEMAINE corpus, only low-quality conversational videos are used, and 54 videos are selected by excluding those with start and end sessions. Inspired by the video down-sampling method for calculating a single feature over a number of video frames in the AVEC 2012 [2], we then select 50 frames arbitrarily from each video. After removing those failed during face and fiducial point detection, a total of 1,027 NVIE and 1,527 FEEDTUM images, as well as 2,474 SEMAINE frames are obtained (the failure rates of Viola–Jones and ASM are 0.2%, 0.4%, and 8.4% respectively on the three databases). Fig. 6 shows samples from the NVIE and FEEDTUM databases for the six basic emotions, and samples of SEMAINE video frames.

4.2. Experimental set-up

10 random subject-independent cross-validations are conducted to evaluate the performance in regressing arousal and valence dimensions. To be specific, all images are first divided into different sets according to the subject identity. Then 10% are randomly selected for the testing set and the other 90% for the training set, the process is repeated 10 times to generate average performance. The performance is evaluated using four parameters: 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. Given N inputs $(\mathrm{x}_i, \mathrm{y}_i)$, $0 < i < N$, where x_i and y_i are the feature vector and the ground truth of the i^{th} input sample respectively. The R^2 statistic can be expressed as:

$$R^2 = 1 - \left[\sum_{i=1}^{N} (y_i - R(x_i))^2 \Big/ \sum_{i=1}^{N} (y_i - \overline{y})^2\right] \tag{2}$$

where $R(x_i)$ is the predicted value for the i^{th} sample, $\overline{y}$ is the mean

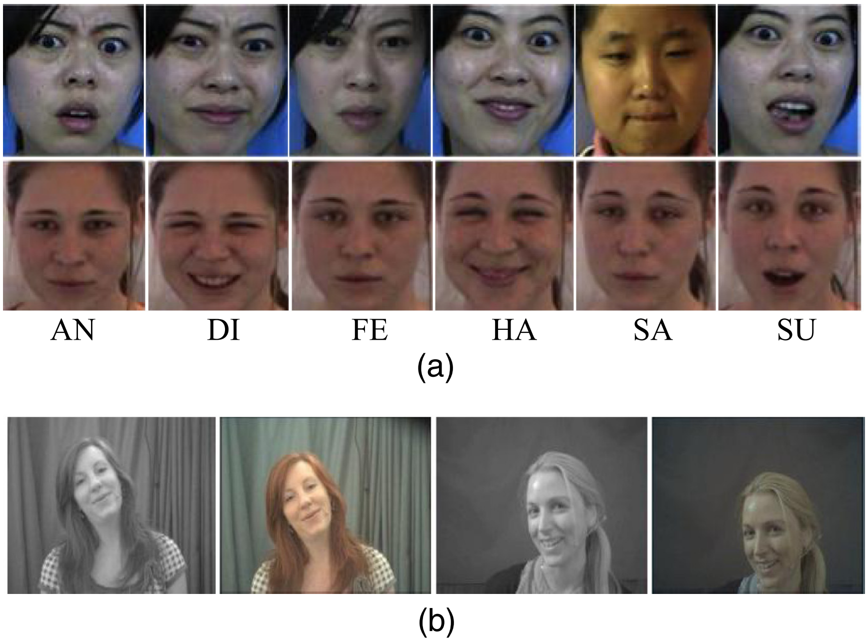


Fig. 6. (a) Samples for six basic emotions. The two rows are for NVIE and FEEDTUM databases, respectively. (b) Samples of SEMAINE video frames.

value of the ground truths. A value of 1 means the prediction results and ground truths are perfectly fitted, while a negative value means the data does not help the prediction.

2) The Pearson correlation coefficient (CC) measures the strength of a linear relationship between two variables. It is defined as the covariance of the two variables (x_i, y_i) divided by the product of their standard deviations:

$$CC = \sum_{i=1}^{N}(x_i - \overline{x_i})(y_i - \overline{y_i}) / \sum_{i=1}^{N}(x_i - \overline{x_i})^2 \sum_{i=1}^{N}(y_i - \overline{y_i})^2. \tag{3}$$

The absolute value of correlation coefficient is less than or equal to 1, where 1 indicates perfect correlation and 0 implies uncorrelated. The larger the coefficient, the stronger is the association between two variables.

3) The mean linear error (MLE) measures the average of the absolute error between the predicted results and the ground truths of the quantity being estimated.

4) The Bhattacharyya distance (BD) measures the similarity of two probability distributions, by taking into account both the mean and covariance of the data. For multivariate Gaussian distributions $p_i = N(m_i, p_i)$, BD is calculated using:

$$D_B = \frac{1}{8}(m_1 - m_2)^T P^{-1}(m_1 - m_2) + \frac{1}{2}\ \ln\left(\frac{DetP}{\sqrt{DetP_1 DetP_2}}\right) \tag{4}$$

where m_i and P_i are the mean and covariance of each distribution, respectively, P is the average of P_1 and P_2 and $DetP$ indicates the determinant of P. A distance close to 0 means that two distributions are similar, and a larger value indicates a bigger difference.

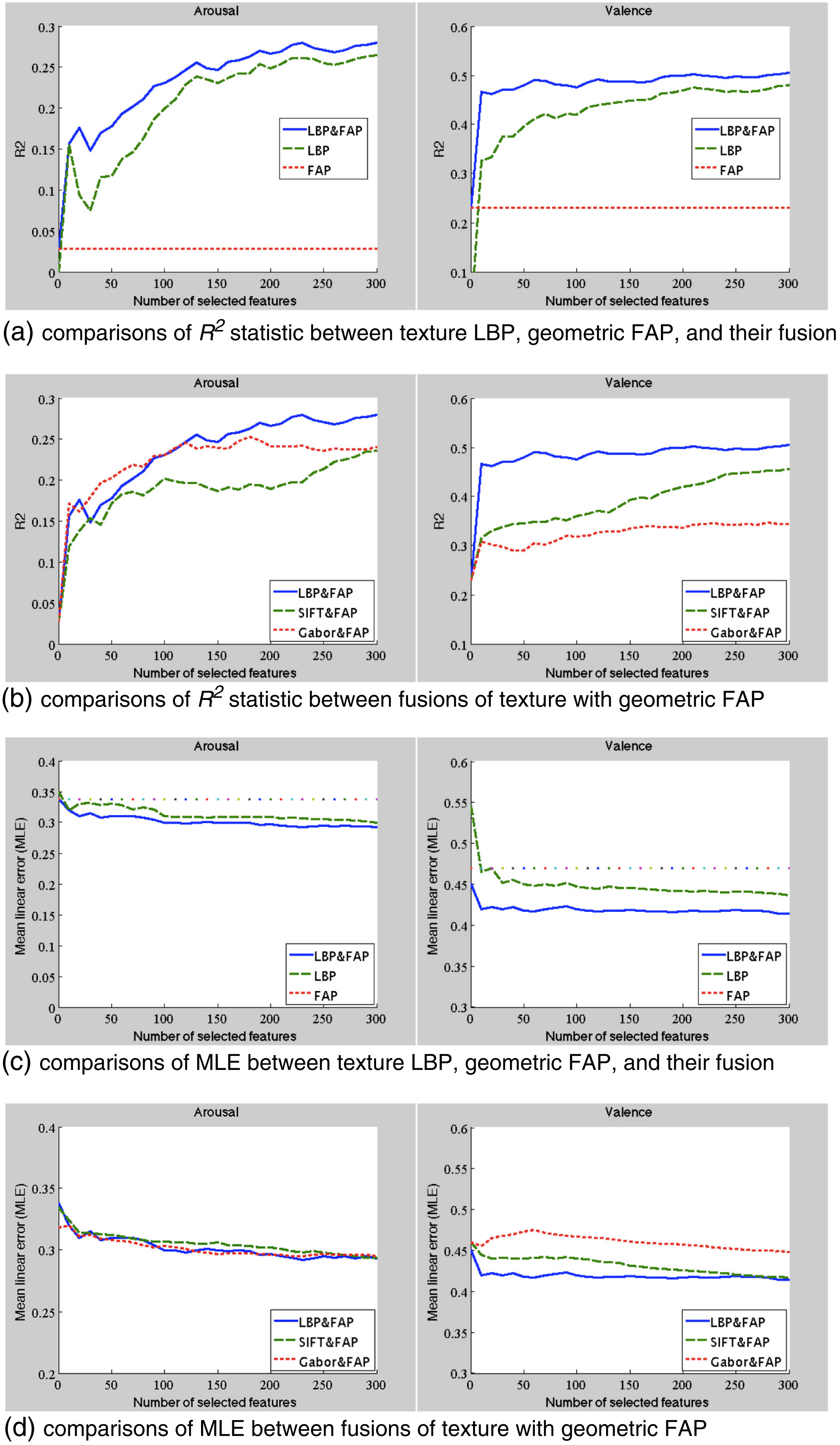
4.3. Regression performance of arousal and valence dimensions using different features[1]

1) Regression performance. Fig. 7 demonstrates the R^2 statistic and the mean linear error (MLE) of the SVR generated arousal and valence values using different lengths of texture features and the 43 FAP features on the NVIE database. 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 (details can be found in [50]). Three types of feature combinations are used: texture feature alone, geometry (FAP) feature alone, and their fusion.

Fusion leads to only small performance improvements over the use of texture alone, but a significantly better performance than using FAP alone. This is observed for all three texture features and agrees with previous findings for the cases of categorized facial expression recognition [51] and posed versus spontaneous facial expression discrimination [55]. Take the R^2 statistic of LBP features as an example. The result obtained using LBP + FAP is 0 to 8.6% higher than using LBP alone and 9.6% to 25.2% higher than using FAP alone when regressing arousal. When regressing valence, LBP + FAP achieves 2.4% to 14.0% higher R^2 statistic values than using LBP alone and 22.1% to 27.5% higher values than using FAP alone respectively. Geometric 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. This is not surprising because the texture features are extracted around geometric feature locations (ASM points) and collectively carry some of this information as relative locations change from one expression to another.

The LBP + FAP combination shows the best overall performance for both arousal and valence. LBP + FAP regression on valence gives the best R^2 and CC, and nearly the best MLE and BD values jointly with SIFT + FAP. The Gabor + FAP combination shows 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, MLE, and CC, while LBP + FAP and SIFT + FAP outperform Gabor + FAP in terms of BD. The highest overall performance using LBP is probably due to its tolerance to illumination variations, shifting of key points from inaccurate ASM detection, and image scale changes [37]. It may be noted that the facial images used here are directly derived from the Viola–Jones face detector without any pro-processing, such as illumination normalization and face alignment. Fig. 8 shows a testing image and its closest match in the train set to the arousal and valence values predicted by the SVR using LBP + FAP features.

[1] A portion of the results have been presented at the DICTA 2011 conference [49].



(a) comparisons of R^2 statistic between texture LBP, geometric FAP, and their fusion

(b) comparisons of R^2 statistic between fusions of texture with geometric FAP

(c) comparisons of MLE between texture LBP, geometric FAP, and their fusion

(d) comparisons of MLE between fusions of texture with geometric FAP

Fig. 7. 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, the fusion of texture LBP and geometric FAP has a much better performance than FAP alone, but there is only a small performance improvement over texture alone (SIFT and Gabor features have similar results). The LBP + FAP combination achieves better overall performance than SIFT + FAP or Gabor + FAP in each case.

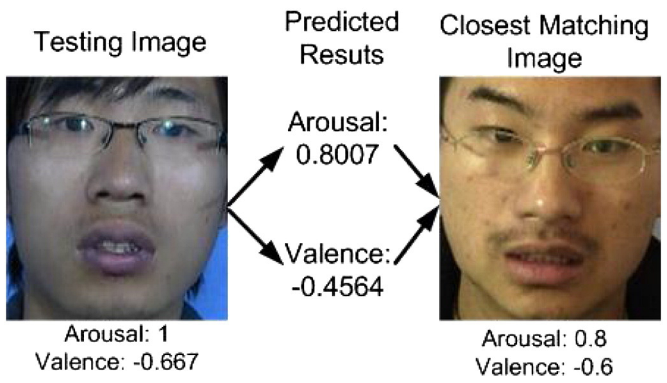


Fig. 8. An example showing the values of arousal and valence predicted by the SVR for a testing image using LBP + FAP and the closest matching train set image for these values. Ground truths of arousal and valence are listed below each image.

There is a performance difference between regression to arousal and regression to 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 (see details in Section 4.4). This may explain the larger MLE values in Table 2 despite having higher correlation (R^2 and CC). Bhattacharya distances are nevertheless lower for valence in general. Better correlation appears to co-occur with higher error. It is worth noting that a result in this evaluation is contrary to the result presented in audio analysis [52] where arousal gets a higher R^2 than valence (0.583 versus 0.281). This result confirms psychological evidence [11] and the result in [28] indicating that visual cues (as opposed to audio) are more indicative of valence than arousal.

Table 2 gives the numerical values of the regression results — R^2, CC, MLE and BD obtained for arousal and valence using the top 100 selected LBP features and 43 FAP features. The best results in each column are boldfaced. The LBP + FAP combination attains the best overall performance for both arousal and valence with R^2 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 in dimensional emotion regression performance.

2) Performance results on SEMAINE video frames. An interesting question is whether 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 arbitrarily selected from the SEMAINE database. Table 3 shows the regression results obtained from this data using 100 LBP 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.

Table 2
Regression results obtained using 100 texture features plus 43 FAP features on the NVIE database. The boldfaced 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
Gabor + FAP	**0.230**	0.497	**0.296**	0.061	0.319	0.570	0.472	0.061
Gabor	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

Table 3
Regression results of video frames on the SEMAINE database.

	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**

3) Performance comparison with previous work. Results obtained using LBP + FAP are compared with those reported in previous work as shown in Table 4. Note that the baseline results on the SEMAINE database in the AVEC 2012 are given in [2], and the results in [25,14] are based on images selected from the VAM Corpus and videos segmented from the SEMAINE database, respectively. In addition, the results in [25] are obtained based on facial expressions, while those in [14] are obtained using audio, video, and event features, individually and in combination (only the best results are reported here).

Table 4 shows that LBP + FAP has comparable performance to previous work, evaluated in terms of CC and MLE. Although it shows a −0.07 CC for arousal, it achieves more than 0.1 higher CC for valence compared to the baseline results in the AVEC 2012 [2] when evaluated on the same SEMAINE database. This may be partially caused by the fact that only 50 frames from each SEMAINE video are used in our paper as opposed to all frames in [2]. LBP + FAP outperforms the work [25] 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 reported in [14,53], LBP + FAP demonstrates better CC, but poorer MLE. The higher MLE using LBP + FAP is, to some extent, due to the fact that the predicted values of arousal and valence are not restricted to a range of [−1, 1], while the previous work compared here sets such a restriction. It also can be seen from Table 4 that audio modality seems to also convey important information for the regression to dimensional representations of emotions, fusion of multiple-modalities helps to

Table 4
Performance comparison with previous work. The boldfaced figures indicate that the proposed method outperforms previous work.

Method	Modality	Dataset	Arousal		Valence	
			CC	MLE	CC	MLE
Our work	Facial expression	NVIE	0.498	0.297	**0.690**	0.415
		SEMAINE	−0.070	0.217	**0.241**	**0.206**
[2]	Facial expression (FCSC test)	SEMAINE	0.077	–	0.134	–
	Facial expression (WLSC test)		0.005	–	0.005	–
	Audio (WLSC test)		0.014	–	0.040	–
[14]	Head motion	SEMAINE	0.204	0.208	0.037	0.258
	Audio + video + event		0.699	0.153	0.165	0.245
[25]	Facial expression	VAM	0.53	0.30	0.45	0.31
[53]	Audio	VAM	0.83	0.15	0.42	0.14
		SAL	0.24	0.28	0.15	0.38

improve the regression, and the dataset also has big impact on the performance.

4.4. Correlation of arousal and valence dimensions to six basic categorized emotions

4.4.1. Spatial distributions of ground truths and prediction results in AV space

Figs. 9 and 10 show the distributions of the ground truths and prediction results for regression of arousal and valence for the six categorized emotions, on the NVIE and FEEDTUM databases, respectively. The results are obtained based on 100 LBP features plus 43 FAPs. Note that multiple ground truths (and prediction results) may be overlapped on the same point in some instances. For all emotions and for both databases, the overall distributions of the predicted results of arousal and valence are similar to those of the ground truths, and both distributions are approximately consistent to those for the six categorized emotions from previous psychological findings as shown in Fig. 1. To be specific, the majority of the predicted results and ground truths for anger, disgust and fear are distributed in the upper-left quadrant, and those for happiness in the upper-right quadrant, while those for sadness and surprise are across the two left quadrants and two upper quadrants, respectively. However, there is also a subset of ground truths that is not predicted accurately by SVR, especially those with negative arousal values (e.g. the results for sadness in Fig. 9). The ellipses of the predicted results are more similar for the six categorized emotions than those of the ground truths, and this is probably because the machine vision system tends to produce similar prediction outputs for all emotion classes even it was trained using inputs from different classes. These ellipses also exhibit small differences in the orientation and size.

Among the six emotions, anger, disgust and happiness appear to achieve more consistent distributions between predicted and ground truth values than the other emotions. The distributions between anger, disgust and fear are highly overlapped; therefore, these emotions tend to have similar predicted values of arousal and valence. As suggested in [54], adding another dimension (i.e. dominance) helps to differentiate these emotions more clearly.

When evaluated on the FEEDTUM database as shown in Fig. 10, the predicted results for the six categorized emotions are distributed primarily in adjacent regions to both the corresponding ground truths and to the predicted results for the NVIE database. This testifies the constancy of the predicted results from SVR across the two databases. However, the results for FEEDTUM are more compactly clustered than both the ground truths and predicted results for NVIE. One possible reason is that facial expressions in FEEDTUM images are not as exaggerated as those in NVIE images. Similar to the results in Fig. 9, arousal has a lower overall shift of mean values than valence.

4.4.2. Shifts between ground truths and prediction results in AV space

Table 5 illustrates shifts of cluster centers between the prediction results and the ground truths of arousal and valence for the six categorized emotions. Among the six emotions, anger, disgust, surprise and happiness appear to have the smallest shifts of cluster centers in the arousal dimension, while it is fear that has the smallest shifts in valence, for both the databases. This is contrary to the results observed in discrete emotion recognition, where fear is normally more difficult for correct recognition than happiness and surprise. On the other hand, the largest shifts in arousal and valence are exhibited respectively for sadness and for happiness and disgust, for both the databases. For all emotions except for sadness and fear, arousal has much lower shifts than valence. This agrees with the results in [14] that arousal has a consistent lower MLE than valence using audio, video and event features, individually and in combinations.

Comparing the shifts of cluster centers between arousal and valence, it can be observed that:

a) For anger, disgust, happiness and surprise, the shift in arousal is smaller than that in valence on both the databases, which means

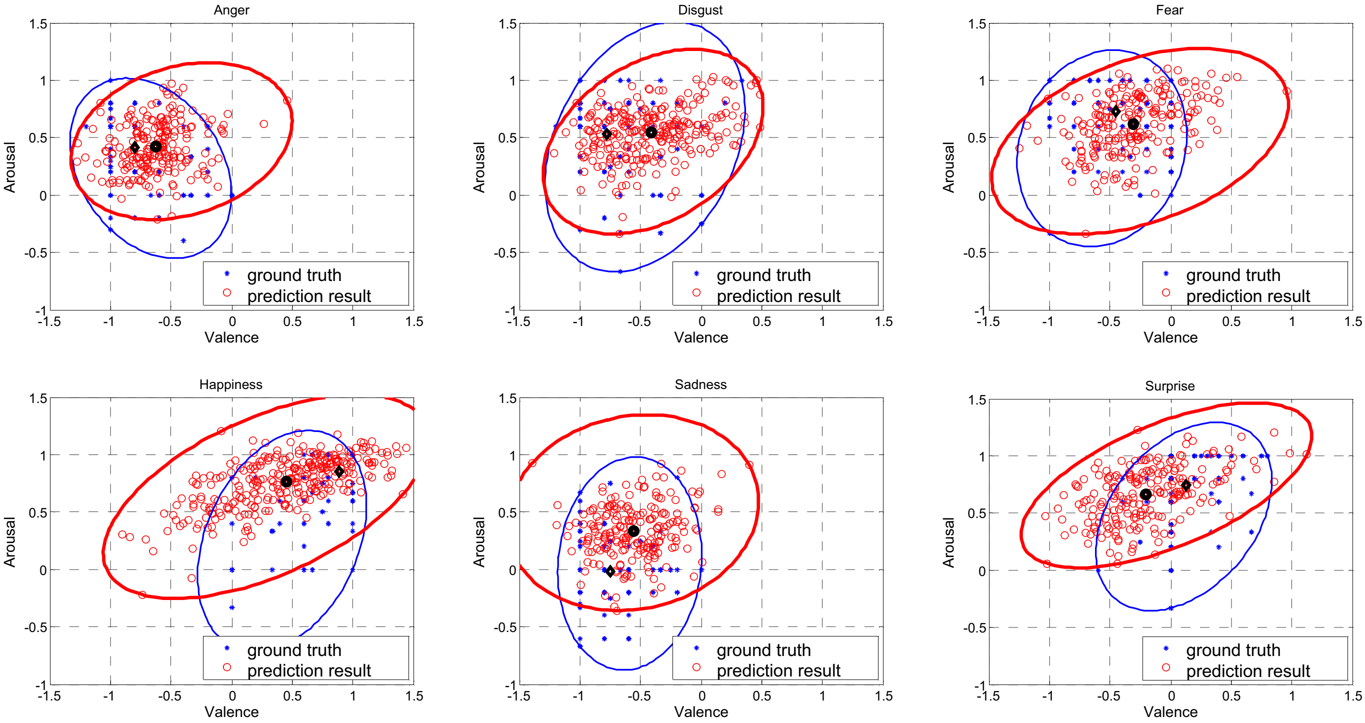


Fig. 9. Distribution of ground truths (blue point) and prediction results (red circle) in the arousal–valence dimensional space for the six basic categorized emotions using 100 LBP and 43 FAP features on the NVIE database. For each emotion, ellipses are drawn to model the periphery (or outside points) of the clusters in the arousal–valence space, and they provide an indication of the maximum distribution region of ground truth or predicted result points in the space. The mean values of these clusters for ground truths and prediction results, respectively, are indicated as black dots. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

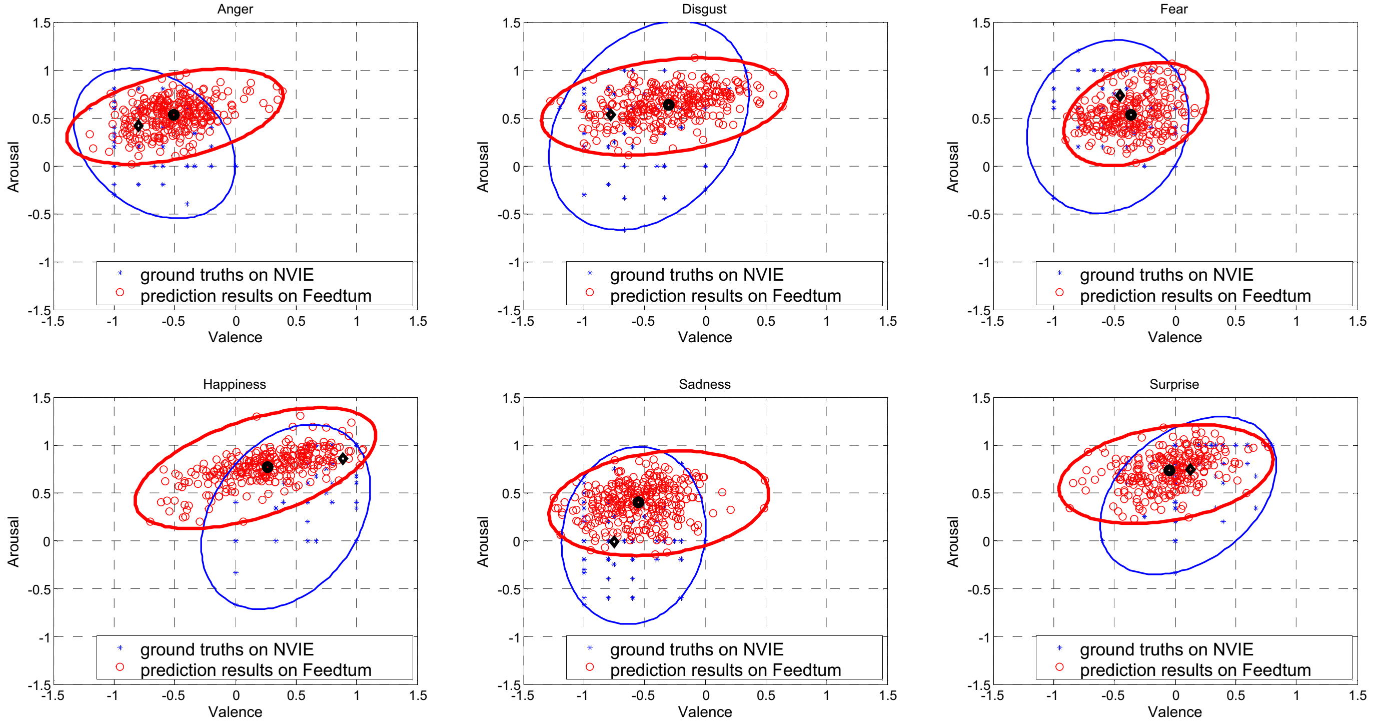


Fig. 10. Distribution of ground truths (blue point) on the NVIE database and prediction results (red circle) on the FEEDTUM database in the AV dimensional space for the six basic categorized emotions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

that the distributions of the regression results are clustered more closely for arousal than those for valence in terms of the cluster center. These emotions are high on an intensity (arousal) scale and are likely to have more scope for variation in the negative–positive valence axis.

b) For sadness, the shift in arousal is larger than that in valence on both the databases, and therefore arousal is more difficult than valence in obtaining higher regression accuracy for sadness. Sadness is negative on the valence scale and has more scope for variation in intensity (arousal).

c) For fear, the shift in arousal is larger than that in valence on the FEEDTUM database, while a contrary result is observed for the NVIE database. This may imply that the regression of fear emotion is prominently dependent on the nature of the data collected for this emotion.

Compared with the results on the NVIE database, there are significant increases in the amount of shifts for the results of arousal and valence on the FEEDTUM database for most of the six emotions. This is probably because, when evaluated on the FEEDTUM database, the ground truths and predicted results come from two different databases, whose images may have big variations in emotional intensity, subject, culture, and occlusion etc. Surprise replaces anger and disgust in producing the least amount of shifts in arousal. This suggests that the data used may have big impact on the amount of the shift of cluster centers for the same emotion.

Fig. 11 gives an impression of feature movements from ground truths to the corresponding predicted results for the two best, two normal and two worst regressed images. The best cases are predicted with nearly no errors, while the errors become significant for the worst cases. The normal cases represent average movements over all images, and have lengths less than 0.5 and different directions. These movements are similar to those obtained in music emotion recognition [52], indicating some consistency in emotion recognition between using facial expressions and using audio features.

4.4.3. Similarity between ground truths and prediction results in AV space (Bhattacharyya distance)

Table 6 shows the confusion matrices for Bhattacharyya distances in arousal and valence between the two clusters for the prediction results and for the ground truths on the NVIE Database, for each of the six categorized emotions. For arousal, anger, disgust, fear, and happiness are easier than sadness and surprise for correct regression as they exhibit the lowest Bhattacharyya distances by themselves, respectively. Sadness and fear contribute the most and the least to the overall regression performance of arousal, respectively, as they have the largest and smallest overall distances among all emotions observed from the 5th and 3rd columns in the results for arousal in Table 6. On the other

Table 5
Shifts of cluster centers in arousal and valence for six basic categorized emotions. The smaller shift between arousal and valence (A, V) is highlighted in bold.

	NVIE database			FEEDTUM database		
	Ground truth (A, V)	Predicted result (A, V)	Shift (A, V)	Ground truth (A, V)	Predicted result (A, V)	Shift (A, V)
AN	(0.416, −0.801)	(0.424, −0.623)	(**0.008**, 0.178)	(0.416, −0.801)	(0.528, −0.510)	(**0.112**, 0.291)
DI	(0.535, −0.781)	(0.543, −0.410)	(**0.008**, 0.371)	(0.535, −0.781)	(0.636, −0.302)	(**0.101**, 0.479)
FE	(0.730, −0.450)	(0.618, −0.306)	(**0.112**, 0.144)	(0.730, −0.450)	(0.534, −0.361)	(0.196, **0.089**)
HA	(0.857, 0.890)	(0.767, 0.452)	(**0.090**, 0.438)	(0.857, 0.890)	(0.770, 0.265)	(**0.087**, 0.625)
SA	(−0.017, −0.749)	(0.335, −0.555)	(0.352, **0.194**)	(−0.017, −0.749)	(0.396, −0.550)	(0.413, **0.199**)
SU	(0.741, 0.129)	(0.662, −0.203)	(**0.079**, 0.332)	(0.741, 0.129)	(0.729, −0.048)	(**0.012**, 0.177)

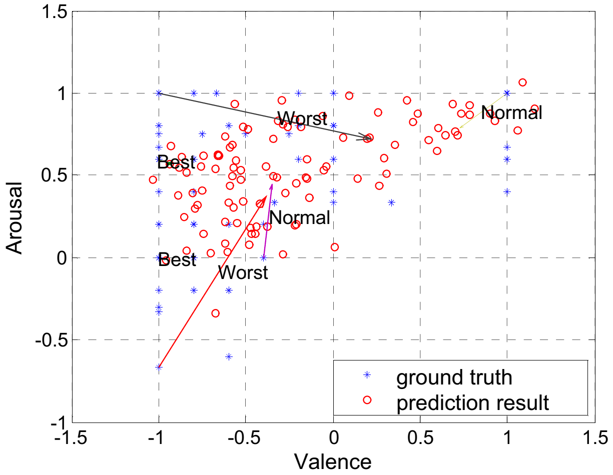


Fig. 11. Feature movements from ground truths to the corresponding prediction results for two best, two normal and two worst regression results.

hand, for valence, fear alone exhibits the lowest distance and it is the easiest one for regression. Happiness and fear contribute the most and the least to the overall regression performance of valence. For most emotions, lower distances are obtained for arousal than valence (i.e. values in diagonal lines), but bigger distance differences between the six emotions are observed for valence than arousal. A bigger difference value normally plays a more important role than the absolute value of the Bhattacharyya distance in distinguishing different emotions, therefore, valence tends to have lower BD, higher R^2 and CC, but lower MLE than arousal for all emotions as shown in Table 2.

Table 7 presents the confusion matrices for Bhattacharyya distances between the two clusters for the prediction results on FEEDTUM and for the ground truths on NVIE. The results are similar to those in Table 6. Sadness and fear are still the biggest and least contributors respectively to the regression performance for arousal, and they are happiness and fear for regression of valence. For most cases for arousal and valence, it is not the emotion in diagonal lines that exhibits the lowest distance among all emotions, but it is still one of the easiest for regression. On the other hand, the distances in diagonal lines are bigger than the corresponding ones in Table 6. This is expected as the predicted results for one database often have a large variation from the ground truths for another database.

4.4.4. Correlation between ground truths and prediction results in AV space (Pearson correlation coefficient — CC)

Table 8 reports the CC values of arousal and valence to the six basic emotions on the NVIE database. Note that the result for the test on the FEEDTUM database is not given here because the ground truths and prediction results are from different databases and not matched in pairs. Overall, arousal exhibits higher correlation with all categorized emotions (except anger) than valence. This is particularly evidenced by the CC value obtained for fear, where valence has a small negative value indicating nearly no correlation, while arousal has a positive value of 0.356 showing strong correlation. However, arousal has only slightly higher correlation than valence for surprise and sadness. Among all emotions, surprise and happiness show the closest correlation with both arousal and valence, whereas anger and sadness have the weakest correlation. Since surprise and happiness are generally expressed more exaggeratedly than anger and sadness, such as an open mouth, it seems that the correlation of arousal or valence dimension to an emotion is also correlated with the level of expression exaggeration of this emotion. This also agrees with the result in [6] which indicated that anger has lower correlation than happiness when predicted in continuous values using 2-D facial point features.

5. Conclusion

This paper evaluates the performance of representing and 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 similar results to those established for categorized facial expression recognition. Fusion of texture and FAP features leads to only small performance improvements over texture alone, but significant improvements over FAP alone, for both arousal and valence. Fusion of LBP and FAP performs the best among all fusion methods. Dimensional emotion regression does not work well for still frames arbitrarily selected from annotated SEMAINE 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.

Correlations between emotion dimensions and categorized emotions are investigated from four aspects, the spatial distribution, shift, similarity (Bhattacharyya distance), and correlation between predicted results and the corresponding ground truths. These are computed for the regression of arousal and valence, using the LBP and FAP fusion method over two databases — NVIE and FEEDTUM. The results demonstrate similar spatial distributions across the two databases for most emotions, and that the distributions are consistent with previous psychological findings, particularly for anger, disgust and happiness. Valence and arousal dimensions behave differently. Higher correlation appears to be accompanied by greater mean error values after regression. For categorized emotions that tend to have close predicted results of arousal and valence, such as anger, disgust and fear, it is advisable to add another dimension (i.e. dominance) to differentiate them more clearly. Smaller shifts of the cluster centers of regressed and ground truth values, are observed for arousal than for valence, for most emotions except for sadness. The amount of the shift for the same emotion may be influenced by the data used. The movements from ground truths to the corresponding predicted results for FER in the AV space were found to be similar to those in music emotion recognition, and visual cues (as opposed to audio) appear to be more indicative of valence than arousal in emotion recognition. The confusion matrices for

Table 6
Confusion matrices for Bhattacharyya distances between clusters for the prediction results and the ground truths on the NVIE database for six basic categorized emotions.The boldfaced values in diagonal lines indicate the similarity between predication results and ground truths for each of six basic emotions, while the values left show the similarity between different emotions. A larger value indicates a bigger difference.

	Arousal						Valence					
	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**0.020**	0.066	0.217	0.420	0.359	0.222	**0.060**	0.037	0.045	4.206	0.047	0.997
DI	0.050	**0.034**	0.075	0.196	0.581	0.087	0.224	**0.149**	0.005	2.237	0.212	0.412
FE	0.093	0.033	**0.017**	0.085	0.696	0.026	0.363	0.263	**0.021**	2.208	0.335	0.308
HA	0.321	0.166	0.035	**0.027**	1.241	0.045	1.519	1.271	0.664	**0.278**	1.502	0.196
SA	0.023	0.092	0.303	0.539	**0.228**	0.299	0.102	0.067	0.023	3.879	**0.081**	0.848
SU	0.129	0.055	0.016	0.072	0.806	**0.027**	0.556	0.442	0.089	1.476	0.531	**0.145**

Table 7
Confusion matrices for Bhattacharyya distances between clusters for the prediction results on FEEDTUM and for the ground truths on NVIE for six basic categorized emotions. The bold-faced values in diagonal lines indicate the similarity between predication results and ground truths for each of six basic emotions.

	Arousal						Valence					
	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**0.095**	0.100	0.159	0.318	0.687	0.179	**0.163**	0.118	0.018	3.954	0.129	0.785
DI	0.177	**0.119**	0.087	0.175	0.949	0.110	0.359	**0.260**	0.025	1.971	0.337	0.279
FE	0.063	0.060	**0.115**	0.257	0.614	0.130	0.405	0.302	**0.037**	3.484	0.354	0.508
HA	0.333	0.187	0.058	**0.064**	1.294	0.076	1.470	1.183	0.521	**0.561**	1.449	0.063
SA	0.345	0.094	0.267	0.494	**0.367**	0.272	0.115	0.077	0.019	3.833	**0.092**	0.818
SU	0.250	0.136	0.041	0.070	1.117	**0.058**	0.996	0.759	0.217	1.630	0.960	**0.061**

Table 8
Correlations (CC) of arousal and valence dimensions to six basic categorized emotions on the NVIE database. The highest values are shown in bold.

	AN	DI	FE	HA	SA	SU
Arousal	0.024	0.336	0.356	0.412	0.146	**0.465**
Valence	0.157	0.213	−0.009	0.353	0.121	**0.441**

Bhattacharyya distances indicate that fear is the least contributor to regressed values of both arousal and valence, while sadness and happiness contribute most to the regression for arousal and valence, respectively. The arousal dimension exhibits higher correlation than valence with all categorized emotions except anger. These results imply that arousal and valence may need to be given different priorities when pursuing the highest regression accuracy for a specific categorized emotion. Among the six categorized emotions, surprise and happiness are the strongest correlated with both arousal and valence, while anger and sadness have the weakest correlation. It seems that the correlation of arousal or valence dimension to an emotion is also correlated with the level of expression exaggeration of this emotion.

Future work will extend the evaluated framework to video by considering the temporal correlations between arousal and valence dimensions simultaneously using classifiers such as the long short-term memory neural network, and incorporating audio features as well.

Acknowledgments

This work is funded by the National Natural Science Foundation of China (Grant No. 61402362) and the Xi'an University of Technology (Grant No. 116-211406).

References

[1] S. Kai, Y. Junqing, H. Yue, H. Xiaoqiang, An improved valence–arousal emotion space for video affective content representation and recognition, Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on, 2009, pp. 566–569.

[2] B. Schuller, M. Valstar, R. Cowie, M. Pantic, AVEC 2012: the continuous audio/visual emotion challenge — an introduction, Proceedings of the 14th ACM international Conference on Multimodal Interaction, ACM, Santa Monica, California, USA, 2012, pp. 361–362.

[3] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, M. Pantic, AVEC 2013: the continuous audio/visual emotion and depression recognition challenge, Proceedings of the 3rd ACM International Workshop on Audio/visual emotion challenge, ACM, Barcelona, Spain, 2013, pp. 3–10.

[4] J.A. Russell, A circumplex model of affect, J. Pers. Soc. Psychol. 39 (1980) 1161–1178.

[5] R.J. Harris, A.W. Young, T.J. Andrews, Morphing between expressions dissociates continuous from categorical representations of facial expression in the human brain, Proc. Natl. Acad. Sci. U.S.A. 109 (51) (2012) 21164–21169.

[6] M.A. Nicolaou, S. Zafeiriou, M. Pantic, Correlated-spaces regression for learning continuous emotion dimensions, Proceedings of the 21st ACM International Conference on Multimedia, ACM, Barcelona, Spain, 2013, pp. 773–776.

[7] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 39–58.

[8] P. Ekman, Strong evidence for universals in facial expressions — a reply to Russells mistaken critique, Psychol. Bull. 115 (1994) 268–287.

[9] P. Ekman, W. Friesen, The Facial Action Coding System: A technique for the measurement of facial movement, Consulting Psychologists Press, Palo Alto, CA, USA, 1978. 274–280.

[10] P. Maja, S. Nicu, F.C. Jeffrey, H. Thomas, Affective multimodal human–computer interaction, Proceedings of the 13th Annual ACM International Conference on Multimedia, ACM, Hilton, Singapore, 2005, pp. 669–676.

[11] H. Gunes, M. Pantic, Automatic, dimensional and continuous emotion recognition, Int. J. Synth. Emot. 1 (2009) 68–99.

[12] H. Gunes, B. Schuller, Categorical and dimensional affect analysis in continuous input: current trends and future directions, Image Vis. Comput. 31 (2013) 120–136.

[13] H. Gunes, M. Nicolaou, M. Pantic, Continuous analysis of affect from voice and face, in: A.A. Salah, T. Gevers (Eds.), Computer Analysis of Human Behavior, Springer, London, 2011, pp. 255–291.

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

[15] M.A. Nicolaou, H. Gunes, M. Pantic, Continuous prediction of spontaneous affect from multiple cues and modalities in valence–arousal space, IEEE Trans. Affect. Comput. 2 (2011) 92–105.

[16] H. Gunes, B. Schuller, M. Pantic, R. Cowie, Emotion representation, analysis and synthesis in continuous space: a survey, Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, 2011, pp. 827–834.

[17] G. Caridakis, K. Karpouzis, M. Wallace, L. Kessous, N. Amir, Multimodal user's affective state analysis in naturalistic interaction, J. Multimodal User Interfaces 3 (2010) 49–66.

[18] M.A. Nicolaou, H. Gunes, M. Pantic, Audio-visual classification and fusion of spontaneous affective data in likelihood space, Pattern Recognition (ICPR), 20th International Conference on, 2010, pp. 3695–3699.

[19] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, M. Pantic, AVEC 2011—the first international audio/visual emotion challenge, Proceedings of the 4th International Conference on Affective computing and intelligent interaction — Volume Part II, Springer-Verlag, Memphis, TN, 2011, pp. 415–424.

[20] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, G. Rigoll, LSTM-modeling of continuous emotions in an audiovisual affect recognition framework, Image Vis. Comput. 31 (2013) 153–163.

[21] G. Ramirez, T. Baltrušaitis, L.-P. Morency, Modeling latent discriminative dynamic of multi-dimensional affective signals, in: S. D'Mello, A. Graesser, B. Schuller, J.-C. Martin (Eds.), Affective Computing and Intelligent Interaction, Springer, Berlin/Heidelberg, 2011, pp. 396–406.

[22] T. Baltrusaitis, N. Banda, P. Robinson, Dimensional affect recognition using Continuous Conditional Random Fields, Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, 2013, pp. 1–8.

[23] E. Sanchez-Lozano, P. Lopez-Otero, L. Docio-Fernandez, E. Argones-Rua, J.L. Alba-Castro, Audiovisual three-level fusion for continuous estimation of Russell's emotion circumplex, Proceedings of the 3rd ACM International Workshop on Audio/visual emotion challenge, ACM, Barcelona, Spain, 2013, pp. 31–40.

[24] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, J. Epps, Diagnosis of depression by behavioural signals: a multimodal approach, Proceedings of the 3rd ACM International Workshop on Audio/visual emotion challenge, ACM, Barcelona, Spain, 2013, pp. 11–20.

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

[26] D. Yangzhou, B. Wenyuan, W. Tao, Z. Yimin, A. Haizhou, Distributing expressional faces in 2-D emotional space, 6th ACM International Conference on Image and Video Retrieval, Amsterdam, 2007, pp. 395–400.

[27] M. Yeasin, B. Bullot, R. Sharma, Recognition of facial expressions and measurement of levels of interest from video, IEEE Trans. Multimedia 8 (2006) 500–508.

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

[29] Y. Chang, C. Hu, R. Feris, M. Turk, Manifold based analysis of facial expression, Image Vis. Comput. 24 (2006) 605–614.

[30] N. Sebe, M.S. Lew, T.S. Huang, C. Shan, S. Gong, P.W. McOwan, Appearance manifold of facial expression, Computer Vision in Human–Computer Interaction, Springer, 2005, pp. 221–230.

[31] Y. Shi, G. van Albada, J. Dongarra, P. Sloot, Y.-s. Shin, Facial expression recognition based on emotion dimensions on manifold learning, Computational Science — ICCS 2007, Springer, Berlin/Heidelberg, 2007, pp. 81–88.

[32] S. Mingli, T. Dacheng, L. Zicheng, L. Xuelong, Z. Mengchu, Image ratio features for facial expression recognition application, IEEE Trans. Syst. Man Cybern. Part B Cybern. 40 (2010) 779–788.

[33] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2004) 137–154.

[34] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models—their training and application, Comput. Vis. Image Underst. 61 (1995) 38–59.

[35] S. Berretti, A.D. Bimbo, P. Pala, B.B. Amor, M. Daoudi, A Set of selected SIFT features for 3D facial expression recognition, Pattern Recognition (ICPR), 2010 20th International Conference on, 2010, pp. 4125–4128.

[36] T. Ojala, M. Pietikainen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern Recogn. 29 (1996) 51–59.

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

[38] J.G. Daugman, Two-dimensional spectral analysis of cortical receptive field profiles, Vis. Res. 20 (1980) 847–856.

[39] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.

[40] I.S. Pandzic, R. Forchheimer, MPEG-4 Facial Animation: The Standard, Implementation and Applications, Wiley, 2002.

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

[42] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1997) 119–139.

[43] C. Shan, T. Gritti, Learning discriminative LBP-histogram bins for facial expression recognition, Proc. British Machine Vision Conference, 2008.

[44] M.A. Hall, Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning, The University of Waikato, Department of Computer Science, 1999 (Doctoral Dissertation).

[45] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, Adv. Neural Info. Proc. Syst. (1997) 155–161.

[46] W. Shangfei, L. Zhilei, L. Siliang, L. Yanpeng, W. Guobing, P. Peng, C. Fei, W. Xufa, A natural visible and infrared facial expression database for expression recognition and emotion inference, IEEE Trans. Multimedia 12 (2010) 682–691.

[47] F. Wallhoff, Facial Expressions and Emotion Database, http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html Technische Universität München, 2006.

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

[49] L. Zhang, D. Tjondronegoro, V. Chandran, Evaluation of texture and geometry for dimensional facial expression recognition, Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on, 2011, pp. 620–626.

[50] L. Zhang, Towards Spontaneous Facial Expression Recognition in Real-world Video, Queensland University of Technology, 2012. (PhD Thesis).

[51] L. Zhang, D. Tjondronegoro, V. Chandran, Facial expression recognition experiments with data from television broadcasts and the World Wide Web, Image Vis. Comput. 32 (2014) 107–119.

[52] Y. Yi-Hsuan, L. Yu-Ching, S. Ya-Fan, H.H. Chen, A regression approach to music emotion recognition, IEEE Trans. Audio Speech Lang. Process. 16 (2008) 448–457.

[53] F. Eyben, M. Wollmer, B. Schuller, OpenEAR — introducing the munich open-source emotion and affect recognition toolkit, Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, 2009, pp. 1–6.

[54] C.E. Osgood, George J. Suci, Percy H. Tannenbaum, The Measurement of Meaning, University of Illinois Press, Urbana, 1957.

[55] L. Zhang, D. Tjondronegoro, V. Chandran, Geometry vs. Appearance for Discriminating between Posed and Spontaneous Emotions, Neural Information Processing, Springer, Berlin / Heidelberg, 2011, pp. 431–440.