Facial Expression Recognition Using Facial Movement Features

Facial Expression Recognition Using Facial Movement Features

Ligang Zhang and Dian Tjondronegoro

Abstract— Facial expression is an important channel for human communication and can be applied in many real applications. One critical step for facial expression recognition (FER) is to accurately extract emotional features. Current approaches on FER in static images have not fully considered and utilized the features of facial element and muscle movements, which represent static and dynamic, as well as geometric and appearance characteristics of facial expressions. This paper proposes an approach to solve this limitation using 'salient' distance features, which are obtained by extracting patch-based 3D Gabor features, selecting the 'salient' patches, and performing patch matching operations. The experimental results demonstrate high correct recognition rate (CRR), significant performance improvements due to the consideration of facial element and muscle movements, promising results under face registration errors, and fast processing time. The comparison with the state-of-the-art performance confirms that the proposed approach achieves the highest CRR on the JAFFE database and is among the top performers on the Cohn-Kanade (CK) database.

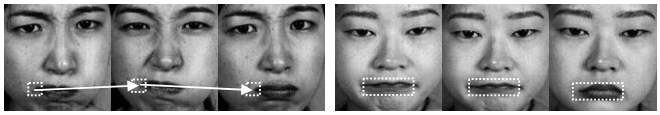
Index Terms— facial expression analysis, feature evaluation and selection, computer vision, Gabor filter, Adaboost.

1 Introduction

Facial expression recognition (FER) has been dramatically developed in recent years, thanks to the advancements in related fields, especially machine learning, image processing and human cognition. Accordingly, the impact and potential usage of automatic FER have been growing in a wide range of applications, including human-computer interaction, robot control and driver state surveillance. However, to date, robust recognition of facial expressions from images and videos is still a challenging task due to the difficulty in accurately extracting the useful emotional features. These features are often represented in different forms, such as static, dynamic, point-based geometric or region-based appearance.

Facial movement features, which include feature position and shape changes, are generally caused by the movements of facial elements and muscles during the course of emotional expression. The facial elements, especially key elements, will constantly change their positions when subjects are expressing emotions. As a consequence, the same feature in different images usually has different positions, as shown in Fig.1 (a). In some cases, the shape of the feature may also be distorted due to the subtle facial muscle movements. For example, the mouth in the first two images in Fig. 1 (b) presents different shapes from that in the third image. Therefore, for any feature representing a certain emotion, the geometric-based position and appearance-based shape normally changes from one image to another image in image databases, as well as in videos. This kind of movement features represents a rich pool of both static and dynamic characteristics of expressions, which play a critical role for FER.

The vast majority of the past work on FER does not take the dynamics of facial expressions into account [1]. Some efforts have been made on capturing and utilizing facial movement features, and almost all of them are video-based. These efforts try to adopt either geometric features of the tracked facial points (e.g. shape vectors [2], facial animation parameters [3], distance and angular [4], and trajectories [5]), or appearance difference between holistic facial regions in consequent frames (e.g. optical flow [6], and differential-AAM [7]), or texture and motion changes in local facial regions (e.g. surface deformation [8], motion units [9], spatiotemporal descriptors [10], animation units [11], and pixel difference [12]). Although achieved promising results, these approaches often require accurate location and tracking of facial points, which remains problematic [13]. In addition, it is still an open question how to learn the grammars in defining dynamic features, and handle ambiguities in the input data [14]. On the other hand, image-based FER techniques provide an alternative way to recognize emotions based on appearance-based features in a single image, and are important for the situation where only several images are



(a) (b)

Fig. 1. Facial movement features. (a) Feature position (left mouth corner) changes. (b) Feature shape (mouth) changes. Facial regions are manually cropped from two subjects "KA" and "KL" on the JAFFE database.

• *The authors are with the Faculty of Science and Technology, Queensland University of Technology, 2 George St, Brisbane, QLD 4000, Australia. E-mail: Ligzhang@gmail.com, dian@qut.edu.au.*

Manuscript received (September, 2, 2010).

available for training and testing. However, to the best of our knowledge, no research has been reported on image-based FER that considers facial movement features.

In this paper, we aim for improving the performance of FER by automatically capturing facial movement features in static images based on distance features. The distances are obtained by extracting 'salient' patch-based Gabor features and then performing patch matching operations. Patch-based Gabor features have shown excellent performance in overcoming position, scale, and orientation changes [15], [16], [17], as well as extracting spatial, frequency and orientation information [18]. They also show a great advantage over the commonly used fiducial point-based Gabor [19], [20], [21], [22], [23], graph-based Gabor [24] and discrete Fourier transform [25] features in capturing regional information. Although other appearance-based features, such as local binary patterns (LBP) [26], [27], [10], Haar [28] and histograms of oriented gradients (HOG) [29], have shown a good performance in FER, they lack the capacity of capturing facial movement features with high accuracy. This is due to the fact that these appearance-based features are based on statistic values (e.g. histogram similarity) extracted from sub-regions; therefore, they produce similar results even when facial features move a bit from the original position. On the other hand, Gabor features have the capacity of accurately capturing movement information, and have been proven as being robust even in the case of face misalignment [30].

The idea of patch matching operations has been used to build features for object recognition [15], [16] and action classification [17], which remain robust when there are changes in position, scale, and orientation. To fit for the purpose of FER, we define the matching area and matching scale to restrict the operations within a suitable space. By matching patch-based Gabor features in this space, multi-distance values are obtained. The minimum distance is chosen as the final feature for emotion classification. In this way, one patch, which varies in its position, scale and shape, still can be captured provided that it is located within the defined matching space. To show the effectiveness of using the proposed distance features, we demonstrate the high performance on two widely used databases, significant improvements due to the consideration of facial movement features, and promising results under face registration errors.

The remainder of the paper is organized as follows. Section 2 describes the proposed framework, while the details of building distance features and 'salient' feature selection are explained in Section 3 and 4 respectively. Section 5 represents the recognition and speed performance, and demonstrates comparison with the state-of-the-art performance. Finally, conclusions are drawn in Section 6.

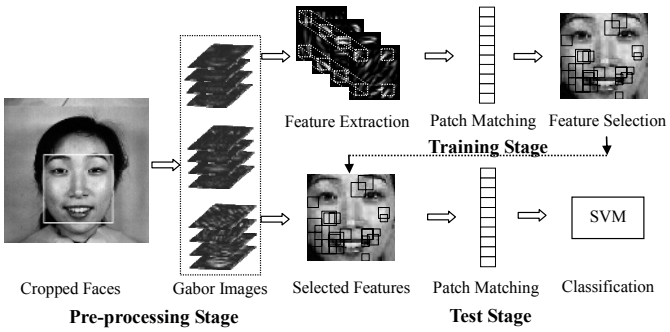


Fig. 2. Proposed framework.

2 Proposed Framework

Fig. 2 illustrates the proposed framework, which is composed of pre-processing, training and test stages. At the pre-processing stage, by taking the nose as the centre and keeping main facial components inclusive, facial regions are manually cropped from database images and scaled to a resolution of 48*48 pixels. No more processing is conducted to imitate the results of real face detectors. Then multi-resolution Gabor images are attained by convolving eight-scale, four-orientation Gabor filters with the scaled facial regions. During the training stage, a whole set of patches is extracted by moving a series of patches with different sizes across the training Gabor images. Then patch matching operation is proposed to convert the extracted patches to distance features. To capture facial movement features, the matching area and matching scale are defined to increase the matching space, whereas the minimum rule is used to find the best matching feature in this space. Based on the converted distance features, a set of 'salient' patches is selected by Adaboost. At the test stage, the same patch matching operation is performed on a new image using the 'salient' patches. The resulting distance features are fed into a multi-class support vector machine (SVM) to recognize six basic emotions, including anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA) and surprise (SU).

The rest of this section gives an introduction of Gabor filters and SVM. The details of building distance features and feature selection are explained in Section 4 and 5 respectively.

In this paper, 2D Gabor filter [31] is adopted and it can be mathematically expressed as:

$$F(x,y) = \exp\left(-\frac{(X^2+\gamma^2Y^2)}{2\sigma^2}\right)\times\cos\left(\frac{2\pi}{\lambda}X\right)$$
$$X = x\cos\theta + y\sin\theta \qquad Y = -x\sin\theta + y\cos\theta \tag{1}$$

Where, θ the orientation, σ the effective width, λ the wavelength, and $\gamma = 0.3$ the aspect ratio. Instead of the widely used five scales, eight scales (5:2:19 pixels) are adopted here to test the results using a larger number of scales. The values of the rest of the parameters are set based on [15] due to the high reported performance. As a result, four orientations (-45°, 90°, 45°, 0°) are used.

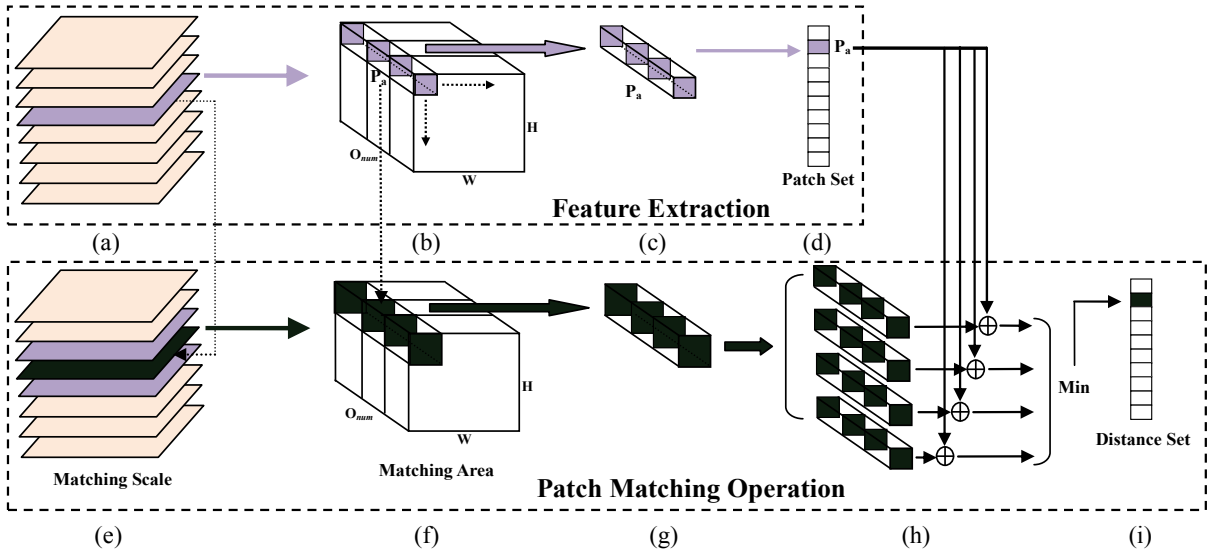


Fig. 3. Building distance features. (a) One scale is selected from eight-scale Gabor images. (b) Patches are extracted across all rows and columns in the selected scale image. (c) One extracted patch P_a. (d) Extracted patch set. (e) Defined matching scale. (f) Defined matching area. (g) One matching area. (h) Distance calculation. (i) Distance feature set.

SVM [32] is one of the most widely used machine learning algorithms for classification problems. This paper directly uses the LIBSVM [33] implementation of SVMs with four different kernels, including linear, polynomial, radial basis function (RBF), and sigmoid. The six-emotion-class problem is solved by the one-versus-rest strategy.

3 Building Distance Features

Fig.3 depicts the two main processes to build distance features: feature extraction and patch matching operation. The feature extraction aims to collect a set of discriminating 3D patches for all emotions, whereas the patch matching operation converts these patches to distance features, which can capture facial movement features.

3.1 Patch-based Feature Extraction

The upper part of Fig. 3 and Fig. 4 show the feature extraction algorithm, which is comprised of four steps: first, all training images are classified into 10 sets. For each emotion, each Gabor scale and each patch size, one Gabor image is randomly selected from all the images of emotion E_k (Fig. 3a). Second, given one patch P_a with a size of P_j*P_j*O_{num}, move this patch across the row and column pixels of this Gabor image (Fig. 3b), a set of 3D patches can be extracted (Fig.4 line-2), one example is shown in Fig. 3c). Third, the matching area and matching scale are recorded (Fig.4 line-3 and -4), details explained in Section 3.2). Finally, a patch set is constituted by combining the extracted patches of all emotions, all scales and all patch sizes (Fig. 3d).

To reduce the feature dimension and increase the processing speed, we only extract a part of all patches by moving the patch P_a with a step. As indicated by the line-

Input: Image set S_i (i=1,…,10); patch size P_j (j=1,…,4); emotion index E_k (k=1,…,6); scale SC_m (m=1,…,8); orientation O_n (n=1,…,4); orientation number O_{num}; image width W, height H (W=H=48).

Output: Extracted patches, matching area and matching scale.

For each emotion E_k, each scale SC_m, each patch size P_j
  Choose one set S_i randomly from 10 sets;
  Choose one image of emotion E_k randomly from S_i;
  $Move\_step = P_j/2$; (1)
  For ih = 1 to (H-P_j+1) by $Move\_step$
    For iw = 1 to (W-P_j+ 1) by $Move\_step$
      Extract *patches* with sizes of P_j*P_j*O_{num}; (2)
      Record the *matching area* (L_x,L_y,R_x,R_y): (3)
      L_x = Max(ih-0.5*P_j,1); R_x = Min(ih+1.5*P_j,H);
      L_y = Max(iw-0.5*P_j,1); R_y = Min(iw+1.5*P_j,W).
      Record the *matching scale* SC_m; (4)
    End
  End
End
Return *patches, matching area* and *matching scale*.

Fig. 4. Pseudo code of building distance features. (1) Defining moving steps. (2) Extracting Patches. (3) Recording matching area. (4) Recording matching scale.

1 in Fig. 4, the moving steps are set to 1, 2, 3 and 4 corresponding to four patch sizes of 2*2*4, 4*4*4, 6*6*4 and 8*8*4. Given 48*48 facial images, eight-scale and four-orientation Gabor filters, the final set contains 148,032 patches.

As indicated and investigated by Zhao and Pietikainen [10], current patch-based approaches only concentrate on the location information of the selected patches, whereby,

one location is shared by all emotions and only the location information is preserved after feature selection. Thus, the useful multi-resolution information contained in these patches is discarded. On the contrary, our approach reserves both the location and multi-resolution information of patches for recognition, resulting in an equal set of patches for each emotion.

3.2 Patch Matching Operation

As shown in the lower part in Fig. 3, the patch matching operation comprises of four steps for each patch and each training image: first, the matching area and matching scale are defined to provide a bigger matching space (Fig. 3e and 3f). Second, the distances are obtained by matching this patch with all patches within its matching space in a training image (Fig. 3h). This step takes two patches as inputs and yields one distance value based on a distance metric. Third, the minimum distance is chosen as the distance feature of this patch in the training image (Fig. 3h). Finally, the distance features of all patches are combined into a final set with 148,032 elements (Fig. 3i).

3.2.1 Matching Area and Matching Scale Definition

The matching area and matching scale are used to accurately capture the position and scale changes caused by facial feature movements. The idea of them stems from the observation that position and scale of one feature do not move or change a lot in different facial images once these images are roughly located by a face detector. Thus, the invariance to position and scale changes can be accomplished by defining a bigger area and a larger scale for each patch when performing patch matching.

Fig. 5 illustrates the matching area *Area* of a patch P_a with a size of P_j*P_j*O_{num}, while Fig. 3e shows the matching scales that are drawn in a gray color. In this paper, the *Area* is set to two times of P_a in width and height, but with the same orientation number O_{num} and centre point. That is *Area* = (2*P_j)*(2*P_j)*O_{num}. The matching scale is the same with that of P_a because the cropped facial regions generally belong to the same scale. However, it is flexible to increase the matching scale in the case of large scale variations.

3.2.2 Distance Metric Definition

The distance metric is used to compute the similarity between two patches. Several metrics have been adopted in previous work, such as Gaussian-like Euclidean [34] and normalized dot-product [17]. In this paper, four metrics, including dense L_1 (DL_1), dense L_2 (DL_2), sparse L_1 (SL_1) and sparse L_2 (SL_2), are used due to the computational simplicity, and they can be mathematically expressed as:

$$DL_1 : \|P_b - P_c\| = \frac{1}{P_j * P_j * O_{num}} \sum_{i=1}^{P_j} \sum_{j=1}^{P_j} \sum_{o=1}^{O_{num}} \left| P_b^{ijo} - P_c^{ijo} \right| \quad (2)$$

$$DL_2 : \|P_b - P_c\| = \frac{1}{P_j * P_j * O_{num}} \sqrt{\sum_{i=1}^{P_j} \sum_{j=1}^{P_j} \sum_{o=1}^{O_{num}} (P_b^{ijo} - P_c^{ijo})^2} \quad (3)$$

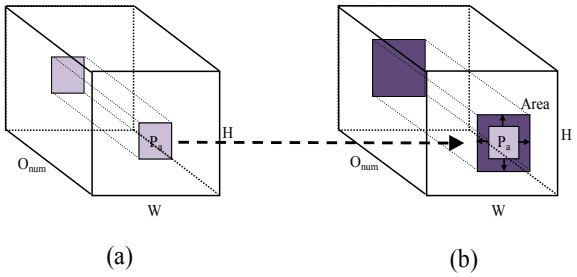


Fig. 5. Matching area. (a) One patch P_a. (b) The corresponding matching area '*Area*' in a new image.

$$SL_1 : \|P_b - P_c\| = \frac{1}{P_j * P_j} \sum_{i=1}^{P_j} \sum_{j=1}^{P_j} (\max \sum_{o=1}^{O_{num}} P_b^{ijo} - \max \sum_{o=1}^{O_{num}} P_c^{ijo}) \quad (4)$$

$$SL_2 : \|P_b - P_c\| = \frac{1}{P_j * P_j} \sqrt{\sum_{i=1}^{P_j} \sum_{j=1}^{P_j} (\max \sum_{o=1}^{O_{num}} P_b^{ijo} - \max \sum_{o=1}^{O_{num}} P_c^{ijo})^2} \quad (5)$$

where, P_b and P_c represent two patches, P^{ijo} represents the pixel values in the i^{th} row, j^{th} column and o^{th} orientation of the patches. The distances are normalized by dividing the number of pixels in the patches. The dense distances can be conceived as taking into account all orientations of each pixel in one patch, whereas the sparse distances only consider the *dominant orientation* [16] of each pixel in one patch. Therefore, they differ in representing all features or only dominant features of all orientations.

4 Patch-based Feature Selection and Analysis

In this section, we use Adaboost for discriminative (called 'salient' here) patch selection on the JAFFE and CK databases. To give a deeper understanding of the selected 'salient' patches, and provide useful information on the design of Gabor filters and feature extraction algorithms, we also present a description on their position, number, size, scale, and overlap distributions.

4.1 Databases

The Japanese female facial expression (JAFFE) database [35] contains 213 gray images of seven facial expressions (six basic + neutral) posed by 10 Japanese females. Each image has a resolution of 256*256 pixels. Each object has three or four frontal face images for each expression and their faces are approximately located in the middle of the images. All images have been rated on six emotion adjectives by 60 subjects.

The Cohn-Kanade AU coded facial expression (CK) database [36] is one of the most comprehensive benchmarks for facial expression tests. The released portion of this database includes 2105 digitized image sequences from 182 subjects ranged in age from 18 to 30 years. 65% are female; 15% are African-American and 3% Asian or Latino. Six basic expressions were based on descriptions of prototypic emotions. Image sequences from neutral to

target display were digitized into 640*480 or 490 pixel arrays with eight-bit precision for gray scale values.

In this paper, all the images of six basic expressions from the JAFFE database are used. For the CK database, 1,184 images that represent one of the six expressions are selected, four images for each expression of 92 subjects. The images are chosen from the last image (peak) of each sequence, then one every two images. The images of 10 subjects in the JAFFE database are classified into 10 sets, each of which includes images of one subject. Similarly, all images in the CK database are classified into 10 similar sets and all images of one subject are included in the same set.

4.2 'Salient' Patch Selection

The feature extraction step produces a feature set containing 148,032 patches. To reduce the feature dimension and the redundant information, it is necessary to select a subset of 'salient' patches. In this paper, the widely used and efficiency proved boosting algorithm - Adaboost [37] is used for 'salient' patch selection.

Since Adaboost was designed to solve two-class problems, in this research, the one-against-rest strategy is used to solve the six-emotion-class problem. The training process stops when the empirical error is below 0.0001 with an initial error of 1. This setting is inspired by the stopping condition in [20] that there is no training error and the generalization error becomes flat. For the training set, the JAFFE database includes all database images, whereas the CK database is only composed of the peak frames.

4.3 Position Distribution of 'Salient' Patches

The position distribution of the 'salient' patches demonstrates the most important facial areas for each emotion. In Fig. 6, the patches are distributed over different Gabor scales, and they are drawn in one scale image for a simply and clear demonstration. Based on this figure, we can see that the positions are distributed differently over six emotions. However, most of these patches for all emotions tend to concentrate on the areas around mouth and eyes. For sad and surprise, the 'salient' patches on JAFFE focus on the eye areas, while those on CK focus on the mouth area. For the rest four emotions, they have similar distributions between two databases. As shown in the second and third rows in Fig. 6, the positions of the 'salient' patches in our work and those of the point-based 'salient' features in [38] for the same emotion tend to focus on the same areas. This suggests that there exist the same 'salient' areas for each emotion regardless of using point-based or patch-based Gabor features. However, the overall number of the 'salient' patches is much less than that of the point-based 'salient' features (177 versus 538).

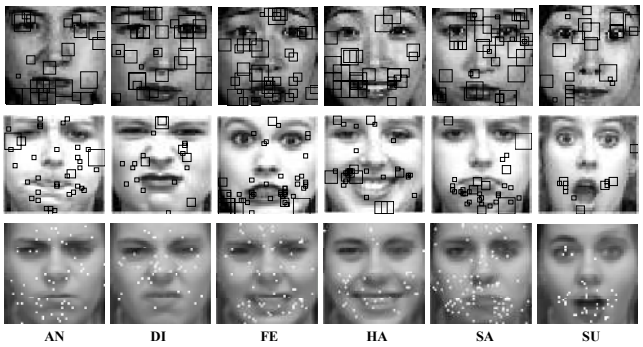


Fig. 6. Position distribution of the selected 'salient' patches for six emotions. The first and second rows show positions of the selected patches in the proposed approach on the JAFFE and CK databases respectively, whereas the third row reveals positions of the selected point-based Gabor features in [38] on the CK database.

4.4 Number and Size Distributions of 'Salient' Patches

The number and size distributions can provide useful hints on the number of patches for different emotions and how to choose suitable patch sizes during feature extraction. Seen from Fig. 7, two databases have a similar overall number of the 'salient' patches. Among six emotions, fear and sad need the largest numbers of patches to achieve the pre-set recognition accuracy, whereas surprise requires the least number. Within four patch sizes, the size 4*4 takes a significant proportion of the overall number of the 'salient' patches.

On the other hand, there are also some differences be-

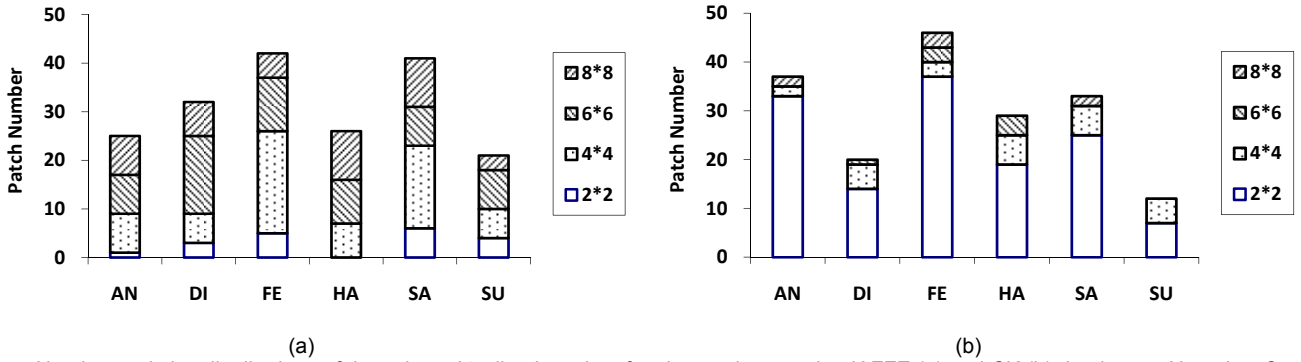


Fig. 7. Number and size distributions of the selected 'salient' patches for six emotions on the JAFFE (a) and CK (b) databases. Note that O_{num} is four for all patch sizes of $P_j*P_j*O_{num}$ and it is not shown.

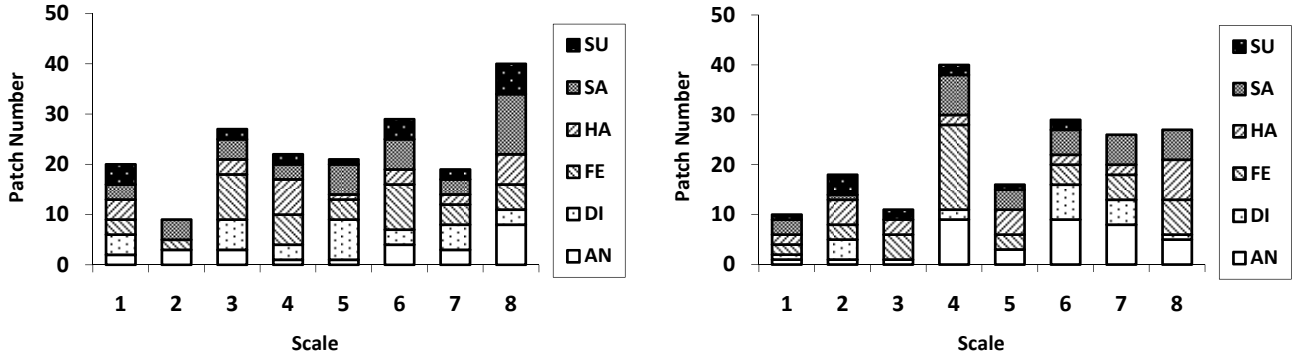


Fig. 8. Scale distribution of the selected 'salient' patches for six emotions on the JAFFE (a) and CK (b) databases.

TABLE 1
OVERLAPPING PATCHES ON JAFFE AND CK DATABASES

	JAFFE	CK
Patch Size	5(4*4); 1(6*6); 2(8*8)	2(2*2); 1(8*8)
Patch Scale	1(3rd); 2(6th); 2(7th); 3(8th)	3(4th)
Emotion Pair	3(AN-AN); 2(DI-DI); 1(FE-FE); 1(FE-SA); 1(SA-SA)	2(AN-AN); 1(AN-DI)
Total Number	8	3

The figures before parentheses stand for the number of the overlapped patches, while content in parentheses indicates patch sizes, scales and emotion pairs. As an instance of emotion pairs, "1(FE-SA)" means one patch is shared by fear and sad.

tween two databases. The number for anger on JAFFE is much less than that on CK, while the number for disgust on JAFFE is much bigger than that on CK. Moreover, four patch sizes are evenly distributed among six emotions on JAFFE, but the patch size 2*2 takes a significant proportion of the overall number of the 'salient' patches on CK. This reflects that emotions in JAFFE images need big sizes of patches to represent useful information, whereas those in CK images only require small sizes of patches. The reason may be that the emotions in CK are most distinct than those in JAFFE.

4.5 Scale Distribution of 'Salient' Patches

The scale distribution is a very important factor for determining the scale number of Gabor filters. Observed from Fig. 8, the 'salient' patches of two databases are unevenly distributed across eight scales. JAFFE emphasizes on the 8th scale and CK focuses on the 4th scale. For both databases, the higher scales (4th to 8th) contain more patches than the lower scales (1st to 3rd). Therefore, the emotional information is distributed across all scales with an emphasis on the higher scales, which confirms Littlewort's argument that a wider range of spatial frequencies, particularly high frequencies, could potentially improve performance [20].

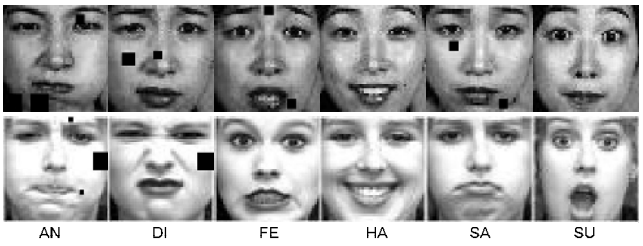


Fig. 9. Position distribution of the overlapping patches for six emotions. The upper row is for JAFFE and the lower row is for CK.

4.6 Overlap Distribution of 'Salient' Patches

Table 1 demonstrates the characteristics of number, size, scale, emotion pair of the overlapping patches, which are selected as 'salient' patches more than one time. As can be seen, the JAFFE database has a larger number of the overlapping patches than the CK database (8 versus 3). As for the patch size, 4*4 dominates the overlapping patches on JAFFE, while 2*2 takes the most part of these patches on CK. With respect to the patch scale, the patches of JAFFE tend to distribute on the *6th*, *7th* and *8th* scales, whereas those of CK are all included by the *4th* scale. The patch size and scale distributions again reveal that JAFFE needs larger patches than CK. For the emotion pair, the majority of the overlapping patches are shared by the same emotion. As shown in Fig. 9, the overlapping patches are mainly distributed over disgust and anger.

5 EXPERIMENTAL RESULTS

In this section, we present the recognition and computational performance of the proposed approach. The performance obtained with and without matching area is also compared. Finally, a performance comparison with previous approaches is conducted.

5.1 Recognition Performance

5.1.1 JAFFE Database

The performance results are obtained by averaging the

TABLE 2
CRRs OF SIX EMOTIONS ON JAFFE DATABASE

	DL_1	DL_2	SL_1	SL_2
Linear	81.52%	**92.93%**	87.50%	88.59%
Polynomial	54.89%	64.13%	45.65%	60.33%
RBF	76.63%	89.67%	82.07%	87.50%
Sigmoid	25.00%	-	26.09%	29.35%

The CRR of DL_2 and sigmoid SVM is not shown.

TABLE 3
CONFUSION MATRIX OF SIX EMOTIONS ON JAFFE DATABASE

	AN	DI	FE	HA	SA	SU	Overall
AN	29	1	0	0	0	0	**96.67%**
DI	2	27	0	0	0	1	**90.00%**
FE	2	0	30	0	0	0	93.75%
HA	1	0	1	29	2	0	93.55%
SA	0	0	1	1	29	0	93.55%
SU	0	1	1	1	0	27	**90.00%**

TABLE 4
CRRs OF SIX EMOTIONS ON CK DATABASE

	DL_1	DL_2	SL_1	SL_2
Linear	90.20%	93.36%	83.67%	86.71%
Polynomial	87.73%	91.22%	65.43%	80.97%
RBF	92.34%	**94.48%**	80.07%	86.26%
Sigmoid	26.46%	-	37.39%	66.67%

The CRR of DL_2 and sigmoid SVM is not shown.

TABLE 5
CONFUSION MATRIX OF SIX EMOTIONS ON CK DATABASE

	AN	DI	FE	HA	SA	SU	Overall
AN	81	1	2	0	9	0	**87.10%**
DI	4	92	2	2	1	1	90.20%
FE	0	4	138	7	0	1	92.00%
HA	0	2	2	203	0	0	98.07%
SA	6	0	2	0	118	3	91.47%
SU	0	0	0	0	0	207	**100%**

correct recognition rate (CRR) of all sets in 10 leave-one-set-out cross-validations. Table 2 shows the results obtained using four SVMs and four distances. From this table, we can see that the proposed approach performs the best with a CRR of 92.93% using DL_2 and linear SVM. Regarding the performance of distances, DL_2 achieves higher CRRs than the other three distances for all SVMs. When L_1 is used, sparse distances outperform dense distances for linear, RBF and sigmoid SVMs. On the contrary, when L_2 is used, dense distances outperform spare distances for all SVMs (note that the CRR of DL_2 and sigmoid SVM is not shown). For both sparse and dense distances, L_2 performs better than L_1 for all SVMs. Among four SVMs, linear and RBF outperform polynomial and sigmoid for all distances. More exactly, the best performance is obtained by linear, which is followed by RBF, whereas sigmoid ranks the least.

Table 3 demonstrates the confusion matrix of six emotions using DL_2 and linear SVM. Observed from this table, disgust and surprise belong to the most difficult facial expressions to be correctly recognized with the same CRR of 90.00%, whereas anger is the easiest one with a CRR of 96.67%. Regarding the misrecognition rate, anger contributes the most; as a result, it has a major negative impact on the overall performance. The emotion that follows in misrecognition rate is fear.

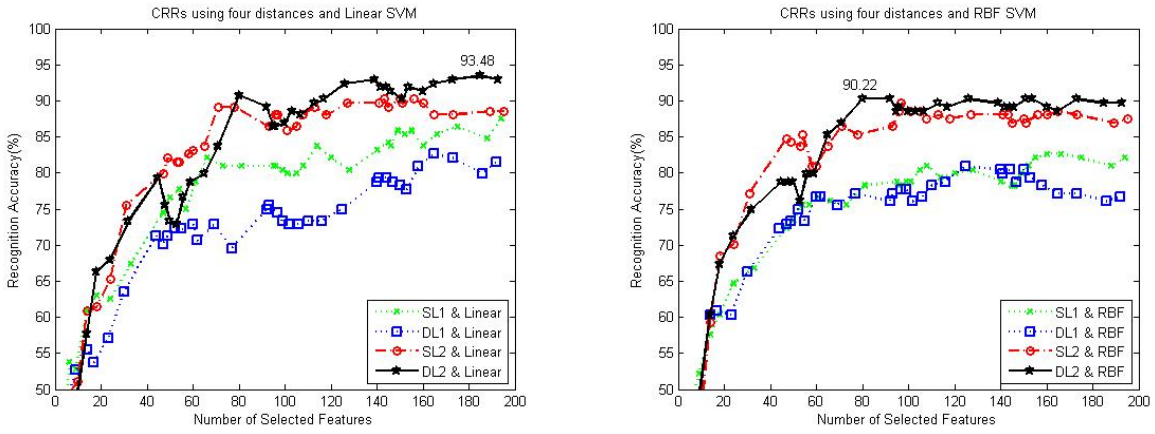
5.1.2 CK Database

The CRRs using four SVMs and four distance metrics are shown in Table 4, in which the proposed approach obtains the highest CRR of 94.48% using DL_2 and RBF SVM. Regarding the performance of distances, DL_2 keeps the highest CRRs for all SVMs (note that the CRR of DL_2 and sigmoid SVM is not shown). Moreover, dense distances have a higher overall performance than sparse distances. This reflects that emotional information in the CK images is distributed over all orientations rather than the dominant orientation of Gabor features. As for SVMs, RBF performs the best for dense distances, while linear performs the best for sparse distances. This confirms with the results in [20] that RBF and linear perform better than polynomial on the CK database.

Table 5 shows the confusion matrix of six emotions using DL_2 and RBF SVM. As can be seen, surprise performs the best with a CRR of 100%, the following one is happy with a CRR of 98.07%. On the other hand, anger is the most difficult facial expression to be correctly recognized with a CRR of only 87.10%. The performance of surprise and anger on CK contrasts with that on JAFFE, in which surprise and anger are the most difficult and easiest emotions respectively. The reason probably is that surprise images on CK are often characterized as an exaggerated "open mouth", while those on JAFFE are normally with a "close or slightly open mouth". This can be seen from Fig. 6 that the selected patches for CK focus on the mouth region, but those for JAFFE are mainly distributed around the eyes regions. Similarly, anger images on JAFFE are better expressed by the selected patches in mouth region than the selected patches are all over the face region those on CK. Among six emotions, anger and sad contribute most to the misrecognition rate.

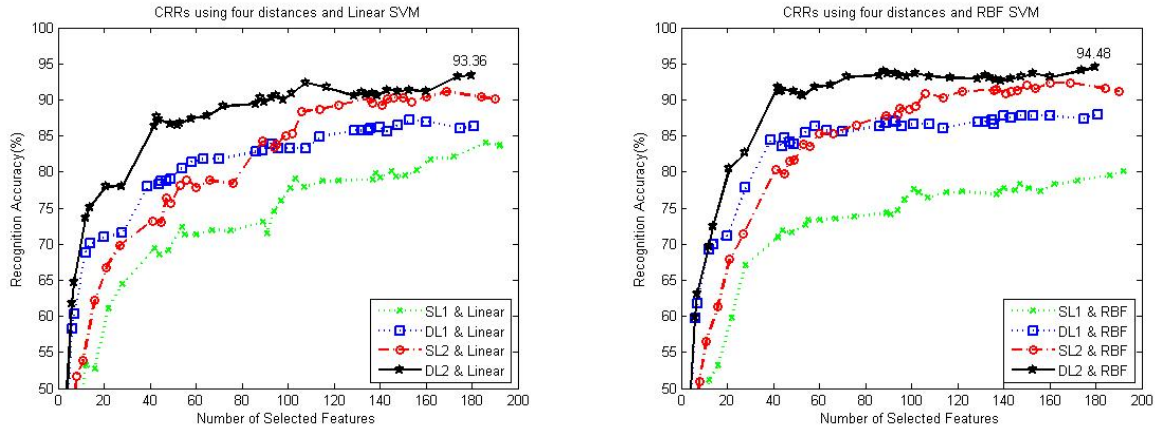
5.2 Performance versus Number of Patches

We also test the relationship between the performance and the number of patches. Table 6 shows the error thresholds used to control the number of the patches selected by Adaboost. These thresholds are set based on our observation that the empirical errors of Adaboost decrease



(a) (b)

Fig. 10. Recognition performance versus the number of patches on the JAFFE database using linear (a) and RBF (b) SVMs.



(a) (b)

Fig. 11. Recognition performance versus the number of patches on the CK database using linear (a) and RBF (b) SVMs.

TABLE 6
ERROR THRESHOLDS USED TO CONTROL THE NUMBER OF PATCHES

Index	Error thresholds
1^{st} -10^{th}	(10, 9, 8, 7, 6, 5, 4, 3, 2, 1) * 0.1
11^{th} -19^{th}	(9, 8, 7, 6, 5, 4, 3, 2, 1) * 0.01
20^{th} -28^{th}	(9, 8, 7, 6, 5, 4, 3, 2, 1) * 0.001
29^{th} -38^{th}	(9, 8, 7, 6, 5, 4, 3, 2, 1, 0) * 0.0001

Note that '1' rejects all features, whereas '0' accepts all features.

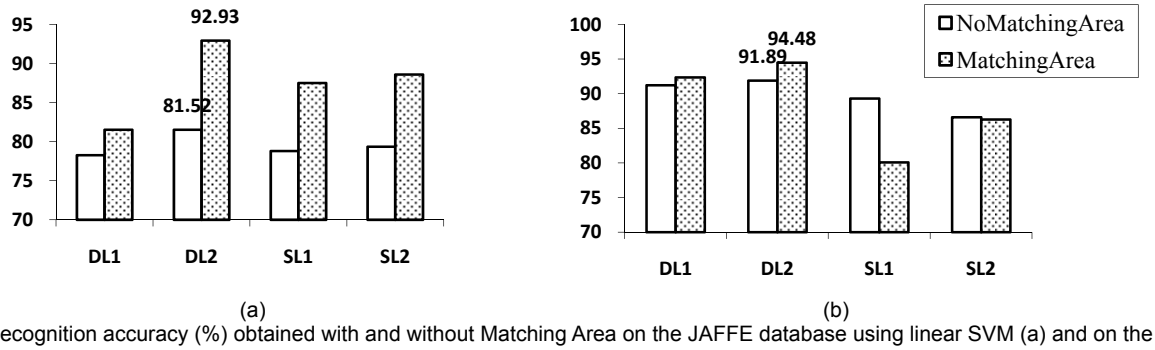
with a factor of 10 and the numbers are evenly distributed between decimal intervals. For instance, the number of errors between 0.01 and 0.02 is similar to that between 0.003 and 0.004. Accordingly, 38 groups of features are obtained by selecting patches with empirical errors bigger than the corresponding error thresholds.

For the JAFFE database, as can be seen from Fig. 10, the proposed approach achieves the highest CRR of 93.48% using DL_2 and linear SVM when the error threshold equals to 0.0001 and the number of patches equals to 185. The overall performance of four distances grows up rapidly at the starting stage, however, it begins to level off when the number of patches exceeds 150 for linear and 80 for RBF. For the overall performance of SVMs, linear performs better than RBF for all distances. Regarding the overall performance of distances, for both linear and RBF, the best performance is achieved by DL_2, which is followed by SL_2. On the other hand, SL_1 and DL_1 rank the last two.

For the CK database, seen from Fig. 11, the proposed approach obtains the highest CRR of 94.48% using DL_2 and RBF SVM when the error threshold is 0 and the number of patches is 180. This implies that a performance improvement still can be achieved using a larger number of patches. Similar to that on JAFFE, the CRR grows up rapidly at the starting stage and L_2 outperform L_1 for both linear and RBF. On the other hand, the CRR reaches the plateau with a quicker speed than that on JAFFE and DL_1 performs better than SL_1. Moreover, the performance difference between linear and RBF is smaller than that on JAFFE.

5.3 Matching Area versus No Matching Area

To evaluate the performance improvement rising from



(a) (b)

Fig. 12. Recognition accuracy (%) obtained with and without Matching Area on the JAFFE database using linear SVM (a) and on the CK database using RBF SVM (b).

the use of facial movement features, we compare the performance of with and without matching area. In the latter case, the distance features are obtained by performing subtraction between two patches at the exact same position. Therefore, the resulting features do not include the information of feature movements. Fig. 12 shows the comparison results obtained when the error threshold of Adaboost is 0. The classifiers of JAFFE and CK are linear and RBF SVMs respectively. As can be seen, for the JAFFE database, the recognition performance of the proposed approach using four distances is greatly boosted due to the use of matching area. There is a CRR increase of 11.41% using DL_2. For the CK database, the CRRs of DL_1 and DL_2 are improved about 2.5% due to the use of matching area, while the performance of SL_1 and SL_2 is not benefited from matching area. Considering the highest CRR of four distances, we can see that taking facial movement features into account helps to improve the recognition performance.

5.4 Perforamnce under Registration Errors

To test the performance of the proposed approach using images with registration errors, we add uniform random noises into the coordinate of the top-left corner and the scale of each face region produced by the widely used Viola-Jones face detector [39]. The noises are controlled so that both the coordinate and the image scale randomly change within a range of [-a%, a%] ($\alpha \subset [1,2,3,4,5,6]$) of face width. In addition, we also include neutral images in the experiment. In videos, the starting and ending frames for an expression can be determined by classifying the current frame into "neutral" or "emotion". In this way, two sets of database images with different levels of noises are created. Fig. 13 shows sample images with 3% and 6% errors from the JAFFE and CK databases. After scaling the simulated images into 48*48, there will be a maximum of a pixels (48*2*a%) changes of position and scale for a level of a% errors.

To handle scale changes, the matching scale is set to include two neighbor scales and the scale itself. In each of the 10 sets cross-validation, non-error images in nine sets are used for training, while the error-simulated images of the one set left are used for testing. This is important for the real situation that only ideally registered images are available for the training. To testify the usefulness of matching scale, we give the results of using both matching area and matching scale (AreaScale), as well as only using matching area (Area). Note that SL_1 is used here. We also compare the results with using point-based 4 orientations and 8 scales Gabor features + Adaboost feature selector + RBF SVM classifier (point-based Gabor).

(a) (b)

Fig. 13. Samples images with errors simulated by uniform random noises ranged [-3%, 3%] (a) and [-6%, 6%] (b) of face width.

Fig. 14 shows the performance of three approaches under the simulated errors. As expected, all approaches suffer decreasing performances using a larger percentage of errors. The two proposed approaches achieve higher overall performances than the approach using point-based Gabor features, for both JAFFE and CK. This again demonstrates the advantage of patch-based Gabor over point-based Gabor features in terms of the performance under face registration errors. Using both matching area and scale (AreaScale) performs better than using matching area only (Area) for both the databases. At the error level of 4%, which can be considered as larger than the errors produced by real face detectors, AreaScale still keeps a CRR of 69.5% and 83.9% for JAFFE and CK respectively. Therefore, the proposed approach achieves promising results under the simulated registration errors.

5.5 Computational Time Performance

Fig. 15 illustrates the average computational time at three stages, including Gabor images (Gab), patch matching (PM), and classification (SVM). The program was developed by Matlab 7.6.0 under a laptop configuration of core duo 1.66GHz CUP and 2GB memory. The proposed approach achieves a speed of 0.1258 seconds per image for the JAFFE database (using DL_2 and linear) and 0.1185 seconds per image for the CK database (using DL_2 and

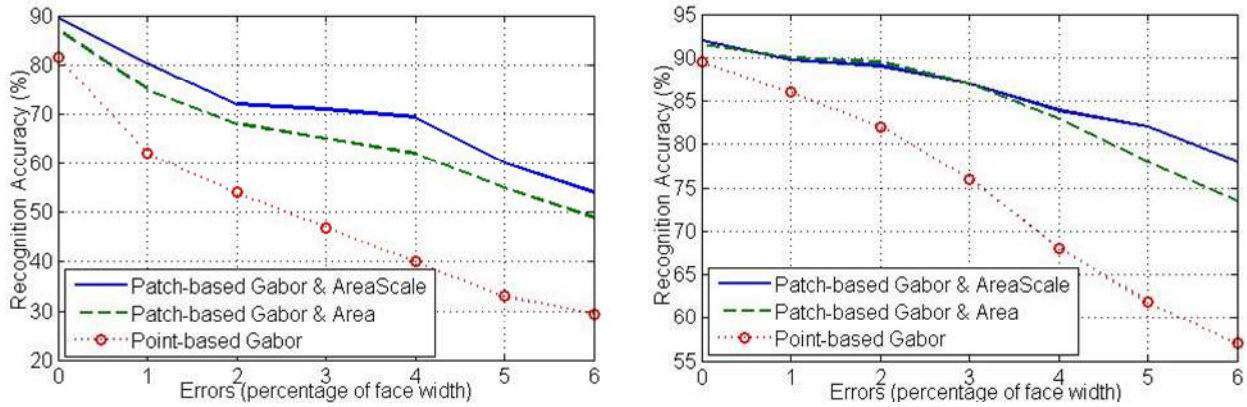


Fig. 14. Recognition performance under face registration errors on JAFFE (a) and CK (b) databases.

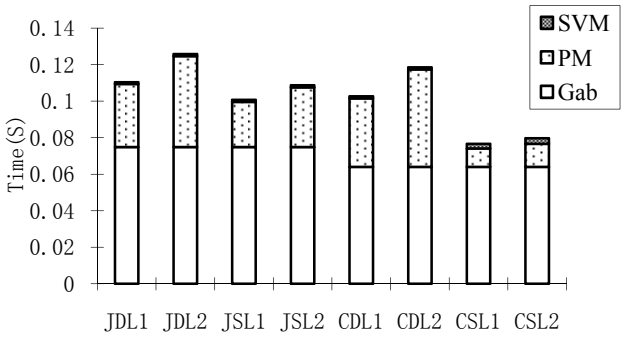


Fig. 15. The used time (in seconds) per image at three stages. 'Gab' and 'PM' indicate the stages of Gabor image and patch matching, 'J' and 'C' indicate using the JAFFE and CK databases, 'D' and 'S' stand for using dense and sparse distances, 'L1' and 'L2' represent L_1 and L_2 distances respectively. The time is obtained using the SVM type with the highest CRR for each distance.

RBF). Thus, a real-time processing is expected once our approach could be developed by time efficient languages, such as C, C++. Among three stages, computing Gabor images takes the biggest proportion of the overall time, while the classification requires the least amount of time.

5.6 Comparison with State-of-the-Art Performance

Table 7 demonstrates the results compared with the reported results of the benchmarked approaches. These approaches are selected because they produced the state-of-the-art performance using a similar testing strategy and the same databases. In [40], only the recognition results of the leave-one-subject-out strategy is used here as this strategy is more similar to our leave-one-set-out cross validations. The recognition results in [41] were obtained by removing two JAFFE images named "KR.SR3.79" and "NA.SU1.79".

As shown in table 7, the proposed approach outperforms all nine benchmarked approaches ([23], [26], [40], [41], [42], [43], [44], [45], [46]) when the JAFFE database is used, and three of four benchmarked approaches ([20], [21], [26], [46]) when the CK database is used. When six

TABLE 7
COMPARISON WITH STATE-OF-THE-ART PERFORMANCE

	Feature	JAFFE	CK
Our approach	patch-based Gabor	**92.93%** (6)	**94.48%** (6)
[20], 2006	Gabor	-	93.3% (7)
[21], 2008	Gabor + Haar	-	93.1% (7)
[23], 2005	Gabor + FSLP	91.0% (7)	-
[26], 2009	boosted-LBP;	81.0%; (7)	95.1%; (6)
	LBP	-	92.6% (6)
[40], 2009	SFRCS	85.92% (7)	-
[41], 2008	WMMC	65.77% (7)	-
[42], 2005	fuzzy integral	83.2% (6)	-
[43], 2006	KCCA	77.05% (6)	-
[44], 2007	KCCA	67.0% (7)	-
[45], 2008	DCT	79.30% (7)	-
[46], 2010	FEETS + PRNN	83.84% (7)	95.87% (5)

The numbers in parentheses stand for the number of the testing facial expressions. Abbreviations: kernel canonical correlation analysis (KCCA); weighted maximum margin criterion (WMMC); discrete cosine transform (DCT); salient feature and reliable classifier selection (SFRCS); feature selection via linear programming (FSLP); face emotion tree structures (FEETS); probabilistic based recursive neural network (PRNN).

emotions are used, the CRR of the proposed approach is 9.73% and 15.88% higher than those in [42] and [43] respectively on the JAFFE database, as well as 1.88% higher than that obtained using LBP features in [26] on the CK database. The result of the proposed approach on the CK database is 0.62% lower than the result obtained using the boosted-LBP features in [26] and 1.39% lower than the result in [46]. However, the work [26] normalized the face based on manually labeled eye locations and improved the results by optimizing the SVM parameters. While the proposed approach does not involve normalization of face regions and uses the default parameters in LIBSVM. Another difference is that the database images in the proposed approach represent bigger emotional intensity than those in [26]. To be specific, the proposed approach collects images using a "every two images from the peak frame" strategy, while [26] just used the three peak frames. Wong and Cho [46] obtained the results based on

five-fold cross validations and five expressions, therefore, it used more training images and classified less emotions compared to our approach.

6 Conclusion and Future Work

This paper explores the issue of facial expression recognition using facial movement features. The effectiveness of the proposed approach is testified by the recognition performance, computational time, and comparison with the state-of-the-art performance. The experimental results also demonstrate significant performance improvements due to the consideration of facial movement features, and promising performance under face registration errors.

The results indicate that patch-based Gabor features show a better performance over point-based Gabor features in terms of extracting regional features, keeping the position information, achieving a better recognition performance, and requiring a less number. Different emotions have different 'salient' areas; however, the majority of these areas are distributed around mouth and eyes. In addition, these 'salient' areas for each emotion seem to be not influenced by the choice of using point-based or using patch-based features. The 'salient' patches are distributed across all scales with an emphasis on the higher scales. For both the JAFFE and CK databases, DL_2 performs the best among four distances. As for emotion, anger contributes most to the misrecognition. The JAFFE database requires larger sizes of patches than the CK database to keep useful information.

The proposed approach can be potentially applied into many applications, such as patient state detection, driver fatigue monitoring, and intelligent tutoring system. In our future work, we will extend our approach to a video-based FER system by combining patch-based Gabor features with motion information in multi-frames. Recent progress on action recognition [47] and face recognition [48] has laid a foundation for using both appearance and motion features.

Acknowledgment

The authors wish to thank Nicki Ridgeway for providing the Cohn-Kanade AU-Coded Facial Expression Database and the providers of the JAFFE database.

References

[1] 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.

[2] T. Yan, C. Jixu, and J. Qiang, "A Unified Probabilistic Framework for Spontaneous Facial Action Modeling and Understanding," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 32, pp. 258-273, 2010.

[3] P. S. Aleksic and A. K. Katsaggelos, "Automatic facial expression recognition using facial animation parameters and multi-stream HMMs," *Information Forensics and Security, IEEE Transactions on,* vol. 1, pp. 3-11, 2006.

[4] D. Hamdi, V. Roberto, S. Albert Ali, and G. Theo, "Eyes do not lie: spontaneous versus posed smiles," in *Proceedings of the international conference on Multimedia* Firenze, Italy: ACM, 2010, pp. 703-706.

[5] T.-H. Wang and J.-J. James Lien, "Facial expression recognition system based on rigid and non-rigid motion separation and 3D pose estimation," *Pattern Recognition,* vol. 42, pp. 962-977, 2009.

[6] 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.

[7] Y. Cheon and D. Kim, "Natural facial expression recognition using differential-AAM and manifold learning," *Pattern Recognition,* vol. 42, pp. 1340-1350, 2009.

[8] F. Tsalakanidou and S. Malassiotis, "Real-time 2D+3D facial action and expression recognition," *Pattern Recognition,* vol. 43, pp. 1763-1775, 2010.

[9] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding,* vol. 91, pp. 160-187, 2003.

[10] G. Zhao and M. Pietikainen, "Boosted multi-resolution spatiotemporal descriptors for facial expression recognition," *Pattern Recognition Letters,* vol. 30, pp. 1117-1127, 2009.

[11] F. Dornaika and F. Davoine, "Simultaneous Facial Action Tracking and Expression Recognition in the Presence of Head Motion," *International Journal of Computer Vision,* vol. 76, pp. 257-281, 2008.

[12] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *International Journal of Human-Computer Studies,* vol. 65, pp. 724-736, 2007.

[13] L. Peng and S. J. D. Prince, "Joint and implicit registration for face recognition," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 1510-1517.

[14] T. Huang, A. Nijholt, M. Pantic, and A. Pentland, "Human Computing and Machine Understanding of Human Behavior: A Survey," in *Artifical Intelligence for Human Computing.* vol. 4451: Springer Berlin / Heidelberg, 2007, pp. 47-71.

[15] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust Object Recognition with Cortex-Like Mechanisms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 29, pp. 411-426, 2007.

[16] J. Mutch and D. G. Lowe, "Multiclass Object Recognition with Sparse, Localized Features," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, pp. 11-18.

[17] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A Biologically Inspired System for Action Recognition," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1-8.

[18] L. Zhen, L. Shengcai, H. Ran, M. Pietikainen, and S. Z. Li, "Gabor volume based local binary pattern for face representation and recognition," in *Automatic Face & Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, 2008, pp. 1-6.

[19] L. Wiskott, J. M. Fellous, N. Kuiger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 19, pp. 775-779, 1997.

[20] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," *Image and Vision Computing,* vol. 24, pp. 615-625, 2006.

[21] H. Y. Chen, C. L. Huang, and C. M. Fu, "Hybrid-boost learning for multi-pose face detection and facial expression recognition," *Pattern Recognition,* vol. 41, pp. 1173-1185, 2008.
[22] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an automotive environment," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 2005, pp. 1085-1088.
[23] G. Guo and C. R. Dyer, "Learning from examples in the small sample case: face expression recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B,* vol. 35, pp. 477-488, 2005.
[24] S. Zafeiriou and I. Pitas, "Discriminant Graph Structures for Facial Expression Recognition," *Multimedia, IEEE Transactions on,* vol. 10, pp. 1528-1540, 2008.
[25] T. Xiang, M. K. H. Leung, and S. Y. Cho, "Expression recognition using fuzzy spatio-temporal modeling," *Pattern Recognition,* vol. 41, pp. 204-216, 2008.
[26] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *Image and Vision Computing,* vol. 27, pp. 803-816, 2009.
[27] Z. Guoying and M. Pietikainen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 29, pp. 915-928, 2007.
[28] P. Yang, Q. Liu, and D. N. Metaxas, "Boosting encoded dynamic features for facial expression recognition," *Pattern Recognition Letters,* vol. 30, pp. 132-139, 2009.
[29] C. Orrite, A. Gañán, and G. Rogez, "HOG-Based Decision Tree for Facial Expression Classification," in *Pattern Recognition and Image Analysis*, 2009, pp. 176-183.
[30] S. Shiguang, G. Wen, C. Yizheng, C. Bo, and Y. Pang, "Review the strength of Gabor features for face recognition from the angle of its robustness to mis-alignment," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, pp. 338-341 Vol.1.
[31] D. Gabor, "Theory of communication," *Institution of Electrical Engineers -- Journal -- Radio and Communication Engineering,* vol. 93, pp. 429-457, 1946.
[32] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, pp. 273-297, 1995.
[33] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines, 2001," *Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm,* 2001.
[34] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 994-1000 vol. 2.
[35] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, 1998, pp. 200-205.
[36] T. Kanade, J. F. Cohn, and T. Yingli, "Comprehensive database for facial expression analysis," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, 2000, pp. 46-53.
[37] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," 1997, pp. 119-139.
[38] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction," in *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03. Conference on*, 2003, pp. 53-53.
[39] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision,* vol. 57, pp. 137-154, 2004.
[40] M. Kyperountas, A. Tefas, and I. Pitas, "Salient feature and reliable classifier selection for facial expression classification," *Pattern Recognition,* vol. 43, pp. 972-986, 2010.
[41] C. Zhengdong, S. Bin, F. Xiang, and Z. Yu-Jin, "Automatic coefficient selection in Weighted Maximum Margin Criterion," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 2008, pp. 1-4.
[42] W. Yuwen, L. Hong, and Z. Hongbin, "Modeling facial expression space for recognition," in *Intelligent Robots and Systems, 2005. (IROS 2005). 2005 IEEE/RSJ International Conference on*, 2005, pp. 1968-1973.
[43] Z. Wenming, Z. Xiaoyan, Z. Cairong, and Z. Li, "Facial expression recognition using kernel canonical correlation analysis (KCCA)," *Neural Networks, IEEE Transactions on,* vol. 17, pp. 233-238, 2006.
[44] Y. Horikawa, "Facial Expression Recognition using KCCA with Combining Correlation Kernels and Kansei Information," in *Computational Science and its Applications, 2007. ICCSA 2007. International Conference on*, 2007, pp. 489-498.
[45] J. Bin, Y. Guo-Sheng, and Z. Huan-Long, "Comparative study of dimension reduction and recognition algorithms of DCT and 2DPCA," in *Machine Learning and Cybernetics, 2008 International Conference on*, 2008, pp. 407-410.
[46] J.-J. Wong and S.-Y. Cho, "A face emotion tree structure representation with probabilistic recursive neural network modeling," *Neural Computing & Applications,* vol. 19, pp. 33-54, 2010.
[47] K. Schindler and L. van Gool, "Action snippets: How many frames does human action recognition require?," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1-8.
[48] A. Hadid and M. Pietikainen, "Combining appearance and motion for face and gender recognition from videos," *Pattern Recognition,* vol. 42, pp. 2818-2827, 2009.

Ligang Zhang is currently pursuing the Ph.D. degree at the Queensland University of Technology, Brisbane, Australia. His research interests include face-related technologies (e.g. face expression recognition and face recognition), affective computing, affective content analysis, pattern recognition and computer vision.

Dian Tjondronegoro is Senior Lecturer in the Faculty of Science and Technology at Queensland University of Technology. He leads Mobile Multimedia Research Group and has published more than 60 refereed articles in the field. His research interests include video analysis, summarization, and visualisation, multi-channel content analysis, mobile applications, and interaction design.