This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2593719, IEEE Transactions on Affective Computing

Facial Expression Recognition in Video with Multiple Feature Fusion

Junkai Chen, Zenghai Chen, Zheru Chi, *Member, IEEE* and Hong Fu,

Abstract—Video based facial expression recognition has been a long standing problem and attracted growing attention recently. The key to a successful facial expression recognition system is to exploit the potentials of audiovisual modalities and design robust features to effectively characterize the facial appearance and configuration changes caused by facial motions. We propose an effective framework to address this issue in this paper. In our study, both visual modalities (face images) and audio modalities (speech) are utilized. A new feature descriptor called Histogram of Oriented Gradients from Three Orthogonal Planes (HOG-TOP) is proposed to extract dynamic textures from video sequences to characterize facial appearance changes. And a new effective geometric feature derived from the warp transformation of facial landmarks is proposed to capture facial configuration changes. Moreover, the role of audio modalities on recognition is also explored in our study. We applied the multiple feature fusion to tackle the video-based facial expression recognition problem under lab-controlled environment and in the wild, respectively. Experiments conducted on the extended Kohn-Kanada (CK+) database and the Acted Facial Expression in Wild (AFEW) 4.0 database show that our approach is robust in dealing with video-based facial expression recognition problem under lab-controlled environment and in the wild compared with the other state-of-the-art methods.

Index Terms—Facial expression recognition; Multiple feature fusion; HOG-TOP; Geometric warp feature; Acoustic feature.

1 Introduction

Facial expression, as a powerful nonverbal channel, plays an important role for human beings to convey emotions and transmit messages. Automatic facial expression recognition (AFEC) can be widely applied in many fields such as medical assessment, lie detection and human computer interaction [1]. AFEC has attracted great interest in the past two decades. However, facial expression analysis is a very challenging task because facial expressions caused by facial muscle movements are subtle and transient [2]. To capture and represent these movements is a key issue to be addressed in facial expression analysis.

Two main streams of facial expressions analysis are widely adopted in the current research and development. One stream is to detect facial actions. The study reported in [3], [4] showed that each facial expression contains a unique group of facial action units. The Facial Action Coding System (FACS), which was first proposed by Ekman and Friesen in 1978 [5] and then enhanced in 2002 [6], is the best known system developed for human beings to describe facial actions. Another stream of facial expression analysis is to carry out facial affect (emotion) recognition directly. Most researchers deal with the recognition task of six universal emotions: happy, sad, fear, disgust, angry and surprise [7].

Many efforts have been made for facial expression recognition. The methodologies used are commonly categorized into appearance based methods and geometry based methods [8]. An appearance based method applies feature descriptors to model facial texture changes. A geometry based method captures facial configurations in which a set of facial fiducial points is used to characterize the face shape.

Previous works mainly focused on static and single face image based facial expression recognition. Recently, facial expression recognition in video has attracted great interest. Compared with a static image, a video sequence can not only provide spatial appearance but also include facial motions and accompanied speech. The key to solve the problem of video based facial expression recognition is to exploit the representation capability of multi modalities (e.g. visual and audio information) and design robust features to effectively characterize the facial appearance and configuration changes caused by facial muscular activities.

To achieve this goal, we propose an effective framework based on multiple feature fusion for facial expression recognition in video. We explore the potentials of visual modalities (face images) and audio modalities (speech) in our study. In addressing visual modalities, we extend the Histograms of Oriented Gradients (HOG) [9] to temporal Three Orthogonal Planes (TOP), inspired by a temporal extension of Local Binary Patterns, LBP-TOP [10]. The proposed HOG-TOP is used to characterize facial appearance changes. We show that HOG-TOP performs as well as LBP-TOP for facial expression recognition. In addition, compared with LBP-TOP, HOG-TOP is more compact and effective to characterize facial appearance changes. Moreover, an effective geometric warp feature derived from the warp transformation of facial landmarks is proposed to capture facial configuration changes. We show that the proposed geometric warp feature is more effective compared with other proposed geometric features [11], [12]. We also ex-

- *Junkai Chen and Zheru Chi are with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong. E-mail: Junkai.Chen@connect.polyu.hk, chi.zheru@polyu.edu.hk*
- *Zenghai Chen is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. E-mail: zenghaichen@ntu.edu.sg*
- *Hong Fu is with the Department of Computer Science, Chu Hai College of Higher Education, Hong Kong. E-mail: hongfu@chuhai.edu.hk*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2593719, IEEE Transactions on Affective Computing

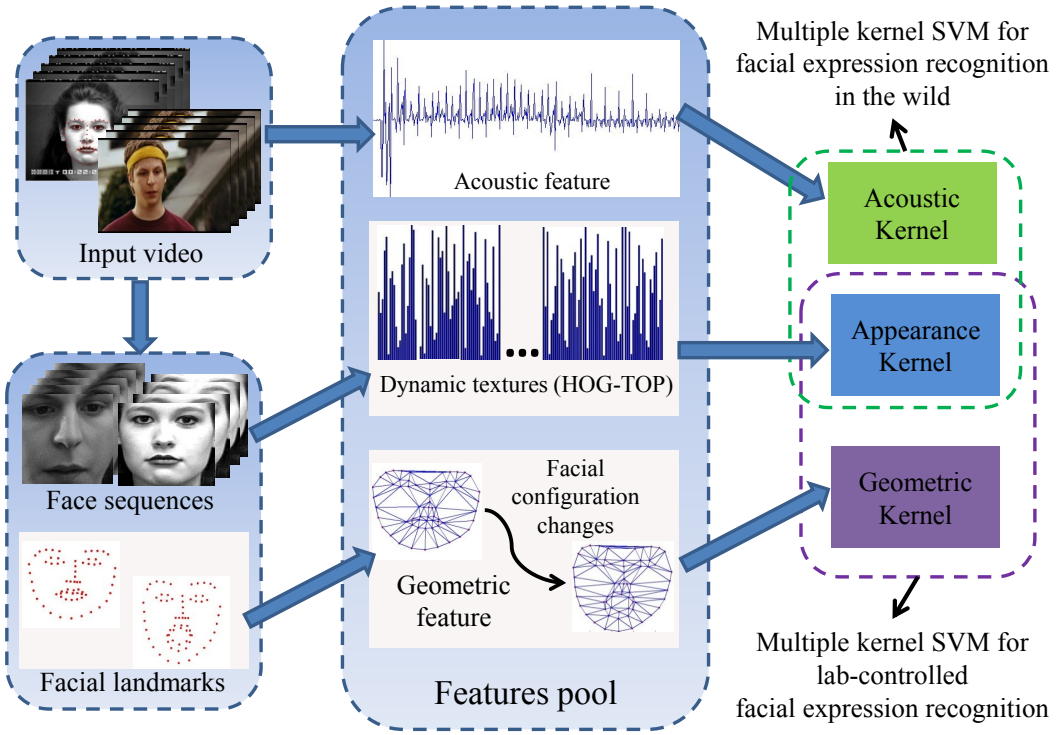

Fig. 1: Block diagram of our proposed framework. Geometric features coupled with dynamic textures (HOT-TOP) are used to deal with lab-controlled facial expression recognition while acoustic features and dynamic textures (HOG-TOP) are fused to tackle facial expression recognition in the wild.

plore the role of audio modalities on affect recognition. We find that audio modalities can provide some complementary information, especially for facial expression recognition in the wild. We further apply a multiple feature fusion method to deal with facial expression recognition under lab-controlled environment and in the wild, respectively. As shown in Fig. 1, geometric features coupled with dynamic textures (HOT-TOP) are used to deal with lab-controlled facial expression recognition. Acoustic features and dynamic textures (HOG-TOP) are fused to tackle facial expression recognition in the wild.

Our contributions are summarized as follows:

1. We develop a framework which can effectively tackle facial expression recognition in video. A multiple feature fusion method is used to deal with facial expression recognition under lab-controlled environment and in the wild, respectively.
2. We propose a new feature descriptor HOG-TOP to characterize facial appearance changes and a new effective geometric feature to capture facial configuration changes.
3. We show that multiple features can make different contributions and can achieve better performance than the individual features applied alone. We also show that multiple feature fusion can enhance the discriminative power of multiple features.

The remainder of the paper is organized as follows. In Section 2, we review some related work. Section 3 presents our proposed approach. Experimental results and discussions are presented in Section 4. The paper is concluded in Section 5.

2 Related Work and Motivation

2.1 Static Image Based Methods

Many researchers apply static image based models to handle facial expression problem. One or several peak frames are usually selected for extracting appearance or geometric features. For instance, the methods reported in [11], [12], [13] applied the facial landmarks to characterize the whole face shape. And the method [14] measured the displacements of several selected candidate fiducial points. Bag of Words (BoW) based on the multi-scale dense SIFT features were applied to represent facial appearance textures in [15]. Local Fisher discriminant analysis (LFDA) was used for feature extraction in [16]. The method [17] applied Gabor filters to extract facial movement features. A novel framework for expression recognition by using appearance features of selected facial patches was proposed in [18]. However, automatically picking out key frames from a video sequence is usually difficult. The methods [19], [20] attempted to classify every frame first and adopt a voting strategy to label the video sequence. It is necessary to extract features from each frame. LBP was applied in [19]. Pyramid of Histograms of Oriented Gradients (PHOG) and Local Phase Quantization (LPQ) features were used in [20].

2.2 Dynamic Texture Based Methods

There exists a drawback for a static image based method: extracting features from an individual frame fails to utilize dynamic information which is important to describe facial motions. Dynamic texture based methods can effectively deal with this problem. Dynamic texture based methods

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2593719, IEEE Transactions on Affective Computing

attempt to simultaneously model the spatial appearance and dynamic motions in a video sequence. Zhao et al. [10] proposed LBP-TOP, a temporal extension of local binary patterns, for facial expression analysis in video. A facial component LBP-TOP was proposed in [21]. A Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) was proposed in [22]. A SpatioTemporal Local Monogenic Binary Pattern (STLMBP) feature descriptor was proposed in [23], [24]. In addition, Long et al. [25] employed Independent Component Analysis (ICA) to learn spatiotemporal filters from videos, and then extracted dynamic textures using the learned filters. Chew et al. [26] employed sparse temporal representation to model the temporal dynamics of facial expressions in video. Li et al. [27] developed a dynamic Bayesian network to simultaneously and coherently represent the facial evolvement at different levels.

2.3 Audiovisual Based Methods

Static image based methods or dynamic texture based methods only rely on visual modalities. However, audio or speech is also important for human beings to convey emotions and intentions. Audio modalities can provide some complementary information in addition to visual modalities. Recently, audiovisual based methods for affect recognition have attracted growing attention from the affective computing community. A number of approaches have been proposed to combine audio and visual modalities for affect recognition (e.g. [28], [29], [30], [31]). A comprehensive survey can be found in [32]. Acoustic features extracted from voice or speech and visual features extracted from face images are combined to tackle this problem. For example, voice and lip activity were used in [33]. Face images and speech were employed in [34]. The methods reported in [35], [36] applied several feature descriptors such as SIFT, HOG, PHOG etc. to encode face images and combined them with acoustic features to recognize facial expression in the wild.

2.4 Motivation

We can see that feature extraction plays a center role on affect recognition in video. Designing an effective feature is important and meaningful. LBP-TOP is widely used for modeling dynamic textures. However, there are two limitations of LBP-TOP. One is the high dimensionality. The size of LBP-TOP coded using a uniform pattern is 59×3 [10]. Moreover, although LBP-TOP is robust to deal with illumination changes, it is insensitive to facial muscle deformations. In this work, we propose a new feature called HOG-TOP, which is more compact and effective to characterize facial appearance changes. More details on HOG-TOP can be found in Section 3.1.

In addition, configuration and shape representations play an important role in human vision for the perception of facial expressions [37]. We believe that previous works have not yet fully exploited the potentials of configuration representations. Characterizing face shape [11], [12] or measuring displacements of fiducial points [14], [38] only are not sufficient to capture facial configuration changes, especially the subtle non-rigid changes. In this work, we introduce a more robust geometric feature to capture facial configuration changes. More discussion on our proposed geometric feature is given in Section 3.2.

Fig. 2: The textures in XY, XT and YT planes.

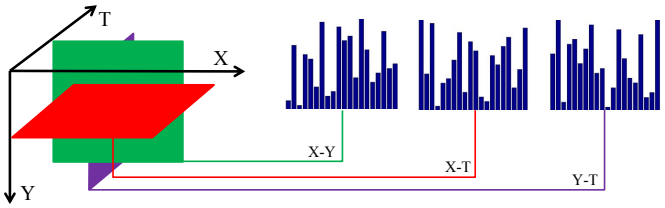

Fig. 3: The HOG from Three Orthogonal Planes (TOPs).

3 Methodology

This section presents the details of our proposed approach. We introduce the three types of features and multiple feature fusion employed in our study.

3.1 Histogram of Oriented Gradients from Three Orthogonal Planes

Histograms of oriented gradients (HOG) [9] were first proposed for human detection. The basic idea of HOG is that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions. HOG is sensitive to object deformations. Facial expressions are caused by facial muscle movements. For example, mouth opening and raised eyebrows will generate a surprise facial expression. These movements could be regarded as types of deformations. HOG can effectively capture and represent these deformations [39]. However, the original HOG is limited to deal with a static image. In order to model dynamic textures from a video sequence with HOG, we extend HOG to 3-D to compute the oriented gradients on three orthogonal planes XY, XT, and YT (TOP), i.e. HOG-TOP. The proposed HOG-TOP is used to characterize facial appearance changes.

A video sequence includes three orthogonal directions, i.e. X, Y, and T (time) directions. The XY plane provides spatial appearance, and XT and YT planes record temporal or motion information. Fig. 2 illustrates the textures extracted from the three orthogonal planes. In our study, we compute the distributions of oriented gradients of each plane and obtain HOG features, namely HOG-XY, HOG-XT and HOG-YT, as shown in Fig. 3. Each point in a video sequence includes three orthogonal neighborhoods lying on XY, XT and YT planes, respectively. We first compute the gradients along X, Y and T directions with a 3×3 Sobel mask. The gradient orientations are defined as $\theta_{XY} = tan^{-1}(G_Y/G_X)$, $\theta_{XT} = tan^{-1}(G_T/G_X)$, $\theta_{YT} = tan^{-1}(G_T/G_Y)$, where G_X, G_Y, and G_T are the gradients along the X, Y and T

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2593719, IEEE Transactions on Affective Computing

directions, respectively. These angles are quantized into K (K is 9 in our work) orientation bins with a range of $0°-360°$ or $0° - 180°$.

We enumerate the appearance of these gradient orientations and obtain a histogram in each plane. The three histograms are concatenated to form a global description with the spatial and temporal features. Fig. 3 shows that the three histograms from the three planes are combined into a single one. The HOG-TOP computation algorithm is shown in Algorithm 1.

Algorithm 1 Compute the HOG-TOP.

Input: Video sequence V, which contains N frames with the same width and height.
Output: The histograms of oriented gradients from three orthogonal plans (HOG-TOP).
Algorithm:
Get the number of frames N, frame width W and height H.
for $t = 2 : N - 1$ **do**
  for $x = 2 : W - 1$ **do**
    for $y = 2 : H - 1$ **do**
      get the local patch in XY, XT, and YT planes.
      $P_{xy} = V(x-1:x+1, y-1:y+1, t)$;
      $P_{xt} = V(x-1:x+1, y, t-1:t+1)$;
      $P_{yt} = V(x, y-1:y+1, t-1:t+1)$;
      Compute the gradients G_X, G_Y, and G_T; and gradient orientations θ_{XY}, θ_{XT}, θ_{YT}. Quantize the θ_{XY}, θ_{XT}, θ_{YT} into one of 9 bins. Get a histogram in each plan, i.e. HOG-XY, HOG-XT and HOG-YT.
    end for
  end for
end for
Normalize the HOG-XY, HOG-XT and HOG-YT respectively. Concatenate the three histograms into a long histogram.

LBP-TOP computes the difference of a pixel with respect to its neighborhood, making LBP-TOP robust in dealing with illumination changes. HOG-TOP computes the oriented gradients of a pixel, which is more sensitive to object deformations [9]. Facial expressions are caused by facial muscle movements, which can be regarded as types of muscle deformations. HOG-TOP is therefore more effective to characterize facial appearance changes than LBP-TOP.

Another advantage of HOG-TOP is the feature dimensionality. Compared with LBP-TOP, the size of HOG-TOP is much smaller than that of LBP-TOP. The size of LBP-TOP coded using a uniform pattern is 59×3 [10], [40]. and the size of HOG-TOP quantized into 9 bins is 9×3, which is much more compact than that of LBP-TOP.

In order to utilize local spatial information, a block-based method is introduced in our study, as shown in Fig. 4. We can divide the image sequence into many blocks and extract the HOG-TOP features from each block. The HOG-TOP features of all the blocks can be concatenated to represent the whole sequence. In our experiments, the face is first cropped from the original image and resized to 128×128. We partition the face image into 8×8 blocks with each block having a size of 16×16. The number of bins is set to 9 with an angle range of $0° - 180°$.

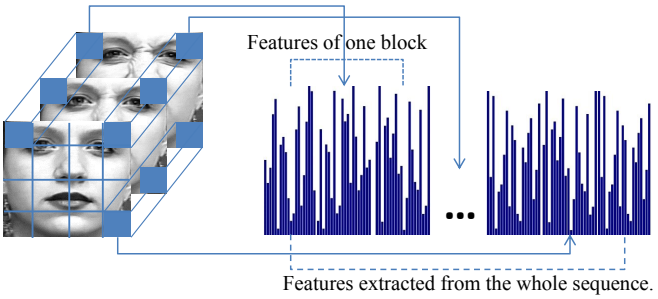

Fig. 4: The HOG-TOP features extracted from each block are concatenated together to represent the whole sequence.

3.2 Geometric Warp Feature

In this section, we introduce a more robust geometric feature namely geometric warp feature, which is derived from the warp transform of the facial landmarks. Facial expressions are caused by facial muscle movements. These movements result in the displacements of the facial landmarks. Here we assume that each face image consists of many sub-regions. These sub-regions can be formed with triangles with their vertexes located at facial landmarks, as shown in Fig. 5. The displacements of facial landmarks cause the deformations of the triangles. We propose to utilize the deformations to represent facial configuration changes.

Facial expression can be considered as a dynamic process including onset, peak and offset. We consider the displacement of the corresponding facial landmarks between onset (neutral face) and peak (expressive face). Given a set of facial landmarks $s = (x_1, y_1, x_2, y_2, ...x_n, y_n)$, where (x_i, y_i) denote the coordinates of the i-th facial landmark. These facial landmarks make up the mesh of a face, as shown in Fig. 5.

As we can see, there are many small triangles in the face, and each triangle is determined by three facial landmarks. Facial muscle movements cause the deformations of the triangles when a neutral face transforms to an expressive face. We consider a pixel (x, y) which lies in a triangle ΔABC belonging to the neutral face and the corresponding pixel (u, v) lies in a triangle $\Delta A'B'C'$ belonging to the expressive face, as shown in Fig. 6. From [41], we know that the pixel (x, y) can be expressed with a linear combination of the three vertexes.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \lambda_1 \begin{bmatrix} x_2 - x_1 \\ y_2 - y_1 \end{bmatrix} + \lambda_2 \begin{bmatrix} x_3 - x_1 \\ y_3 - y_1 \end{bmatrix} \quad (1)$$

And the coefficients λ_1, λ_2 can be obtained as

$$\lambda_1 = \frac{(x - x_1)(y_3 - y_1) - (y - y_1)(x_3 - x_1)}{(x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1)} \quad (2)$$

$$\lambda_2 = \frac{(x_2 - x_1)(y - y_1) - (y_2 - y_1)(x - x_1)}{(x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1)} \quad (3)$$

The point (u, v) in the triangle $\Delta A'B'C'$ of the expressive face can be defined with the three vertexes and λ_1, λ_2,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2593719, IEEE Transactions on Affective Computing

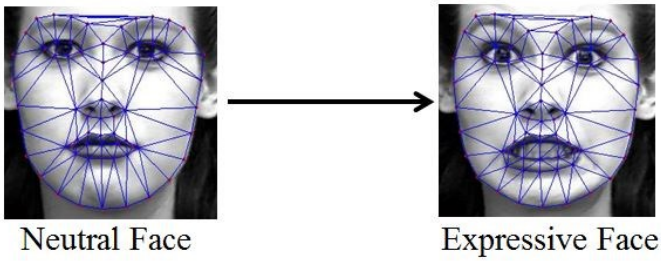

Fig. 5: Facial landmarks describe the shape of a face.

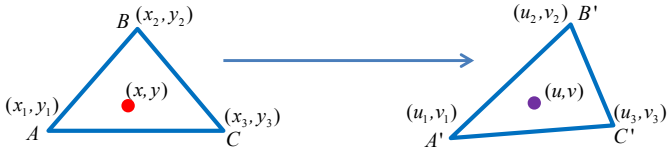

Fig. 6: A pixel (x, y) in a triangle ΔABC of the neutral face is transformed to another pixel (u, v) in a triangle $\Delta A'B'C'$ of the expressive face.

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} + \lambda_1 \begin{bmatrix} u_2 - u_1 \\ v_2 - v_1 \end{bmatrix} + \lambda_2 \begin{bmatrix} u_3 - u_1 \\ v_3 - v_1 \end{bmatrix} \tag{4}$$

Combining Eq. (2) with Eq. (3), Eq. (4) can be rewritten as:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{bmatrix} \tag{5}$$

Each pair of triangles between the neutral face and the expressive face can define a unique transform and each affine transform is determined by 6 parameters $a_1, a_2, ..., a_6$. We compute the 6 parameters for each warp transform and concatenate all the parameters as a long global feature vector, which is used to characterize facial configuration changes. We will show by experiments that the proposed geometric warp feature is more effective than the other geometric features [11], [14], [38].

3.3 Acoustic Feature

Visual modalities (face images) and audio modalities (speech) can both convey the emotions and intentions of human beings. Audio modalities also provide some useful clues for affect recognition in video. For instance, with voice signal, the method [42] proposed an enhanced autocorrelation (EAC) feature for emotion recognition in video.

One successful acoustic feature extraction is to obtain the time series of multiple paralinguistic descriptors and then using pooling operations on each time series to extract feature vectors. Schuller et al. [43] showed how to compute the acoustic features by taking 21 functionals of 38 low level descriptors and their first regression coefficients. The 38 low-level descriptors shown in Table 1 are first extracted and smoothed by simple moving average low-pass filtering. After that, 21 functionals are employed and 16 zero-information features are eliminated. Finally, two single features: the number of onsets (F0) and turn duration are added. A total of 1,582 acoustic features are extracted from each video. These acoustic features include energy/spectral Low Level Descriptors (LLD) (top 6 items in Table 1) and voice related LLD (bottom 4 items in Table 1).

TABLE 1: Acoustic features: 38 low level descriptors along with their first regression coefficients and 21 functionals [43].

Descriptors	Functionals
PCM loudness	Position max./min.
MFCC (0-14)	Arithmetic Mean
log Mel Freq. Band (0-7)	skewness, kurtosis
LSP Frequency (0-7)	lin. regression coeff.
F0	lin. regression error
F0	Envelope quartile
Voicing Prob.	quartile range
Jitter local	percentile
Jitter consec. frame pairs	percentile range
Shimmer local	up-level time

We explore the representation ability of acoustic features for affect recognition in our study. Experiments show that audio modalities (speech) can provide useful complementary information in addition to visual modalities. The visual features coupled with acoustic features can achieve better performance for facial expression recognition in the wild.

3.4 Multiple Feature Fusion

Features from different modalities can make different contributions. Traditional SVM concatenates different features into a single feature vector and built a single kernel for all these different features. However, constructing a kernel for each type of features and integrating these kernels optimally can enhance the discriminative power of these features. The study in [44] showed that using multiple kernels with different types of features can improve the performance of SVM. A multiple kernel SVM is designed to learn both the decision boundaries between data from different classes and the kernel combination weights through a single optimization problem [45].

Given a training set with labeled samples $D = \{(\mathbf{x}_i, y_i)|\mathbf{x}_i \in \mathbf{R}^n, y_i \in \{-1, 1\}\}_{i=1}^N$. A decision line is obtained by solving the following primal optimization problem,

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{1}{2} \|\mathbf{w}\|^2 \\ &s.t. \quad y_i(\mathbf{w}\mathbf{x}_i + b) \geq 1, i = 1, 2, ...N \end{aligned} \tag{6}$$

In general, we solve the dual form of the primal optimization problem. The dual formulation of the traditional single kernel SVM optimization problem is given by

$$\begin{aligned} &\max_{\alpha} \left[\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K_{ij} \right] \\ &s.t. \quad \sum_{i=1}^{N} \alpha_i y_i = 0, 0 \leq \alpha_i \leq C \end{aligned} \tag{7}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2593719, IEEE Transactions on Affective Computing

where K_{ij} is the kernel matrix, and $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, here $k(\cdot,\cdot)$ is the kernel function and $\mathbf{x}_i, \mathbf{x}_j$ are the feature vectors.

Multiple kernel fusion applies a linear combination of multiple kernels to substitute for the single kernel. In our study, we adopt the formulation proposed in [46] in which the kernel is actually a convex combination of basis kernels:

$$\begin{aligned} K_{ij} &= \sum_{i=1}^{M} \beta_m k_m(\mathbf{x}_i, \mathbf{x}_j) \\ s.t. \quad & \beta_m \geq 0, \quad \sum_{i=1}^{M} \beta_m = 1 \end{aligned} \tag{8}$$

We apply a multiple kernel fusion framework to deal with facial expression recognition under lab-controlled environment and in the wild, respectively, as shown in Fig. 1. HOG-TOP and acoustic feature are optimally fused to handle the problem of facial expression recognition in the wild, while HOG-TOP and geometric warp feature are combined to tackle the problem of facial expression recognition under lab-controlled environment.

In the followings, we detail how to find an optimal combination of HOG-TOP and acoustic feature for facial expression recognition in the wild. It can be easily extended to the problem of facial expression recognition under lab-controlled environment.

We denote the dynamic texture HOG-TOP as $\mathbf{x}$ and acoustic feature as $\mathbf{z}$, then we have

$$K_{ij} = \beta k_1(\mathbf{x}_i, \mathbf{x}_j) + (1-\beta) k_2(\mathbf{z}_i, \mathbf{z}_j) \tag{9}$$

with $0 \leq \beta \leq 1$, where K is the kernel matrix, $k_1(\cdot,\cdot), k_2(\cdot,\cdot)$ are the basis kernels. The basis kernels could be linear kernel, radial basis function (RBF) kernel and polynomial kernel, etc. We need to learn the kernel weight β and coefficients α. In our study, we construct a linear kernel for each type of feature and build a two-step method to search for the optimal values of β and α. We set two nested iterative loops to optimize both the classifier and kernel combination weights. In the outer loop, we adopt the grid search to find the kernel weight β. In the inner iteration, a solver of SVM (LIBSVM [47] is used in our work) is implemented by fixing the kernel weight β to find the coefficients α. Then given a new sample which contains visual feature HOG-TOP $\mathbf{x}$ and acoustic feature $\mathbf{z}$, the predict label y can be obtained by

$$y = sgn(\sum_{i=1}^{N} y_i \alpha_i (\beta k_1(\mathbf{x}_i, \mathbf{x}) + (1-\beta) k_2(\mathbf{z}_i, \mathbf{z})) + b) \tag{10}$$

In our work, the one-vs-one method is employed to deal with the multiclass-SVM problem and we adopt the max-win voting strategy to do the classification. Finally, the β value and α values with the highest overall classification accuracy in the validation data set are obtained as the optimal kernel weight and coefficients.

4 Experimental Results and Discussions

4.1 Data sets

In order to evaluate our methods, we conduct the experiments on three public data sets: the Extended Cohn-Kanade (CK+) data set [11], GEMEP-FERA 2011 data set [19] and the Acted Facial Expression in Wild (AFEW) 4.0 data set [48]. We first give a brief description of the three data sets.

The **Extended Cohn-Kanade (CK+) data set** contains 593 image sequences from 123 subjects. The face images in the database are lab-controlled. The image sequences vary in duration from 10 to 60 frames. In total, 327 of 593 image sequences have emotion labels and each is categorized into one of the following seven emotion classes: anger (An), contempt (Co), disgust (Di), fear (Fe), happiness (Ha), sadness (Sa) and surprise (Su). Each image sequence changes from the onset (the neutral frame) to the peak (the expressive frame). In addition, the X-Y coordinates of 68 facial landmark points were given for each image in the database. The landmark points of key frames within each video sequence were manually labelled, while the remaining frames were automatically aligned using the AAM fitting algorithm [41].

The **GEMEP-FERA 2011 data set** contains 289 sequences of 10 actors, who are trained by a professional director. It is divided into a training set of 155 sequences and a test set of 134 sequences. Each sequence is categorized into the following five emotions: anger (An), fear (Fe), happiness (Ha), relief (Re) and sadness (Sa). Only the training set provides emotion labels. This database is more challenging than the CK+ database, since there are head movements and gesture variations in image sequences.

The **Acted Facial Expression in Wild (AFEW) 4.0 data set** includes video clips collected from different movies which are believed to be close to real world conditions. The database splits into a training set, a validation set and a test set. There are 578 video clips in the training set. The validation and test sets have 383 video clips and 407 video clips, respectively. Each video clip belongs to one of the seven categories: anger (An), disgust (Di), fear (Fe), happiness (Ha), neutral (Ne), sadness (Sa), and surprise (Su). This database provides original video clips and aligned face sequences. They applied the model proposed in [49] to extract the faces from video clips and aligned the faces. Different from the CK+ and GEMEP-FERA 2011 data sets, facial expressions in AFEW 4.0 are more natural and spontaneous. The variations in illumination, pose and background in image sequences increase the complexity of facial expression analysis.

Fig. 7 shows the selected image sequences from the three databases. The first row is the face images from the CK+ database, which are frontal-view and lab-controlled faces. The middle row shows the images from the GEMP-FERA 2011 database; there exist head movements and gesture variations. The bottom row is an image sequence from AFEW 4.0 database. We can see that the background is complex and there exist illumination changes and pose variations.

4.2 Feature Extraction

In our experiments, three types of features are employed, namely HOG-TOP, geometric warp feature and acoustic feature.

In extracting HOG-TOP from image sequences, each face image is first cropped and resized to 128×128. The resized face image is then partitioned into 8×8 blocks with a size of 16×16. The bin number is set to 9 with an angle range

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2593719, IEEE Transactions on Affective Computing

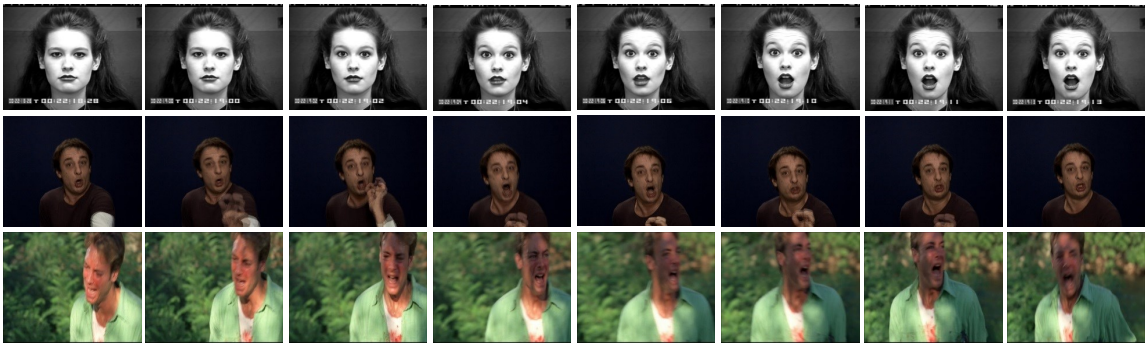

Fig. 7: The selected image sequences from the three databases. From top to bottom: CK+, GEMEP-FERA2011 and AFEW 4.0.

of $0^\circ - 180^\circ$. In each block, we can obtain a HOG-TOP with a dimension of $3 \times 9 = 27$. We then concatenate the HOG-TOP of the 8×8 blocks into a long feature vector with a dimension of $3 \times 9 \times 8 \times 8 = 1728$.

Facial landmarks are used to compute the geometric warp features. We compute the warp transform of facial landmarks between the neutral face and an expressive face. Each face contains 68 facial landmarks. These facial landmarks divide the face into many non-overlap sub regions by Delaunay triangulation. In our work, we take 109 pair of triangles (the smallest number of triangles available in face images). Each pair of triangles between the neutral face and an expressive face can define a unique transform and each affine transform is determined by six parameters (see Section 3.2). The warp transform coefficients are finally concatenated as a feature vector of $6 \times 109 = 654$ elements to represent the geometric warp feature.

The acoustic features with a length of 1582 used in our work are provided by the database [34], [48]. The acoustic features are extracted by applying the open-source Emotion Affect Recognition (openEAR) toolkit [50] backend OpenSMILE [51].

4.3 Experimental Results

4.3.1 A Comparison of HOG-TOP and LBP-TOP

We first compare the performance of HOG-TOP proposed in our work with LBP-TOP proposed in [10]. When we compute the LBP-TOP features, we take the general settings adopted in most reported works. The resized face image is partitioned into 4×4 blocks. The LBP-TOP is coded with a uniform pattern. The LBP-TOP histogram of each block is a feature vector of $3 \times 59 = 177$ elements. The length of the feature vector consists of 4×4 blocks is $3 \times 59 \times 4 \times 4 = 2832$.

There are 327 image sequences with emotion labels belonging to 118 subjects in the CK+ database. We follow the protocol proposed in [11] and take the leave-one-subject-out cross validation strategy. Each time the samples from one subject are used for testing and the remaining samples from all other subjects are used for training. In order for each subject to be evaluated once, we carry out 118 validations. The classification accuracy acquired on the CK+ database by using two types of features is shown Table 2.

We also compare the performance of the two features in the GEMEP-FERA 2011 database. Since only the emotion labels of training set are publicly available, we do the evaluation on the training set. There are seven subjects in the training set. We adopt the leave-one-subject-out strategy and carry out seven cross validations. Table 3 shows the performance obtained by applying the two features.

As for the AFEW 4.0 database, we utilize the training set to train an SVM classifier and test the classifier on the validation set. The database provided a baseline method [34] which employed LBP-TOP to represent the dynamic textures of the video sequence and trained an SVM with non-linear RBF kernel for emotion classification. The accuracy acquired on the AFEW 4.0 database by applying two types of features is shown in Table 4.

We use the overall accuracy to evaluate the performance.

TABLE 2: The classification accuracy of LBP-TOP and HOG-TOP on the CK+ database (%).

	LBP-TOP	HOG-TOP
Anger	75.6	88.9
Contempt	88.9	66.7
Disgust	93.2	94.9
Fear	80.0	76.0
Happiness	98.5	95.6
Sadness	78.6	67.9
Surprise	92.8	97.6
Overall	89.3	89.6

TABLE 3: The classification accuracy of LBP-TOP and HOG-TOP on the GEMEP-FERA 2011 database (%).

	LBP-TOP	HOG-TOP
Anger	56.2	43.7
Fear	26.7	36.7
Joy	58.1	61.3
Relief	51.6	54.8
Sad	74.2	74.2
Overall	53.6	54.2

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2593719, IEEE Transactions on Affective Computing

TABLE 4: The classification accuracy of LBP-TOP and HOG-TOP on validation set of the AFEW 4.0 database (%).

	LBP-TOP	HOG-TOP
Neutral	19.0	58.7
Anger	50.0	73.4
Disgust	25.0	22.5
Fear	15.2	4.3
Happiness	57.1	60.3
Sadness	16.4	4.9
Surprise	21.7	2.2
Overall	30.6	35.8

The overall accuracy is defined as

$$O_{acc} = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K} m_{nk}}{\sum_{n=1}^{N}\sum_{k=1}^{K} M_{nk}} \quad (11)$$

where K is the number of classes, N is the number of cross validation folds, m_{nk} is the number of correctly predicted samples of the k-th class in the n-th fold, and M_{nk} denotes the total samples of the k-th class in the n-th fold. The classification rate of each individual facial expression (k-th class) is $\frac{\sum_{n=1}^{N} m_{nk}}{\sum_{n=1}^{N} M_{nk}}$.

From the experimental results, we can see that the overall classification accuracy obtained by using HOG-TOP on the CK+ database and GEMEP-FERA 2011 database is 89.6% and 54.2%, respectively. It is competitive with the result of 89.3% and 53.6% obtained by applying LBP-TOP on the two databases. While the overall classification rate of HOG-TOP on the AFEW 4.0 database is 35.8%, which is better than 30.6% obtained by using LBP-TOP, meaning that HOG-TOP is more robust in capturing the subtle facial appearance changes in the wild. In addition, HOG-TOP with a length of 1728 is more compact than LBP-TOP with a length of 2832. We further examine the confidence intervals (the variances across cross-validation folds) of two feature sets. Since each fold in the CK+ database contains several samples only (118 folds (subjects) all together), the variance across cross-validation folds is very large and therefore it is not very meaningful to report the variances for this data set. On the other hand, the training set and validation set are fixed in the AFEW 4.0 data set and therefore no variance cross the folds for this data set can be reported. We only compare the variances of the two feature sets on the GEMEP-FERA 2011 database. The variances of HOG-TOP and LBP-TOP are 12.7% and 12.6%, respectively, which are comparable with each other. We also compare the computational speeds of the two features under the 64-bit Win 7 operating system with a Core i7 CPU. We computed the two features with Matlab 8.2. The computation time depends on the block size and sequence duration. With the same block size (16×16) and sequence duration (11 frames), the computation time of HOG-TOP and LBP-TOP is 0.027s and 0.042s, respectively, showing the computational efficiency of HOG-TOP.

TABLE 5: The comparison results of different geometric features on the CK+ database (%). (GWF is our proposed geometric warp feature).

	GWF	[11]	[14]	[38]
Anger	86.7	35.0	75.6	62.2
Contempt	94.4	25.0	–	72.2
Disgust	96.6	68.4	88.2	86.4
Fear	36.0	21.7	76.0	56.0
Happiness	98.5	98.4	97.1	91.3
Sadness	75.0	4.0	89.3	39.3
Surprise	96.4	100.0	98.7	95.2
Overall	89.0	66.7	87.5	79.2

TABLE 6: The classification accuracy obtained by using four different feature sets on the CK+ database (%).

	HOG-TOP	Geometric Feature	Hybrid Feature I	Hybrid Feature II
Anger	88.9	86.7	95.6	100.0
Contempt	66.7	94.4	94.4	94.4
Disgust	94.9	96.6	94.9	96.6
Fear	76.0	36.0	52.0	84.0
Happiness	95.6	98.5	98.5	100.0
Sadness	67.9	75.0	78.6	78.6
Surprise	97.6	96.4	96.4	98.8
Overall	89.6	89.0	91.4	95.7

4.3.2 Facial Expression Recognition Under Lab-controlled Environment

We develop a model which combines HOG-TOP and geometric warp to handle the problem of facial expression recognition under lab-controlled environment. We evaluate the following different feature sets: geometric warp feature, dynamic appearance feature (HOG-TOP), hybrid feature I and hybrid feature II. Hybrid feature I denotes the feature vector of concatenating HOG-TOP and geometric warp feature directly and hybrid feature II is the optimal combination of the HOG-TOP and geometric warp feature.

We first compare our proposed geometric warp feature with the other geometric features on the CK+ data set. All the methods take the leave-one-subject-out cross validation. The comparison results are shown in Table 5. The method [11] applied a set of facial landmarks to characterize the face shape. The relative distance of eight selected fiducial points are measured to represent the geometric feature in [14]. The shifts of the facial landmarks between the neutral face and the expressive face are computed to represent the geometric feature in [38]. We can see that our proposed geometric warp feature achieves a superior performance compared with the other geometric features, meaning that the geometric warp feature is more effective to capture facial configuration changes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2593719, IEEE Transactions on Affective Computing

	An	Co	Di	Fe	Ha	Sa	Su
An	0.89	0.02	0.02	0.02	0.00	0.04	0.00
Co	0.17	0.67	0.00	0.05	0.05	0.00	0.06
Di	0.01	0.00	0.95	0.02	0.00	0.00	0.02
Fe	0.00	0.00	0.04	0.76	0.08	0.12	0.00
Ha	0.00	0.00	0.00	0.01	0.96	0.00	0.03
Sa	0.21	0.00	0.00	0.07	0.00	0.68	0.04
Su	0.00	0.01	0.00	0.00	0.00	0.01	0.98

(a)

	An	Co	Di	Fe	Ha	Sa	Su
An	0.87	0.00	0.06	0.00	0.00	0.07	0.00
Co	0.00	0.94	0.00	0.00	0.00	0.06	0.00
Di	0.01	0.00	0.97	0.02	0.00	0.00	0.00
Fe	0.00	0.00	0.04	0.36	0.28	0.04	0.28
Ha	0.00	0.00	0.00	0.01	0.99	0.00	0.00
Sa	0.14	0.00	0.07	0.04	0.00	0.75	0.00
Su	0.00	0.02	0.00	0.00	0.00	0.02	0.96

(b)

	An	Co	Di	Fe	Ha	Sa	Su
An	0.96	0.00	0.04	0.00	0.00	0.00	0.00
Co	0.00	0.94	0.00	0.00	0.00	0.00	0.06
Di	0.03	0.00	0.95	0.00	0.02	0.00	0.00
Fe	0.00	0.00	0.04	0.52	0.24	0.04	0.16
Ha	0.00	0.00	0.00	0.01	0.99	0.00	0.00
Sa	0.14	0.00	0.03	0.04	0.00	0.79	0.00
Su	0.00	0.02	0.00	0.00	0.00	0.02	0.96

(c)

	An	Co	Di	Fe	Ha	Sa	Su
An	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Co	0.00	0.94	0.00	0.00	0.00	0.00	0.06
Di	0.01	0.00	0.97	0.00	0.00	0.02	0.00
Fe	0.00	0.00	0.04	0.84	0.12	0.00	0.00
Ha	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Sa	0.21	0.00	0.00	0.00	0.00	0.79	0.00
Su	0.00	0.01	0.00	0.00	0.00	0.00	0.99

(d)

Fig. 8: The confusion matrices obtained by using the four feature sets on the CK+ database: (a) HOG-TOP, (b) geometric feature, (c) hybrid feature I and (d) hybrid feature II. (An: Anger, Co: Contempt, Di: Disgust, Fe: Fear, Ha: Happiness, Sa: Sadness and Su: Surprise).

We further evaluate hybrid feature I and hybrid feature II with the leave-one-subject-out cross validation on the CK+ database and compare the performance with that obtained by applying geometric feature and HOG-TOP alone. Table 6 shows the classification accuracy obtained by the four different feature sets. Fig. 8 shows the confusion matrices of using the four different feature sets. We can see that the emotions "disgust", "happiness" and "surprise" have higher classification rates than the other emotions, indicating that these three emotions are easier to distinguish than the others. We also note that hybrid feature I (91.4%) and hybrid feature II (95.7%) outperform the geometric warp feature (89.0%) and HOG-TOP (89.6%) applied individually. We can conclude that different features (hybrid feature I) can provide complementary information and multiple feature fusion (hybrid feature II) can further enhance the discriminative ability of the combined features .

We also compare our method with the other methods. All the methods we compared follow the baseline method [11] and take the leave-one-subject-out cross validation. The methods [11], [12] combined geometric feature and appearance feature and trained an SVM to perform the classification. In [21], a weighted component-based feature descriptor to extract dynamic appearance feature was utilized and multiple kernel learning was applied to train the SVM for recognition. A sparse temporal representation classifier was proposed for facial expression recognition in [26]. The method in [24] applied spatiotemporal local monogenic binary pattern (STLMBP) feature to handle the problem of facial expression recognition.

As can be seen in Table 7, the HOG-TOP (89.6%) and geometric feature (89.3%) proposed in our method can achieve a competitive performance compared with SPTS+CAPP [11] (88.38%), CLM [12] (82.4%) and STLMBP [24] (88.4%). It demonstrates the effectiveness of our proposed features. The hybrid feature II as the optimal combination of HOG-TOP and geometric feature achieves a superior performance compared with the other methods tested, showing the effectiveness of the multiple feature fusion.

TABLE 7: Performance comparison with other methods on CK+ database.

Method	Accuracy (%)
HOG-TOP	89.6
Geometric Feature	89.3
Hybrid Feature II	**95.7**
SPTS+CAPP [11]	88.4
CLM [12]	82.4
STLMBP [24]	88.4
STR [26]	94.9
CFD [21]	93.2

TABLE 8: The classification accuracy obtained by using four feature sets on validation set of the AFEW 4.0 database (%).

	HOG-TOP	Acoustic Feature	Hybrid Feature I	Hybrid Feature II
Neutral	58.7	57.1	65.1	69.8
Anger	73.4	64.1	75.0	76.6
Disgust	22.5	15.0	12.5	17.5
Fear	4.3	26.1	8.7	15.2
Happiness	60.3	34.9	57.1	63.5
Sadness	4.9	14.7	13.1	9.8
Surprise	2.1	0.0	4.4	2.1
Overall	35.8	32.9	37.6	40.2

4.3.3 *Facial Expression Recognition in the Wild*

HOG-TOP and acoustic feature are fused to tackle the problem of facial expression recognition in the wild. We first evaluate our method on the validation set. Four feature sets are explored: HOG-TOP only, acoustic feature only, hybrid feature I and hybrid feature II. Hybrid feature I concatenates the HOG-TOP and acoustic feature directly. Hybrid feature II is the optimal combination of the HOG-TOP and acoustic feature.

Table 8 shows the classification accuracy obtained by applying four different feature sets. The corresponding confusion matrices are shown in Fig. 9. We can see that the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2593719, IEEE Transactions on Affective Computing

	An	Di	Fe	Ha	Ne	Sa	Su
An	0.73	0.05	0.06	0.02	0.14	0.00	0.00
Di	0.23	0.23	0.08	0.13	0.33	0.03	0.00
Fe	0.52	0.11	0.04	0.11	0.17	0.04	0.00
Ha	0.05	0.06	0.08	0.60	0.19	0.02	0.00
Ne	0.16	0.03	0.05	0.10	0.59	0.08	0.00
Sa	0.31	0.13	0.00	0.08	0.43	0.05	0.00
Su	0.26	0.02	0.07	0.09	0.50	0.04	0.02

(a)

	An	Di	Fe	Ha	Ne	Sa	Su
An	0.64	0.03	0.05	0.13	0.11	0.02	0.03
Di	0.18	0.15	0.00	0.18	0.35	0.10	0.05
Fe	0.33	0.02	0.26	0.13	0.15	0.02	0.09
Ha	0.24	0.03	0.11	0.35	0.27	0.00	0.00
Ne	0.08	0.10	0.06	0.14	0.57	0.03	0.02
Sa	0.02	0.15	0.11	0.23	0.30	0.15	0.05
Su	0.17	0.09	0.13	0.17	0.35	0.09	0.00

(b)

	An	Di	Fe	Ha	Ne	Sa	Su
An	0.75	0.06	0.02	0.02	0.11	0.03	0.02
Di	0.23	0.13	0.00	0.23	0.38	0.05	0.00
Fe	0.41	0.11	0.09	0.15	0.20	0.04	0.00
Ha	0.11	0.03	0.06	0.57	0.19	0.02	0.02
Ne	0.08	0.14	0.00	0.06	0.65	0.06	0.00
Sa	0.21	0.23	0.03	0.07	0.33	0.13	0.00
Su	0.15	0.11	0.07	0.15	0.48	0.00	0.04

(c)

	An	Di	Fe	Ha	Ne	Sa	Su
An	0.77	0.06	0.02	0.02	0.11	0.03	0.00
Di	0.28	0.18	0.00	0.15	0.33	0.08	0.00
Fe	0.39	0.09	0.15	0.15	0.20	0.02	0.00
Ha	0.08	0.05	0.06	0.63	0.16	0.02	0.00
Ne	0.10	0.08	0.00	0.08	0.70	0.05	0.00
Sa	0.21	0.18	0.00	0.07	0.44	0.10	0.00
Su	0.22	0.04	0.07	0.15	0.50	0.00	0.02

(d)

Fig. 9: The confusion matrices obtained by using the four feature sets on the validation set of AFEW 4.0 database: (a) HOG-TOP, (b) acoustic feature, (c) hybrid feature I and (d) hybrid feature II. (An: Anger, Di: Disgust, Fe: Fear, Ha: Happiness, Ne: Neutral, Sa: Sadness and Su: Surprise).

TABLE 9: Performance comparison with other methods on the test set of the AFEW 4.0 database.

Method	Accuracy (%)
Hybrid Feature II (our method)	**45.2**
LBP-TOP + Voice [34]	33.7
Lip activity + Voice [33]	35.3
STLMBP [23]	41.5
EAC [42]	40.1
ELM [52]	44.2

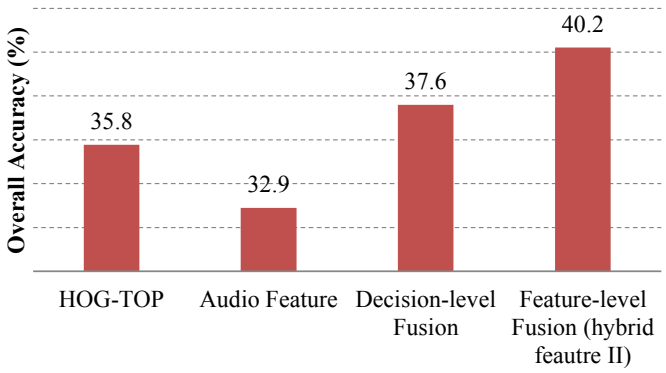

Fig. 10: The comparison results of different methods on the validation set of AFEW 4.0 database.

classification rates are much lower than the results shown in Table 6. Different from facial expressions under lab-controlled environment in which the actors or subjects can pose distinguished facial expressions, the facial expressions in the wild may be more subtle. The factors including head movements, pose variations etc. also increase classification difficulties. And sometimes, several facial expressions in the wild may appear together, which makes a facial expression to be confused with other expressions. We observe that the classification rate of emotion "surprise" is the lowest. From the confusion matrices shown in Fig. 9, we find the emotion "surprise" is mostly misclassified as emotions "anger", "happiness" and "neutral". The emotions "anger" and "neutral" have higher recognition accuracies than the other emotions. Hybrid feature I and hybrid feature II outperform the HOT-TOP and acoustic feature used individually, indicating that two feature sets are complementary with each other. Hybrid feature II achieves a superior performance compared with hybrid feature I, demonstrating that the effectiveness of the multiple feature fusion in dealing with the facial expression recognition problem in the wild.

We further apply hybrid feature II which achieves the best performance on the validation set (the kernel weights of HOG-TOP and acoustic feature are 0.73 and 0.27) to evaluate the test set. The overall recognition accuracy on the test set is 45.2%. Table 9 shows the results compared with the other methods. The baseline method [34] combined LBP-TOP and the acoustic feature. Lip activity was incorporated with voice in [33] to tackle the emotion recognition problem. The method [23] used dynamic textures only and the method [42] applied the voice only. The method [35] employed audiovisual feature for emotion recognition. We can see that our method (45.2%) improves significantly compared with the baseline method [34] and the method [33], with an improvement of about 11% and 10%, respectively. Our method is also better than [23] (41.5%) and EAC [42] (40.1%). Compared with the method [52] (44.2%), our performance is still competitive. Moreover, we applied our method to participate in the second emotion recognition in the wild challenge (EmotiW 2014) [34] and achieved the second runner-up award.

4.3.4 *Decision-level fusion vs Feature-level fusion*

The multiple feature fusion applied in our work is a kind of feature-level fusion method. Another technique, namely decision-level fusion, is also widely used in computer vision community to deal with multiple sets of features. In a preliminary study, we explore the effectiveness of the two techniques for facial expression recognition in video. A preliminary experiment is conducted on AFEW 4.0 database. As we have mentioned above, we employ the one-vs-one technique to tackle the multiclass-SVM problem, and use max-win voting strategy to conduct the classification. For decision-level fusion, we first apply the HOG-TOP and acoustic feature separately and then saved the predict results,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2593719, IEEE Transactions on Affective Computing

TABLE 10: The parameters used for extracting HOG-TOP.

Image Size	Block Size	Overlap	Blocks
96×96	12×12	No	8×8
96×96	24×24	Half	7×7
128×128	16×16	No	8×8
128×128	32×32	Half	7×7

i.e. the number of votes for each class of the two features, respectively. After that, we add the votes obtained by each individual feature together and based on the combined votes, we carry out the max-win voting strategy again to make the final decision. The overall classification rate is computed as the performance of decision-level fusion method.

Fig. 10 shows the results of the different methods tested. The overall accuracy of HOG-TOP, acoustic feature and feature-level fusion (hybrid feature II) shown in Fig. 10 is the same as shown in Table 8. From the experimental results, we find that feature-level fusion outperforms the decision-level fusion method, although they both utilize the same multiple sets of features. We also find that the improvement acquired by decision-level fusion over individual features sets is not as significant as that achieved by feature-level fusion.

4.3.5 The Effect of Block Size on HOG-TOP

We further explore the representation ability of HOG-TOP with different block sizes, from 12×12 to 32×32. Table 10 shows the parameters used for extracting HOG-TOP. The blocks with a small size (12 and 16) are not overlapped and large blocks (24×24 and 32×32) are half overlapped. We employ HOG-TOP on the CK+ database and the AFEW 4.0 database. Experiment results are shown in Tables 11 and 12.

We can see that the HOG-TOP with various block sizes achieve the similar overall accuracy. We can therefore conclude that HOG-TOP is robust to scales. We can further examine the experimental results. For facial expressions under lab-controlled environment (Table 11), HOG-TOP with a small size (12) is more effective to recognize the facial expressions "fear" and "contempt" which have subtle facial muscle activities. A small block size is more robust to capture local subtle appearance changes than a large block size. For facial expressions in the wild (Table 12), HOG-TOP with a large size (24 and 32) achieves a superior performance for facial expression "surprise", indicating that HOG-TOP with a large block size is more robust to distinguish this expression from others in the wild. Table 12 also shows that HOG-TOP with various block sizes outperforms the LBP-TOP (30.6%) for facial expression recognition in the wild.

4.4 Discussion

From the experimental results reported above, we can see that our proposed framework can efficiently handle the problem of facial expression recognition in video. Facial expressions under lab-controlled environment are different from those in the wild which are more natural and spontaneous. We propose two approaches to tackle the two different facial expression recognition problems. The two approaches both apply HOG-TOP, indicating that facial appearance plays an important role for both facial expression recognition problems. Compared with LBP-TOP, HOG-TOP is more compact and effective to characterize facial appearance changes. Facial configuration changes also provide useful clues for facial expression analysis. The facial landmarks can be located exactly on a face image under lab-controlled environment, representing the facial configuration changes caused by facial muscle movements. We propose a new effective geometric feature based on warp transform of facial landmarks and the proposed geometric warp feature is robust to capture facial configuration changes. On the other hand, it is very challenging to locate facial landmarks on face images in the wild. However, the speech also plays an important role on affect recognition. Instead of using geometric feature, acoustic feature is employed for facial expression recognition in the wild. Experimental results show that different features can make different contributions to facial expression recognition and the multiple feature fusion can enhance the discriminative ability of the multiple features. We also note that for facial expression recognition in the wild, although our method outperforms the baseline method, the performance is in general not as good as that in facial expression recognition under lab-controlled environment. Facial expression recognition in the wild is much more challenging and it will be one of our future research focuses.

TABLE 11: The performance of HOG-TOP with various block sizes on the CK+ database (%).

	12×12	24×24	16×16	32×32
Anger	84.4	80.0	88.9	84.4
Contempt	72.2	66.7	66.7	66.7
Disgust	93.2	93.2	94.9	96.6
Fear	88.0	80.0	76.0	72.0
Happiness	95.7	95.7	95.7	97.1
Sadness	64.3	64.3	67.9	60.7
Surprise	96.4	96.4	97.6	97.6
Overall	89.3	87.8	89.6	88.7

TABLE 12: The performance of HOG-TOP with various block sizes on the validation set of AFEW 4.0 database (%).

	12×12	24×24	16×16	32×32
Neutral	54.0	38.1	58.7	46.0
Anger	71.9	71.9	73.4	73.4
Disgust	20.0	15.0	22.5	20.0
Fear	6.50	15.2	4.30	13.0
Happiness	57.1	63.5	60.3	60.3
Sadness	4.90	9.80	4.90	9.80
Surprise	2.20	13.0	2.20	8.70
Overall	34.2	35.2	35.8	36.0

5 Conclusion

Video based facial expression recognition is a challenging and long standing problem. In this paper, we exploit the potentials of audiovisual modalities and propose an effective framework with multiple feature fusion to handle this problem. Both the visual modalities (face images) and audio modalities (speech) are utilized in our study. A new feature descriptor called Histogram of Oriented Gradients from Three Orthogonal Planes (HOG-TOP) is proposed to extract dynamic textures from video sequences to characterize facial appearance changes. Experiments conducted on three public databases (CK+, GEMEP-FERA 2011, AFEW4.0) have shown that HOG-TOP performs as well as a widely used feature LBP-TOP in representing dynamic textures from video sequences. Moreover, HOG-TOP is more effective to capture subtle facial appearance changes and robust in dealing with facial expression recognition in the wild. In addition, HOG-TOP is more compact. In order to capture facial configure changes, we introduce an effective geometric feature deriving from the warp transform of the facial landmarks. Realizing that voice is another powerful way for human beings to transmit message, we also explore the role of speech and employ the acoustic feature for affect recognition in video. We applied the multiple feature fusion to deal with facial expression recognition under lab-controlled environment and in the wild. Experiments conducted on two facial expression datasets, CK+ and AFEW 4.0, demonstrate that our approach can achieve a promising performance in facial expression recognition in video.

Acknowledgments

The work reported in this paper was partially supported by a research grant from the Natural Science Foundation of China (NSFC) grant (Project Code: 61473243).

References

[1] R. A. Calvo and S. D'Mello, "Affect Detection An Interdisciplinary Review of Models, Methods, and Their Applications," *IEEE Transactions on Affective Computing*, vol. 1, pp. 18-37, 2010.

[2] Y. l. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 97-115, 2001.

[3] K. Scherer and P. Ekman, "Handbook of Methods in Nonverbal Behavior Research," *UK: Cambridge Univ. Press*, 1982.

[4] J. F. Cohn and P. Ekman, "Measuring facial action," 2005.

[5] P. Ekman and W. V. Friesen, "Facial Action Coding System: A Technique for the Measurement of Facial Movement," *Consulting Psychologists Press*, 1978.

[6] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial Action Coding System: The Manual on CD ROM. A Human Face," 2002.

[7] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, pp. 169-200, 1992.

[8] S. Z. Li and A. K. Jain, "Handbook of face recognition," *springer*, 2011.

[9] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886-893.

[10] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 915-928, 2007.

[11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+)_A complete dataset for action unit and emotion-specified expression," *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 94-101.

[12] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, and S. Sridharan, "Person-independent facial expression detection using constrained local models," *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 2011, pp. 915-920.

[13] S. Taheri, P. Turaga, and R. Chellappa, "Towards view-invariant expression analysis using analytic shape manifolds," *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 2011, pp. 306-313.

[14] A. Saeed, A. Al Hamadi, R. Niese, and M. Elzobi, "Effective geometric features for human emotion recognition," *IEEE 11th International Conference on Signal Processing (ICSP)*, 2012, pp. 623-627.

[15] K. Sikka, T. Wu, J. Susskind, and M. Bartlett, "Exploring bag of words architectures in the facial expression domain," *in Computer Vision-ECCV Workshops and Demonstrations*, 2012, pp. 250-259.

[16] Y. Rahulamathavan, R. C. W. Phan, J. A. Chambers, and D. J. Parish, "Facial Expression Recognition in the Encrypted Domain Based on Local Fisher Discriminant Analysis," *IEEE Transactions on Affective Computing*, vol. 4, pp. 83-92, 2013.

[17] L. Zhang and D. Tjondronegoro, "Facial expression recognition using facial movement features," *IEEE Transactions on Affective Computing*, vol. 2, pp. 219-229, 2011.

[18] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Transactions on Affective Computing*, vol. 6, pp. 1-12, 2015.

[19] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 2011, pp. 921-926.

[20] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 2011, pp. 878-883.

[21] X. Huang, G. Zhao, M. Pietikainen, and W. Zheng, "Expression Recognition in Videos Using a Weighted Component-Based Feature Descriptor," in *Proceedings of the 17th Scandinavian conference on Image analysis*, 2011, pp. 569-578.

[22] T. R. Almaev and M. F. Valstar, "Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition," in *Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 356-361.

[23] X. Huang, Q. He, X. Hong, G. Zhao, and M. Pietikainen, "Improved Spatiotemporal Local Monogenic Binary Pattern for Emotion Recognition in The Wild," in *ACM International Conference on Multimodal Interaction*, 2014, pp. 514-520.

[24] X. Huang, G. Zhao, W. Zheng, and M. Pietikainen, "Spatio temporal Local Monogenic Binary Patterns for Facial Expression Recognition," *IEEE Signal Processing Letters*, vol. 19, pp. 243-246, 2012.

[25] F. Long, T. Wu, J. R. Movellan, M. S. Bartlett, and G. Littlewort, "Learning spatiotemporal features by using independent component analysis with application to facial expression recognition," *Neurocomputing*, vol. 93, pp. 126-132, 2012.

[26] S. W. Chew, R. Rana, P. Lucey, S. Lucey, and S. Sridharan, "Sparse Temporal Representations for Facial Expression Recognition," in *Advances in Image and Video Technology*, 2012, pp. 311-322.

[27] Y. Li, S. Wang, Y. Zhao, and Q. Ji, "Simultaneous Facial Feature Tracking and Facial Expression Recognition," *IEEE Transactions on Image Processing*, vol. 22, pp. 2559-2573, 2013.

[28] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Emotion Recognition in the Wild with Feature Fusion and Multiple Kernel Learning," in *ACM International Conference on Multimodal Interaction*, 2014, pp. 508-513.

[29] S. E. Kanou, C. Pal, X. Bouthillier, P. Froumenty, ?. Gl?ehre and R. Memisevic, et al., "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 543-550.

[30] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild," in *ACM International Conference on Multimodal Interaction*, 2014, pp. 494-501.

[31] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3687-3691.

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

[33] F. Ringeval, S. Amiriparian, F. Eyben, K. Scherer, and B. Schuller, "Emotion Recognition in the Wild Incorporating Voice and Lip Activity in Multimodal Decision-Level Fusion," in *ACM International Conference on Multimodal Interaction*, 2014, pp. 473-480.

[34] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion Recognition In The Wild Challenge 2014: Baseline, Data and Protocol," in *ACM International Conference on Multimodal Interaction*, 2014, pp. 461-466.

[35] B. Sun, L. Li, T. Zuo, Y. Chen, G. Zhou, and X. Wu, "Combining Multimodal Features with Hierarchical Classifier Fusion for Emotion Recognition in the Wild," in *ACM International Conference on Multimodal Interaction*, 2014, pp. 481-486.

[36] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 517-524.

[37] A. Martinez and S. Du, "A model of the perception of facial expressions of emotion by humans_ Research overview and perspectives," *The Journal of Machine Learning Research*, vol. 13, pp. 1589-1608, 2012.

[38] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Dynamic texture and geometry features for facial expression recognition in video," in *IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 4967-4971.

[39] T. Gritti, C. Shan, V. Jeanne, and R. Braspenning, "Local features based facial expression recognition with face registration errors," in *8th IEEE International Conference on Automatic Face & Gesture Recognition*, 2008, pp. 1-8.

[40] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 971-987, 2002.

[41] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, pp. 135-164, 2004.

[42] S. Meudt and F. Schwenker, "Enhanced Autocorrelation in Real World Emotion Recognition," in *ACM International Conference on Multimodal Interaction*, 2014, pp. 502-507.

[43] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers and C. A. Müller, et al., "The INTERSPEECH 2010 paralinguistic challenge," in *INTERSPEECH*, 2010, pp. 2794-2797.

[44] M. Gönen and E. Alpaydin, "Multiple Kernel Learning Algorithms," *The Journal of Machine Learning Research*, vol. 12, pp. 2211-2268, 2011.

[45] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the Kernel Matrix with Semi-Definite Programming," *The Journal of Machine Learning Research*, vol. 5, pp. 27-72, 2004.

[46] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491-2521, 2008.

[47] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)* ,vol. 2, 2011.

[48] Abhinav Dhall, Roland Goecke, Simon Lucey, and T. Gedeon, "A semi-automatic method for collecting richly labelled large facial expression databases from movies," *IEEE Multimedia*, 2012.

[49] X. Zhu and D. Ramanan, "Face detection, pose estimation and landmark localization in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition* , 2012, pp. 2879-2886.

[50] F. Eyben, M. Wollmer, and B. Schuller, "OpenEAR - Introducing the munich open-source emotion and affect recognition toolkit," in *3rd International Conference onAffective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1-6.

[51] Florian Eyben, Martin Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459-1462.

[52] H. Kaya and A. A. Salah, "Combining Modality-Specific Extreme Learning Machines for Emotion Recognition in the Wild," in *ACM International Conference on Multimodal Interaction*, 2014, pp. 487-493.

Junkai Chen received his Bachelor and Master degrees from Northwestern Polytechnic University and Xi'an Jiaotong University in 2010 and 2013, respectively. He is currently a Ph.D. candidate with Department of Electronic and Information Engineering, The Hong Kong Polytechnic University. His research interests include pattern recognition, computer vision and machine learning.

Zenghai Chen received his B.Eng. in Department of Electronics and Communication Engineering from Sun Yat-set University, Guangzhou in 2009, and Ph.D. degree in Department of Electronic and Information Engineering from The Hong Kong Polytechnic University, Hong Kong in 2014. He is now a postdoctoral fellow in School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include pattern recognition, computer vision and machine learning.

Zheru Chi received his B.Eng. and M.Eng. degrees from Zhejiang University in 1982 and 1985, respectively, and his Ph.D. degree from the University of Sydney in March 1994, all in electrical engineering. Between 1985 and 1989, he was on the Faculty of the Department of Scientific Instruments at Zhejiang University. He worked as a Senior Research Assistant/Research Fellow in the Laboratory for Imaging Science and Engineering at the University of Sydney from April 1993 to January 1995. Since February 1995, he has been with the Hong Kong Polytechnic University, where he is now an Associate Professor in the Department of Electronic and Information Engineering. Since 1997, he has served on the organization or program committees for a number of international conferences. He was an associate editor for IEEE Transactions on Fuzzy Systems between 2008 and 2010, and is currently an editor for International Journal of Information Acquisition. His research interests include image processing, pattern recognition, and computational intelligence. Dr Chi has authored/co-authored one book and 11 book chapters, and published more than 190 technical papers.

Hong Fu received her Bachelor and Master degrees from Xian Jiao tong University in 2000 and 2003, and Ph.D. degree from the Hong Kong Polytechnic University in 2007. She is now an Associate Professor in Department of Computer Science, Chu Hai College of Higher Education, Hong Kong. Her research interests include eye tracking, computer vision, pattern recognition, and artificial intelligence.