Pattern Recognition 48 (2015) 3407–3416

Contents lists available at ScienceDirect

Pattern Recognition

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

A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences

Xijian Fan, Tardi Tjahjadi *

School of Engineering, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, United Kingdom

ARTICLE INFO

Article history:
Received 13 October 2014
Received in revised form
31 March 2015
Accepted 27 April 2015
Available online 8 May 2015

Keywords:
Histogram of gradients
Facial expression
Optical flow
Feature extraction

ABSTRACT

Facial expression causes different parts of the facial region to change over time and thus dynamic descriptors are inherently more suitable than static descriptors for recognising facial expressions. In this paper, we extend the spatial pyramid histogram of gradients to spatio-temporal domain to give 3-dimensional facial features and integrate them with dense optical flow to give a spatio-temporal descriptor which extracts both the spatial and dynamic motion information of facial expressions. A multi-class support vector machine based classifier with one-to-one strategy is used to recognise facial expressions. Experiments on the CK+ and MMI datasets using leave-one-out cross validation scheme demonstrate that the integrated framework achieves a better performance than using individual descriptor separately. Compared with six state of the art methods, the proposed framework demonstrates a superior performance.

1. Introduction

The automated recognition of facial expressions has been a widely researched topic in recent years due to its wide range of applications such as surveillance, human–computer interaction and data-driven animation. Other motivations include advancements in related research in face detection [1], tracking and recognition [2], as well as new developments in feature extraction algorithms and machine learning [3]. Six prototypical facial expressions were first formalised in [4], namely anger, disgust, fear, happiness, sadness and surprise. Although much progress has been made since then, accurate recognition of facial expressions is still a challenging problem due to the subtlety, complexity and variability of facial expressions [5,6].

Most existing works on facial expression recognition focus on analysing and extracting facial features in a single image or one frame in an image sequence, i.e., recognition of static expression. Previous methods have mainly concentrated on attempting to capture expressions through either action units [5,7] or via discrete frame extraction techniques [8]. All of these methods require either manual selection of facial features in order to determine where the particular changes in the facial region occur, or the subjective thresholding for feature selection. This means that any classification is highly dependent on subjective information in the form of a threshold or other a priori knowledge.

A facial expression involves a dynamic process, and the dynamic information such as the movement of facial landmarks and the change in facial shape contains useful information that can represent a facial expression more effectively. Thus, it is important to capture such dynamic information so as to recognise facial expressions over the entire video sequence. Previous recognition methods on video sequences tend to only focus on the movement of facial landmarks, not analysing the variation of facial shape. In this paper, we utilise two types of dynamic information to enhance the recognition: a novel spatio-temporal descriptor based on the pyramid histogram of gradients (PHOG) [9] to represent changes in facial shape, and dense optical flow to estimate the movement (displacement) of facial landmarks. We view an image sequence as a spatio-temporal volume, and use temporal information to represent the dynamic movement of facial landmarks associated with a facial expression. In this context, we extend PHOG descriptor representing spatial local shape to spatio-temporal domain to capture the changes in local shape of facial sub-regions in the temporal dimension to give 3-dimensional (3D) facial component sub-regions of forehead, mouth, eyebrow and nose. We refer this descriptor as PHOG_three orthogonal planes (PHOG_TOP). By combining PHOG_TOP and dense optical flow of the facial region, we exploit the fusion of discriminant features for classifying and thus recognising facial expressions.

The main contributions of this paper are: (a) a framework that integrates the dynamic information extracted from variation in facial

* Corresponding author. Tel.: +44 24 76523126; fax: +44 24 76418922.
E-mail address: t.tjahjadi@warwick.ac.uk (T. Tjahjadi).

http://dx.doi.org/10.1016/j.patcog.2015.04.025
0031-3203/© 2015 Elsevier Ltd. All rights reserved.

shape and movement of facial landmarks, (b) PHOG_TOP 3D facial features, (c) a means of fusing weighted PHOG_TOP with dense optical flow, and (d) an analysis on the contribution of different facial subregions using the proposed framework.

This paper is organised as follows. Previous related work is presented in Section 2. Section 3 presents PHOG_TOP, the dense optical flow descriptor, and the fusion of these descriptors. The proposed facial expression recognition framework and the experimental results are, respectively, presented in Sections 4 and 5. Finally, Section 6 concludes the paper.

2. Related work

There are two main approaches to recognising facial expressions: (1) recognition based on facial action coding system (FACS) [10] action units (AUs) and (2) direct content-based recognition (non-AU). The weakness of the AU based approach is that errors in the AU classification affect the recognition rate. Thus, the framework proposed in this paper adopts the non-AU approach.

A typical facial expression recognition system comprises three modules: image pre-processing, facial feature extraction, and facial expression classification. The feature extraction module is important and thus numerous methods for facial features extraction have been proposed. These methods can either be appearance-based or geometric-based methods. The features extracted using either approach should minimise intra-class variation of facial expressions, while maximising inter-class variations.

In the appearance-based approach, transformations and statistical methods are used to determine the feature vectors that represent textures and are thus simple to implement. Gabor wavelets [11] and local binary patterns (LBPs) [12] are two representative feature vectors of such approach that describe the local appearance models of facial expressions. Gabor magnitudes are commonly adopted as features as they are robust to misalignment of corresponding image features. However, computing Gabor filters has a high computational cost, and the dimensionality of the output can be large, especially if they are applied to a wide range of frequencies, scales and orientations of the image features. The LBP descriptor is a histogram where each bin corresponds to one of the different possible binary patterns representing a facial feature, resulting in a 256-dimensional descriptor. However, it has been shown that some of the patterns are more prone to encoding noise. The most popular LBP is the uniform LBP [13]. Zhao and Pietikainen [14] proposed a method which extends LBP to spatio-temporal domain so as to utilise the dynamic information, which results in a significant improvement in the recognition rate. One drawback of appearance-based approach is that it is difficult to generalise appearance features across different persons.

In the geometric-based approach, shape and position information of facial landmarks or region are extracted to represent the face geometry [11,15–17]. For example, the method in [11] uses the geometric position of 34 fiducial points as facial features to represent the face geometry. In image sequences, optical flow analysis has been applied to detect the movements of facial components which are qualified by measuring the geometric displacement of facial feature points between two consecutive frames [18,19]. Although geometric-based methods are sensitive to noise, and require accurate tracking of facial features, geometric features alone can provide sufficient information for efficient recognition of facial expressions.

Histogram of gradients (HOG) [20] was originally developed for person detection and object recognition. In [21], HOG descriptors are extracted from face image using a dense grid, and are used for face recognition. The PHOG proposed in [9] is an extension of HOG and is used to represent the local shape of facial region. However, all these methods only analyse individual frames of a video sequence, i.e., not taking the dynamics of a facial expression into account.

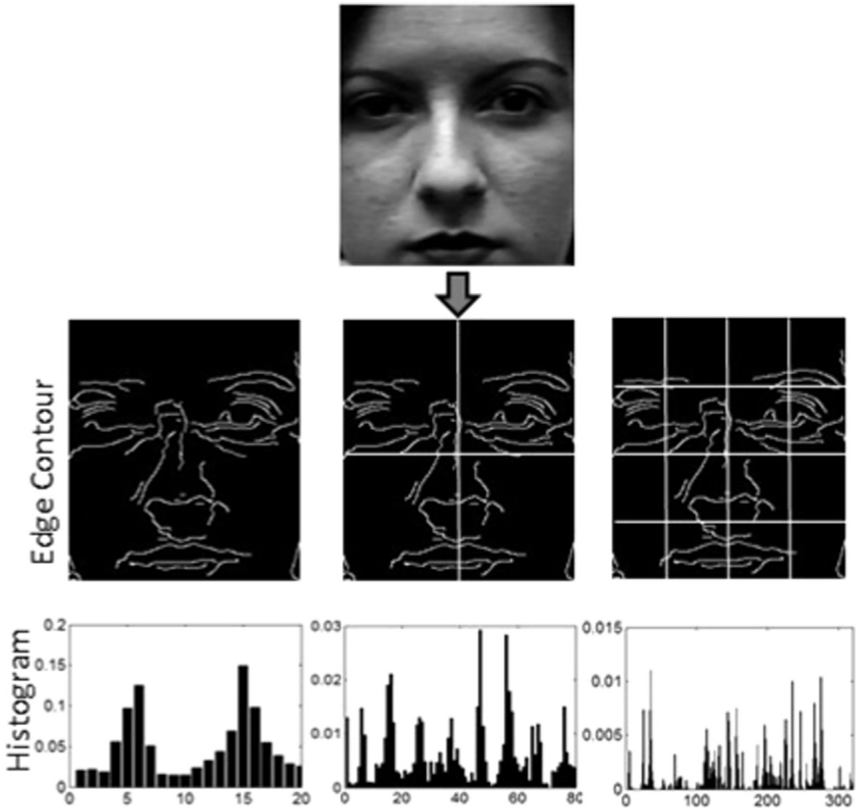

Fig. 1. PHOG descriptor of a face.

Many techniques have been proposed for classification of facial expressions. Support vector machines (SVM) [7] classifier has been shown to be effective in recognising different expressions. Though new kernels have been presented, the linear, polynomial and radial basis function (RBF) kernels are the most widely used.

3. Dynamic features

Our framework uses dynamic features. This takes the form of PHOG_TOP and dense optical flow, combined to give a robust and accurate recognition of facial expressions.

3.1. PHOG_TOP descriptor

PHOG [9], originally designed for object classification, contains the local shape information of an image and spatial layout of this shape. This descriptor was inspired by HOG [20] and the image pyramid representation [22]. In essence, the PHOG is a descriptor based on edge information. More specifically, edge contours of an image are extracted at different pyramid resolution level, and occurrences of gradient orientation of edges are counted to construct a gradient histogram. The PHOG descriptor is obtained by concatenating the histograms from selected pyramid levels. An example of a PHOG descriptor of a face is shown in Fig. 1.

Facial expression is usually performed dynamically, thus its dynamic information is essential for its recognition. Motivated by the temporal extension of LBP [12], we propose a spatial-temporal descriptor by concatenating the PHOG of three orthogonal planes *XY*, *XT* and *YT* to give PHOG_TOP, taking into account the co-occurrence statistics in these three planes. The *XY* plane is used to extract the local spatial information, and the *XT* and *YT* planes are used to extract temporal information. We consider a video sequence as a stack of *XY* slices in the temporal dimension, and similarly for *XT* and *YT* slices but in the *Y* and *X* dimensions, respectively. The spatio-temporal PHOG over each slice in three orthogonal axes (i.e., XY_PHOG, XT_PHOG, and YT_PHOG) are separately obtained and then combined. Take XY_PHOG for instance, we compute PHOG descriptor in every single image from a video sequence, i.e., along the temporal axis. First, the Canny edge detector is employed to capture the edge information. The image region is divided into a set of spatial grid by repeatedly doubling the number of divisions along each axis. Thus, the grid at resolution level l has 2^l cells along each dimension.

The orientation of gradient for each grid at each resolution level are computed using a Sobel mask without Gaussian smoothing, as the smoothing decreases the performance of classification [9]. The histogram of edge orientations within an image sub-region is quantised into *K* bins, where *K* is set to 20 as in [9]. In order to reflect the contribution of each edge, a weight proportional to its magnitude is added. Each bin in the histogram represents the occurrences of edges that have orientations within a certain angular range. We use the range [0, 360] to take into account all orientations. The PHOG descriptor for a slice (image) is obtained by concatenating all the vectors at each pyramid resolution. The final PHOG_XY is obtained by averaging the PHOG features over all slices in the temporal dimension. The creation of the PHOG descriptor is illustrated in Fig. 2.

The PHOG is normalised to sum to unity. Consequently, level 0 is represented by a *K* vector corresponding to *K* bins of the histogram, level 1 by a 4*K*-vector, and the normalised PHOG_XY descriptor of the entire sequence is the vector

$$\text{PHOG\_}\mathbf{XY}_{\text{Seq}} = \frac{\sum_k \text{PHOG\_}\mathbf{XY}_k}{k} \tag{1}$$

with dimensionality

$$\text{Dim}_{xy} = K \sum_{l \in L} 4^l, \tag{2}$$

where *L* denotes the number of levels, which is set to 2 to prevent over fitting of the edge contours over the grid. PHOG_TOP is a concatenation of the descriptors of three planes (PHOG_XY$^{\text{Seq}}$, PHOG_XT$^{\text{Seq}}$, and PHOG_YT$^{\text{Seq}}$, resulting in

$$\text{PHOG\_}\mathbf{TOP} = \{\text{PHOG\_}\mathbf{YT}_{\text{Seq}}, \text{PHOG\_}\mathbf{XT}_{\text{Seq}}, \text{PHOG\_}\mathbf{XY}_{\text{Seq}}\} \tag{3}$$

of dimensionality $\text{Dim}_{xy} + \text{Dim}_{xt} + \text{Dim}_{yt}$.

Similar to [23], we segment an image sequence into a number of 3D facial component sub-regions (forehead, mouth, eyebrow and nose) to enhance the spatial information. However, unlike in [23] we do not subdivide the video sequence in the temporal dimension. This is because the length of the video sequences (as used in our experiments which start with the neutral expression and end with the peak phase, i.e., with the most significant motion) extracted from extended CK dataset (CK+) [24] and MMI dataset [29] is relatively short. Fig. 3 shows the four facial sub-regions, and the 3D sub-region of mouth in video sequences. PHOG_TOP is applied to the entire face video sequence and four different sub-regions.

3.2. Dense optical flow descriptor

Optical flow captures the dynamic information in a video sequence. For our framework, optical flow is used to track facial points in a video sequence and compute the displacement which represents the movement of corresponding points between two consecutive frames. Instead of tracking some facial fiducial points (landmarks), we track dense facial points uniformly distributed on a grid placed on the central facial region. The merit of the dense optical flow is that the points on the grid are tracked as one entity. As a result, the global displacement of these considered points is obtained, which is used to generate the feature vector.

The Lucas–Kanade optical flow algorithm [25] has been used to estimate the displacement of facial feature points. It is a gradient-based method for motion estimation, and approximates the motion between two consecutive frames. Given two consecutive frames I_{t-1} and I_t, for a point $p = (x, y)^T$ in I_{t-1}, if the optical flow is $d = (u, v)^T$ then the corresponding point in I_t is $p + d$, where *T* is the transpose operator. The algorithm finds *d* which minimises the match error between the local appearances of two corresponding points. A cost function $e(d)$ is defined for the local area $R\ (p)$, i.e.,

$$e(d) = \sum_{x \in R(p)} w(x)(I_t(x+d) - I_{t-1}(x))^2, \tag{4}$$

where $w(x)$ is a weights window, which assigns larger weight to pixels that are closer to the central pixel as these pixels give more importance than others. Optimising (4) gives the solution

$$d = G^{-1}H \tag{5}$$

where

$$G = \sum_{x \in N(p)} w(x) \nabla I_t (\nabla I_t)^T \tag{6}$$

$$H = \sum_{x \in N(p)} w(x) \nabla I_t \Delta \tag{7}$$

$$\Delta I = I_{t-1} - I_t \tag{8}$$

$$\nabla I_t = \frac{dI_t}{dx}. \tag{9}$$

The efficiency of computing the dense optical flow depends on the grid size. Since a grid with small cells leads to high computation, we

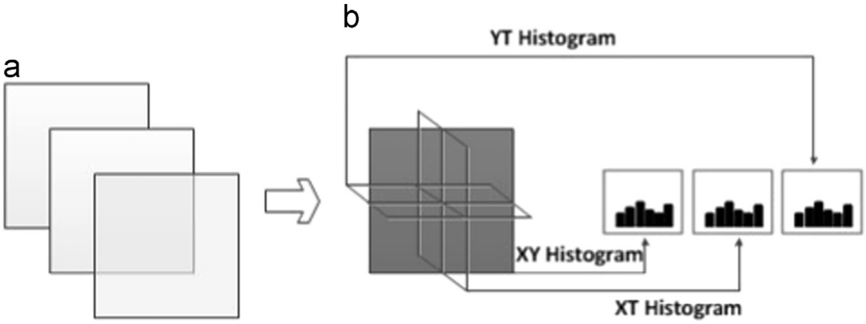

Fig. 2. Three planes in spatio-temporal domain for extracting TOP features, and the histogram concatenated from three planes: (a) original image and (b) the x–y, y–t, and x–t planes, and the concatenation of resulting histograms into a single feature set.

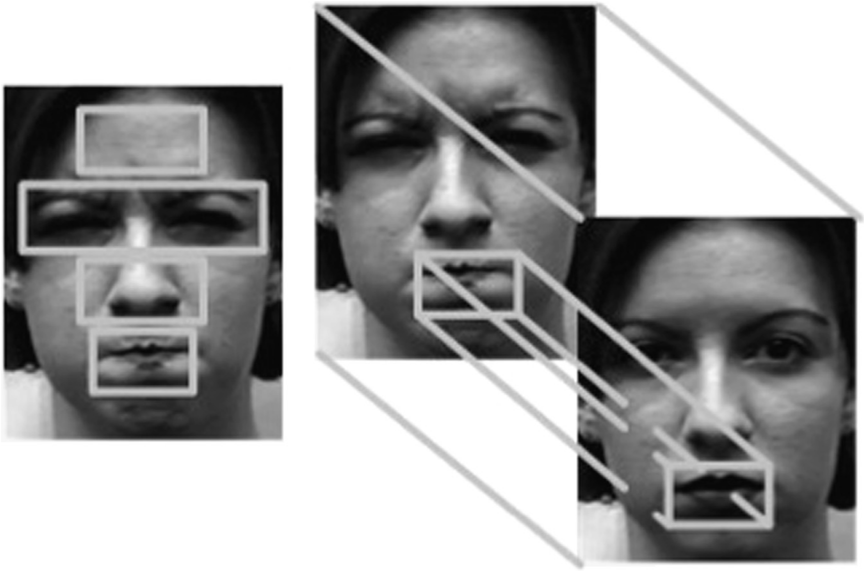

Fig. 3. (Left) Four facial sub-regions and (right) face video sequence with 3D mouth sub-region.

use a grid of 20 × 15 placed on the central face region, and its vertices are the facial points to be tracked. The displacement vector of two consecutive frames is then computed on these 300 points, which reduces the computation significantly. The right image of Fig. 4 shows the trajectories of tracked points from the previous frame of a neutral facial expression.

The corresponding inter-frame displacement vectors are added to obtain the global displacement vectors corresponding to each point during an expression (i.e., from a neutral face to the peak phase of the expression). More formally, let $(p_1, \ldots, p_S)$ be the respective spatial image positions of a facial feature point (i.e., one of the grid vertices) p at frames $1, \ldots, S$, where S is the number of frames in the video sequence. The global displacement vector of point p is

$$v_p = \sum_{i=1}^{S-1} v_{pi} = \sum_{i=1}^{S-1} (p_{i+1} - p_i), \tag{10}$$

where

$$\| v_p \| = \sqrt{p_x^2 + p_y^2}, \quad \theta_p = \tan^{-1}\left(\frac{p_y}{p_x}\right), \tag{11}$$

and p_x and p_y are, respectively, the x and y components of v_p.

The normalisation of the dense optical flow is slightly different from that of PHOG_TOP, where we do not detect the landmarks over the entire video sequence. Instead, we only locate the landmarks of the first frame of each sequence, and introduce a reference vector for normalisation. More specifically, the two inner eye corner points are used to define a reference vector $\overline{nv}$ and its angle θ with the horizontal axis $\overline{w} = (1, 0)^T$, i.e.,

$$\| \overline{nv} \| = \sqrt{nv_x^2 + nv_y^2} \tag{12}$$

$$\theta = \cos^{-1}\left(\frac{\overline{nv} \cdot \overline{w}}{\| \overline{nv} \| \, \| \overline{w} \|}\right). \tag{13}$$

These are used to normalise the global displacement vectors and to correct the orientation angles of these vectors. Finally, the global displacement of any feature point is

$$d_{\text{global}} = \frac{\| v_p \|}{\| \overline{nv} \|}. \tag{14}$$

Its normalised angle is computed by adding it to $\theta_{\overline{p}}$ (i.e., the orientation of vector $\overline{p}$ is the positive or negative value of θ) to give

$$\theta_{\text{global}} = \theta + \theta_{\overline{p}}. \tag{15}$$

Two normalised features are thus obtained for each of the tracked feature points.

A global feature vector v_{global} of 600 components, i.e., the dense optical flow, is then created for each video sequence, containing the pair of displacement features of the points, i.e.,

$$v_{\text{global}} = \left[\| p_1 \|, \theta_{p_1}, \| p_2 \|, \theta_{p_2}, \ldots, \| p_R \|, \theta_{p_R} \right], \tag{16}$$

where $\| p_i \|$ and θ_{p_i}, respectively, denote the normalised modulus and the normalised angle of the vector p_i of accumulated displacement of a given feature point, and R denotes the number of tracked points.

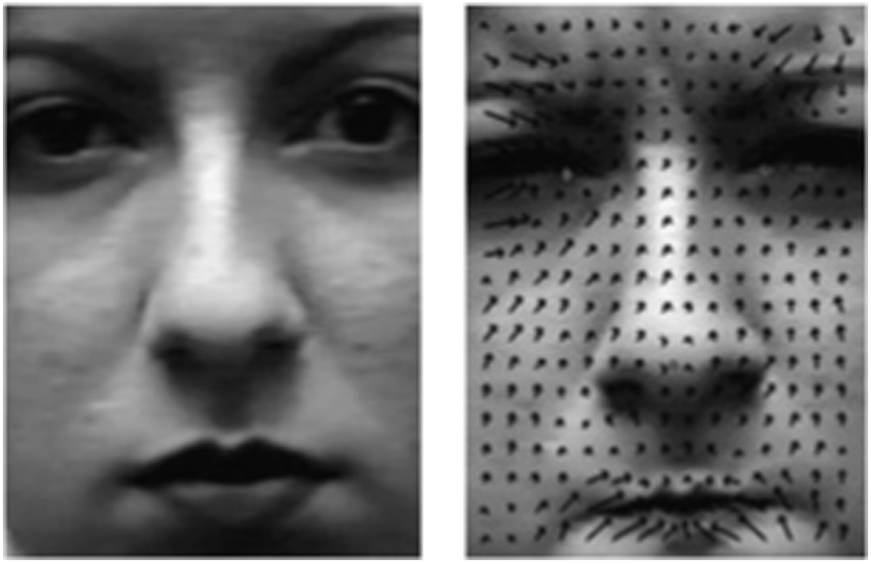

Fig. 4. (Left) A neutral image and (right) with the dense optical flow superimposed.

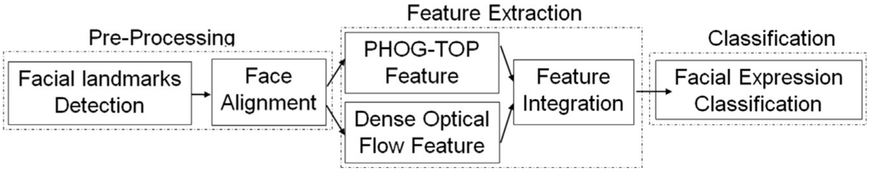

Fig. 5. Framework for the facial expression recognition.

3.3. Integration of descriptors

The integration of the normalised descriptors of a video sequence is

$$\begin{aligned} f_{Int} &= \{f_{motion}, f_{spatial}\} \\ &= \{\omega_1 \cdot \mathrm{PHOG\_}\mathbf{YT}_{\mathrm{Seq}},\ \omega_2 \cdot \mathrm{PHOG\_}\mathbf{XT}_{\mathrm{Seq}}, \\ &\quad \omega_3 \cdot v_{\mathrm{global}},\ \omega_4 \cdot \mathrm{PHOG\_XY}_{\mathrm{Seq}}\}, \end{aligned} \tag{17}$$

where ω_i denote the weights for the different descriptors and represent the contribution of the descriptors to the integrated descriptor. Note that if every weight is set to 1 then the integration is simply the concatenation of all descriptors. The analysis of the selection of weights is presented in Section 5.

The set of the integrated feature vectors, corresponding to all considered video sequences, is divided into two disjoint sets: training and test.

4. Framework for facial expression recognition

The proposed framework comprises three phases: pre-processing, feature extraction and classification as shown in Fig. 5. The pre-processing includes facial landmark detection and face alignment, where face alignment is applied to reduce the effect of variation in head pose and scene illumination to give a better recognition performance.

The local evidence aggregated regression (LEAR) [26] is employed to detect facial landmarks over every frame of a video sequence, and the locations of the detected eyes are then used to align any in-plane rotation, where the angle of two eyes in each frame is rotated at their centre to line up to the horizontal axis. The two eyes and nose tip are used to scale and crop the image into a 160 × 240 rectangular region of interest containing the central face region. In the cropped image, the x coordinate of the centre of two eyes are the centre in the horizontal direction, while the y coordinate of the nose tip locates the lower third in the vertical direction.

The feature extraction phase includes the generation of PHOG_-TOP and v_{global}. The details are presented in Section 3.

SVM have been widely applied to facial expression recognition due to its following properties: (1) its ability to work with high-dimensional data and (2) a high generalisation performance without the need of a priori knowledge, even when the dimension of the input space is very large. The linear, polynomial, and RBF kernels have been shown in [27] to achieve good performance in facial expression recognition. Thus, to achieve an effective classification of facial expressions, we apply the classical SVM classifier with the RBF kernel. In this paper, we use grid-search and 10-fold cross-validation [28] to estimate the kernel parameter. The parameter that achieves the best cross-validation accuracy is selected.

SVM are a binary classifier, however, the classification of facial expressions is a multiclass classification problem. Thus, the binary SVM classifiers need to be combined for recognition of multiple classes [28]. Two strategies are commonly used: one-versus-one and one-versus-all. In the one-versus-all strategy, the classifier with the highest output assigns the class. In the one-versus-one case, every classifier assigns test data to one of the two classes, the vote for the assigned class is increased by one vote, and finally the class with the most votes determines the classification. One-versus-one strategy is used in our framework due to the simplicity of its implementation and its robustness in classification as follows. Suppose C classes are classified using binary classifiers, $C(C-1)/2$ binary classifiers are built from all pairs of distinct classes. Sometimes, more than one expressions (i.e., classes) obtain the most number of votes. In this case the features extracted from image sequences of one of the two classes, say class m, are averaged to obtain a template representing that class:

$$T_m = \frac{\Sigma_i^M f_{i,m}}{M}, \tag{18}$$

where $f_{i,m}$ denotes the ith feature belonging to class m. During the classification, a simple nearest neighbour classifier is used and the feature belonging to the test data h_1 is classified as the nearest class template, i.e.,

$$T_m : D(h_1, T_m) < D(h_1, T_n), \tag{19}$$

where $D(.,.)$ is distance function, and m and n are the indices of the two classes with the most votes.

5. Experiments

5.1. Facial expression datasets

The extended CK dataset (CK+) [24] is the most widely used data for evaluating facial expression recognition methods, and is publicly available. This dataset contains 593 image sequences of seven basic facial expressions (namely, anger, contempt, disgust, fear, happiness, sadness and surprise). These expressions were performed by 120 subjects. The age of the participants ranges from 18 to 30 years, 65% of them are female, 81% are Euro-American, 13% are Afro-American, and 6% of other racial groups. Each frame of the image sequences is 640 × 480 or 640 × 490 pixels with an 8-bit greyscale. The video sequences vary in duration (i.e., 10–60 frames) and incorporate the onset (i.e., the neutral frame) to peak phase of the facial expression. We used 327 image sequences of seven expressions, where we replaced the expression neutral with Contempt. The top row of Fig. 6 shows sample images of a subject expressing six expressions. Table 1 shows the occurrences of the various expression classes in CK+ dataset.

The MMI dataset [29] is another well-known dataset, which comprises video sequences including both posed and spontaneous expressions. These expressions were performed by 19 subjects (44% female), with age ranging from 19 to 62, and of European, Asian and South American ethnicity. The subjects performed 79 expressions including the six basic facial expressions with neutral frame at the start of each sequence. Every video frame is at 720 × 576 spatial resolution. We converted the original frames into 8-bit greyscale images for our experiments, and extracted the sub-sequence from the neutral frame to the peak phase. In our experiments, 203 image sequences labelled as one of the six basic facial expressions are selected from the MMI dataset. The bottom row of Fig. 6 shows sample images of the six basic expressions.

Since the MMI dataset is generated in a different way to the CK+ dataset, and may contain larger pose variation, we use slightly different pre-processing to align the data. For dense optical flow, the nose tip is used to detect and subtract the effect of head motion. For PHOG_TOP, since the sequences selected for our experiments are frontal view with no out-of-plane head pose which can cause self-occlusions of the eyes region (which are used for our alignment process), we applied the same pre-processing method as for CK+ dataset.

5.2. Experimental results

Three sets of experiments are performed on the CK+ dataset (with smaller head motion than in MMI dataset) to evaluate the performance of the proposed framework and to compare its performance with two state of the art facial expression recognition methods. Leave-sequence-out cross-validation scheme is employed as follows. One video sequence is selected for testing and the remaining video sequences are used for training, which guarantees that a sequence selected for testing is not in the training set and consequently sequence independence is realised in our experiments.

The first set of experiments investigates whether spatio-temporal PHOG performs better than using only one histogram from either *XY*, *YT* or *XT* plane. Cross-validation was applied 327 times on the selected 327 video sequences. The recognition rate of classifying all expressions using PHOG from each individual plane, i.e., PHOG_XY $_{\mathrm{Seq}}$, PHOG_XT$^{\mathrm{Seq}}$ and PHOG_YT$^{\mathrm{Seq}}$), and their combination, i.e., PHOG_TOP, is summarised in Fig. 7.

Fig. 7 shows that the recognition rate using the non-weighted (i.e., with $\omega_{\mathrm{i}}=1$ for $i=1, 2, 3, 4$) combination of features is the highest. Also, using the features from *YT* plane gives the better performance than from the other two planes, and the features which represent the variation in shape horizontally from the *XT* plane give the lowest rate. This means: (1) the dynamic motion features (i.e., feature from the YT plane) plays more important role than spatial features in recognising facial expressions and (2) the vertical variations in shape (e.g, the opening of mouth) are more significant than horizontal variations.

We combine the features extracted from the three planes on the assumption that all components have equal contribution in the recognition. However, from the aforementioned analysis, not all

Table 1
Number of image sequences (subjects) for each expression in the CK+ dataset.

Expression	CK+	MMI
Anger	45	32
Disgust	59	28
Fear	25	28
Happiness	69	42
Sadness	28	32
Surprise	83	41
Contempt	18	0
Total	327	203

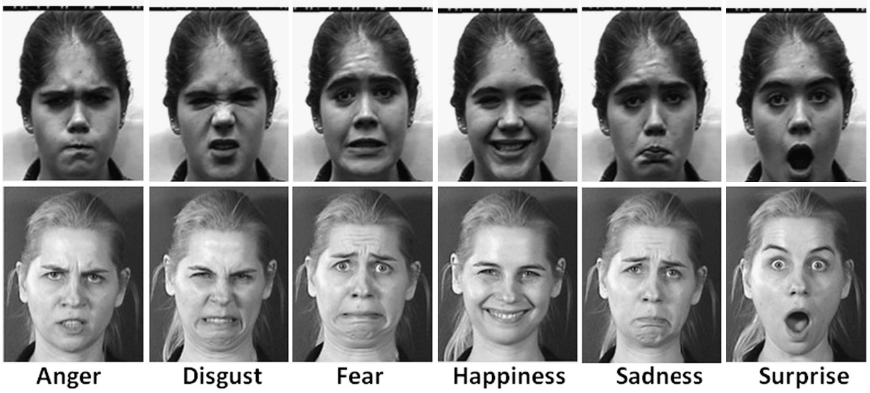

Fig. 6. Example of expressions from CK+ dataset (top row) and MMI dataset (bottom row). Each image is the frame with the most expressive face in a video sequence.

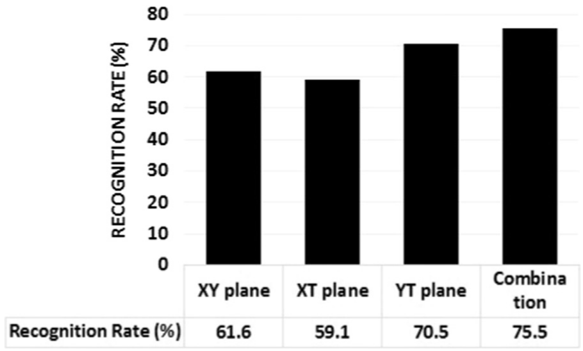

Fig. 7. Recognition rates of all expressions PHOG from either *XY*, *XT*, *YT* or their combination, i.e., PHOG_TOP.

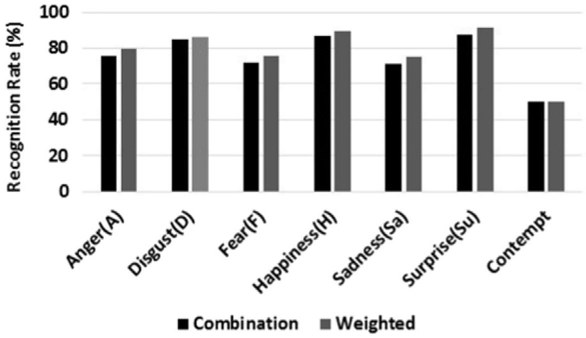

Fig. 8. Recognition rates of all expressions in using combination of non-weighted and weighted PHOG_TOP.

Table 2
Multiclass SVM results of PHOG_TOP from the whole face on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation.

Expression	A	D	F	H	Sa	Su	C
Anger (A)	**36**	**1**	**3**	**0**	**3**	**0**	**2**
Disgust (D)	**5**	**51**	**0**	**0**	**0**	**2**	**1**
Fear (F)	**2**	**0**	**19**	**0**	**3**	**1**	**0**
Happiness (H)	**1**	**1**	**2**	**62**	**0**	**1**	**2**
Sadness (Sa)	**1**	**0**	**3**	**1**	**21**	**1**	**1**
Surprise (Su)	**1**	**0**	**2**	**2**	**1**	**76**	**1**
Contempt (Su)	**3**	**3**	**0**	**2**	**1**	**0**	**9**

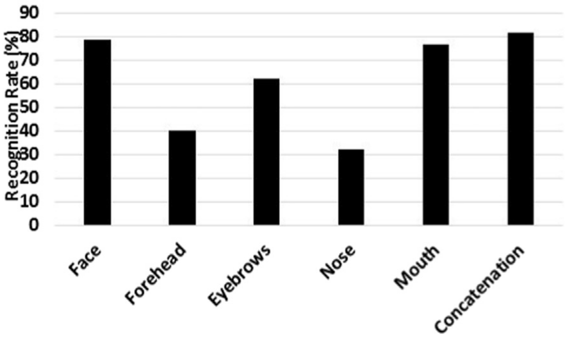

Fig. 9. Discriminant power of facial sub-regions in recognising six expressions using PHOG_TOP.

features are equally important, and some of them contain more useful information than others. A weighting strategy is thus introduced to improve the performance in recognition.

In order to determine the appropriate weights, the recognition rates achieved by using features from the three planes separately are analysed so that the features that give better recognition are identified and are allocated bigger weight, i.e.,

$$f_{weighted} = \omega_i f_i \tag{20}$$

where ω_i and f_i are the weight of the ith plane and PHOG extracted from the ith plane, respectively.

The weighting strategy is as follows. First, given the three recognition rates achieved using features from the three planes separately, we obtain $R = [R_1, R_2, R_3]$ from the output of the SVM classifier, where the lower the rate the smaller contribution the feature has. The weight vector is

$$\omega_i = \frac{\| R_i \|}{\| R \|}, \tag{21}$$

where $\| \cdot \|$ is L_2 norm, R_i denotes the average recognition rate using features from the ith plane, and $R = [R_1, R_2, R_3]$.

The recognition rates achieved using combination of non-weighted and weighted features are shown in Fig. 8. The use of the combination of weighted features resulted in better performance for sadness and surprise, and similar performance for disgust and contempt. This is because the variations in shape in sadness and surprise are more significant.

The second set of experiments evaluates the discriminant power of the different facial regions (i.e., how discriminative different facial regions are) and determines how much the PHOG_TOP extracted from different facial regions contribute to the six expressions and contempt. Four facial sub-regions (namely eyebrows, forehead, nose and mouth) are extracted and are used to compute the PHOG_TOP. The concatenation of these facial sub-regions is also taken into account. Table 2 shows that the classification of PHOG_TOP from the whole facial region achieves good recognition rate (i.e., $\geq 85\%$) for disgust, happiness and surprise, but less for fear and sadness. This is due: (a) the number of samples for fear and sadness are relatively small, (b) happiness and surprise were performed with much more distinguishable movement of facial landmarks, and (c) fear and sadness were poorly performed that it is difficult even for human to distinguish them. Fig. 9 shows the average classification of six expressions using PHOG_TOP for the considered sub-regions and their concatenation. The results show that among the sub-regions the mouth provides the best recognition rate. This is due to the more-differentiated movement of the mouth muscles (and consequently the corresponding facial points) for each facial expression. Fig. 9 also shows the concatenation of the four sub-regions provides better recognition than using the whole face. This is because the local shape information in both spatial and temporal domains is enhanced, while the other information is removed.

The third set of experiments evaluates the effectiveness of combining spatial shape information with dynamic motion features. Tables 3–5, respectively, show the results obtained in using dense optical flow, PHOG_TOP and the proposed framework with multi-class SVM classifier, including the overall classification performance of multi-class SVM, where the dark grey shade indicates the correct recognition result of each emotion while the lighter grey shade indicates the recognition result of each emotion misclassified with the maximal error. The three tables show that the combination of PHOG_TOP and dense optical flow feature is more accurate than using individual features separately in recognising facial expressions.

It is difficult to make a quantitative comparison between the state of the art facial expression recognition methods due to the different pre-processing and experiment strategies involved. Since the methods of Eskil and Benli [30] and Lucey et al. [24] were evaluated on the CK+ dataset using leave-one-out cross-strategy,

Table 3
Multiclass SVM results in using dense flow optical flow on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation. Dark grey shade – correct recognition result of each emotion, and lighter grey shade – recognition result of each emotion misclassified with the maximal error.

Expression	A	D	F	H	Sa	Su	C	%
A	33	4	**0**	**3**	4	**0**	**1**	**73.3**
D	4	50	**1**	**1**	**0**	**0**	**3**	**84.8**
F	**3**	**1**	9	5	**2**	**3**	**2**	**36.0**
H	**1**	**0**	2	63	**1**	**0**	2	**91.3**
Sa	4	**0**	**2**	**2**	14	**3**	**3**	**50.0**
Su	**1**	**1**	**0**	**1**	7	72	**1**	**86.8**
C	6	**0**	**2**	**1**	**1**	**0**	8	**44.4**

Table 4
Multiclass SVM results in using PHOG_TOP on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation. Dark grey shade – correct recognition result of each emotion, and lighter grey shade – recognition result of each emotion misclassified with the maximal error.

Expression	A	D	F	H	Sa	Su	C	%
A	38	**1**	3	**0**	**2**	**0**	**1**	**84.4**
D	4	53	**0**	**0**	**0**	**1**	**1**	**89.8**
F	**1**	**0**	21	**0**	3	**0**	**0**	**84.0**
H	**0**	**1**	2	64	**0**	**0**	2	**92.8**
Sa	**2**	**0**	3	**0**	21	**1**	**1**	**75.0**
Su	**0**	**0**	**1**	2	**1**	79	**0**	**95.2**
C	4	**2**	**0**	**2**	**1**	**0**	9	**50.0**

Table 5
Multiclass SVM results in using the proposed framework on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation. Dark grey shade – correct recognition result of each emotion, and lighter grey shade – recognition result of each emotion misclassified with the maximal error.

Expression	A	D	F	H	Sa	Su	C	%
A	40	**0**	**1**	**0**	3	**0**	**1**	**88.9**
D	2	55	**1**	**0**	**0**	**0**	**1**	**93.2**
F	**0**	**1**	20	**1**	3	**0**	**0**	**80.0**
H	**1**	**0**	2	65	**0**	**1**	**0**	**94.2**
Sa	**2**	**0**	3	**0**	22	**0**	**1**	**78.5**
Su	**0**	**0**	**0**	**1**	3	79	**0**	**95.2**
C	3	**1**	**1**	**2**	**1**	**0**	10	**55.6**

Table 6
Comparative evaluation of the proposed framework with 2 methods using leave-subject-out cross-validation.

Study	Methodology	Recognition rate
Eskil and Benli [30]	SVM Adaboost	**76.8** **76.3**
Lucey et al. [24]	SVM (shape) SVM (appearance) SVM (combined)	**50.4** **66.7** **83.3**
Proposed	SVM	**83.7**

it is possible to compare the proposed framework with these methods as shown in Table 6. Table 6 shows that the proposed framework achieves an average recognition rate for all seven facial expressions of 83.70%, which outperforms the dynamic method of Eskil and Benli [30] and the static method of Lucey et al. [24]. The confusion matrices in Tables 7–9, respectively, show the recognition rate in using the proposed features of dense optical flow, PHOG_TOP and the combined optical flow and PHOG_TOP (i.e., the proposed framework), where the number in a parentheses denotes the difference in recognition rate in using the proposed features and those used in Lucey at al [24] (i.e., shape, appearance and combined shape and appearance, respectively), where a positive number indicates the proposed features performs better and a negative number indicate worse. Tables 7 and 8, respectively, show the use of dense optical flow and PHOG_TOP achieves significantly better performance than the use of shape and appearance in five of the seven facial expressions. Table 9 shows the combined use of dense optical flow and PHOP_TOP is significantly better in three expressions, slightly worse in two expressions and significantly worse in one expression. However the average recognition performance of the proposed framework is slightly better as shown in Table 6. We can also conclude that the combination of spatial local shape information and dynamic features improves the recognition.

The MMI dataset is used to provide quantitative comparisons with the methods developed by Shan et al. [27], Fang et al. [31], AAM [31], and ASM [31]. For the experiments, we performed 10-fold cross-validation. The average recognition rates are shown in Table 10. The performance of the proposed framework on MMI dataset is worse than that on the CK+ dataset. This is because there are fewer data for training and there are larger changes in head pose in MMI dataset. However, the proposed framework outperforms all the other four methods.

To evaluate the across-dataset performance of the proposed framework, we extracted the integrated features from CK+ dataset as training data, and then use the MMI dataset for testing. The recognition result (in Table 10) shows the generalisation performance across datasets is much lower (58.7% on the MMI database). Thus, we conclude that the current expression classification trained on a single dataset under controlled environment gives good performance only on that dataset.

The computational complexity is analysed with the view of assessing the potential of our proposed framework for real-time application. The processing time is approximated to be the time needed for preprocessing, extracting the dense optical flow and PHOG_TOP, and classification per image frame of the video sequence. The processing time (measured using the computer system clock) is estimated using OpenCV 2.4.3 in Microsoft Visual Studio 2010 Express Edition environment on an Intel(R) Core (TM) i7-3770 CPU @ 3.40 GHz with 16 GB RAM running on Windows 7 operating system. The average processing time per image is under 350 ms and 520 ms for CK+ and MMI datasets, respectively.

6. Conclusion

This paper presents a facial expression recognition framework which uses PHOG_TOP and dense optical flow. The framework which comprises pre-processing, feature extraction and multi-class SVM-based feature classification achieves better performance than two state of the art methods on CK+ dataset and four other methods on MMI dataset. The average recognition rate is 83.7% on the CK+ dataset and 74.3% on the MMI dataset. The expressions of happiness and surprise are easier to be distinguished than other facial expressions, and the results on these two expressions demonstrate the capability of the proposed framework. Nevertheless, the recognition rate of the expressions of contempt (55.6%) and sadness (78.5%) on the CK+ dataset are lower because these two expressions are often misclassified as anger and fear. A limitation of the proposed framework is its generalisation to other datasets. This is because different datasets are generated under different environments (e.g., with varying illumination and head pose).

The proposed framework is able to classify the expressions of happiness and surprise accurately, but encountered some difficulty

Table 7
Confusion matrix in using dense flow optical flow on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation. The value in a parentheses denotes the difference in recognition rate in using dense optical flow and in using shape (of Lucey et al. [24]). Dark grey shade – correct recognition result of each emotion, and lighter grey shade – recognition result of each emotion misclassified with the maximal error.

Expression	A	D	F	H	Sa	Su	C
A	73.3(38.3)	**8.9**	**0.0**	**6.7**	**8.9**	**0.0**	**2.2**
D	**6.8**	84.8(16.4)	**1.7**	**1.7**	**0.0**	**0.0**	**5.1**
F	**12.0**	**4.0**	36.0(14.3)	**20.0**	**8.0**	**12.0**	**8.0**
H	**1.4**	**0.0**	**2.9**	91.3(−7.1)	**1.4**	**0.0**	**2.9**
Sa	**14.3**	**0.0**	**7.1**	**7.1**	50.0(46.0)	**10.7**	**10.7**
Su	**1.2**	**1.2**	**0.0**	**1.2**	**8.4**	72(−28.0)	**1.2**
C	**33.3**	**0.0**	**11.1**	**5.6**	**5.6**	**0.0**	44.4(19.0)

Table 8
Confusion matrix in using PHOG_TOP on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation. The value in a parentheses denotes the difference in recognition rate in using PHOG_TOP and in using appearance (of Lucey et al. [24]). Dark grey shade – correct recognition result of each emotion, and lighter grey shade – recognition result of each emotion misclassified with the maximal error.

Expression	A	D	F	H	Sa	Su	C
A	84.4(14.4)	**2.2**	**6.7**	**0.0**	**4.4**	**0.0**	**2.2**
D	**6.8**	53(41.7)	**0.0**	**0.0**	**0.0**	**1.7**	**1.7**
F	**4.0**	**0.0**	84.0(62.3)	**0.0**	**12.0**	**0.0**	**0.0**
H	**0.0**	**1.4**	2.9	64.0(−36.0)	**0.0**	**0.0**	**2.9**
Sa	**7.1**	**0.0**	**10.7**	**0**	75.0(15.0)	**3.6**	**3.6**
Su	**0**	**0**	**1.2**	**2.4**	**1.2**	95.2(−0.8)	**0**
C	**22.2**	**11.1**	**0.0**	**11.1**	**5.6**	**0.0**	50.0(28.1)

Table 9
Confusion matrix in using the proposed framework on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation. The value in a parentheses denotes the difference in recognition rate in using the combined features of dense optical flow and PHOG_TOP and in using the combined features of shape and appearance (of Lucey et al. [24]). Dark grey shade – correct recognition result of each emotion, and lighter grey shade – recognition result of each emotion misclassified with the maximal error.

Expression	A	D	F	H	Sa	Su	C
A	88.9(13.9)	**0.0**	**2.2**	**0.0**	**6.7**	**0.0**	**2.2**
D	**3.4**	93.2(−1.5)	**1.7**	**0.0**	**0.0**	**0.0**	**1.7**
F	**0.0**	**4.0**	80.0(14.8)	**4.0**	**12.0**	**0.0**	**0.0**
H	**1.4**	**0.0**	**2.9**	94.2(−5.6)	**0.0**	**1.4**	**0**
Sa	**7.1**	**0.0**	**10.7**	**0.0**	78.5(10.5)	**0.0**	**3.6**
Su	**0.0**	**0.0**	**0.0**	**1.2**	**3.6**	95.2(−0.8)	**0**
C	**16.7**	**5.6**	**5.6**	**11.1**	**11.1**	**0.0**	55.6(−28.8)

Table 10
Comparative evaluation of the proposed framework using MMI dataset.

Study	**Recognition rate**
Shan et al. [27]	**54.45**
AAM in [31]	**62.38**
ASM in [31]	**64.35**
Fang et al. [31]	**71.56**
Proposed framework	**74.30**
Train: CK+ Test: MMI	**58.70**

in recognising the other expressions. This is because the datasets used in this paper are small and some of the expressions are poorly represented. One way to address this is to either balance the dataset, or acquire more data. Another limitation of the proposed framework is that it does not consider occlusions (e.g., subjects wearing glasses and with varying hair styles), which will be addressed in our future work. Other aspects that could be useful are the use of more contextual information (e.g., body movements) to achieve better performance and a better understanding of human emotions. The incorporation of contextual information will also be investigated in our future work.

Conflict of interest

None.

Acknowledgements

The authors would like to thank China Scholarship Council/ Warwick Joint Scholarship (Grant no. 201206710046) for providing the funds for this research.

References

[1] P. Viola, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, 2002, pp. 511–518.

[2] H. Fang, N. Costen, From rank-*n* to rank-1 face recognition based on motion similarity, in: Proceedings of the British Conference on Machine Vision, 2009, pp. 1–11.

[3] C. Cornelis, M.D. Cock, A.M. Radzikowska, Vaguely quantified rough sets, in: Proceedings of the International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, vol. 4482, Springer, Toronto, Canada, 2007, pp. 87–94.

[4] P. Ekman, W. Friesen, Constants across cultures in the face and emotion, J. Pers. Soc. Psychol. 17 (2) (1971) 124–129.

[5] M. Pantic, L.J.M. Rothkrantz, Automatic analysis of facial expressions: the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1424–1445.

[6] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognit. 36 (1) (2003) 259–275.

[7] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Recognising facial expression: machine learning and application to spontaneous behavior, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2005, pp. 568–573.

[8] Y.-I. Tian, T. Kanade, J. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2002) 97–115.

[9] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: Proceedings of the International Conference on Image and Video Retrieval, 2007, pp. 401–408.

[10] J.F. Cohn, A. Zlochower, J. Lien, T. Kanade, Feature-point tracking by optical flow discriminates subtle differences in facial expression, in: Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 396–401.

[11] Z. Zhang, M. Lyons, M. Schuster, S. Akamatsu, Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer, in: Proceeding of International Conference on Automatic Face and Gesture Recognition, 1998, pp. 454–459.

[12] G. Zhan, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expression, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.

[13] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[14] G.Y. Zhao, M. Pietikainen, Dynamic texture recognition using local binary pattern with an application to facial expression, IEEE Trans. Pattern Anal. Mach. Intell. 2 (6) (2007) 915–928.

[15] Y. Zhang, Q. Ji, Active and dynamic information fusion for facial expression understanding from image sequences, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 699–714.

[16] M. Pantic, Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences, IEEE Trans. Syst. Man Cybern. 36 (2) (2006) 433–449.

[17] M.F. Valstar, I. Patras, Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data, in: IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2005 pp. 76–84.

[18] Y. Yacoob, L.S. Davis, Recognizing human facial expression from long image sequences using optical flow, IEEE Trans. Pattern Anal. Mach. Intell. 18 (6) (1996) 636–642.

[19] M. Yeasin, B. Bullot, R. Sharma, From facial expression to level of interests: a spatio-temporal approach, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, vol. 2, 2004, pp. 922–927.

[20] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 886–893.

[21] O. Deniz, G. Bueno, J. Salido, Face recognition using histograms of oriented gradients, Pattern Recognit. Lett. 32 (12) (2011) 1598–1603.

[22] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene Categories, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2169–2178.

[23] Z. Li, J. Imai, M., Kaneko, Facial-component-based bag of words and PHOG descriptor for facial expression recognition, in: IEEE International Conference on Systems, Man and Cybernetics, 2009, pp. 1353–1358.

[24] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, Extended Cohn–Kande Dataset (CK+): a complete facial expression dataset for action unit and emotion-specified expression, Paper Presented at the Third IEEE Workshop on CVPR for Human Communicative Behaviour Analysis, 2010, pp. 94–101.

[25] B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proceedings of the 7th International Joint Conference on Artificial Intelligence, vol. 2, 1981, pp. 674–679.

[26] B. Martinez, M.F. Valstar, X. Binefa, M. Pantic, Local evidence aggregation for regression based facial point detection, IEEE Trans. Pattern Anal. Mach. Intell. 35 (5) (2013) 1149–1163.

[27] C. Shan, S. Gong, P. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (2009) 803–816.

[28] Chih-Wei Hsu, Chih-Jen Lin, A comparison of methods for multiclass support vector machines, IEEE Trans. Neural Netw. 13 (2) (2002) 415–425.

[29] M. Pantic, M. Valstar, R. Radermaker, L. Maat, Web-based database for facial expression analysis, in: Proceedings of the 13th ACM International Conference on Multimedia, 2005, pp. 317–321.

[30] M.T. Eskil, K. Benli, Facial expression recognition based on anatomy, Comput. Vis. Image Underst. 119 (2014) 1–14.

[31] H. Fang, N.M. Parthaláin, J. Aubrey, K.L. Tam, R. Borgo, L. Rosin, W. Grant, D. Marshall, M. Chen, Facial expression recognition in dynamic sequences: an integrated approach, Pattern Recognit. 47 (3) (2014) 1271–1281.

Xijian Fan received B.Sc. in Information and Communication Technology from Nanjing University of Posts and Telecommunications, China, and M.Sc. in Computer Information and Science from Hohai University, China, in 2008 and 2012, respectively. He is currently pursuing Ph.D. in Engineering at the University of Warwick, UK. His research interests include image processing and facial expression recognition.

Tardi Tjahjadi received B.Sc. in Mechanical Engineering from University College London in 1980, and M.Sc. in Management Sciences in 1981 and Ph.D. in Total Technology in 1984 from UMIST, UK. He has been an associate professor at Warwick University since 2000 and a reader since 2014. His research interests include image processing and computer vision.