Pattern Recognition 71 (2017) 249–263

Contents lists available at ScienceDirect

Pattern Recognition

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

Multi-gradient features and elongated quinary pattern encoding for image-based facial expression recognition

S.A.M. Al-Sumaidaee [a,b,*], M.A.M. Abdullah [a,c], R.R.O. Al-Nima [a,d], S.S. Dlay [a], J.A. Chambers [a]

[a] ComS²IP Research Group, School of Electrical and Electronic Engineering, Newcastle University, England, UK
[b] Al-Mustansiriya University, Baghdad, Iraq
[c] University of Ninevah, Mosul, Iraq
[d] Technical College of Mosul, Mosul, Iraq

ARTICLE INFO

Article history:
Received 6 December 2016
Revised 9 May 2017
Accepted 1 June 2017
Available online 3 June 2017

Keywords:
Texture feature analysis
Spontaneous facial expression
Multi gradient magnitude and angle images
Elongated quinary pattern
Multi classifier system
Multi-class SVM classifier

ABSTRACT

In this paper we propose a novel texture feature extraction method for posed and spontaneous image based facial expression recognition. The kernel Sobel filter is used with eight masks to derive the gradient components for each pixel in the image. Two types of gradient images are extracted for different directions denoted as *xy* and *lr*. The robust Elongated Quinary Pattern (EQP) descriptor is then used to quantize neighborhood local gradients around each point using five discrimination levels. We next divide each encoded image into a number of blocks and concatenate the local histogram features of each image individually. In order to boost the performance, we adopt a Multi Classifier System (MCS) to combine all scores of the encoded images based upon a multi-class Support Vector Machine (SVM) classifier. Experimental results show a significant improvement over previous approaches in the average recognition accuracy when using the spontaneous Moving Faces and People (MFP) database. In addition, the proposed method outperformed state-of-the-art methods when applied to the posed CK database with a recognition performance of 99.36% in the case of seven classes and 99.72% without the neutral class.

1. Introduction

Several methods can be used to express and reflect the mode or emotions of humans such as voice, gestures or facial expressions. Facial expression is one of the most accurate nonverbal communication methods. Facial expressions are commonly categorized by seven classes such as anger, disgust, fear, happy, sad and surprise along with the neutral state [1]. In the last two decades, the study of facial expression recognition has been employed in various Human Computer Interface (HCI) applications. A set of muscle movements known as Facial Action Units (FAUs) was defined to make the facial expression recognition more accurate. Several researchers in facial expression recognition have used the FAUs for extracting the facial features, others utilized direct approaches such as geometric or appearance based methods [2,3]. However, using the FAUs has several drawbacks, such as revealing unwanted information that may confuse the classifier [4]. Moreover, it requires more time for coding every frame, making it unreliable for real-time application [5].

Generally speaking, there are two types of methods used to distinguish between different facial expression approaches, namely geometric and appearance based methods. A number of researchers have used geometric based information within the FAUs to calculate features [6]. It has been reported that the geometric based methods have high complexity as they require accurate and reliable algorithms to detect different facial expression changes. Hence, the geometrical based methods appear unsuitable for use in many real-time applications [7,8]. Appearance based methods can be used for extracting the features either from the whole image or at fiducial point locations [9]. Appearance based methods seem more reliable to analyze and extract facial expressions compared with geometric methods since they only involve convolving different types of filters with the facial image, for example, convolving a set of Gabor filter banks with a facial image at a number of scales and orientations [10]. However, they generally require more time in the training process to achieve satisfactory recognition results. On the other hand, other important studies have focused on extracting the facial expression features using local descriptors, such Local Binary Patterns (LBPs) which have low computational time and can significantly improve the recognition accuracy [7,11]. Furthermore, additional studies have combined local descriptors with complimentary descriptors, such as the work in [12] or previous work in [13] where the recognition performance

* Corresponding author.
E-mail addresses: s.a.m.al-sumaidaee@ncl.ac.uk, saadoon_groupvet@yahoo.com (S.A.M. Al-Sumaidaee), m.a.m.abdullah@ncl.ac.uk (M.A.M. Abdullah), satnam.dlay@ncl.ac.uk (S.S. Dlay), Jonathon.Chambers@ncl.ac.uk (J.A. Chambers).

http://dx.doi.org/10.1016/j.patcog.2017.06.007
0031-3203/

is improved significantly. All the above works have been applied on different static or dynamic types of posed databases such as the Cohn-Kanade (CK) or Multimedia Understanding Group (MUG) databases that were collected under tightly controlled conditions [14].

Recently, spontaneous facial expression recognition has been gaining more attention than posed facial expression recognition. Spontaneous facial expression recognition is more challenging than posed facial expression recognition due to several factors. Firstly, spontaneous facial expressions can be less expressive due to the slight changes in the facial muscle actions as demonstrated in [15]. Secondly, some spontaneous facial expressions have overlapped geometric and appearance features. For example, the subject may show a sad expression instead of a fear expression inside a fear video [16]. Thirdly, the time period to express each facial expression is different from one expression to others. Some videos require more time to reveal the expressions compared to others, as explained in [17,18]. Finally, it is not necessary for the expressive images to appear at the end of the spontaneous video as in the case with a posed database. Some expressive image(s) may occur at the start, middle or toward the end of the video. Therefore, in real-time facial expression recognition systems, a robust method is needed to detect the apex expressions frames in the video [19]. All these issues will affect the performance of the best classifiers and make them fail to attain high classification rates as explained in [16].

In this paper, we focus on overcoming the major difficulties associated with spontaneous facial expression problems which are mentioned above and achieve a high recognition accuracy. We therefore propose a robust method for spontaneous facial expression recognition by utilizing different gradient images of magnitudes and angles at different orientations while extracting Elongated Quinary Pattern (EQP) features. To obtain the scores, these features are then fed through multi-class SVM classifiers. Finally, to boost the performance of the final scores, the output scores of each classifier are fused together using a Multi Classifier System (MCS) method and the final decision is then determined. In summary, our contributions are as follows:

- Deriving multi-gradient features for both magnitude and angle components using eight Sobel masks. Using multi-gradient image features rather than texture image features to help to isolate high gradient regions such as wrinkles, bulges, furrows or edges in different orientations. Utilizing the EQP descriptor to provide more consistency in uniform and near uniform regions to extract more gradient features (magnitude and direction) in different directions.
- Proposing the Multi Classifier System (MCS) based on the multi-class SVM classifier which fuses the output scores of different histogram features of multi-gradient features based on the EQP descriptor. Reducing the False Negative (FN) error for each class by decreasing the inter-class distance and increasing the intra-class distance which improves the overall accuracy.
- Performing evaluations on the spontaneous Moving Faces and People (MFP) [20] and posed CK databases which shows superior performance and outperformed the current state-of-the-art methods.

2. Related works

There are two trends in studying spontaneous facial expression recognition, the first one focuses on detecting the micro movements and determining the most expressive images inside each class (e.g. apex changes), while the second, focuses on designing an expression recognition system by increasing and enhancing the features (i.e. geometric and/or appearance) that are extracted from images to enhance the classification process.

For the first trend, Moilanen *et al.* [21] proposed a simple method for spotting the rapid movements from video using the CASME II database. The histograms of the LBP descriptor are calculated for each frame in the video after dividing it into smaller blocks. The Chi-Square (χ^2) distance is used to threshold the greatest difference (i.e. peak movement) between each frame and the average feature frames of all the video sequence. Their method achieves *Recall* of 0.5171. Davison *et al.* [22] introduced a new method for micro-movement detection in the CASME II video database. The Histogram of Oriented Gradients (HOGs) [23] was calculated as a feature of each frame in the video. Then the χ^2 distance was used to measure the dissimilarity of each frame over the selected interval. The proposed method improves the sensitivity *Recall* to 0.8429 for micro-facial movements. The main drawback is represented by the high computation time required for evaluation of feature descriptors which makes them unsuitable to be used with high speed videos.

For the second trend, Wan and Aggarwal [19] conducted a study in spontaneous facial expression recognition using the MFP database. This work improved the overall accuracy by increasing the training features and the probability distance inside each class (i.e. it increased the sensitivity inside each class); 13 annotators were employed to choose the most expressive image inside a video sequence. However, their method has a high complexity as it depends on a geometrical Active Appearance Model (AAM) for detection of the landmark facial points of each image. Happy *et al.* [24] established a new spontaneous facial expression database that contains a number of Indian male and female contributers. Different methods were used for feature extraction, such as the Local Gabor Binary Pattern (LGBP) or Pyramid of Histogram of Gradients (PHOGs) with different types of classifiers. Experimental results yielded that the maximum accuracy is equal to 86.46% and 67.75% when using LGBP and PHOG features respectively. However, it is reported that there will be variations in the results due to the small number of samples in each class and the imbalanced sample distribution in this database. Takahashi *et al.*[25] performed a study on facial expression recognition using HOG features and a multi-layer quaternion neural network classifier and applied it on Japanese male samples. It was shown that this classifier has the ability to attain 50% recognition rate in the person-dependent experiment. At the same time, Donia *et al.* [26], performed a study to show the effectiveness of using the gradient components (i.e. magnitude and direction) with the spontaneous facial expressions using the HOG histogram bins features, which was originally exploited to detect edges and local shape information [27]. Experimental results on static images and video attained 95% and 80% recognition accuracy, respectively.

With the intention of increasing the overall accuracy in posed facial expression databases and extracting more expression features using the gradient image, several researchers tried to improve the HOG histogram bin features. In 2012, Ahmed [28] proposed to use the Gradient Directional Pattern (GDP) features in facial expression recognition. His method improved the overall recognition accuracy as compared to some types of works that used different appearance based methods. After that, Ahmed and Hossain [8] proposed to use the LTP descriptor on the gradient magnitude image, in which wrinkles, spots or edges are highlighted. This combination method named as the Gradient Local Ternary Pattern (GLTP), improved the overall accuracy to 97.2% in 6 class facial expression recognition when applied on the CK database.

In the next section we introduce the methods used to extract the texture feature using Local Binary Pattern (LBP), Local Ternary Pattern (LTP) and Elongated Quinary Pattern (EQP) descriptors.

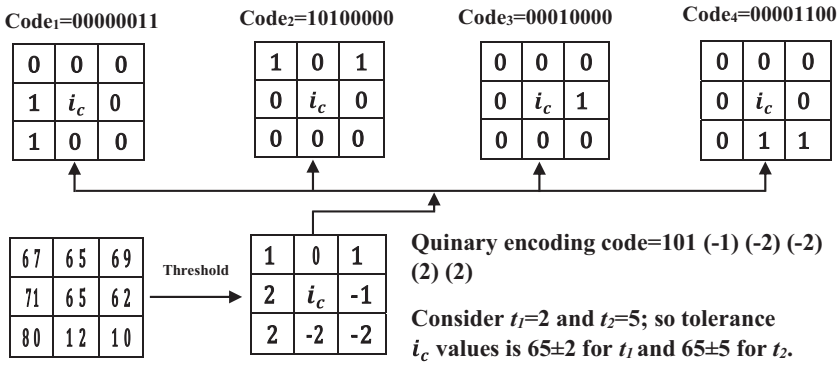

Fig. 1. The basic quinary encoding process after splitting it into four LBP descriptors.

3. LBP variants as texture descriptors

Ojala *et al.* [29] invented the LBP operator which is used widely in face analysis and facial expression recognition. This local descriptor uses two discrimination levels to encode the micro level information such as spots, edges or curves by quantizing the neighbouring gray levels of any local area around a pixel and change its value with the binary code value. The LBP descriptor has less complexity and is less affected by some types of illumination which thereby improved both the accuracy and reduced the classifier time. However, the LBP only has one threshold value, as such it is affected by some types of noises, particularly in near-uniform regions. Tan and Triggs [30] addressed this problem by increasing the LBP discrimination through extending it to three level discrimination. Hence, the gray-levels (i_p) in a zone of threshold (t) with width $\pm\, t$ about the center pixel (i_c) are quantized to zero, whereas ones that are above ($i_c + t$) are quantized to +1 and ones that are bellow ($i_c - t$) are quantized to -1, and it was named the LTP descriptor. The LTP coding involves 0, 1 and -1 values, so, it can be split into two binary codes (e.g. positive and negative), which enables it to calculate the LBP histogram features for each of the LTP splitter codes. Despite t being a user specified threshold which gives the LTP resistance to noise, it is unable to deal with gray level transformations.

Nanni *et al.* [31] suggested a novel method which extends the LBP to work with five values (-2, -1, 0, +1, +2) of encoding named the quinary descriptor. Therefore, the gray-levels (i_p) in a zone of two threshold values (i.e. $\pm\, t_1$ and $\pm\, t_2$) are encoded. Hence, the gray-levels in a zone of $\pm\, t_1$ about the center pixel (i_c) are quantized to 0, whereas those above or equal to ($i_c + t_1$) and ($i_c + t_2$) are quantized to +1, and those above or equal to ($i_c - t_2$) and ($i_c - t_1$) are quantized to -1. Similarly, the gray-levels above or equal to ($i_c + t_2$) are quantized to +2, otherwise they are quantized to -2. The quinary encoding method is more robust than LTP in terms of sensitivity in the uniform and near-uniform regions. We can express the quinary encoding descriptor by a function of five discrimination levels specified by four thresholds i_p, i_c, t_1 and t_2 as:

$$d(i_p, i_c, t_1, t_2) = \begin{cases} +2 & i_p \geq i_c + t_2, \\ +1 & i_c + t_1 \leq i_p < i_c + t_2, \\ 0 & i_c - t_1 \leq i_p < i_c + t_1, \\ -1 & i_c - t_2 \leq i_p < i_c - t_1, \\ -2 & \text{otherwise}. \end{cases} \quad (1)$$

The quinary descriptor is split into four different types of LBP descriptors according to the following $d_k(x)$ function as:

$$d_k(x) = \begin{cases} 1 & x = k, \\ 0 & otherwise, \end{cases} \quad (2)$$

where $k \in \{-2, -1, +1, +2\}$. Fig. 1 represents the basic operation example of the quinary encoding descriptor after splitting into four LBP descriptors. A number of features was added to modify the performance of the quinary encoding and make it able to exploit the important information that has an anisotropic structure form [27,32]. These modifications were achieved by changing the circular neighborhood topology to different quadratic neighborhood topologies such as an ellipse, parabola or hyperbola. In addition, the number of neighborhood points (P) has been changed to 10 or 18. Moreover, different numbers of point locations have been used (β).

It is reported that using an elliptic neighborhood topology with 5 levels of scales to encode the difference of the local gray-levels outperformed different encoding methods, and is denoted the Elongated Quinary Pattern (EQP) [31].

4. Proposed method

We propose a robust method for spontaneous and posed facial expression recognition. We first derive the multi gradient components of the magnitude and angle images in the *xy* and *lr* directions. Then we propose a new method to encode the gradient images using the EQP descriptor, which enables us to extract more information near the uniform regions. After that, we utilize the local histogram features of the new descriptors, along with the PCA technique to reduce features. Finally, we propose an MCS method based on the multi-class SVM classifier for the combination of scores.

4.1. The derivation of multi-gradient features for different directions

In this paper we have adopted the Sobel masks in eight directions for generating gradient magnitude and angle images in different directions. The selection of the Sobel masks may increase the stability near the non-uniform regions, particularly edges that may be affected with illumination changes or noise. Moreover, it requires less calculation time with better resolution compared to other masks such as the Robert masks. With the aim of extracting the major facial expression features, we used the effect of any perpendicular Sobel masks on the image to generate the multi gradient features for both magnitude and angle components at different orientations.

To derive the gradient relation of the magnitude and angle components of each pixel in the image, the image is first convolved with two perpendicular components from the kernel of the eight Sobel masks. For example, Fig. 2 represents 4 perpendicular components of the eight Sobel masks that may be convolved with any 3×3 region of an image. To find the gradient relations of both magnitude and angle in the *xy* direction of each pixel in the image, we first convolved the east (e) and north (n) Sobel masks di-

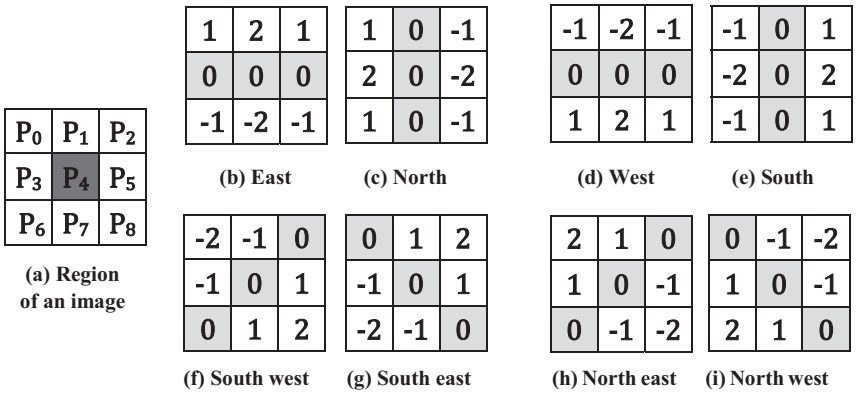

Fig. 2. (a) 3 × 3 region of an image. The perpendicular components of the Sobel masks: (b-c) Horizontal-Vertical. (d-e) Vertical-Horizontal. (f-g) South west-South east. (h-i) North east-North west.

rections Fig. 2 (b-c) with each pixel around (P_4) Fig. 2 (a) as:

$$G_x(e) = (p_0 + 2p_1 + p_2) - (p_6 + 2p_7 + p_8), \tag{3}$$

where, $G_x(e)$ represents the gradient of the (P_4) pixel in the horizontal direction. Similarly, $G_y(n)$ represents the gradient of the (P_4) pixel in the vertical direction as:

$$G_y(n) = (p_0 + 2p_3 + p_6) - (p_2 + 2p_5 + p_8). \tag{4}$$

Then we can calculate the gradient magnitude and angle components of the (P_4) pixel in the east and north direction as follows:

$$GM_{xy}(e\_n) = |G_x(e)| + |G_y(n)|, \tag{5}$$

where, $GM_{xy}(e\_n)$ is the gradient magnitude component in the *xy* direction, and:

$$\theta_1(x, y) = GD_{xy}(e\_n) = tan^{-1}(G_y(n)/G_x(e)), \tag{6}$$

is the gradient angle component in the *xy* direction, which is equal to $\theta_1(x, y), \theta_1 \in (-\pi, \pi]$.

In the same way, we convolve the west (*w*) and south (*s*) Sobel masks with each pixel around (P_4) Fig. 2 (d-e). We find the other two gradients of the (P_4) pixel in the horizontal and vertical directions. This is not going to affect the calculation of the magnitude and angle gradients in the *xy* direction as these components represent the mirror form only, because:

$$G_x(e) = -G_x(w), \tag{7}$$

and

$$G_y(n) = -G_y(s). \tag{8}$$

Similarly, to derive the gradient magnitude and angle relations of the (P_4) pixel in the image in the left right (*lr*) direction, the south east (*se*) and south west (*sw*) Sobel masks directions are convolved with each pixel around (P_4) Fig. 2 (f-g) as:

$$G_l(se) = (p_1 + 2p_2 + p_5) - (p_3 + 2p_6 + p_7), \tag{9}$$

where, $G_l(se)$ represents the gradient of the (P_4) pixel in the left diagonal mask direction, and G_r represents the gradient of the (P_4) pixel in the right diagonal mask direction as:

$$G_r(sw) = (p_5 + 2p_8 + p_7) - (2p_0 + p_1 + p_3). \tag{10}$$

In the same way, we can convolve the north west (*nw*) and north east (*ne*) Sobel masks with each pixel around (P_4) Fig. 2 (h-i). We find the other two gradients of the (P_4) pixel in the *lr* direction. This does not affect the calculation of the magnitude and angle gradients in the *lr* direction as they represent the mirror forms only, because:

$$G_l(se) = -G_l(nw), \tag{11}$$

and

$$G_r(sw) = -G_r(ne). \tag{12}$$

Then we can calculate the gradient magnitude and angle of the (P_4) in the south east and south west direction as follows:

$$GM_{lr}(se\_sw) = |G_l(se\_sw)| + |G_r(sw)|, \tag{13}$$

where, $GM_{lr}(se\_sw)$ is the gradient magnitude component in the *lr* direction, while the gradient angle component in the *lr* direction can be calculated as:

$$\theta_2(l, r) = GD_{lr}(se\_sw) = tan^{-1}(G_l(se)/G_r(sw)), \tag{14}$$

and $\theta_2 \in (-\pi, \pi]$. Therefore, (5)-(6) and (13)-(14) represent the derivation of the gradient magnitude and angle components of each pixel in the image in the perpendicular planes of the *xy* and *lr* directions.

Finally, to achieve better invariance to small variation in the image we suggest to add π to θ_1 and θ_2 when their values are negative. Therefore, θ_1 and θ_2 are within the range $[0, \pi]$ [25]. Fig. 3, represents the input facial expression image with its correspondent gradient components in the *xy* and *lr* directions.

4.2. Feature extraction using the EQP descriptor based on the multi-gradient images

For facial expression recognition, there are several distinctive aspects of gradient features over texture feature that makes them more convenient to use in the process of feature extraction. This is because the gradient images have a high reliability to isolate the consistent regions such as high or smooth textures. In addition, the patterns in gradient images have robustness to different types of illumination variations [8,33].

In our method, we propose to use the two gradient types of each image, magnitude and angle in the *xy* and *lr* directions. The diversity in the gradient features enabled us to extract more consistent features for a specific region. The gradient magnitude features in the high texture regions such as edge or curves are more consistent and stable. Therefore, the ability to isolate the high frequency regions from smooth regions is high. Furthermore, the gradient angle features have the ability to explore different types of micro features such as wrinkles, bulges or subtle changes at different angles $[0, \pi]$. Thus using multi-gradient features represents the most important tool in facial expression recognition as it reduces the overlap in the features among different facial expressions, particularly in the spontaneous database type.

To boost the performance of the feature extraction in uniform and near uniform regions of the gradient images and to exploit the

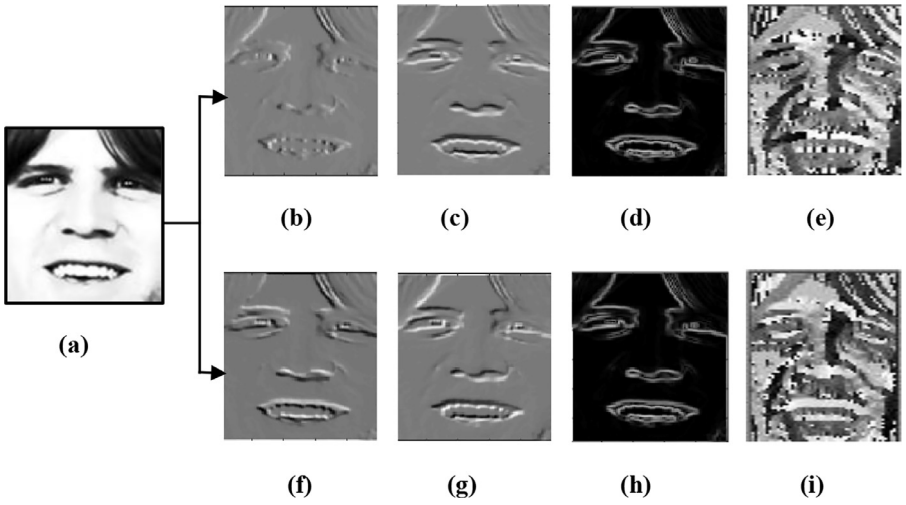

Fig. 3. Facial expression image and its gradient based components. (a) Happy cropped expression image. (b) The gradient image in the vertical direction G_y. (c) The gradient image in the horizontal direction G_x. (d) The gradient magnitude image GM_{xy} in the xy direction. (e) The gradient angle image GD_{xy} in the xy direction. (f) The gradient image in the left direction G_l. (g) The gradient image in the right direction G_r. (h) The gradient magnitude image GM_{lr} in the lr direction. (i) The gradient angle image GD_{lr} in the lr direction.

important information that has an anisotropic structure, we applied the EQP descriptor on both magnitude and angle gradients of the image at different directions (e.g. xy and lr). The EQP descriptor encodes each pixel in the gradient image according to the neighbor pixels using two threshold values (t_1 and t_2) to reduce the sensitivity of the gradient pixels in the important regions of the facial images, particularly near the uniform regions.

Therefore, we can rewrite (1) which represents the five encoding levels of the EQP descriptor in terms of the two gradients directions of the magnitude (GM_{xy} or GM_{lr}) and angle (GD_{xy} or GD_{lr}) components as:

$$d_{GMEQP}(GM_p, GM_c) = \begin{cases} +2 & GM_p \geq GM_c + t_2, \\ +1 & GM_c + t_1 \leq GM_p < GM_c + t_2, \\ 0 & GM_p - t_1 \leq GM_p < GM_c + t_1, \\ -1 & GM_c - t_2 \leq GM_p < GM_p - t_1, \\ -2 & \text{otherwise}, \end{cases} \quad (15)$$

where, GM_p represents the gradient magnitudes of any of the neighbor points that are surrounding the gradient magnitude (GM_c) of the center pixel (x_c, y_c); t_1 and t_2 represent the user specified thresholds that are chosen empirically to be 10 and 18 respectively and the GMEQP is generated from the EQP descriptor after applying it on the gradient magnitude components of the xy and lr directions respectively. Similarly, the d_{GDEQP} can be written as:

$$d_{GDEQP}(GD_p, GD_c) = \begin{cases} +2 & GD_p \geq GD_c + t_2, \\ +1 & GD_c + t_1 \leq GD_p < GD_c + t_2, \\ 0 & GD_p - t_1 \leq GD_p < GD_c + t_1, \\ -1 & GD_c - t_2 \leq GD_p < GD_p - t_1, \\ -2 & \text{otherwise}, \end{cases} \quad (16)$$

where, GD_p represents the gradient angle of any of the neighbor points that are surrounding each gradient angle GD_c of the center pixel (x_c, y_c); t_1 and t_2 represent the user specified thresholds that are chosen empirically to be 1 and 18 respectively, and the GDEQP is generated from the EQP descriptor after applying it on the gradient angle components of the xy and lr directions respectively.

Inevitably, using five encoding levels has an order of 5^P, where P is equal to 8 or 10, which will increase the number of GMEQP or GDEQP patterns. To reduce these patterns we adopted the Tan and Triggs approach [30] by splitting each proposed descriptor into four LBPs and then calculating the concatenated histogram features among them. Fig. 4 represents a basic example of selecting a 3×3 region from the expression image and then splitting into four different LBP codes. These codes are obtained after applying the basic EQP descriptor for encoding the two gradient images (GM and GD) at multi directions (e.g. xy and lr). We can express the GMEQP descriptors for the xy and lr directions as:

$$GM_{xy}EQP = \sum_{P=0}^{P-1} d_n(d_{GM_{xy}EQP}(P)).2^P; n = 0, \quad (17)$$

where, P is the number of the neighborhood points, and,

$$GM_{lr}EQP = \sum_{P=0}^{P-1} d_n(d_{GM_{lr}EQP}(P)).2^P; n = 1. \quad (18)$$

In the same way, we can express the GDEQP descriptors for the xy and lr directions as:

$$GD_{xy}EQP = \sum_{P=0}^{P-1} d_n(d_{GD_{xy}EQP}(P)).2^P; n = 2, \quad (19)$$

$$GD_{lr}EQP = \sum_{P=0}^{P-1} d_n(d_{GD_{lr}EQP}(P)).2^P; n = 3. \quad (20)$$

We can split each one of the above descriptors into four different types of LBP descriptor as:

$$d_{n_k}(u) = \begin{cases} 1 & u = k, \\ 0 & \text{otherwise}, \end{cases} \quad (21)$$

where, $n = (0, \cdots, 3)$ and $k \in \{-2, -1, +1, +2\}$.

Fig. 5, represents the effect of using the robust EQP descriptor after applying it on the multi-gradient images at different magnitude and angle components.

4.3. Histogram representation of multi-gradient features using EQP descriptor

We now have two new types of descriptors, each descriptor has two gradient components in the xy and lr directions (i.e. $GM_{xy}EQP$, $GM_{lr}EQP$, $GD_{xy}EQP$ and $GD_{lr}EQP$). These descriptors will generate four encoded images in the xy and lr directions when applied on

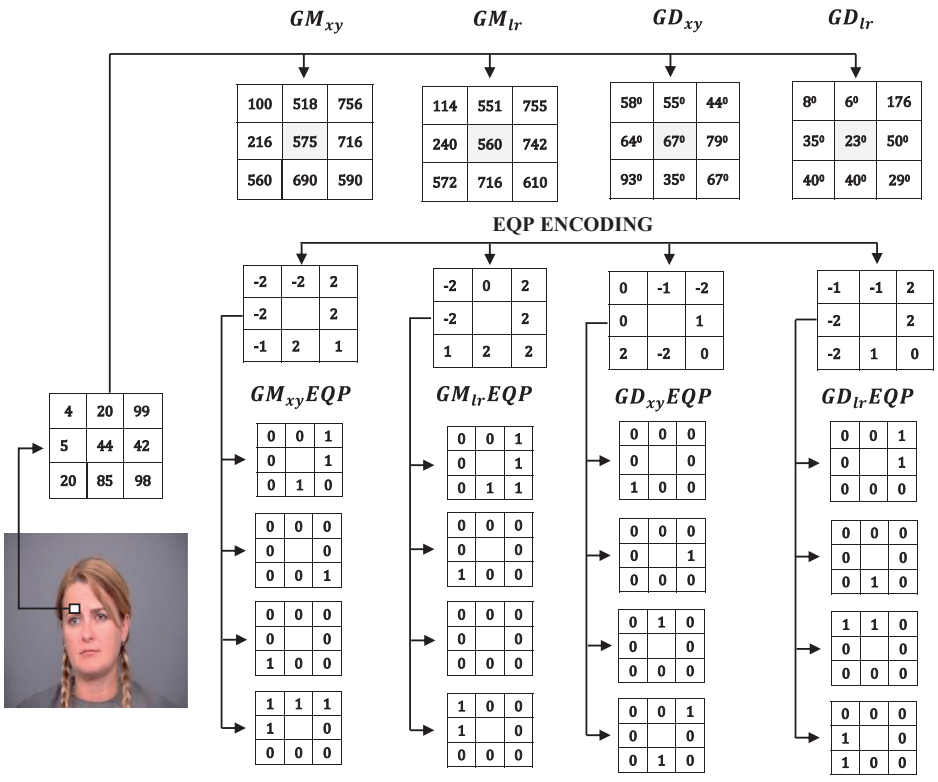

Fig. 4. The proposed method of encoding and splitting the GMEQP and GDEQP descriptors at different directions.

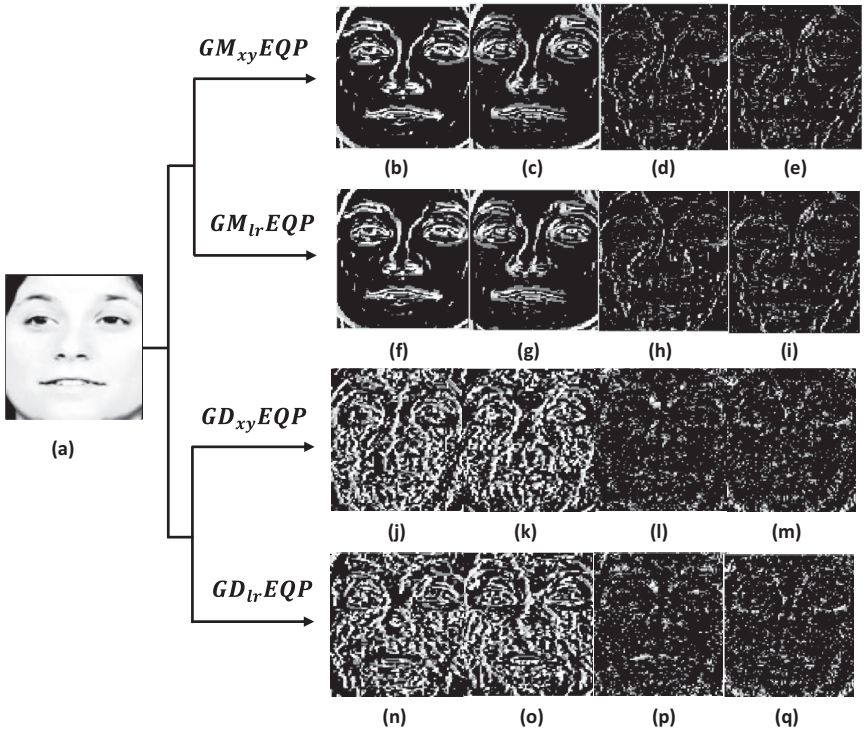

Fig. 5. Multi gradient images after applying the EQP encoding method at different directions. (a) Surprise cropped image. (b-e) GM_{xy}EQP. (f-i) GM_{lr}EQP. (j-m) GD_{xy}EQP. (n-q) GD_{lr}EQP; columns (b-n), (c-o), (d-p) and (e-q) correspond respectively to the four outputs from the splitting process represented in Fig. 1.

the facial expression image, as shown in Fig. 5. The histogram features of the first two encoded images are computed as:

$$H_{GM_{xy}EQP}(\tau_h) = \sum_{i=1}^{m}\sum_{j=1}^{n} f(GM_{xy}EQP(i,j), \tau_h); h = 0, \quad (22)$$

$$H_{GM_{lr}EQP}(\tau_h) = \sum_{i=1}^{m}\sum_{j=1}^{n} f(GM_{lr}EQP(i,j), \tau_h); h = 1, \quad (23)$$

similarly, the histogram features of the other encoded images are computed as follows:

$$H_{GD_{xy}EQP}(\tau_h) = \sum_{i=1}^{m}\sum_{j=1}^{n} f(GD_{xy}EQP(i,j), \tau_h); h = 2, \quad (24)$$

$$H_{GD_{lr}EQP}(\tau_h) = \sum_{i=1}^{m}\sum_{j=1}^{n} f(GD_{lr}EQP(i,j), \tau_h); h = 3, \quad (25)$$

$$f(\nu, \tau_h) = \begin{cases} 1 & \nu = \tau_h, \\ 0 & \text{otherwise.} \end{cases} \quad (26)$$

Here, $h = (0, \cdots, 3)$, $\tau_h \in \{-2, -1, +1, +2\}$ and f is the function of the histogram depending on the value of τ_h. However, the histogram information that is generated from the whole encoded image does not express the local spatial information of these images, since it only represents the occurrence frequencies. In order to deal with this problem and to extract more local and efficient histogram information, we divided each encoded image into a number of non-overlapped regions (blocks), such as 7×6, then we calculated the local occurrence frequencies of each block individually. By doing this, we incorporated the occurrence of micro-level information at different facial regions. Then, we concatenated all histogram features which are generated from the encoded images individually. Fig. 6 represents four vectors of the histogram local features that are generated from encoded images at different directions after applying GMEQP and GDEQP descriptors on a cropped input image.

In order to reduce the vector size, we applied the Principal Component Analysis (PCA) method to extract the essential features that have high discriminating capacity to be suitable for the classifier. These features represented by the significant Principal Components (PCs) retained 98% of the PCA energy (e.g. PC_0, PC_1, PC_2 and PC_3) [19] as illustrated in Fig. 6.

4.4. The MCS method based upon the multi-class SVM classifier

The SVM is a binary decision classifier (supervised machine learning technique) which is used to separate two classes (e.g. positive and negative samples) with a margin (C) of the hyperplane. The test data can be classified as: Consider a training data set (TR) with number of labels M; $TR = \{(X_i, Y_i, i = 1, \cdots, M)\}$, where $X_i \in R^n$ and $Y_i \in \{-1, +1\}$, so:

$$f(x) = sign\left(\sum_{i=1}^{M} \alpha_i Y_i K(X_i, X) + b\right), \quad (27)$$

here, α_i represent the Lagrange multipliers of dual optimization problem (e.g. $\alpha_i > 0$ for the maximum separation margin) and b is the bias and K is the kernel function. The value of K can be selected according to the data type used with the SVM classifier. There are different kernels which can be used with SVM such as Linear, Polynomial, Radial Basis Function (RBF), Histogram Intersection (HI) or Sigmoid. To improve the performance of the proposed method, a suitable kernel function for the SVM should be selected carefully. As our method utilizes the histogram features of the GMEQP and the GDEQP, the HI kernel would suit the requirement of such histogram features [34,35]. Therefore, the HI kernel has been employed to carry out the experiments. In addition, the RBF kernel function is also used to achieve fair comparisons with other works that employed the same kernel. It is important to optimize the (C) parameter through a grid search strategy by selecting an appropriate (C) soft or hard margin value. After that, we can carry out the subject-independent cross-validation.

Because the facial expression recognition is considered as a multi-class problem (i.e. Cl), where Cl represents the number of classes, a multi-class SVM is suitable for this purpose. Therefore, several methods can be used with the multi-class problems such as One Versus One (OVO), One Versus All (OVA) or Directed Acyclic Graph (DAG). In this work, we adopted the OVO method to classify the seven facial expressions, because it has less training time compared to the OVA method as explained in [36].

It is reported that combining different types of classifiers would give better prediction results [37]. In our method, we used the Multi Classifier System (MCS). In this system the output prediction scores of the encoded images are combined based on a multi-class SVM classifier to generate a final decision later. There is an important reason for using this method; as the proposed method produces four vectors of the reduced features PC_0, PC_1, PC_2 and PC_3, the first two vectors have diverse magnitude information in the *xy* and *lr* directions, while the other two have complementary angle information in the *xy* and *lr* directions which have been encoded based on the robust EQP descriptor. Hence, such a combination will integrate and boost the scores of the different features that are extracted from the same local regions of the facial expression image. Due to the above mentioned point, we used the MCS method to obtain the final prediction scores. First, we used the OVO multi-class SVM classifier method to classify 7 class of the facial expression for each vector of trained/tested features (e.g. PC_0, PC_1, PC_2 and PC_3) individually. The probability of the prediction scores (Sc) for each multi-class SVM classifier is:

$$Sc_i^R(PC_c); i = (1, \cdots, l), \quad (28)$$

where, R is a multi-class SVM classifier, $R \in (1, \cdots, 4)$, $c \in (0, \cdots, 3)$ and l is the number of the classes.

Then, to find the final prediction scores, that are generated from the combination schemes of the individual multi-class SVM scores, the sum rule operation is utilized. The sum-rule operation is less sensitive to the effect of noise than other rules such as product, mean or minimum as reported in [38]. Therefore, the output score is combined with the sum-rule for every class and then assigned the class label with a maximum score for the input vectors of PCs:

$$FSc_j = Comb(Sc_i^1(PC_0), \cdots\cdots, Sc_i^4(PC_3)); i \neq j; i, j \in \{1, \cdots, l\}, \quad (29)$$

where, *Comb* is the sum rule function and FSc_j is the final prediction of the class labels j. The final j labels are more stable than other i labels that are generated individually from each multi-class SVM classifier as shown in (28) and illustrated in the Fig. 7. It is worth highlighting that the score combination method outperforms the feature combination because score fusion produces better overall recognition accuracy as the results are based on the average energy of the encoded image. In addition, the score fusion has a lower complexity compared to feature fusion as the vector size is smaller [39]. In the next section, we analyze and evaluate our method using two different types of posed and spontaneous databases.

5. Experiments and results

To assess the performance of our proposed method under different conditions, we ran two experiments, the first one uses the spontaneous (involuntary) MFP database. While the second uses the posed CK database. Although our proposed method works with different types of illumination conditions, we pre-processed the images by applying the histogram equalization in order to achieve a robust illumination normalization. After that, we applied the face detection method based upon the Viola-Jones algorithm as in [40] to crop each image to 112×96 pixels. To evaluate the performance of our work under different setups, we adopted two methods for dividing our databases:

1. Hold-out method: The database is divided randomly into two sets; a completely separate training set and a hold-out test

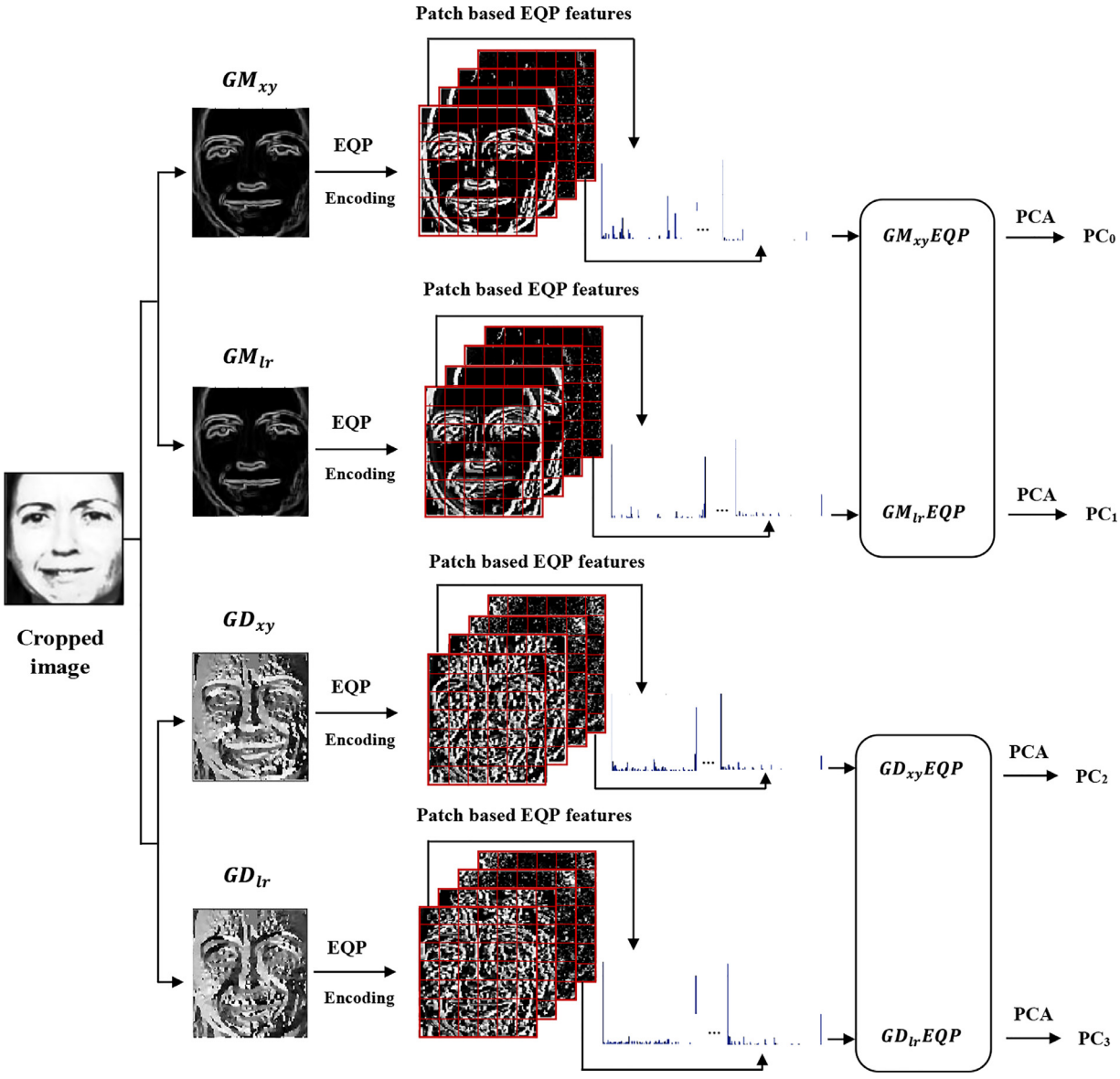

Fig. 6. The block diagram of the proposed method for feature extraction.

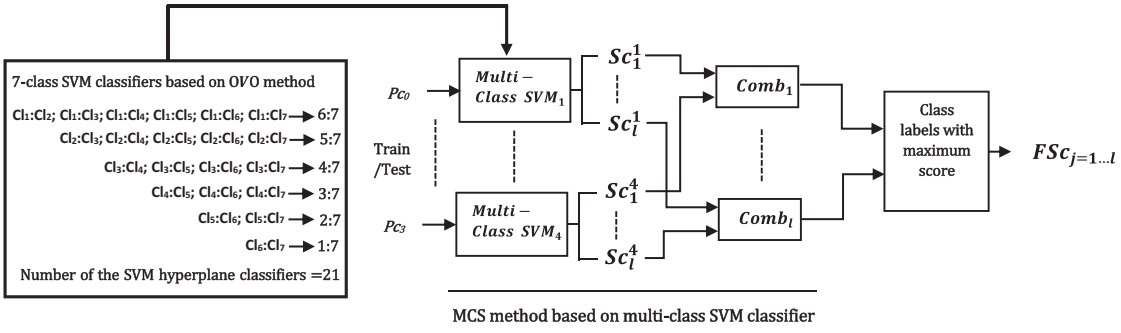

Fig. 7. Proposed final prediction scores generation using MCS method based on multi-class SVM classifier.

set in a ratio of 3/4 and 1/4, respectively. Since the training process is carried out without using the test samples, this will lead to good unbiased estimation in the prediction of the test class labels and increase the recognition accuracy. The performance of this method depends on the size of the database that is used in the evaluation process, and it is better when the number of samples in the database is high [41]. The evaluation of the recognition accuracy is calculated at one iteration and does not represent the average accuracy. Therefore, to evaluate the average accuracy of our method without loss of generalization, we also adopted the k-cross validation method.

2. k-cross validation method: this is an important tool to measure the performance of a system through evaluating the average accuracy at different iterations (e.g. k = 10). More specifically, we randomly divided the subjects' database into "k" equal size groups. After partition, one of these groups is left out as a test set and the remaining becomes a training set. This operation is repeated iteratively for each group and the recognition accuracy is calculated. Finally, the average accuracy of the overall iteration is calculated. This method is more similar to the Leave One Subject Out method (LOSO), which has been used to evaluate the average recognition accuracy by dividing the database according to the number of subjects (n), in which "n" is equal to the "k". Unfortunately, even though the LOSO method has less biasing error compared to the k-cross validation method, it has some drawbacks. One of the big drawbacks is represented by the time consumed in the validation processes. In addition, the test set used at each iteration in the LOSO is too small, because it is represented by the subject itself, and this will increase the variance error at each iteration. For these reasons, and to make a trade-off between the bias and variance errors, we selected to use the k-cross validation method to evaluate our two database types [2,10,42–44].

5.1. Experiment one (spontaneous MFP database)

In this experiment, we tested the performance of our method using the MFP database. The MFP database contains 309 subjects, each subject has a small video sequence (e.g. 10 min). Each sequence of video represents a specific type of gesture emotions collected at different subtle conditions of the spontaneous behaviors. The age range of subjects is over 18–75 years; 249 out of 309 subjects are women (80.58%), 8 subjects of women are wearing glasses (3.21%), 196 subjects of women having a single image view (78.71%) and 45 subjects of women having multiple image views (18.07%). 60 subjects are men (19.41%), 3 subjects of men are wearing glasses (5%), 55 subjects of men having a single image view (91.66%) and 2 subjects of men having multiple image views (3.33%). As we explained in Section 1, there are several challenging characteristics for any spontaneous database. Unfortunately, the MFP database is a challenging spontaneous database. To overcome most of these challenges and to increase the probability of the image selection (i.e. increasing the Recall of the system), while decreasing the distance among the selected images of each class, the expression images have been chosen manually (i.e. the most expressive images inside each video). Therefore, we used the opinions of 7 expert annotators in the selection of images from all videos in the MFP database. Then we utilized the voting among opinions of annotators to select the suitable image labels of each class. Here, we selected the wining labels if they corresponded with more than half of the opinions of annotators. Finally, we selected 2016 (288 × 7) annotated images for this experiment.

In the next subsections, we present an analysis for different parameters that affect the recognition performance.

5.1.1. Best blocks number

In this experiment, we divided each encoded image to a number of blocks, to determine the best number of blocks required to obtain the best accuracy. In our method we reported that dividing the encoded image into 5 × 5 of non-overlapped blocks may lead to a variation in the overall recognition rate. This happened because dividing the face images into 5 × 5 blocks will result in big blocks and hence lead to reduction in the ability of the descriptors to extract the micro-level information of the small curves, bulges or furrows within the important regions in the face, as explained in [7,12].

Table 1
Recognition accuracy (%) on the MFP database using different numbers of regions.

Block number	7 class %	6 class %	Features type	#features
Whole image	80.95	79.62	Histogram	372
5 × 5	89.08	90.51	PCA	168
7 × 6	**94.64**	**95.37**	**PCA**	**168**
8 × 8	91.47	93.05	PCA	168
10 × 10	88.70	88.90	PCA	168

On the other hand, dividing the encoded image into 8 × 8 or 10 × 10 of non-overlapped blocks may lead to divergence in the recognition rate. The degradation in the recognition rates comes as a result of the small size of each block, and will lead to reduce the ability of descriptors to isolate the subtle changes of the micro-level patterns. In addition, more time is required to extract the facial features. Empirical results show that dividing the encoded image into 7 × 6 of non-overlapped blocks represents the best effective range of block number, because it results in high recognition accuracy as shown in Table 1.

5.1.2. Effective PCs

In order to attain a high classification rates, we tested different numbers of PC coefficients. Hence, we tested the performance of our method with various numbers of PC coefficients (ranging from 42–216) for each vector of the encoded images (e.g. PC_0, PC_1, PC_2 and PC_3). Then, the effect of varying the PC numbers on the recognition performance was investigated before and after applying the MCS method based on the multi-class SVM classifier, as proposed in (29) and explained clearly in Table 2. Fig. 8, illustrates the effect of different ranges of the PCs on the recognition accuracy of 7 and 6 class respectively. Obviously, the selection of the range of PCs (126–168) is more efficient for different classes of the facial expression recognition, since it achieves a high recognition rate of 94.64% in the 7 class and 95.37% in the 6 class prototypes, respectively. In addition, the difference in the recognition accuracy between the 7 and 6 class is changed slightly. This happened due to the ability of our method in reducing the effect of the neutral image on each class.

5.1.3. Optimal encoding parameters

We used different gradient images that are encoded based upon variant LBP descriptors such as LBP, LTP and EQP to extract the local features. To examine the recognition rates using different encoding methods, we empirically selected different parameters that boost the encoding process and achieve a high recognition accuracy. Therefore, we chose to use $P = 10$ for encoding points that surround each center pixel, this will increase the occurrence numbers of the uniform patterns in the histogram features to 92 bins rather than 58 bins plus 1 bin for all non-uniform patterns. In addition, we selected to use an ellipse neighbourhood topology for representing points that surround each center pixel rather than a circle topology in order to extract anisotropic information that cannot be exploited by the circular topology. Moreover, we selected to use the semi major and the semi minor axis length of ellipse r_1 and r_2 to be between 2 and 3, since it was reported in [12] that increasing or decreasing the radius value of the LBP descriptor deteriorates the performance of the descriptor. Also, we tested different β angles for the ellipse topology (i.e. 0^0, 45^0, 90^0 and 135^0), and we found that β with 45^0 results with high recognition rate when associated with different types of descriptors, particularly with the EQP descriptor.

To provide consistent near-uniform regions in the gradient images, we utlized three values of the ternary code descriptor (i.e.

Table 2
The overall accuracy of different descriptors using the optimal encoding parameters.

Method	Description	Parameter values	7 class	6 class
$GM_{xy}LBP_P^{u,r_1,r_2}$	$GM_{xy}(e\_n)$ based LBP descriptor	neighborhood:Ellipse $P=10$, $r_1=3$, $r_2=2$ & $\beta=45.^0$	66.46%	66.9%
$GM_{lr}LBP_P^{u,r_1,r_2}$	$GM_{lr}(se\_sw)$ based LBP descriptor	neighborhood:Ellipse $P=10$, $r_1=3$, $r_2=2$ & $\beta=45.^0$	62.7%	65.51%
$GD_{xy}LBP_P^{u,r_1,r_2}$	$GD_{xy}(e\_n)$ based LBP descriptor	neighborhood:Ellipse $P=10$, $r_1=3$, $r_2=2$ & $\beta=45.^0$	67.06%	67.82%
$GD_{lr}LBP_P^{u,r_1,r_2}$	$GD_{lr}(se\_sw)$ based LBP descriptor	neighborhood:Ellipse $P=10$, $r_1=3$, $r_2=2$ & $\beta=45.^0$	66.46%	66.66%
MCS based SVM	**Combination: Sum**	**Kernel:HI & C=0.25**	**83.92**%	**84.25**%
$GM_{xy}LTP_P^{u,r_1,r_2}$	$GM_{xy}(e\_n)$ based LTP descriptor	neighborhood:Ellipse $P=10$, $r_1=3$, $r_2=2$, $\beta=45^0$ & $t=10$.	89.68%	89.81%
$GM_{lr}LTP_P^{u,r_1,r_2}$	$GM_{lr}(se\_sw)$ based LTP descriptor	neighborhood:Ellipse $P=10$, $r_1=3$, $r_2=2$, $\beta=45^0$ & $t=10$.	88.89%	89.58%
$GD_{xy}LTP_P^{u,r_1,r_2}$	$GD_{xy}(e\_n)$ based LTP descriptor	neighborhood:Ellipse $P=10$, $r_1=3$, $r_2=2$, $\beta=45^0$ & $t=10$.	69.24%	69.91%
$GD_{lr}LTP_P^{u,r_1,r_2}$	$GD_{lr}(se\_sw)$ based LTP descriptor	neighborhood:Ellipse $P=10$, $r_1=3$, $r_2=2$, $\beta=45^0$ & $t=10$.	73.02%	76.16%
MCS based SVM	**Combination:Sum**	**Kernel:HI & C=0.25**	**92.26**%	**93.98**%
$GM_{xy}EQP_P^{u,r_1,r_2}$	$GM_{xy}(e\_n)$ based EQP	neighborhood:Ellipse $P=10$, $r_1=3$, $r_2=2$, $\beta=45^0$, $t_1=10$ & $t_2=18$.	91.27%	91.67%
$GM_{lr}EQP_P^{u,r_1,r_2}$	$GM_{lr}(se\_sw)$ based EQP descriptor	neighborhood:Ellipse $P=10$, $r_1=3$, $r_2=2$, $\beta=45^0$, $t_1=10$ & $t_2=18$.	92.26%	92.82%
$GD_{xy}EQP_P^{u,r_1,r_2}$	$GD_{xy}(e\_n)$ based EQP descriptor	neighborhood:Ellipse $P=10$, $r_1=3$, $r_2=2$, $\beta=45^0$, $t_1=1$ & $t_2=18$.	77.58%	78.24%
$GD_{lr}EQP_P^{u,r_1,r_2}$	$GD_{lr}(se\_sw)$ based EQP descriptor	neighborhood:Ellipse $P=10$, $r_1=3$, $r_2=2$, $\beta=45^0$, $t_1=1$ & $t_2=18$.	81.2%	82.18%
MCS based SVM	**Combination:Sum**	**Kernel:HI & C=0.25**	**94.64**%	**95.37**%

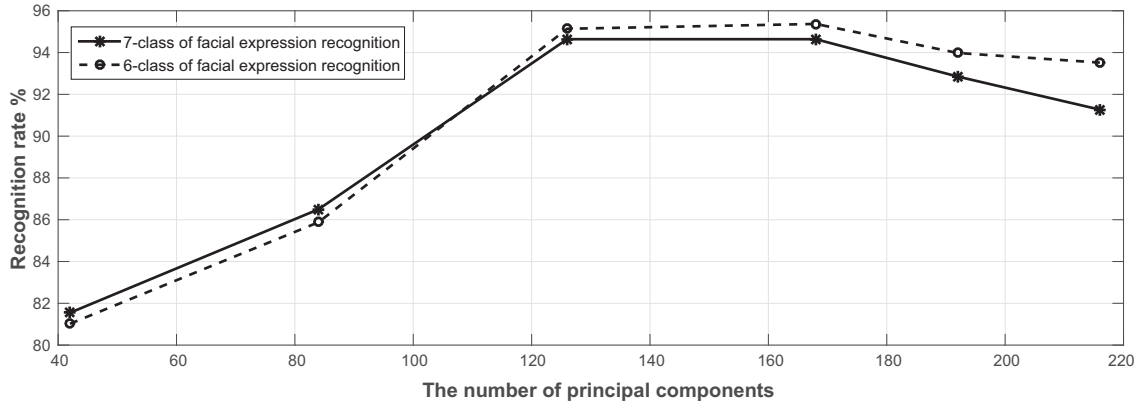

Fig. 8. Recognition accuracy versus different PCs numbers for 7 and 6 class prototypic expression.

LTP), which uses the threshold value t as a base in the encoding process. For further improvement in the uniformity of regions within different gradient images, we suggested to use five values of the quinary code descriptor (i.e. EQP), which has two threshold values t_1 and t_2 as a base in the encoding process. In previous work by Nanni *et al.* [31] it has been found that the performance in terms of accuracy is not affected greatly by the selection of t, t_1 and t_2. In our work, we also chose our threshold values t, t_1 and t_2 so as to maximize accuracy. We chose $t=10$ when we applied the LTP on both gradients of the magnitude and angle images, $t_1=10$ and $t_2=18$ when the EQP is applied on the gradient magnitude images, $t_1=1$ and $t_2=18$ when we applied the EQP on the gradient angle images. We also tried different values over the range: $5-10$ for t, $1-12$ for t_1 and $8-20$ for t_2 but saw little change in performance.

On the other hand, there are important orthogonal features that may affect and increase the recognition accuracy. Table 2 shows high recognition rates when the two gradient magnitude images

GM_{xy} and GM_{lr} (i.e. orthogonal amplitude features extracted in the *xy* and *lr* directions) are encoded based upon variant LBP descriptors. While less recognition accuracy is achieved when the other two gradient angle images GD_{xy} and GD_{lr} (i.e. orthogonal angle features extracted in the *xy* and *lr* directions) are encoded based upon variant LBP descriptors. Despite the extracted features of the GD_{xy} and GD_{lr} decreasing the recognition accuracy when encoded with variant LBP descriptors, they represent other orthogonal features that may be used together with gradient magnitude features to feed the separate classifiers with the multi-gradient of magnitude and angle features (i.e. diversity property).

Finally, it is worth mentioning that using the MCS method based on the multi-class SVM classifier increases the recognition accuracy by feeding it with different types of encoding information using different gradient images at different directions.

Table 2 compares the overall recognition accuracy of different types of the gradient images that are encoded by variant LBP descriptors using the MCS method based on the multi-class SVM classifier.

5.1.4. System evaluation

In order to measure the performance and to calculate the average recognition accuracy, we used Confusion Matrices (CMs) to obtain a full description of the recognition accuracy of each facial expression. The columns of the CM contain information about the actual class, while the rows contain the obtained labels through classification. The main diagonal containing the number of the facial expressions that are classified correctly are called True Positive (TP). Other locations in the CM correspond to misclassification errors. As mentioned earlier, two methods are used for dividing the databases namely: Hold-out and 10-cross validation.

Firstly, we used Hold-out method in which we split the MFP database (e.g. 2016 images) in a ratio of 1/4 in the testing and 3/4 in the training processes. So, we used 504 image as a test of input data, and 1512 image as training data. Hence, we used 1728 images to express the 6 class prototypical expression plus 288 images to express the neutral expression from the overall images that are selected and voted by annotators which is equal to 2016 images. Experimental results showed a superior recognition rate compared to other works as the recognition rate is 94.64% in 7 class and 95.37% in the 6 class respectively. A slight change in the overall accuracy is noticed (0.73%) when the neutral images are removed.

Secondly, to reduce the variability in the recognition accuracy, the 10-cross validation is utilized. To obtain the equal 10-folds of the MFP database, we selected 1960 images randomly. The size of each test fold is equal to 196 images (around 10% of the data of each class) and 1764 images for the 9 training folds. There are 10 iterations, in each iteration we train the SVM with 9 folds while the remaining fold is left for testing. This operation is performed 10 times for each fold while omitting the last fold at each iteration. Then the overall accuracy is measured using the average of all 10 iterations. The HI as a kernel function is selected with soft margin value ($C = 0.25$). Tables 3 and 4 represent a full description of the average recognition accuracy using 10-cross validation. The obtained result of the overall accuracy showed a high improvement of 80.31% in 7 class and 81.37% in the 6 class respectively. In addition, the recognition accuracy increased slightly by 1.064% when removing the effect of the neutral image with high performance ability in classifying for the difficult expressions such as sad expression. The obtained results in the Tables 3 and 4 reflect clearly the ability of the proposed method in reducing the effect of the neutral expression.

The neutral expression represents a major problem that causing FN errors with different facial expressions inside each class, since the neutral image contains the identity feature which may overlap with different facial expressions and causing errors. In addition, the proposed method outperforms all the related works that used the MFP database. The superiority of our method is represented by exploring the subtle changes of the different spontaneous facial expressions and extracted it. Tables 3 and 4, show the performance of the proposed method in terms of CMs using 7 and 6 classes, respectively.

Table 3
The CM of 7 class expression recognition (%) using MCS method based on the multi-class SVM classifier applied for recognizing spontaneous MFP database using 10-cross validation method.

Expressions	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	**83.57**	3.93	2.14	0.0	4.64	4.64	1.07
Disgust	5.36	**72.86**	5.71	6.79	2.5	3.21	3.57
Fear	8.21	6.43	**74.64**	0.36	3.21	0.71	6.43
Happy	0.0	3.21	0.36	**95.36**	0.36	0.0	0.71
Neutral	2.5	0.0	2.14	0.0	**84.64**	5.0	5.71
Sad	6.79	3.57	5.0	1.43	8.57	**69.64**	5.0
Surprise	2.86	2.86	2.50	1.79	5.71	2.86	**81.43**

Table 4
The CM of 6 class expression recognition (%) using MCS method based on the multi-class SVM classifier applied for recognizing spontaneous MFP database using 10-cross validation method.

Expressions	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**85.36**	3.93	2.86	0.0	6.43	1.43
Disgust	5.71	**74.29**	7.14	6.07	3.93	2.86
Fear	9.29	6.43	**73.93**	0.36	2.50	7.50
Happy	0.0	3.21	0.71	**95.0**	0.0	1.07
Sad	8.21	3.93	4.64	0.71	**76.07**	6.43
Surprise	3.21	2.86	2.86	2.14	5.36	**83.57**

In order to validate the average recognition accuracy using 10-cross validation, we utilized different RBF kernel parameters for the SVM classifier, in which we selected to use γ equal to 0.1 with hard margin value equal to C=1000. It is noted that a slight improvement in the overall recognition accuracy which is equal to 82.96% in 7 class and 82.86% in 6 class, respectively.

5.2. Experiment two (posed CK database)

The proposed method not only works with the spontaneous database (MFP) but also it works with posed database such as the CK. The CK database consists of 100 of University students (18–30 years old) from different ethnicities. Around 65% were female, 15% were African-American samples and 3% were Asian/Latino origins. We selected 408 image sequences from 96 subjects; each subject has a short video sequence (e.g. 10 min), which represents one label from the basic six expressions. Hence, for the 6 class prototypical expression, we selected the last three images to be the most expressive image frames taken from each sequence and therefore obtained 1224 expression images. We added the first image from each sequence to express the 408 neutral images which resulted in 1632 images for the 7 class prototypical expression.

First, we adopted the Hold-out method in which we divided our database (e.g. 1632 images) in a ratio of 1/4 in the testing and 3/4 in the training processes. Then, we selected 406 images out of 408 images, around (25%) are used as a test of input data (i.e. 58 images/expression), and 1224 images (75%) are used as a training data (i.e. 175 images/expression). Experimental results showed high recognition rate of 98.13% in the 7 class case and 99.63% in the 6 class case respectively.

Second, to evaluate the average of the recognition accuracy, we adopted the 10-cross validation method. To obtain the equal size of the 10-folds of the CK database, we selected 1610 images randomly. The size of each test fold is equal to 161 images and 1449

Table 5
The CM of 7 class expression recognition (%) using MCS method based on the multi-class SVM classifier applied for recognizing posed CK database using 10-cross validation method.

Expressions	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	**100**	0.0	0.0	0.0	0.0	0.0	0.0
Disgust	0.0	**99.44**	0.0	0.0	0.56	0.0	0.0
Fear	0.56	0.0	**99.44**	0.0	0.0	0.0	0.0
Happy	0.0	0.0	0.0	**98.89**	0.0	1.11	0.0
Neutral	0.0	0.0	0.0	0.0	**95.65**	4.35	0.0
Sad	0.56	0.0	0.0	0.0	1.11	**98.33**	0.0
Surprise	0.0	0.0	0.0	0.0	0.56	0.0	**99.44**

Table 6
The CM of 6 class expression recognition (%) using MCS method based on the multi-class SVM classifier applied for recognizing posed CK database using 10-cross validation method.

Expressions	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**100**	0.0	0.0	0.0	0.0	0.0
Disgust	0.0	**100**	0.0	0.0	0.0	0.0
Fear	1.11	0.0	**98.89**	0.0	0.0	0.0
Happy	0.0	0.0	0.0	**100**	0.0	0.0
Sad	0.56	0.0	0.0	0.0	**99.44**	0.0
Surprise	0.0	0.0	0.0	0.0	0.0	**100**

images for the 9 training folds. These results are obtained after using HI as a kernel function with soft margin value C=0.25. The results show a high recognition rate of 99.36% in 7 class and 99.72% in the 6 class respectively. Better results are obtained again when removing the effect of the neutral image.

Tables 5 and 6, show the performance of the proposed method in terms of CMs using 7 and 6 classes, respectively.

5.3. Time analysis

Although our method outperforms previous state-of-the-art techniques, this is achieved at the price of the increased complexity which we next quantify. This complexity is represented by the time consumed during the extraction of the orthogonal features of each pixel in the input image. In addition, using the MCS method based upon the multi-class SVM may increase the consumed time in the training processes.

Table 7 shows the comparisons in the time among different gradient images that are encoded based upon the variant LBP descriptors for extracting the feature of the single facial expression image and the average consumed time in the training and testing processes of the spontaneous MFP database. It is clearly shown that using gradient images such as GM or GD based upon the EQP descriptor for extracting the feature may increase the consumed time slightly compared to other descriptors.

The time in Table 7 was produced by using unoptimized MATLAB R2014b code on a desktop with dual Core CPU running at 2.4 GHz. It is worth noting that the consumed time of every descriptor in Table 7 is considered for the *xy* or *lr* direction only.

Table 7
The consumed time in seconds for extraction the feature of single facial expression image and the average consumed time in the training and testing phases using the MFP database.

Descriptor	Feature extraction time	Training time	Test time
GMLBP	0.24	0.125	0.0135
GDLBP	0.46	0.138	0.0153
GMLTP	0.42	0.252	0.0270
GDLTP	0.62	0.284	0.0315
GMEQP	**0.51**	**0.450**	**0.0495**
GDEQP	**0.72**	**0.531**	**0.0522**

5.4. Comparisons with related works

In terms of superiority, we made a fair comparison in the overall recognition accuracy among different methods in the state-of-the-art that used spontaneous MFP and posed CK databases as shown in Table 8. The performances of these methods are cited from the original references directly. Wan and Aggarwal in [16] and [19] produced relatively low recognition rates which are equal to 70.8% and 72.27% in the 7 class case using the MFP database. This relatively low recognition rate comes from the failure in detecting the challenging expressions such as anger, disgust and fear where the TP rates are 34.2%, 60.4% and 40.4%, respectively. This is because the inclusion of the neutral expression results in high error rates with some types of these expressions (20.4% of FN). To highlight the value of our approach, we used the MFP database and compared the performance with the relevant approaches that were used previously to classify posed facial expression, such as LDP [7], LDN [11], GDP [28] and GMLTP [8]. We found that the LDP, LDN and GDP methods produced low recognition rates due to the limitation of these descriptors in detecting the texture features of the spontaneous facial expression. This is possibly due to overlap in feature space when considering the spontaneous facial expression image or the limited expressivity in the spontaneous facial image compared with a deliberate (posed) facial expression image. However, the GMLTP descriptor improved the recognition rates of the spontaneous facial expression recognition slightly higher than the LDP, LDN and GDP methods to 73.64% and 75.78% for the 7 and 6 class cases, respectively, since the LTP descriptor is encoded on the gradient magnitude image in the *xy* direction. Unfortunately, the effect of the neutral image is still high. Nonetheless, our method outperforms all the aforementioned methods through improving the overall recognition rates to 80.31% and 81.37% for the 7 and 6 classes, respectively. In addition, we reduced the effect of the neutral image significantly in the different classes.

On the other hand, we tested and compared the performance of our approach and the aforementioned approaches that were utilized to classify the posed CK database such as LDP [7], LDN [11], GDP [28] and the GLTP in [8]. In addition, the comparison included the works of [16] and [45]. Although some good results were reported in the 6 class facial expression, the recognition accuracy is reduced in the 7 class recognition. For example, the recognition accuracy of methods [11] and [8] are dropped down by 5% and 5.5%, respectively, when the neutral images are added. It can be seen from Table 8 that our method achieved better performance compared to the aforementioned works where the overall recognition rates are 99.36% and 99.72% for the 7 and 6 class cases, respectively.

For all the aforesaid methods, we can say that the key development of our approaches superiority is represented by constructing and feeding of separate classifiers for the orthogonal gradient fea-

Table 8
Comparison with different methods that are used the spontaneous MFP and the posed CK databases. The references cited in this table are [7,8,11,16,19,28,45].

	Method	7 class %	6 class %	Measure	Image selection	Database type	Subjects	Images
[7]	LDP	69.13	69.64	7-fold	Manual annotation	Spontaneous MFP database	309	2016
[11]	LDN	60.66	63.39	10-fold				
[28]	GDP	44.64	46.19	10-fold				
[8]	GLTP	73.64	75.78	10-fold				
[16]	ML based KNN	70.8	n/a	10-fold				
[19]	Robust ML	72.27	n/a	10-fold	Manual/Automatic annotation			
Ours	Multi-gradient Features and EQP Encoder	94.64	95.37	Hold-out	Manual annotation			
		80.31	81.37	10-fold				
[7]	LDP	93.4	96.4	7-fold	Manual annotation	Posed CK database	96	1632
[11]	LDN	94.8	99.2	10-fold				
[28]	GDP	91.6	95.9	10-fold				
[8]	GLTP	91.7	97.2	10-fold				
[16]	ML based KNN	89.4	n/a	10-fold				
[45]	Bag of Words and PHOG Descriptors	n/a	96.33	LOSO				
Ours	Multi-gradient Features and EQP Encoder	98.13	99.63	Hold-out				
		99.36	99.72	10-fold				

tures (magnitude and angle) at different orientations using the five levels encoding method.

6. Conclusions

A robust method for facial expression recognition capable of working with both spontaneous and posed images was proposed. In this paper, we applied the Sobel filter of eight masks on the facial images and derived the gradient components of each pixel in the image. Therefore, each pixel in the facial image could be represented by four values, the first two values, represented the magnitude components, the other two represented angle components from $[0,\pi]$ directions. The diversity property of these components enabled us to extract multiple magnitude and angle features of each pixel in the facial image. Here, we extracted the gradient of magnitude and angle images in different directions such as: GM_{xy}, GM_{lr}, GD_{xy} and GD_{lr}. Moreover, to provide more consistency in the uniform and near uniform regions, we utilized the EQP descriptor on each point in the gradient image. Finally, we adopted the MCS based on the multi-class SVM classifier to combine the different scores that are generated after applying the EQP encoder for each image individually. Our proposed approach has been validated on the spontaneous MFP and posed CK databases and achieved promising results compared to the state-of-the-art. The proposed method improved the ability of solving major problems that come with the subtle changes in the spontaneous database and alleviate it, particularly the overlapped problem in some correlated expressions.

In our future work we will design an automatic facial expression recognition system that selects the most expressive images in each video automatically. This will reduce the dubious images in the training set of each class and increase the sensitivity (*Recall*) of the system.

Acknowledgments

The first three authors would like to thank the Ministry of Higher Education and Scientific Research (MoHESR) in Iraq for supporting their work. Also, the first author would like to thank Dr. Loris Nanni for his help.

References

[1] P. Ekman, Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life, Macmillan, 2007.
[2] I. Kotsia, S. Zafeiriou, I. Pitas, Texture and shape information fusion for facial expression and facial action unit recognition, Pattern Recognit. 41 (3) (2008) 833–851.
[3] M.F. Valstar, M. Pantic, Fully automatic recognition of the temporal phases of facial actions, IEEE Trans. Syst. Man Cybern. Part B (Cybern.) 42 (1) (2012) 28–43.
[4] B. Fasel, F. Monay, D. Gatica-Perez, Latent semantic analysis of facial action codes for automatic facial expression recognition, in: Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, ACM, 2004, pp. 181–188.
[5] G.C. Littlewort, M.S. Bartlett, K. Lee, Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain, in: Proceedings of the 9th International Conference on Multimodal Interfaces, ACM, 2007, pp. 15–21.
[6] T.F. Cootes, G.J. Edwards, C.J. Taylor, A comparative evaluation of active appearance model algorithms., in: BMVC, 98, 1998, pp. 680–689.
[7] T. Jabid, M.H. Kabir, O. Chae, Robust facial expression recognition based on local directional pattern, ETRI J. 32 (5) (2010) 784–794.
[8] F. Ahmed, E. Hossain, Automated facial expression recognition using gradient-based ternary texture patterns, Chi. J. Eng. 2013 (2013) 8–15.
[9] T.H. Zavaschi, A.S. Britto, L.E. Oliveira, A.L. Koerich, Fusion of feature sets and classifiers for facial expression recognition, Expert Syst. Appl. 40 (2) (2013) 646–655.
[10] W. Gu, C. Xiang, Y. Venkatesh, D. Huang, H. Lin, Facial expression recognition using radial encoding of local Gabor features and classifier synthesis, Pattern Recognit. 45 (1) (2012) 80–91.
[11] A.R. Rivera, J.R. Castillo, O.O. Chae, Local directional number pattern for face analysis: face and expression recognition, IEEE Trans. Image Process. 22 (5) (2013) 1740–1752.
[12] T. Wu, N. Butko, P. Ruvolo, J. Whitehill, M. Bartlett, J.R. Movellan, Multilayer architectures for facial action unit recognition, IEEE Trans. Syst. Man Cybern. Part B (Cybern.) 42 (4) (2012) 1027–1038.
[13] S. Al-Sumaidaee, S. Dlay, W. Woo, J. Chambers, Facial expression recognition using local Gabor gradient code-horizontal diagonal descriptor, in: 2nd IET International Conference on Intelligent Signal Processing 2015 (ISP), 2015, pp. 1–6.
[14] H.-Y. Chen, C.-L. Huang, C.-M. Fu, Hybrid-boost learning for multi-pose face detection and facial expression recognition, Pattern Recognit. 41 (3) (2008) 1173–1185.
[15] P.J. Naab, J.A. Russell, Judgments of emotion from spontaneous facial expressions of new guineans., Emotion 7 (4) (2007) 736.
[16] S. Wan, J. Aggarwal, A scalable metric learning-based voting method for expression recognition, in: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, IEEE, 2013, pp. 1–8.
[17] J.F. Cohn, K.L. Schmidt, The timing of facial motion in posed and spontaneous smiles, Int. J. Wavelets Multiresolut. Inf. Process 2 (02) (2004) 121–132.
[18] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 39–58.

[19] S. Wan, J. Aggarwal, Spontaneous facial expression recognition: a robust metric learning approach, Pattern Recognit. 47 (5) (2014) 1859–1868.
[20] A.J. O'Toole, J. Harms, S.L. Snow, D.R. Hurst, M.R. Pappas, J.H. Ayyad, H. Abdi, A video database of moving faces and people, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 812–816.
[21] A. Moilanen, G. Zhao, M. Pietikäinen, Spotting rapid facial movements from videos using appearance-based feature difference analysis, in: Pattern Recognition (ICPR), 2014 22nd International Conference on, IEEE, 2014, pp. 1722–1727.
[22] A.K. Davison, M.H. Yap, C. Lansley, Micro-facial movement detection using individualised baselines and histogram-based descriptors, in: Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on, IEEE, 2015, pp. 1864–1869.
[23] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1, IEEE, 2005, pp. 886–893.
[24] S.L. Happy, P. Patnaik, A. Routray, R. Guha, The Indian spontaneous expression database for emotion recognition, 2015.
[25] K. Takahashi, S. Takahashi, Y. Cui, M. Hashimoto, Remarks on computational facial expression recognition from hog features using quaternion multi-layer neural network, in: International Conference on Engineering Applications of Neural Networks, Springer, 2014, pp. 15–24.
[26] M.M. Donia, A.A. Youssif, A. Hashad, Spontaneous facial expression recognition based on histogram of oriented gradients descriptor, Comput. Inf. Sci. 7 (3) (2014) 31.
[27] N.-S. Vu, A. Caplier, Enhanced patterns of oriented edge magnitudes for face recognition and image mmatching, IEEE Trans. Image Process. 21 (3) (2012) 1352–1365.
[28] F. Ahmed, Gradient directional pattern: a robust feature descriptor for facial expression recognition, Electron. Lett. 48 (19) (2012) 1203–1204.
[29] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[30] X. Tan, B. Triggs, Enhanced local texture feature sets for face recognition under difficult lighting conditions, IEEE Trans. Image Process. 19 (6) (2010) 1635–1650.
[31] L. Nanni, A. Lumini, S. Brahnam, Local binary patterns variants as texture descriptors for medical image analysis, Artif. Intell. Med. 49 (2) (2010) 117–125.
[32] S. Liao, A.C. Chung, Face recognition by using elongated local binary patterns with average maximum distance gradient magnitude, in: Asian Conference on Computer Vision, Springer, 2007, pp. 672–679.
[33] H.-T. Nguyen, Contributions to Facial Feature Extraction for Face Recognition, Université Grenoble Alpes, 2014 Ph.D. thesis.
[34] A. Barla, F. Odone, A. Verri, Histogram intersection kernel for image classification, in: Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on, 3, IEEE, 2003, pp. III–513.
[35] B. Jiang, B. Martinez, M.F. Valstar, M. Pantic, Decision level fusion of domain specific regions for facial action recognition, in: Pattern Recognition (ICPR), 2014 22nd International Conference on, IEEE, 2014, pp. 1776–1781.
[36] C.-W. Hsu, C.-J. Lin, A comparison of methods for multiclass support vector machines, IEEE Trans. Neural Netw. 13 (2) (2002) 415–425.
[37] L.A. Alexandre, A.C. Campilho, M. Kamel, On combining classifiers using sum and product rules, Pattern Recognit. Lett. 22 (12) (2001) 1283–1289.
[38] J. Kittler, M. Hatef, R.P. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.
[39] M.A. Abdullah, S.S. Dlay, W.L. Woo, J.A. Chambers, A novel framework for cross-spectral iris matching, IPSJ Trans. Comput. Vis. Appl. 8 (1) (2016) 9.
[40] P. Jones, P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: University of Rochester. Charles Rich, Citeseer, 2001.
[41] R. Kohavi, et al., A study of cross-validation and bootstrap for accuracy estimation and model selection, in: IJCAI, 14, Stanford, CA, 1995, pp. 1137–1145.
[42] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, 6, Springer, 2013.
[43] P.-N. Tan, et al., Introduction to data mining, 2006.
[44] Y. Sun, G. Wen, Cognitive facial expression recognition with constrained dimensionality reduction, Neurocomputing 230 (2017) 397–408.
[45] Z. Li, J.-i. Imai, M. Kaneko, Facial expression recognition using facial-component-based bag of words and PHOG descriptors, J. Inst. Image Inf. Telev. Eng. 64 (2) (2010) 230–236.

S. A. M. Al-Sumaidaee received the B.Sc. and M.Sc. degrees in Computer Engineering from the University of Technology in 1995 and 2001, respectively. Since 2003, he worked as a Lecturer at Department of Computer and Software Engineering, College of Engineering, at Al-Mustansiriya University. Currently, he is studying toward the Ph.D degree in the School of Electrical and Electronic Engineering at Newcastle University, UK. His research interests are in the fields of facial expression recognition, analysis and image processing. Mr. Al-Sumaidaee is a student member of IET.

M. A. M. Abdullah received the B.Sc. and M.Sc. degrees in Computer Engineering in 2008 and 2010, respectively. During 2010, he worked as a research assistant in Electrical and Electronic Engineering department at University of Liverpool, UK. Currently, he is studying toward the Ph.D degree in the School of Electrical and Electronic Engineering at Newcastle University, UK. His research interests are in the fields of pattern recognition and image processing with emphasis on iris recognition.

R. R. O. Al-Nima received the B.Sc. and M.Sc. degrees in Technical Computer Engineering in 2000 and 2006, respectively. During 2006, he worked as an Assistant Lecturer in the Technical College of Mosul, Iraq. In 2011, he obtained the Lecturer scientific title in the same college. Currently, he is studying toward the Ph.D degree in the School of Electrical and Electronic Engineering at Newcastle University, UK. His research interests are in the fields of pattern recognition, security, artificial intelligence and image processing. Mr. Al-Nima is a student member of IET.

S. S. Dlay received the B.Sc. (Hons.) degree in Electrical and Electronic Engineering and the Ph.D from the Newcastle University, UK in 1979 and 1983 respectively. He is currently a full professor with the School of Electrical and Electronic Engineering at Newcastle University. He has published over 250 research papers ranging from biometrics and security, biomedical signal processing and implementation of signal processing architectures. He serves on many editorial boards and has played an active role in numerous international conferences in terms of serving on technical and advisory committees as well as organizing special sessions. Prof. Dlay is a College Member of the EPSRC.

J. A. Chambers received the Ph.D degree in signal processing from the Imperial College London, UK, in 1990. He currently heads the Communications, Sensors, Signal and Information Processing Group in the school of Electrical and Electronic Engineering at Newcastle University, UK. He has published more than 400 conference and journal articles. His research interests include adaptive and blind signal processing and their applications. Prof. Chambers is a Fellow of the Royal Academy of Engineering, UK, and the Institution of Electrical Engineers. He is also a Fellow of the IET and Institute of Mathematics and its Applications (IMA).