A Dynamic Framework Based on Local Zernike Moment and Motion History Image for Facial Expression Recognition

Xijian Fan, Tardi Tjahjadi

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

Abstract

A dynamic descriptor facilitates robust recognition of facial expressions in video sequences. The current two main approaches to the recognition are basic emotion recognition and recognition based on facial action coding system (FACS) action units. In this paper we focus on basic emotion recognition and propose a spatio-temporal feature based on local Zernike moment in the spatial domain using motion change frequency. We also design a dynamic feature comprising motion history image and entropy. To recognise a facial expression, a weighting strategy based on the latter feature and sub-division of the image frame is applied to the former to enhance the dynamic information of facial expression, and followed by the application of the classical support vector machine. 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-arts methods, the proposed framework demonstrates a superior performance.

Keywords: Zernike moment, facial expression, motion history image, entropy, feature extraction.

1. Introduction

In recent years facial expression recognition has become a popular research topic [1, 2, 3]. With the recent advances in robotics, and as robots interact more and more with human and become a part of human living and work space, there is an increasing requirement that robots are able to understand human emotions via a facial expression recognition system [4]. Facial expression recognition system also plays a significant role in Human-Computer Interaction (HCI) [5], which has helped to create meaningful and responsive HCI interfaces. It has also been widely used in behavioural study, video games, animations, safety mechanism in auto-mobile, etc. [6].

Discriminative and robust features that represent facial expressions are important for effective recognition of facial expressions, and how to obtain them is still a challenging problem. Recent methods that address this problem can be categorised into global-based methods and local-based methods. It has been shown that local-based methods (e.g., based on Gabor wavelets using grid points) achieve better performance than the global-based ones (e.g., based on eigenfaces, Fisher's discriminant analysis, etc.) [7]. Gabor wavelet results in good performance due to its locality and orientation selectivity. However, its computational complexity

requiring high computational time makes it unsuitable for real-time applications. Local Binary Pattern (LBP) descriptor which is based on the histogram of local patterns also achieves a promising performance [8].

Shape as a geometric-based representation is crucial for interpreting facial expressions. However, current state-of-the-art methods only focus on a small subset of possible shape representation, e.g., point-based methods that represent a face using the locations of several discrete points. Noting that image moments can describe simple properties of a shape, e.g., its area (or total intensity), its centre and its orientation, Zernike moments (ZMs) have been used to represent a face and facial expressions in [9, 10]. Zernike moments are rotation invariant features, which can be used to address in-plane head pose variation. In the field of facial expression recognition, rotation invariant LBP and uniform LBP [11] have also been used to overcome the rotation problem. In [12], Quantised Local Zernike Moment (QLZM) is used to describe the neighbourhood of a face sub-region. The Local Zernike moments have more discriminant power than other image features, e.g., local phase-magnitude histogram(H-LZM), cascaded LZM transformation (H-LZM^2) and local binary pattern (LBP) [13].

Since a facial expression involves a dynamic process, and the dynamics contain information that represents a facial expression more effectively, it is important to capture such dynamic information so as to recognise facial expressions over the entire video sequence. Recently, there has been more effort on modelling the dynamics of a facial expression sequence. However, the modelling is still a challenging problem. Thus, in this paper, we focus on analysing the dynamics of facial expression sequences. First, we extend the spatial domain QLZM descriptor into spatio-temporal domain, i.e., Motion Change Frequency based QLZM (QLZM_MCF), which enables the representation of temporal variation of expressions. Second, we apply optical flow to Motion History Image (MHI) [14], i.e., (optical flow based MHI) MHI_OF, to represent spatial-temporal dynamic information (i.e., velocity).

We utilise two types of features: a spatio-temporal shape representation, QLZM_MCF, to enhance the local spatial and dynamic information, and a dynamic appearance representation, MHI_OF. We also introduce an entropy-based method to provide spatial relationship of different parts of a face by computing the entropy value of different sub-regions of a face. The main contributions of this paper are: (a) QLZM_MCF; (b) MHI_OF; (c) an entropy-based method for MHI_OF to capture the motion information; and (d) a strategy integrating QLZM_MCF and entropy to enhance spatial information.

The rest of the paper is organised as follows. Previous related work is presented in Section 2. Section 3 presents QLZM_MCF, the method using MHI_OF and entropy, and the intergration of the two dynamic features. The framework and the experimental results are respectively presented in Section 4 and Section 5. Finally, Section 6 concludes the paper.

2. Related Work

The two main focuses in the current research on facial expression are basic emotion recognition and recognition based on facial action coding system (FACS) action units (AUs). The most widely used facial expression descriptors for recognition and analysis are the six prototypical expressions of Anger, Disgust, Fear, Happiness, Sadness and Surprise [15]. The most widely used facial muscle action descriptors are AUs [1]. With regard to basic emotion recognition, geometric-based features and appearance-based features are most widely used.

Geometric-based methods rely on the locations of a set of fiducial facial points [16, 17], a connected face mesh [18, 19], or the shapes of face components [20]. The commonly used geometric representation is facial points, which represent a face by concatenating the x and y coordinates of a number of fiducial points. Alternative shape representations include the distances between facial landmarks, distance and angle that represent the opening/closing of the eyes and mouth, and groups of points that describe the state of the cheeks. Although it has been shown that shape representation plays a vital role for analysing facial expressions, they have not been exploited to their full potential [12].

Image moments can be categorised into geometric moments, complex moments and orthogonal moments. Although easy to use, the large values of geometric moments are their main limitations leading to numerical instabilities and sensitivity to noise. Complex moments are defined similarly to geometric and have been used to describe the shape of a probability density function and to measure the mass distribution of a body. Hu moments exhibit translation, rotation and scaling invariance, and has been applied in many areas [21]. Orthogonal moments are projections of a function onto a polynomial basis. ZMs employ complex Zernike polynomials as its moment basis set [22], and have been used to recognise facial expressions [23]. The rotation invariance of Zernike-based facial features is discussed in [9, 10]. QLZM is used in [12] for recognising facial expressions. However, ZM has its shortcomings, namely it is a low level histogram representation which ignores the spatial relations (i.e., configure information) among the different facial parts. Also, ZMs only describe the texture information in each frame of an image sequence, and do not capture any dynamic information. In this paper, we address these two limitations by extending QLZM to spatio-temporal in order to extract dynamic information, and introducing an entropy to incorporate spatial relations.

The appearance-based methods try to find a more effective and robust way to represent appearance feature including skin motion and texture changes (i.e., deformation of skin) such as bulges, wrinkle and furrows. Transformations and statistical methods are used to determine the feature vectors that represent textures and are thus simple to implement. Gabor wavelets [24] and LBPs [25] are two representative feature vectors of such an approach that describe the local appearance models of facial expressions. Gabor magnitudes 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. LBP 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. The most popular LBP is the uniform LBP [26]. LBP has been extended to spatio-temporal domain so as to utilise the dynamics information, which results in a significant improvement in recognition rate [27]. One drawback of appearance-based approach is that it is difficult to generalise appearance features across different persons.

A Dynamic Texture (DT) is a spatially repetitive, time-varying visual pattern that forms an image sequence with certain temporal stationarity [28]. MHI applied to the recognition of DT can be used to address the problem of facial expression recognition [29]. MHI decomposes motion-based recognition by first describing where there is motion (i.e., the spatial pattern) and then describing how the object is moving [14], where the temporal information can be retained by eliminating one dimension. One of the advantages of MHI is that a range of times may be encoded in a single frame, and in this way, the MHI spans the time scale of the human motion. In MHI, the intensity value of each image pixel denotes the recent movement, ignoring the speed of the movement. However, speed can be used to distinguish the movement of some facial parts (e.g., opening of mouth and raising of eyebrows) and the movements caused by changes of in-plane head pose or relatively stable facial parts (e.g., cheek, nose, forehead, etc.) during facial expressions. Optical flow has been used to capture the velocity of movement at pixels in an image, but by computing the changes in pixel intensities between two consecutive frames it does not accurately describe the entire video sequence. We address the limitations of MHI and optical flow by combining them so as to incorporate speed and to enable more distinct representations of movement of different facial parts.

Entropy-based methods extract intensity information of image pixels, and have been applied for face recognition. For example, Cament et al. [30] combined entropy-like weighted Gabor features with the local normalisation of Gabor features. Chai et al. [31] introduced the entropy of a facial region, where a low entropy value means the probabilities of different intensities are different, and a high value means the probabilities are the same. They used the entropy of each of the equal-size blocks of a face image to determine the number of sub-blocks within each block. Inspired by [31], we use entropy in the proposed MHI_OF as follows. Since the intensity value of each pixel in MHI represents a movement, the high intensity values denoting large movement will result in high entropy value, and vice versa.

3. Feature extraction

3.1. Motion History Image

MHI can be considered as a two-component temporal template, a vector-valued image where each component of each pixel is some function of the motion at that pixel location. The MHI $H_\tau(x, y, t)$ is computed

from an update function $\Psi(x, y, t)$, i.e.,

$$H_\tau(x, y, t) = \begin{cases} \tau, & \Psi(x, y, z) = 1 \\ \max(0, H_\tau(x, y, t) - \delta), & \text{otherwise} \end{cases} \tag{1}$$

where (x, y, t) is the spatial coordinates (x, y) of an image pixel at time t (in terms of image frame number), the duration τ determines the temporal extent of the movement in terms of frames, and δ is the decay parameter. $\Psi(x, y, z)$ is defined as

$$\Psi(x, y, z) = \begin{cases} 1, & D(x, y, t) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $D(x, y, t)$ is a binary image comprising pixel intensity differences of frames separated by temporal distance Δ, i.e.,

$$D(x, y, z) = |I(x, y, t) - I(x, y, t \pm \Delta)| \tag{3}$$

and $I(x, y, t)$ is the intensity value of pixel with coordinates (x, y) at the tth frame of the image sequence. The duration τ and the decay parameter δ have an impact on the MHI image. If τ is smaller than the number of frames, then the prior information of the motion in its MHI will be lost. For example, when $\tau = 10$ for a sequence with 19 frames, the motion information of the first 9 frame will be lost if the value of $\delta = 1$. On the other hand, if the temporal duration is set at very high value compared to the number of frames, then the changes of pixel value in the MHI is less significant. The MHI of a sequence from the Extended CK dataset (CK+) [32] is shown in Fig. 1.

Figure 1: Example of images from sequences (left and middle) and its MHI (right).

3.2. Optical Flow Algorithm

Optical flow descriptor can represent the velocity of a set of individual pixels in an image, which capture their dynamic information. We employ optical flow descriptor in our framework to exploit velocity information.

The Lucas-Kanade method is one of most widely-used method for optical flow computation [33], which solves basic optical flow equation for all pixels in their local neighbourhood by using the least squares criterion. Given two consecutive image 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 the d which minimises the match error between the local appearances of two corresponding points. A cost function is defined for the local area $R(p)$, i.e., [33]

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

where $w(x)$ is a weights window, which assigns larger weight to pixels that are closer to its central pixel as these pixels are considered to contain more important information than those further away.

3.3. Optical Flow based MHI (MHI_OF)

In [34], Tsai et al. proposed a representation that incorporates both optical flow and a revised MHI for action recognition, which can better describe local movements. Since a video sequence of facial expression involves local movements of different facial parts, we consider applying this representation into spatio-temporal facial expression recognition. As according to [34], we compute the optical flow between two consecutive frames and obtain the optical flow image where the intensity of each pixel represents the magnitude of the optical flow descriptor. The higher values denote the faster movement of facial points. We define MHI_OF of a sequence as

$$M(x, y, t) = d(x, y, t) + M(x, y, t - 1) * \bar{\tau} \tag{5}$$

where $\bar{\tau}$ is another decay parameter, and

$$d(x, y, t) = \begin{cases} a * d(x, y, t) + b & d(x, y, t) > \mathrm{T} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

a and b are scale factors, and T is a threshold which is used to remove small movements, while retaining large movements of some fiducial points (e.g., eyebrows, lips, etc.). Scale factors are used because the optical flow descriptor is not significantly large for the movements of points in two consecutive frames. In our experiments, the original optical flow $d(x, y, t)$ is magnified by a scale factor a of 10 with a starting value b of 20, and the threshold T is set to 1. A large value of the decay parameter $\bar{\tau}$ creates a slow decrement of the accumulated motion strength, and the long-term history of the motion is recorded in the resulting MHI_OF image. A small value of $\bar{\tau}$ gives an accelerated decrement of motion strength, and only the recent short-term movements are retained in the MHI_OF image. Fig. 2 illustrates optical flow based MHI for some facial expressions.

3.4. Entropy

The entropy of a discrete random variable X with possible values $\{x_0, x_1, ...x_2, x_N\}$ can be defined as [35]

$$E(X) = -\sum_{i=0} p(x_i) \times log_2(p(x_i)), \tag{7}$$

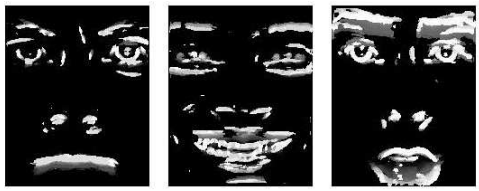

Figure 2: Optical flow based MHI for Anger, Happiness and Surprise (from left to right).

where $p(.)$ is the probability function. For a grey-level face image, the intensity value of each pixel varies from $0-255$, and the possibility of a particular value occurring is random and varies depending on the pattern of different face images. Considering a face image with dimension $H \times W$ having a total of $M = H \times W$ pixels, the probability of a particular intensity value x_i occurring in the image is $p(x_i) = n_i/M$, where n_i is the number of occurrences of x_i among the M pixels. In this case, considering $\Sigma_i n_i = M$, the entropy of the image can be expressed as

$$E(X) = log_2 M - \frac{1}{M} \times \sum_{i=0}^{255} n_i \times log_2(n_i). \tag{8}$$

It is shown in [37] that certain facial regions contain more important information for recognising facial expressions than others. For example the regions of mouth and eyes that produce more changes than those of nose and forehead during an expression have more contribution towards the recognition. Also, as can be seen from the leftmost and middle columns of Fig. 3, different facial regions in MHI have different intensity levels due to the distance and speed of movements during an expression. Thus, introducing a weight function which allocates different weights to different facial regions will improve recognition. Instead of setting weights empirically based on the observation, we utilise entropy to determine the weights as it is expected that the entropy at different facial regions will differ significantly due to pixel intensity variation at these regions.

The size of the training samples in practice is often not large enough to cover all the possible values of pixels in MHI. To address this sparse problem, we divide the possible 256 intensity levels into several sections to form intensity divisions. For a 2-dimensional (2D) matrix $X = (x_{ij})_{H \times W}$, let $\chi = \{t_1, t_2, ..., t_K\}$ be the sorted set of all possible K intensity values that exist in X where $t_1 < t_2 < t_3 ... < t_K$ and K is the

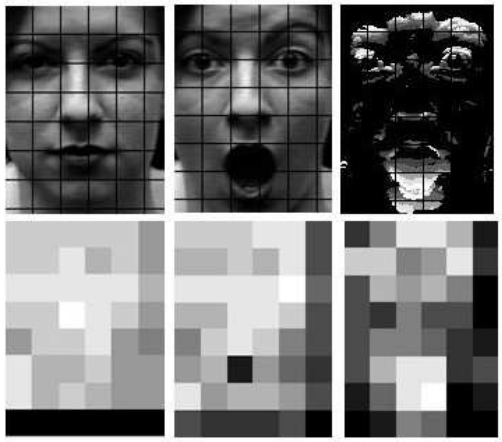

Figure 3: Example image entropies: (left column) neutral image and its entropy; (middle colulmn) surprise image and its entropy; and (right column) MHI of surprise image and its entropy. Lighter shades denote larger entropy values.

number of the distinct intensity values. The process of division is

$$x_{ij} = \begin{cases} x_{t_1}, & t_1 \leq x_{ij} \leq t_2 \\ x_{t_2}, & t_2 \leq x_{ij} \leq t_3 \\ x_{t_3}, & t_3 \leq x_{ij} \leq t_4, \\ & . \\ & . \\ & . \\ x_{t_k}, & t_{K-1} \leq x_{ij} \leq t_K. \end{cases} \tag{9}$$

To compute the weight function, we divide the MHIs with size of $H \times W$ into several non-overlapping sub-regions. The 2D spatial histogram of the intensity values x_{t_k} on each sub-region of X is

$$h_k = \{h_k(p,q) | 1 \leq p \leq P, 1 \leq q \leq Q\}, \tag{10}$$

where p, $q \in \mathbb{Z}^+$, $P \times Q$ is the size of sub-regions, and $h_k(p,q) \in [0, \mathbb{Z}^+]$ is the number of occurrences of the intensity values x_{t_K} in the spatial grid located on the image sub-region of $[(p-1)\frac{H}{P}, p\frac{H}{P}] \times [(q-1)\frac{W}{Q}, q\frac{W}{Q}]$. In forming 2D spatial histogram h_k of intensity values x_{t_k}, the aspect ratio of the original image is maintained on spatial grids. In this way, spatial characteristics of pixels are retained when forming the 2D spatial histogram.

The entropy value on each sub-region of the 2D spatial histogram is computed for intensity values x_{t_k}

using

$$S(p,q) = -\sum_{k=1}^{K} p(h_k(p,q))log_2 p(h_k(p,q)), \tag{11}$$

where $p(h_k(q,p))$ is the possibility of particular intensity value x_{t_k} in the spatial grid located on the image sub-region of $[(p-1)\frac{H}{M}, p\frac{H}{M}] \times [(q-1)\frac{W}{N}, q\frac{W}{N}]$.

The normalisation process is implemented using

$$\omega(p,q) = (s(p,q) - s_{\min})/(s_{\max} - s_{\min}) \tag{12}$$

to convert the range of weights into $0-1$, where $s_{\min}$ and $s_{\max}$ are respectively the maximum value and the minimum of the entropy values over all sub-regions. The computed weights of each subregion on MHI_OF are as the final weight features

$$\text{enMHI\_OF} = \{\omega(1,1), \omega(1,2), ...\omega(1,q), ..., \omega(p,q)\}. \tag{13}$$

The MHI_OF using entropy representation is shown in the rightmost column of Fig. 3.

3.5. Local Zernike Moment

ZMs of an image is computed by decomposing the image onto a set of complex orthogonal basis on the unit disc $x^2 + y^2 \leq 1$ called Zernike polynomials. The Zernike polynomials are defined as [12]

$$V_{nm}(\rho,\theta) = V_{mn}(\rho cos\theta, \rho sin\theta) = R_{nm}(\rho)e^{jm\theta}, \tag{14}$$

where n is the order of the polynomial and m is the number of iterations such that $|m| < n$ and $n - |m|$ is even. R_{mn} are the radial polynomials, i.e.,

$$R_{mn}(\rho) = \sum_{s=0}^{n-|m|} \frac{(-1)^s \rho^{(n-2s)}(n-s)!}{s!(\frac{n+|m|}{2} - s)!(\frac{n-|m|}{2} - s)!}, \tag{15}$$

where ρ and θ are the radial coordinates. A ZM of a face image $I(x,y)$ consisting of a real and an imaginary components is [12]

$$Z^I_{nm} = \frac{n+1}{\pi} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} I(x,y) V^*_{mn}(\rho_{xy}, \theta_{xy}) \Delta\bar{x} \Delta\bar{y}, \tag{16}$$

where x and y are the image coordinates mapped to the range $[-1,+1]$, $\rho_{xy} = \sqrt{\bar{x}^2 + \bar{y}^2}$, $\theta_{xy} = tan^{-1}\frac{\bar{x}}{\bar{y}}$ and $\Delta\bar{x} = \Delta\bar{y} = 2/N\sqrt{2}$.

Since a local descriptor represents the discontinuities and texture of an image effectively, QLZM is proposed in [12] using non-linear encoding and pooling, where non-linear encoding facilitates the relevance of low-level features by increasing their robustness against image noise, while pooling is exploited to deal with the problem of small geometric variation. Non-linear encoding is carried out on complex-valued local

ZMs using binary quantization, which converts the real and imaginary parts of each ZM coefficient into binary values using signum functions. Such coarse quantisation increases compression and encodes each local block with a single integer. Since features along borders may fall out of the local histogram, they are down-weighted in pooling using a Gaussian window peaked at the centre of each subregion. A second partitioning is also applied to account for the down-weighted features, where a higher emphasis is placed on features down-weighted at the first partitioning. The final QLZM feature is constructed by concatenating all local histograms, and the length of extracted correspond to two parameters: the number of moment coefficient K_1 and the size of the grid M, which are computed by

$$2^{2K_1} \times [M^2 + (M+1)^2], \tag{17}$$

where for moment order n, K_1 is computed using the function of moment order n

$$K_1(n) = \begin{cases} \dfrac{n(n+2)}{4} & \text{if } n \text{ is even} \\ \dfrac{(n+1)^2}{4} & \text{if } n \text{ is odd.} \end{cases} \tag{18}$$

The process of generating QLZM is illustrated in Fig. 4.

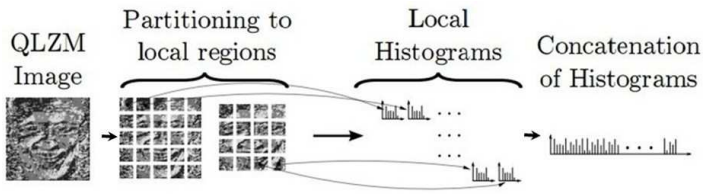

Figure 4: QLZM based facial representation framework.

3.6. Extension to spatio-temporal

QLZM of a 2D image incorporating local spatial textural information has been shown to achieve good facial expression recognition rate [12]. In this paper, we incorporate dynamic information by applying a Motion Change Frequency (MCF) for spatial QLZM, and propose a spatio-temporal descriptor QLZM_MCF. Suppose we have a QLZM sequence where each image frame has been transformed by using QLZM, and the subregions of each QLZM frame are denoted as $Q_{p,q}(i,t)$, where t is the image frame number in the sequence and i denotes the local pattern from a subregion (m,n) of each QLZM image. For each pattern i in subregions (p,q), its positive change sequence $\text{pos}_{p,q}(i,t)$, $t = 1,2,...,T-1$ is defined as

$$pos_{p,q} = \begin{cases} 1 & Q_{p,q}(i,t+1) - Q_{p,q}(i,t) > T_s * Q_{p,q}(i,t) \\ 0 & \text{otherwise,} \end{cases} \tag{19}$$

where T_s is a threshold. Similarly, its negative change sequence is defined as

$$neg_{p,q} = \begin{cases} 1 & Q_{p,q}(i,t+1) - Q_{p,q}(i,t) < -T_s * Q_{p,q}(i,t) \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

Also, we define the unchanged sequence as

$$unc_{p,q} = \begin{cases} 1 & \mid Q_{p,q}(i,t+1) - Q_{p,q}(i,t) \mid \leq T_s * Q_{p,q}(i,t) \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

T_s is an adjustable parameter which affects the performance of the proposed framework. If T_s is set too large then some movements between two consecutive frames might be ignored, while if T_s is set too small then small movements, e.g., due to subtle head pose are detected. In our experiments, T_s is set to 0.1. The QLZM_MCF on each subregion (p, q) is the combination of three changes of the pattern i, i.e.,

$$\begin{aligned} \text{QLZM\_MCF}_{p,q} = \{&\text{QLZM\_MCF}_{p,q}(i,1), \\ &\text{QLZM\_MCF}_{p,q}(i,2), \\ &\text{QLZM\_MCF}_{p,q}(i,3)\} \end{aligned} \tag{22}$$

where

$$\begin{aligned} \text{QLZM\_MCF}_{p,q}(i,1) &= \sum_{t=1}^{T-1} pos(i,t)/(T-1) \\ \text{QLZM\_MCF}_{p,q}(i,2) &= \sum_{t=1}^{T-1} neg(i,t)/(T-1) \\ \text{QLZM\_MCF}_{p,q}(i,3) &= \sum_{t=1}^{T-1} unc(i,t)/(T-1). \end{aligned} \tag{23}$$

The final QLZM_MCF feature is obtained by concatenating all $\text{QLZM\_MCF}_{p,q}$ on each region.

3.7. Fusion using weighting function

Given two different types of facial features, an efficient way to combine them is to concatenate the two features to give

$$f_{\text{FUSION}} = (\text{enMHI\_OF}, \text{QLZM\_MCF}), \tag{24}$$

where enMHI_OF and QLZM_MCF are the two features.

Another combination scheme is also introduced to combine the two features by applying enMHI_OF feature as weight function in pooling during the generation of QLZM. Specifically, we use the same strategy of subregion division on the input image of MHI and QLZM, and the threshold based on enMHI_OF is introduced to each subregion of QLZM image to determine which subregions are removed or retained to

compute spatial-temporal QLZM_MCF. If the enMHI_OF value of a subregion is larger than the threshold, the subregion at the same location in the QLZM image is retained for further processing, otherwise the subregion is removed. The threshold function is defined as

$$R_{p,q} = \begin{cases} R_{p,q} & \text{enMHI\_OF}_{p,q} > T_{en} \\ \text{remove} & \text{otherwise,} \end{cases} \tag{25}$$

where T_{en} is the threshold to be set and $\text{enMHI\_OF}_{p,q}$ is the value of enMHI_OF in subregion (p, q). This scheme is required because subregions with larger enMHI_OF value indicating more significant motion (thus making larger contribution to recognition) should be allocated larger weights, while subregions with smaller enMHI_OF indicating little motion (thus making no or little contribution to recognition) should be allocated smaller weights or removed. The integrated feature is $f_{\text{WeightedFUSION}}$, and the dimension of the feature is $3 \times 2^{2K} \times N_s$, where N_s is the number of selected subregions obtained by the thresholding.

3.8. Dimensionality reduction using 2D PCA

Principal Component Analysis (PCA) is widely used in facial expression recognition for reducing the dimensionality of feature space. It aims to extract decorrelated features out of possible correlated features using a linear mapping function. Under controlled head-pose and imaging conditions, these features capture the statistical structure of facial expressions. 2D PCA has been shown to be superior to PCA in terms of more accurate estimation of covariance matrices and reduced computational complexity for feature extraction by operating directly on 2D matrices instead of 1-dimensional vectors [39], i.e., it is not necessary to convert the 2D image into 1D feature prior to feature extraction. Given L training samples, i.e., $G_1, G_2, .., G_L$, the scatter matrix S is [39]

$$S = \frac{1}{L} \sum_{i=1}^{L} (G_i - M)^T \times (G_i - M), \tag{26}$$

where $M = (1/L) \sum_{i=1}^{L} G_i$. Since there are at most $L-1$ eigenvectors of S with non-zero eigenvalues, N eigenvectors (where $N < L-1$) are randomly chosen from the set of $L-1$ eigenvectors, i.e., $(e_1, e_2, ...e_{L-1})$, with the largest eigenvalues used to construct L subspaces $R_k{}_{K=1}^{L}$. The nth eigenvector with zero eigenvalue is discarded in order to reduce the dimensionality of the feature space while preserving discriminatory information. Thus, 2D PCA is adopted in this paper.

4. Facial expression recognition framework

Fig. 5 outlines the proposed framework which comprises pre-processing, feature extraction and classification. The pre-processing includes facial landmark detection and face alignment, where face alignment is applied to reduce the effects of variation in head pose and scene illumination. We use the local evidence aggregated regression [38] to detect facial landmarks over each frame, where the locations of detected eyes

and nose are used for face alignment including scaling and cropping. The aligned face images are the size of 200×200, where the x coordinate of the centre of the 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. Since the dimensionality of the features is high, following the feature extraction as in Section 3 a dimension reduction technique is applied to obtain a more compact representation. Different classifiers may lead to different recognition performance. We use support vector machines (SVM) that has been widely used and shown to be effective in recognising facial expressions.

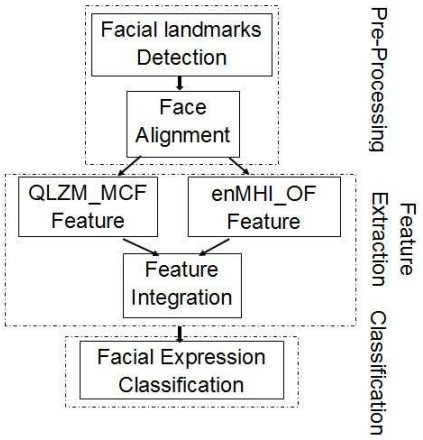

Figure 5: The proposed framework.

5. Experiments

5.1. Facial expression datasets

We use the Extended CK dataset (CK+) as it is widely used for evaluating the performance of facial expression recognition methods and thus facilitates comparison of performances. The dataset includes 327 image sequences of six basic expressions (namely Anger, Disgust, Fear, Happiness, Sadness and Surprise) and a non-basic emotion expression (namely Contempt), performed by 118 subjects. Each image sequence from this dataset has various number of frames and starts with the neutral state and ends with the peak phase of a facial expression. We use standard leave-one-out cross-validation scheme to evaluate the performance of the proposed framework by computing the average recognition rate. One sequence corresponding to an expression is chosen for testing and the remaining sequences of the same expression are used for training. We run the proposed recognition system 327 times on the selected image sequences, and averaged all recognition rates to obtain the final rates.

We also use MMI [36], a publicly available dataset, which includes both posed and spontaneous facial expression sequences. 203 sequences which are labelled as one of six basic expressions are selected, and all

selected sequences are converted into 8-bit grey-scale images with only the sub-sequences from start frame to the frame with the peak expression phase included.

5.2. Experimental results

The first experiment aims to investigate the effectiveness of the enMHI_OF feature, and is conducted on the CK+ dataset. As the performance of enMHI_OF might rely on the size of sub-regions and the number of grey levels represented by K, we conducted our experiment using different sizes and K. Table 1 shows that better performances are achieved using divided grey levels (i.e., using K=4, 10, 20) than using the entire 256 grey levels. Also, using sub-regions with size 20×20 gives the best performances.

Table 1: Recognition rate of enMHI_OF using several combinations of different grey levels and block sizes on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation.

	20×20	10×10	8×8	5×5
K=4	74.31	70.33	71.55	67.28
K=10	75.84	75.84	70.63	72.78
K=20	**76.14**	75.53	72.48	74.92
K=256	73.40	70.94	71.55	73.09

The second experiment compares the performance difference between the spatial and spatio-temporal features. The recognition rates in using spatial QLZM and the spatio-temporal QLZM_MCF which employs dynamic information are summarised in Fig. 6. We also compare the use of MHI_OF and MHI. Since using MHI image as input of the classifier may lead to higher dimensionality, we use histogram computation to represent MHI and MHI_OF. The recognition rates in using MHI and MHI_OF are shown in Fig. 7. As can be seen from Fig. 7, the performance in using MHI_OF is better than in using MHI. These two figures show that QLZM_MCF and MHI_OF outperform the spatial QLZM and MHI, respectively, although the performance of both MHI and MHI_OF are less than satisfactory.

The third experiment investigates the effectiveness of concatenating QLZM_MCF with enMHI_OF in the proposed framework. Table 2, Table 3, Table 4 and Table 5 respectively show the results using two individual features separately, the simple fusion strategy f_{FUSION} and the proposed fusion strategy $f_{\text{WeightedFUSION}}$. The overall recognition rates using all four features (QLZM_MCF, enMHI_OF, feature using simple fusion strategy and feature using proposed weighting fusion strategy) are shown in Table 6. The tables show the framework using the simple fusion strategy of two features performs better than using individual feature separately, and the proposed fusion strategy achieves the best performance. In Table 6, we compare the proposed feature with the method of Eskil et al. [43], the static method of Lucey et al. [32] and our previous

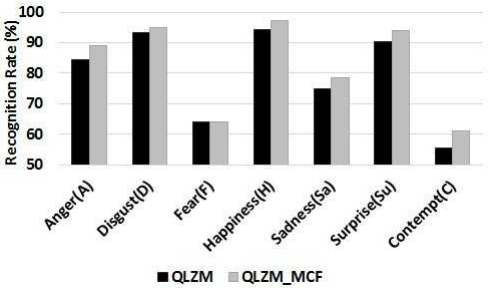

Figure 6: Recognition rates of all expressions using QLZM and QLZM_MCF on CK+ dataset.

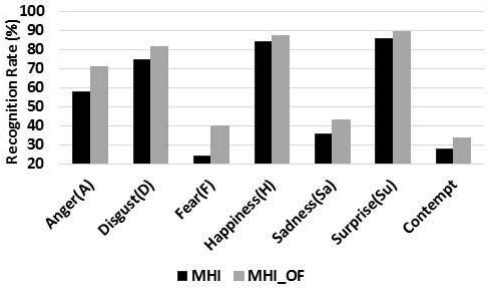

Figure 7: Recognition rates of all expressions using MHI and MHI_OF on CK+ dataset.

work [44], which shows the fused feature achieves an average recognition rate of 88.30% for all seven facial expressions, and outperforms the other methods. Thus, we can also conclude that the combination of two dynamic features improves the recognition rate.

Table 2: Recognition rate (in term of percentage of true positive, true negative, ect.) of QLZM_MCF on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation.

	A	D	F	H	Sa	Su	C
Anger(A)	**89.9**	2.2	0	0	4.4	0	4.4
Disgust(D)	1	**94.9**	1.7	0	0	0	1.7
Fear(F)	4.0	8.0	**64.0**	4.0	12.0	0	4.0
Happiness(H)	1.4	0	0	**97.1**	0	1.4	0
Sadness(Sa)	0	3.6	7.1	0	**78.6**	3.6	7.1
Contempt(Su)	0	0	1.2	2.4	1.2	**94.0**	1.2
Contempt(C)	0	5.6	16.7	11.1	5.6	0	**61.1**

Table 3: Recognition rate (in term of percentage of true positive, true negative, ect.) of enMHI_OF on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation.

	A	D	F	H	Sa	Su	C
Anger(A)	**73.3**	4.4	6.7	6.7	6.7	0	2.2
Disgust(D)	5.1	**84.8**	3.4	0	3.4	0	3.4
Fear(F)	8.0	0	**40.0**	16.0	20.0	4.0	12.0
Happiness(H)	0	0	4.3	**89.9**	2.9	0	2.9
Sadness(Sa)	10.7	0	21.4	3.6	**42.9**	7.1	14.3
Surprise(Su)	1.2	0	2.4	2.4	1.2	**90.4**	2.4
Contempt(C)	11.1	0	27.8	5.6	11.1	5.6	**38.9**

Table 4: Recognition rate (in term of percentage of true positive, true negative, ect.) of using simple fusion strategy on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation.

	A	D	F	H	Sa	Su	C
Anger(A)	**86.7**	2.2	2.2	2.2	6.7	0	0
Disgust(D)	3.4	**94.9**	1.7	0	0	0	0
Fear(F)	8.0	4.0	**72.0**	0	4.0	4.0	8.0
Happiness(H)	0	0	0	**95.7**	8.0	4.0	0
Sadness(Sa)	3.6	0	10.7	0	**78.6**	0	7.1
Surprise(Su)	0	0	1.2	2.4	1.2	**95.2**	0
Contempt(C)	5.6	0	22.2	5.6	5.6	5.6	**55.6**

Table 5: Recognition rate (in term of percentage of true positive, true negative, ect.) of the proposed fusion strategy on classification of six facial expressions of the CK+ dataset and contempt with leave-sequence-out cross-validation.

	A	D	F	H	Sa	Su	C
Anger(A)	**91.1**	2.2	6.7	0	0	0	0
Disgust(D)	3.4	**96.7**	0	0	0	0	0
Fear(F)	4.0	4.0	**80.0**	4.0	0	4.0	4.0
Happiness(H)	0	1.4	0	**98.6**	0	0	0
Sadness(Sa)	3.6	0	0	3.6	**89.3**	3.6	0
Surprise(Su)	0	0	0	0	1.2	**97.6**	1.2
Contempt(C)	5.6	0	11.1	5.6	5.6	0	**72.2**

Table 6: The overall recognition rates of the four spatio-temporal features on the CK+ dataset.

Feature	Recognition rate
Lucey et al [32]	50.4
Eskil et al [43]	76.8
Our previous work [44]	83.7
QLZM_MCF	82.6
enMHI_OF	65.7
simple fusion strategy	82.6
proposed weighting fusion strategy	88.3

We also conducted an experiment on the MMI dataset, comparing the proposed framework with the method that uses LBP and SVM [37], and the methods in [45] and [44] that are evaluated using the same classification strategy of 10-fold cross-validation. The average recognition rates are shown in Table 7. The table shows that the proposed framework outperforms all the other five methods. The result for LBP was obtained by using different samples to those used in [37], and using the same strategy of classification introduced in [45] which is also used in [44] and the proposed method.

Table 7: Comparative evaluation of the proposed framework on the MMI dataset.

Study	**Methodology**
LBP [37]	54.5
AAM [45]	62.4
ASM [45]	64.4
Fang in [45]	71.6
Our previous work [44]	74.3
Proposed weighting fusion strategy	79.8

Although CK+ and MMI are two of the most widely used datasets for evaluating facial expression recognition methods, they are both collected in a strict controlled settings with near frontal poses, consistent illumination and posed expressions. The recent and more challenging datasets of AFEW and SFEW [46] provide platforms for researchers to create, extend and test their methods on a common benchmarked data. Since the proposed framework recognises facial expression on video sequence which treat a sequence as an entity, we use AFEW which are used for EmotiW 2014 for our experiments [47]). AFEW is a dynamic temporal facial expressions data corpus extracted from movies with realistic real world environment. It was collected on the basis of Subtitles for Deaf and Hearing impaired (SDH) and Closed Caption (CC) for searching expression-related content and extracting time stamps corresponding to video clips which represent some meaningful facial motion. The database contains a large age range of subjects from 1-70 years, and the subjects in the clips have been annotated with attributes like Name, Age of Actor, Age of Character, Pose, Gender, Expression of Person and the overall Clip Expression. There are a total of 957 video clips in the database labelled with six basic expressions anger, disgust, fear, happy, sad, surprise and the neutral. To compare with the baseline method of EmotiW 2014 [47], we modified the proposed framework slightly, where we use the pre-processing methods (face detection and alignment) provided by the baseline method.

We used the training samples for training, and the validation samples for performance evaluation. Table 8 shows the recognition rate using the proposed framework on AFEW dataset. The overall recognition rate of the proposed framework on the validation set is 37.63%, which is higher than the 33.15% achieved by the video only baseline method. Unlike the experiments on CK+ dataset, the surprise expression is much

more difficult to be recognised. This is because sometimes the surprise expression might not be acted exaggeratedly (i.e., the openness of mouth) in real situations. Also, the overall recognition rate is much lower than on the CK++ and MMI dataset. This is because numerous frames from the AFEW sequences were captured under poor light condition, have large pose or occlusion, and the expressions are not always from neutral to peak expression.

Table 8: Recognition rate (in term of percentage of true positive, true negative, ect.) of the proposed strategy on classification of six basic facial expressions and neutral expression of the AFEW dataset

	A	D	F	H	N	Sa	Su
Anger(A)	**65.6**	7.8	4.7	1.6	9.4	6.3	4.7
Disgust(D)	15.0	**22.5**	7.5	12.5	17.5	10.0	15.0
Fear(F)	21.7	13.0	**20.6**	15.2	15.2	8.7	6.5
Happiness(H)	3.2	7.9	7.9	**63.5**	9.5	6.3	1.6
Neutral(N)	3.2	6.3	14.3	9.5	**49.2**	9.5	7.9
Sadness(Sa)	8.2	11.5	14.8	14.8	26.2	**21.3**	3.3
Surprise(Su)	13.0	10.9	17.4	13.0	19.6	4.3	**21.7**

6. Conclusion

This paper presents a facial expression recognition framework using enMHI_OF and QLZM_MCF. The framework which comprises pre-processing, feature extraction followed by 2D PCA and SVM classification achieves better performance than most of the state-of-art methods on CK+ dataset and MMI dataset. Our main contributions are three folds. First, we proposed a spatio-temporal feature based on QLZM. Second, we applied optical flow in MHI to obtain MHI_OF feature which incorporates velocity information. Third, we introduced entropy to employ the spatial relation of different facial parts, and designed a strategy based on entropy to integrate enMHI_OF and QLZM_MCF. The proposed framework performs slightly worse in distinguishing the three expressions of Fear, Sadness and Contempt, thus how to design a better feature to represent these expressions will be part of our feature work. Also, since an expression usually occurs along with the movement of shoulder and hands, it might be useful to exploit these information in our recognition system.

When applying a facial expression recognition framework in real situations, computation speed might be a factor to be considered. In some case, the increase in speed may result in a decrease in recognition performance. How to design a framework for facial expression recognition which increases the computational speed without any degradation in the recognition rate remains a challenge.

7. 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] M. Pantic, L. Rothkrantz, Automatic analysis of facial expressions: the state of art, IEEE Transaction on Pattern Analysis and Machine Intelligence 22(12) (2000) 1424-1445.

[2] M. Pantic, L. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction, in: Proceeding of the IEEE, vol. 91, 2003, pp. 1370-1390.

[3] Y. Tian, T. Kanade, J. Cohn, Handbook of face recognition, Springer, 2005 (Chapter 11. Facial expression recognition).

[4] T. Tojo, Y. Matsusaka, T. Ishii , and T. Kobayashi, A conversational robot utilizing facial and body expressions. In Systems, Man, and Cybernetics, 2000 IEEE International Conference on, vol. 2, 2000, pp. 858-863.

[5] R. Cowie,, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, Emotion recognition in human-computer interaction, Signal Processing Magazine, IEEE 18(1) (2001) 32-80.

[6] B. Fasal, J. Luettin, Automatic facial expression analysis: a survey. Pattern Recognition 23 (2003) 259-275.

[7] B. Heisele, P. Ho, J. Wu, and T. Poggio, Face recognition: component-based versus global approaches, Computer Vision and Image Understanding 91 (2003) 6-21.

[8] T. Ahonen, A. Hadid, M. Pietikainen, Face recognition with local binary patterns, in: European Conference on Computer Vision (ECCV), 2004.

[9] A. Ono, Face recognition with Zernike moments, Systems and Computers in Japan, 34(10) (2003) 26-35.

[10] C. Singh, N. Mittal, and E Walia, Face recognition using Zernike and complex Zernike moment features, Pattern Recognition and Image Analysis, 21 (2011) 71-81.

[11] S. Moore, and R. Bowden, Local binary patterns for multi-view facial expression recognition. Computer Vision and Image Understanding, 115(4) (2011) 541-558.

[12] E. Sariyanidi, H. Gunes, M. Gokmen, and A. Cavallaro, Local Zernike Moments representations for facial affect recognition, in Porc. IEEE Int'l Conf. Image Processing, 2012, pp. 585-588.

[13] E. Saryanidi, V. Dal, S.C. Tek, B. Tunc, and M. Gökmen, (2012,). Local Zernike Moments: A new representation for face recognition. In 2012 19th IEEE International Conference on Image Processing (2012) 585-588.

[14] A. F. Bobick and J. W. David, The recognition of human movement using temporal templates, IEEE Trans, Pattern Analysis and Machine Intelligence, 23(3) (2001) 257-267.

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

[16] M. Pantic, L. Rothkrantz, Facial action recognition for facial expression analysis from static face images, IEEE Trans. Systems, Man and Cybernetics, 34(3) (2004) 1449-1461.

[17] M. Pantic, I. Patras, Dynamic of facial expressions - recognition of facial actions and their temporal segments from face profile image sequences, IEEE Trans. Systems, Man and Cybernetics, 36(2) (2006) 433-449.

[18] S. Gokturk, J. Bouguet, C. Tomasi, and B. Girod. Model-based face tracking for view-independent facial expression recognition, In Proc. IEEE Conf. Face and Gesture Recognition, 2002, pp. 272-278.

[19] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and T. Huang, Learning bayesian network classfiers for facial expression recognition both labeled and unlabeled data. In IEEE Conf. Comp. Vision and Pattern Recognition, vol. 1, 2003, pp. 595-601.

[20] Y. Tian, T. Kanade, and J. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. Pattern Analysis and Machine Intelligence, 23(2) (2001) 1-19.

[21] M. K. Hu, Visual pattern recognition by moment invariants, IRE Transctions on Information Theory, 8(2) (1962) 179-187.

[22] M. R. Teague, Image analysis via the general theory of moments*, J. Opt. Soc. Am. 70, (1980) 920-930.

[23] A. Khontanzad and Y. H. Hong, Roatation invariant image recognition using features selected via a systematic method, Pattern Recognition 23 (1990) 1089-1101.

[24] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multilayer, Proceeding of International Conference on Automatic Face and Gesture Recognition, (1998) pp. 454-459.

[25] G. Zhan and M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expression. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6) (2007) 915-928.

[26] T. Ojala, M.Pietikainen, and T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7) (2002) 971-987.

[27] G.Y. Zhao and M. Pietikainen, . Dynamic texture recognition using local binary pattern with an application to facial expression, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2(6) (2007) 915-928.

[28] D. Chetverikov and R. Peteri, A brief survey of dynamic texture description and recognition. Proc. Conf. Computer Recognition Systems, vol. 5 (2005) pp. 17-26.

[29] S. Koelstra, M. Pantic and I. Patras, A dynamic texture-based approach to recognition of facial actions and their temporal models, IEEE transactions on Pattern Analysis and Machine Intelligence, 32(11) (2010) 1940–1954.

[30] L.A. Cament, L.E. Castillo, J.P. Perez, F.J. Galdames, C.A. Perez, Fusion of Local Normalization and Gabor Entropy Weighted Features for Face Identification, Pattern Recogn. 47 (2014) 568577.

[31] Z. Chai, H. Mendez-Vazquez, R. He, Z. Sun, and T. Tan, Semantic pixel sets based local binary patterns for face recognition, In Computer VisionACCV (2012) pp. 639-651, Heidelberg.

[32] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). 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, pp. 94-101.

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

[34] Tsai, D. M., Chiu, W. Y., and Lee, M. H. (2015). Optical flow-motion history image (OF-MHI) for action recognition. Signal, Image and Video Processing, 9(8) pp. 1897-1906.

[35] R. Balian, (2004). Entropy, a Protean concept". In Dalibard, Jean. Poincar Seminar 2003: Bose-Einstein condensation - entropy. Basel: Birkhuser. pp. 119144.

[36] M. Pantic, M.F. Valstar, R. Rademaker, L. Maar, Web-based database for facial expression analysis, in: Proceedings of the IEEE International Conference on Multimedia and Expo, Amsterdam. The Netherlands, 2005, pp. 317-321.

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

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

[39] J. Yang, D. Zhang, A.F. Frangi, and J.Y. Yang. Two-dimensional PCA: a new approach to appearance-based face representation and recognition, IEEE Trans. Pattern Analysis and Machine Intelligence, 26(1) (2004) 131-137.

[40] L. A. Jeni, J. Girard, J. Cohn, and F. De La Torre, Continuous AU intensity estimation using localized, sparse facial feature space, in Proc. IEEE Int'l Conf. Affective Computing and Intelligent Interaction, 2013.

[41] Y. Zhu, F. De la Torre, J. Cohn, and Y. -J Zhang, Dynamic cascades with bidirectional boostrapping for action unit

detection in spontaneous facial behavior, IEEE Trans. Affective Computing, 2(2) (2011) 79-91.

[42] Bosch, A., Zisserman, A., and Munoz, X. (2007). Representing shape with a spatial pyramid kernel, In Proceedings of the International Conference on Image and Video Retrieval, pp. 401-408.

[43] Eskil, M.T., and Benli, K. (2014). Facial expression recognition based on anatomy, Computer Vision and Image Understanding, vol. 119, pp. 1-14.

[44] FAN, X. and Tjahjadi, T. (2015). A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences, Pattern Recognition, 48(11) 3407-3416.

[45] Fang, H., Parthalin, N.M., Aubrey, J., Tam, K.L., Borgo, R., Rosin, L., Grant, W., Marshall, D., and Chen, M. (2014). Facial expression recognition in dynamic sequences: An integrated approach, Pattern Recognition, 47(3), 1271-1281.

[46] A. Dhall, R. Goecke, S. Lucey and T. Gedeon, Collecting large, richly annotated facial-expression databases from movies, IEEE Multimedia 2012, pp. 34-41

[47] A. Dhall, R. Goecke, J. Joshi, K. Sikka and T. Gedeon, Emotion Recognition In The Wild Challenge 2014: Baseline, Data and Protocol, Proceedings of the 16th International Conference on Multimodal Interaction, 2014, pp. 461-466.