IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 34, No. 5, October, 2020, pp. 585-594
Journal homepage: http://iieta.org/journals/ria

Dynamic Features Based on Flow-Correlation and HOG for Recognition of Discrete Facial Expressions

Shivangi Anthwal*, Dinesh Ganotra

Department of Applied Science and Humanities, Indira Gandhi Delhi Technical University for Women, Kashmere Gate, Delhi, 110006, India

Corresponding Author Email: shivangianthwal@hotmail.com

https://doi.org/10.18280/ria.340508

Received: 24 July 2020
Accepted: 9 October 2020

Keywords:
optical flow, HOG, facial expression recognition, emotion interpretation, multi-class support vector machine

ABSTRACT

Facial expressions are the most preeminent means of conveying one's emotions and play a significant role in interpersonal communication. Researchers are in pursuit of endowing machines with the ability to interpret emotions from facial expressions as that will make human-computer interaction more efficient. With the objective of effective affect cognition from visual information, we present two dynamic descriptors that can recognise seven principal emotions. The variables of the appearance-based descriptor, *FlowCorr*, indicate intra-class similarity and inter-class difference by quantifying the degree of correlation of optical flow associated with the image pair and each pre-designed template describing the motion pattern associated with different expressions. The second shape-based descriptor, dyn-HOG, finds the HOG values of the difference image derived by subtracting neutral face from emotional face, and is demonstrated to be more discriminative than previously used static HOG descriptors for classifying facial expressions. Recognition accuracies with multi-class support vector machine obtained on the CK+ and KDEF-dyn datasets are competent with the results of state-of-the-art techniques and empirical analysis of human cognition of emotions.

1. INTRODUCTION

Facial expressions are powerful form of non-verbal communication signals that reflect one's intentions and emotional state. Automated facial expression recognition (FER) has been widely investigated in the last two decades due to its potential applications in human-computer interaction and smart environments. Based on cross-cultural psychological research [1] it was concluded that six fundamental emotions were expressed in similar fashion universally, regardless of culture, gender, and race. These pan cultural expressions were classified as anger, disgust, fear, happiness, sadness and surprise (Figure 1). Further heuristic experiments to analyse the universality [2] concluded that there was no difference in facial muscle movements associated with a specific expression for both sighted and visually impaired implying expressions of emotions are innate and not learned visually or culturally. The list of universal expressions of emotions was subsequently expanded with the inclusion of expression of contempt [3].

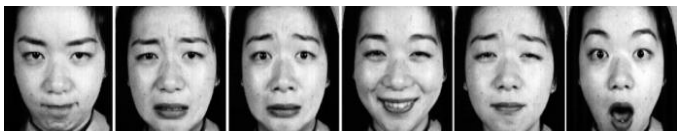

Figure 1. Six pan-cultural expressions: anger, disgust, fear, happiness, sadness, surprise [4]

Automated recognition of these principal emotions has been demonstrated to be useful in expansive range of domains such as affective video summarisation [5], ambient assisted living [6] interactive video gaming [7]. In pursuit of quantifying performance of different approaches for recognition of these emotions conveyed through expressive images or videos, different databases [4, 8-10] having subjects from diverse ethnicities and belonging to different age groups have been proposed in the last few years. The foremost step followed by FER methods is to take images from a dataset followed by locating and cropping facial region in those images. Subsequently, relevant features that aid in characterization are extracted followed by categorisation of the emotion conveyed in the image or video by an adequate classifier. With the objective of endowing machines with emotional intelligence, we present appearance and shape based dynamic feature descriptors based on optical flow correlation and dynamic Histogram of Oriented Gradient (HOG), respectively, to identify the seven principal emotions via facial expressions. The salient features of the work presented and its contribution to the literature will be summarised in the next section.

The rest of the paper is structured such that Section 2 gives a brief overview of the different techniques, including the state-of-the-art, in the field of identifying facial expression by exploiting visual information and delineates the concept of optical flow and HOG, highlighting their significance and contribution in recognizing facial expressions from images. The proposed descriptors are described systematically in Section 3 which also details the entire methodology and the summary of experimental results obtained is discussed in Section 4. Conclusion and future directions are outlined in Section 5.

2. RELATED WORK

2.1 Descriptive features employed for facial modelling

The broad categories of descriptive features presented hitherto for reliable analysis of facial behaviour are derived from the facial action units (AUs), appearance and geometry.

2.1.1 Facial action unit based features

With the objective of decoding facial behaviour, the Facial Action Coding System (FACS) was devised that interpreted a wide spectrum of expressions using different AUs of lower and upper face [11]. To detect and classify eight individual AUs and 4 AU combinations in lower face region in image sequences devoid of head motion, Donato et al. [12] employed Gabor wavelet representation with independent component analysis and reported recognition rates on par with those attained by FACS trained human experts. Li et al. [13] employed a dynamic Bayesian network for their facial activity recognition system that had two principal steps: constructing facial activity model offline and measuring as well as inferring motion of facial components online. By utilizing training data and subjective domain knowledge, the model characterizing facial activity semantically was designed offline. During the online recognition step, facial feature points were tracked to get the objective measurements of facial motions, i.e., AUs used to infer the nature of the facial activities.

2.1.2 Facial geometry- or appearance-based features

Apart from facial AUs, other features employed to characterise facial expressions are archetypally either geometry based or appearance based. Geometric feature based methods [14, 15] analyse shape and location of facial components such as eyes, mouth, nose and eyebrows. They employ feature vectors that characterize the geometry of face. Appearance features [16, 17] are descriptive of the textures of the facial features such as wrinkles and skin folds that may be suitable to represent the muscle movement for an expression. Happy and Routray [18] integrated both the features in a model that used local binary patterns for analysing texture information extracted from salient facial patches. Salmam et al. [19] combined a CNN architecture representing appearance features and a Deep Neural Network (DNN) framework based on geometric features to demonstrate that integrating features increased the efficiency of FER method.

2.2 Methods for analysis of the described features: Dynamic and static methods

Depending on whether temporal information is employed or not, FER methods may be dynamic or static, respectively. While static methods decode the emotion conveyed from a single frame displaying the momentary appearance of the expression, dynamic methods generally model the temporal development of facial components and the correlation among them between successive frames. Some of the earlier techniques that utilised facial dynamics to analyse expressions used dynamic Bayesian networks [20] and hidden Markov models [21]. For the given input neutral and emotional images, Mlakar and Potocnik [22] computed histogram of oriented gradient based feature descriptors characterizing the differences between the two input frames. Wehrle et al. [23] observed in their experiments that in a video clip advancing from neutral to emotional, the reference state of the subject was known and concluded that dynamic methods employing temporal information led to a better perception of facial expressions. However, for the situations with unavailability of a neutral face, static techniques for FER have an edge over dynamic techniques. To circumvent this, "average human face" was suggested as an effective alternative representation of neutral face in a dynamic model [24]. Hitherto, there is no consensus about one technique being superior to the other comprehensively. In this work, a dynamic FER approach is presented that employs optical flow for quantification of facial motion. A brief discussion of flow based approaches for motion analysis proposed in the literature is presented in the next sub-section.

2.3 Optical flow and facial expression classification

Computation of optical flow, the relative displacement of image grey values in a temporally changing image sequence, is an essential step in dynamic scene analysis. For the frames portraying neutral and apex of an expression (Figures 2a and 2b), optical flow field can be estimated to generate a two dimensional vector field (Figure 2c) indicating velocity and direction of each pixel representing the apparent facial motion pattern between consecutive frames. At each pixel, flow is depicted by a vector, whose orientation describes the direction of the flow and the length represents the flow magnitude.

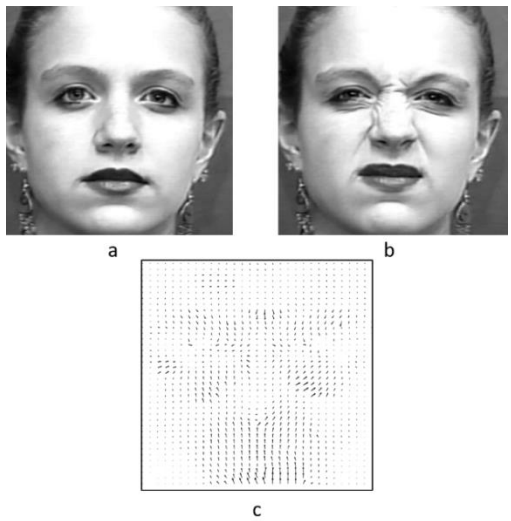

Figure 2. a. Neutral face; **b.** Face with an expression of disgust; **c.** Corresponding optical flow

The use of optical flow in recognizing expressions was pioneered by Mase [25] whose work focused on studying the motion of facial muscles using top-down and bottom-up approach. Four prototypical expressions surprise, happiness, disgust and anger were analyzed in his work. Essa and Pentland [26] added a temporal dimension to the FACS model and developed FACS+ model to allow spatio-temporal modelling of the expressions. They superimposed a mesh on the facial region and tracked its corners based on optical flow for all the images thus capturing facial changes occurring during an expression. Yacoob and Davis [27] classified the six fundamental expressions by computing optical flow and analyzing the rigid and non-rigid motions associated with different emotions. To detect and classify six upper facial actions in image sequences devoid of head motion, Bartlett et al. [28] fed the results of a hybrid model that combined holistic

spatial analysis with optical flow and analysis of local features such as wrinkles to a feed forward neural network. Niese et al. [29] integrated geometric features with transient optical flow features an accomplished classification of five expressions using artificial neural networks and support vector machines. Pu et al. [30] computed displacement between neutral and peak expression frame active appearance model features using optical flow and classified the seven discrete expressions using twofold random forest classifier. Shojaeilangari et al. [31] extracted features dependent only on relative movements of different facial components and developed a pose-invariant robust optical flow descriptor. Zhao et al. [32] introduced accumulative optical flow as a novel feature descriptive of the motion information between the frames. They classified expressions using 3D CNN using both static and dynamic features as input. As evident from the discussion, diverse flow based approached have been proposed previously and demonstrated the effectiveness of optical flow in recognition of facial expressions from the available visual information. In the next section, we propose a novel flow based scheme for effective characterization of facial expressions from input images.

2.4 HOG and facial expression classification

Histogram of Oriented Gradient descriptor was introduced by Dalal and Triggs [33] for human detection in images. The descriptor characterizes shape related information by counting occurrences of gradient orientation in localized portions of an image. Since then it has been employed in diverse frameworks for expression classification. Kumar et al. [34] developed a framework robust to scale and pose variation by extracting HOG features only from active facial patches subsequently feeding them to support vector machine for classification. Wang et al. [35] fused HOG features with Weber local descriptor and classified facial expressions using chi-square distance with nearest neighbour technique. Nigam et al. [36] computed HOG features in discrete wavelet domain for effective characterization of facial expressions. Zeng et al. [37] presented a novel deep architecture wherein they extracted from emotional images HOG features and concatenated them with local binary pattern and grey levels and learnt the discriminative high dimensional features using sparse auto-encoders. Figure 3 depicts a visualization of HOG features being discriminative in categorizing various expressions.

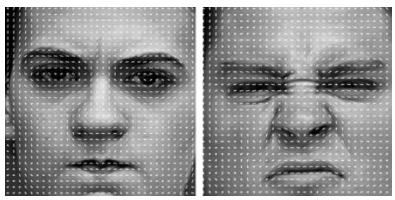

Figure 3. HOG visualization for the expressions of anger and disgust

The figure gives an understanding of how the values of the HOG features vary at each facial location for the two facial displays associated with anger and disgust thus substantiates the superlative capability of the features in discriminating the different facial emotion displays. It is to be noted that these two emotions are quite often confused with each other [9], but the difference is clearly visible with HOG based modelling.

2.5 Facial expression recognition: State-of-the-art

To facilitate extraction of discriminative features for FER, Gan et al. [38] presented an intricate multiple attention network that analysed FEs by simulating coarse-to-fine visual attention. Their framework coupled region-aware sub-net for detecting critical facial regions with expression recognition sub-net for learning comprehensive discriminative features. Most of the state-of-the-art techniques have employed single or multiple stream Convolutional Neural Networks (CNN) for obtaining descriptive features for embodying appearance or morphology of face during the display of an emotion. Wei et al. [39] integrated the 2D coordinates of face key points as low-level empirical features with CNN extracted high-level self-learning features. In addition, they deployed small filters for different convolutional layers to reduce the number of free CNN parameters. Qin et al. [40] derived phase and magnitude features of the histogram equalized expression keyframes with Gabor wavelet transform. They employed a 2-channel CNN for feature training and classification and attained high FER accuracies on CK+ dataset. To assess the performance of the proposed descriptor, its recognition accuracy was compared with all the state-of-the-art techniques when evaluated on CK+ dataset. A performance comparison in terms of recognition accuracies with the state-of-the-art- techniques was carried out and is discussed in Experimental Results.

In this work, the authors present appearance- and shape-based dynamic feature descriptors to identify the seven principal emotions via facial expressions. The research results are promising and the work is original and can prospectively contribute to FER literature. The key points of the work can be summarised as follows:

(1) A novel flow-correlation based descriptor that indicates with its variables inter-class and intra-class degree of similitude is introduced and employed to embody appearance of seven principal expressions. By comparing the derived flow between neutral and emotional faces with pre-designed templates using flow based metrics, the feature is obtained and demonstrated to be more effective than dense horizontal and vertical optical flow values for classification of facial expressions under same experimental settings. The innovation in this work stems from the way it has addressed the complex interplay amongst the seven different facial displays. To the best knowledge of the authors, so far, there is no other work in the literature that had modelled facial emotion displays with the information derived from similarities and differences amongst the seven principal emotions this comprehensively. It is to be noted that highly competent results were attained using only two templates for each facial emotion display making the algorithm computationally less intensive.

(2) For characterisation of shape-related information for facial display modelling, a dynamic version of the static HOG descriptor is presented by computing HOG values of "difference image" that gives the difference between emotional and neutral face and is demonstrated to be more discriminative than its static counterpart for categorisation of expressions under same experimental settings. At the time of writing, no other existing research work the literature has presented an approach similar to this for description of facial features. Most researchers prefer employing HOG to describe the face for an effective modeling of its various components. However, this article demonstrates a superiority of the proposed dyn-HOG feature as a more efficacious descriptive feature for the modelling of the inherently dynamic process of

unfolding of a facial expression.

(3) Results attained on Extended Cohn Kanade (CK+) [8] and dynamic Karolinska Directed Emotional Faces (KDEF-dyn) [9] datasets were promising and were compared with state-of-the-art techniques and empirical analysis of human cognition. The assessment of concordance of the results obtained by different techniques with human cognition has been largely overlooked in the literature. While, a high accuracy on a dataset validates the proposed features, it is imperative to understand how well do the results match-up with the cognitive or perceptive abilities of humans. Drawing this accordance is necessitated by the fact that contemporary and futuristic systems based on artificial intelligent technologies try to emulate human cognitive abilities. Thereby, a high concordance between the results of the proposed method with human cognition substantiates the usefulness of the proposed features in vision-based affective intelligent interfaces and architectures. In this work, the authors have addressed this prominent gap in the literature and tried to find how well-accorded are the results obtained by the proposed descriptors with the ability of humans to perceive and discriminate different emotional displays.

3. FEATURES AND METHODOLOGY

3.1 Database description

For validating efficiency of the proposed descriptors, images were taken from CK+ [8] and KDEF-dyn [9] datasets that are overviewed systematically in Table 1.

For the present work, only the image sequences labelled with either of principal emotion labels (Table 2) were used for evaluation by roughly dividing them into two equal mutually exclusive sets of training and test subsets, ensuring a subject independent evaluation. The experiment was conducted twice with random selection of training and test subsets. An average accuracy of all experiments was computed eventually.

Table 1. Description of CK+ and KDEF-dyn datasets

CK+	KDEF-dyn
593 coloured and greyscale image sequences from subjects displaying a face progressing from neutral to emotion. Only 327 have been provided with one of the seven expression labels viz. anger (an), contempt (co), disgust (di), fear (fe), happiness (ha), sadness (sa), surprise (su)	Coloured video-clips with 40 subjects conveying emotions of anger, contempt, disgust, fear, happiness, sadness, surprise from zero intensity to peak intensity in each clip based on KDEF [10] facial images generated by dynamic morphing between emotional face and corresponding neutral face.
Subjects of diverse ethnicities, 81%, Euro-American, 13% Afro-American, and 6% other groups with their age ranging from 18 to 50 years	Caucasian subjects with their age ranging between 20 to 30 years
A static background shown along with the subject	Background and non-face region darkened

Table 2. Number of test image sequences/video-clips for each emotion

Emotion	An	Co	Di	Fe	Ha	Sa	Su
CK+	45	18	59	25	69	28	83
KDEF-dyn	40	-	40	40	40	40	40

3.2 Image preprocessing: Face localisation

In an image with a subject displaying an emotion, the background information is not needed for the recognition of the expression. Thus, the region containing relevant information i.e. the facial region in the images were tracked using Viola and Jones algorithm [41]. This technique employs histogram of oriented gradients, local binary patterns and Haar-like features with cascaded classifiers trained by boosting. It was utilized due to its speed and precision in locating nearly frontal facial region in the image. The tracked faces were eventually cropped and resized to 256×256 greyscale image (Figure 4). In KDEF-dyn, the background information was already darkened so face detection was not required.

Figure 4. Locating face in the image and cropping it

3.3 Training and testing stages for *FlowCorr* descriptor

From the training sequences, first frame from each sequence was used as neutral frame and the labelled peak expression frame as emotional frame. For an image pair portraying a known expression, the optical flow was extracted from the images using a variational method proposed previously Brox et al. [42]. The flow similarity feature *FlowCorr* was computed for different image pairs and their corresponding expression labels were fed to the classifier for training.

3.3.1 Extraction of the dynamic *FlowCorr* descriptor

An optical flow field is a suitable representation of appearance changes that occur during portrayal of an expression. However, using flow values at each spatial location to describe the appearance change makes the descriptor size large and increase the computational burden on the classifier. We present in this work the descriptor *FlowCorr* which computes the flow at each point and determines its degree of correlation with each motion template (ground truth flow) representing different expressions with flow based error metrics. The computed error/correlation values characterize the inter-class and intra-class relation of the flow with ground truth flow for discrete expression classes and are adequate to characterize the flow, thus reducing the size of the descriptor. We describe the method for computation of the descriptor for the image pair I(x,y,t) representing the neutral facial image and I(x+u,y+v,t+1) the emotional image with corresponding optical flow (u,v).

3.3.2 Computation of optical flow

In this work, a variational framework proposed by Brox et al. [42] for the computation of optical flow has been employed due to its excellent robustness to noise and illumination variation. Let the image intensity be described by the function I(x,y,t) at spatial position (x,y) and time t. Consider the temporal displacement between consecutive frames Δt = 1. Then the image intensity at time t+1 will be given by I(x+u,y+v,t+1) with x+u and y+v denoting the new pixel position. Thus, computation of optical flow requires estimation of a flow field where for each pixel the vector $\mathbf{w} = (u, v, 1)^T$ is determined. Variational methods are global methods in the sense that they operate over the whole image domain. They recover the flow as the minimizer of a suitable functional that is a weighted sum of a data term and a regulariser or smoothness term expressing the model assumptions.

$$E(u, v) = E_{data} + \alpha E_{smooth} \quad (1)$$

The data fidelity term or data term E_{data} represents the consistency between the optical flow and the input images. Usually it is based on the premise of temporal invariance of certain image features. Smoothness term E_{smooth} reflects upon the fact that neighbouring pixels belong to the same surface and hence undergo a similar motion. α is the regularisation parameter signifying the weight of the smoothness term. In this work, variational model proposed by Brox et al. [42] with the following constraints has been used.

Grey value constancy assumption or Brightness Constancy Assumption (BCA): Since the pioneering work for flow computation [43], it has been presumed that grey value of a pixel does not change with displacement. Brightness constancy constraint was deduced from the observation that grey value of a small region remains unaltered despite the change in its position in different frames.

$$I(x, y, t) = I(x + u, y + v, t + 1) \quad (2)$$

Taylor expansion of (2) yields the optical flow constraint

$$I_x u + I_y v + I_t = 0 \quad (3)$$

Gradient Constancy Assumption (GCA): The Grey value constancy term is complemented with gradient constancy to develop an algorithm robust to varying illumination.

$$\nabla I(x, y, t) = \nabla I(x + u, y + v, t + 1) \quad (4)$$

The data term is given by an energy functional that minimizes deviations from the above two assumptions. The global deviations over the entire image shall be measured by:

$$E_{data}(u, v) = \int (|I(\mathbf{x} + \mathbf{w}) - I(\mathbf{x})|^2 + \gamma|\nabla I(\mathbf{x} + \mathbf{w}) - \nabla I(\mathbf{x})|^2)\mathbf{dx} \quad (5)$$

where, $\mathbf{x} = (x, y, t)^T$ and $\mathbf{w} = (u, v, 1)^T$ and γ is the weight between the two assumptions.

Smoothness assumption: The fundamental smoothness term minimizes square of magnitude of flow gradient and penalizes flow discontinuities, i.e. high variations in u and v to attain a smooth flow field. The spatial smoothness term needed to be minimized can be given as:

$$E_{smooth}(u, v) = \int (|\nabla u|^2 + |\nabla v|^2)\mathbf{dx} \quad (6)$$

Thus, the entire function to be minimized is given by:

$$E(u, v) = \int [(|I(\mathbf{x} + \mathbf{w}) - I(\mathbf{x})|^2 + \gamma|\nabla I(\mathbf{x} + \mathbf{w}) - \nabla I(\mathbf{x})|^2) + (|\nabla u|^2 + |\nabla v|^2)]\ \mathbf{dx} \quad (7)$$

A solution (u,v) that minimizes E(u,v) should satisfy Euler Lagrange equations. After the standard discretization for derivatives by finite difference method, the resultant system of equations is solved with successive over relaxation iterations.

3.3.3 Designing motion templates representing ground truth

A primary step in this work was computing motion templates i.e. the ground truth flow corresponding to each expression embodying the ideal motion pattern arising when a face goes from neutral to emotional. The ideal motion pattern cannot be represented by the flow field associated with the progressive affective sequence of a single subject. Despite their generic nature, the intensity with which expressions are manifested usually varies for different individuals (Figure 5). Moreover, there is a possibility of discontinuous motion field due to inconsistent illumination, occlusion or head pose variation. To alleviate these effects, mean flow field was computed. A mean optical flow field of five different subjects portraying the same expression was assumed to represent the ground truth flow associated with the specific expression (Figure 6). Ground truth flow associated with each expression was required so as to have a reference with which the correlation of the optical flow computed for given image pair could be characterized using a well-defined metric. For each expression, two such motion templates were computed.

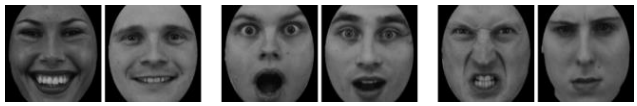

Figure 5. Variation in expression of happiness for different individuals [9]

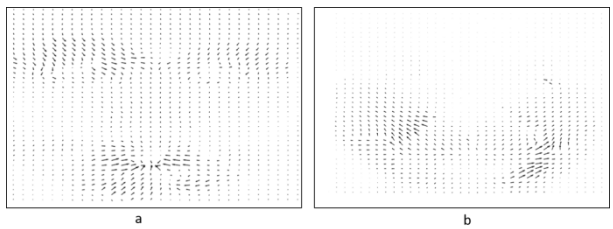

Figure 6. Sample motion template i.e. ground truth flow for (a) anger (b) happiness

3.3.4 Optical flow correlation metrics

To quantify resemblance between the computed optical flow and the ground truth flow, optical flow performance measures angular error (AE) and endpoint error (EE) values [44] were computed and concatenated with 2D correlation coefficients (Eq. (10)) between resultant horizontal flow and horizontal ground truth flow (Corr2Du) and resultant vertical flow and vertical ground truth flow (Corr2Dv).

AE between the optical flow vector (u,v) and the ground truth (u_{GT}, v_{GT}) is the angle in the three dimensional space between (u,v,1.0) and $(u_{GT}, v_{GT}, 1.0)$. A low value of AE

indicates that the resultant flow's direction is well matched with the direction of ground truth. If N is the total number of pixels in the image and $\sum_{\Omega}$ represents the summation over entire image domain Ω. Then,

$$\text{AE} = {}^{1}/_{\text{N}} [\sum_{\Omega} \cos^{-1}(\frac{(1.0 + \text{u} \times \text{u}_{\text{GT}} + \text{v} \times \text{v}_{\text{GT}})}{\sqrt{1.0 + \text{u}^2 + \text{v}^2}\sqrt{1.0 + \text{u}_{\text{GT}}^2 + \text{v}_{\text{GT}}^2}})] \quad (8)$$

EE (per pixel) can be defined as a measure of the absolute difference between the magnitudes of ground truth and resultant optical flow. Low value of EE indicates their magnitudes are comparable.

$$\text{EE} = {}^{1}/_{\text{N}} (\sum_{\Omega} \sqrt{(\text{u} - \text{u}_{\text{GT}})^2 + (\text{v} - \text{v}_{\text{GT}})^2}) \quad (9)$$

2D correlation coefficient between images A (mean of elements $\overline{\text{A}}$) and B (mean of elements $\overline{\text{B}}$) with m rows and n columns is given as:

$$\text{Corr2D} = [\sum_{\text{m}} \sum_{\text{n}} \frac{\text{A}_{\text{mn}} - \overline{\text{A}}}{\sqrt{\sum_{\text{m}} \sum_{\text{n}} (\text{A}_{\text{mn}} - \overline{\text{A}})^2}} \frac{\text{B}_{\text{mn}} - \overline{\text{B}}}{\sqrt{\sum_{\text{m}} \sum_{\text{n}} (\text{B}_{\text{mn}} - \overline{\text{B}})^2}}] \quad (10)$$

The values of Corr2D are computed for resultant horizontal flow and horizontal ground truth flow, Corr2Du and resultant vertical flow and vertical ground truth flow, Corr2Dv. Lower error values and high positive correlation indicate higher degree of similarity between the computed optical flow and the ground truth. Consider the image pair in Figure 2(a and b). The optical flow between the two images is depicted in Figure 2c. Table 3 displays the error and correlation values obtained when the aforementioned flow field was compared with the ground truth flow corresponding to anger, disgust and happiness.

Table 3. AE, EE, and Corr2D values for the three cases

	GT: anger	GT: disgust	GT: happiness
AE	7.39	6.05	9.64
EE	0.13	0.11	0.17
Corr2Du	0.15	0.54	0.86
Corr2Dv	0.57	0.86	0.56

GT: Ground Truth

The error and correlation coefficient values were determined for the computed optical flow and the ground truth flow associated with each expression. After concatenating all such error and correlation values, a feature vector *FlowCorr* was generated that had AE, EE, and 2D correlation coefficient values for the computed flow and ground truth flow corresponding to each expression.

The features or variables of this optical flow similarity based descriptor, reflect the degree of similitude between the expression depicted in the given images and each of the seven expressions.

From the remaining test sequences, first frame from each sequence was used as neutral frame and the labelled peak expression frame as emotional frame. In the testing stage, for the given test image pair, optical flow was found. For the resultant flow, *FlowCorr* descriptor was computed by comparing the flow with each of the pre-designed templates. The feature descriptors computed for the image pairs were fed to the classifier and for each pair, one of the expression labels was generated as output.

3.4 Training and testing stages for dyn-HOG descriptor

From the training sequences, first frame from each sequence was used as neutral frame and the labelled peak expression frame as emotional frame. For an image pair portraying a known expression, *dyn-HOG* was computed for different image pairs and their corresponding expression labels were fed to the classifier for training.

From the remaining test sequences, first frame from each sequence was used as neutral frame and the labelled peak expression frame as emotional frame. In the testing stage, for the given test image pair, dyn-HOG feature was computed as discussed previously. The features computed for the image pairs were fed to the classifier and for each pair, one of the expression labels was generated as output.

3.4.1 Computation of the dyn-HOG descriptor

For the given image pair with neutral image **N** and emotion image **E**, the difference image **D** was computed as the difference **E-N**. The difference image **D** was divided into small spatial regions or "cells" of size 8×8 pixels and histogram of gradient direction or orientation is computed for pixels of the cell. Each cell was discretized as angular bins as per gradient orientation. Feature vectors that fused 9 bin histograms across each block region (2×2 cells) formed the final representation (Figure 7).

Figure 7. a. Neutral face **N**; **b**. Face with an expression of disgust **E**; **c**. Corresponding difference image **D**; **d.** dyn-HOG visualization for **D**

The performance of the descriptors explained above for the task of FER, using the classifier described in the next section, will be discussed in Experimental Results.

3.5 Classification stage

After the feature vectors were computed, they were required to be fed to a classifier for recognizing the expression. In this study, a *Multi-class Support Vector Machine* [45] was employed. A linear Support Vector Machine (SVM) is essentially based on the notion of determining a suitable hyperplane i.e. a decision boundary that divides the dataset into two distinct classes (Figure 8). The distance between the nearest data point from a given class and the hyperplane is known as margin. A number of hyperplanes to classify the data might exist, but the most appropriate is the one that has the largest margin between the two classes, leading to a greater probability of classifying test data correctly. Error correcting output codes (ECOC) model segregates the task of multi-class classification into several binary classification problems. To train a model for k distinct labels, ECOC uses k(k-1)/2 linear SVM models with one-versus-one coding where for each

binary learner one class is taken to be positive and the other to be negative, ignoring the rest. In the end the class with maximum positive votes is assigned to the test data.

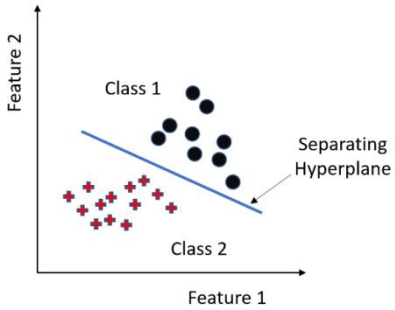

Figure 8. Hyperplane separating two classes in SVM

4. EXPERIMENTAL RESULTS

This section discusses the recognition performance of the proposed descriptors *FlowCorr* and dyn-HOG in comparison with optical flow and HOG on two public facial expression datasets.

4.1 Recognition performance of *FlowCorr* descriptor

Table 4 and Table 5 compare the recognition accuracy of the *FlowCorr* descriptor with the typical horizontal and vertical dense flow values fed as features to the classifier under the same experimental framework on CK+ and KDEF-dyn datasets respectively. As evident from the comparison, the proposed descriptor overall outperforms the dense flow values for recognizing facial expressions. For CK+, under same settings, except for the expression for *disgust* and *surprise*, all the other facial expressions were identified with higher accuracy with *FlowCorr* descriptor as compared to dense flow.

Table 4. Recognition accuracies (in percent) of the *FlowCorr* descriptor and the typical dense flow values when trained and tested on CK+ dataset

Emotion	An	Co	Di	Fe	Ha	Sa	Su
Hor Flow	75.76	74.07	91.67	77.78	91.18	82.22	89.02
Ver Flow	92.42	100	96.67	83.33	97.06	92.22	96.75
FlowCorr	99.24	100	93.89	98.61	99.02	98.89	96.34

Hor Flow: Horizontal flow field; Ver Flow: Vertical flow field

Table 5. Recognition accuracies (in percent) of the *FlowCorr* descriptor and the typical dense flow values when trained and tested on KDEF-dyn dataset

Emotion	An	Di	Fe	Ha	Sa	Su
Hor Flow	58.33	45.83	60.00	62.50	65.00	76.67
Ver Flow	75.00	78.33	73.33	90.00	81.67	70.83
FlowCorr	57.50	67.50	65.00	80.83	79.17	88.33

Hor Flow: Horizontal flow field; Ver Flow: Vertical flow field

Normalised confusion matrices for classification on CK+ with multi-class SVM classifier is displayed in Table 6. The expression of *fear* being misclassified as *sadness* by multi-class SVM is a common error that even humans are bound to commit [9].

In most of the cases it was observed that the error in the recognition surfaced due to slight ambiguity in the enacted expression. The facial images with error cannot be shared here due to terms of the CK+ dataset.

Table 6. Normalized confusion matrix for multi-class SVM model using *FlowCorr* descriptor

	An	*Co*	*Di*	*Fe*	*Ha*	*Sa*	*Su*
An	***99.24***	0.00	0.00	0.00	0.00	0.76	0.00
Co	0.00	***100***	0.00	0.00	0.00	0.00	0.00
Di	3.33	0.00	***93.89***	0.00	1.11	1.67	0.00
Fe	0.00	0.00	0.00	***98.61***	1.39	0.00	4.17
Ha	0.00	0.00	0.00	0.98	***99.02***	0.00	0.00
Sa	1.11	0.00	0.00	0.00	0.00	***98.89***	0.00
Su	0.00	0.00	0.00	1.22	0.00	2.44	***96.34***

4.2 Recognition performance of dyn-HOG Descriptor

Table 7 and Table 8 compare the classification performance of the dyn-HOG descriptor with the static HOG under the same experimental framework on CK+ and KDEF-dyn dataset respectively. The proposed descriptor slightly outperforms the HOG feature descriptor and thus may be visualized as an improved characterization of shape changes occurring during portrayal of facial expressions.

Table 7. Recognition accuracies (in percent) of the *dyn-HOG and HOG* descriptors when trained and tested on CK+ dataset

Emotion	An	Co	Di	Fe	Ha	Sa	Su
HOG	86.36	100	98.33	73.61	98.53	93.33	99.19
dyn-HOG	93.94	100	96.67	100	98.53	90.00	98.37

Table 8. Recognition accuracies (in percent) of the *dyn-HOG and HOG* descriptors when trained and tested on KDEF-dyn dataset

Emotion	An	Di	Fe	Ha	Sa	Su
HOG	83.33	80.83	85	85.83	75.00	70.83
dyn-HOG	84.17	86.67	70.83	92.50	90.00	91.67

Table 9 shows the comparison between the classification accuracy achieved on CK+ database of the proposed descriptors with other techniques.

It is to be noted that as per the scope of this work, the model performance is gauged with emotion display recognition accuracy as the chief parameter. The presented descriptors outperform most of the recently proposed state-of-the-art FER techniques in terms of recognition accuracies, confirming the greater suitability of the features proposed in this work to be employed in affect-sensitive interfaces and technologies that entail highly-precise emotion quantification. The features are proposed as improved versions of the existing features presented previously. They are validated as "improved version" with their higher accuracy on the different datasets. Furthermore, the proposed appearance- and shape-based features are novel, effective, and can be readily computed without any lengthy and intensive computations. The *FlowCorr* feature can also be easily extended by using more templates for a prospective improvement in facial expression recognition and discrimination.

Table 9. Expression recognition accuracies for seven classes of different FER models evaluated on CK+ database

Research Work	Accuracy (in percent)
Fan and Tjahjadi (2019) [46]	92.5
Meena et al. (2020) [47]	92.9
Wei et al. (2020) [39]	94.4
Makhmudkhujaev et al. (2019) [48]	94.5
Cheng and Zhou (2020) [49]	96.0
Chen and Hu (2020) [50]	96.3
Gan et al. (2020) [38]	96.3
De la torre et al. (2015) [51]	96.4
Qin et al.(2020) [40]	96.8
Dyn-HOG (with multi-class SVM)	96.8
Salmam et al. (2019) [19]	96.9
Li et al. (2020) [52]	97.4
Zhao et al. (2018) [32] with optical flow	97.5
FlowCorr (with mult-iclass SVM)	***98.0***
Sadeghi and Raie (2019) [53]	98.2

Table 10 encapsulates percentage accuracy comparison for each emotion of the presented descriptors with a perceptual study of human interpretation of basic emotions conducted by Calvo et al. [9] on KDEF-dyn dataset. In the study, participants were asked to view dynamic video clips and identify the emotion portrayed by the subject in the clip. Inevitably, human judgment surpasses the results by the proposed method. However, the overall results are comparable and in-line with human cognition.

Table 10. Expression recognition accuracies (in percent) for six classes of different FER methods evaluated on KDEF-dyn database

Emotion	*Calvo et al. [9]*	*FlowCorr*	*dyn-HOG*
Anger	91.7	57.5	84.2
Disgust	77.8	67.5	86.7
Fear	68.6	65.0	70.8
Happiness	98.5	80.8	92.5
Sadness	80.7	79.2	90.0
Surprise	93.7	88.3	91.7

5. CONCLUSION

In this article, we introduced two dynamic descriptors for recognition of seven facial expressions viz. *anger, contempt, disgust, fear, happiness, sadness, surprise*, from facial image pairs of a neutral face and an emotional face. The first descriptor quantified similarity between the flow pattern associated with the image pair and the precomputed ground truth flow for each expression. Whereas, the second descriptor gave the HOG values of the difference of the two images. The descriptors were demonstrated to carry discriminative information and successfully accomplished the task of facial expression categorization. Recognition accuracies with multi-class SVM classifier on CK+ dataset was on par with most of state-of-the-art techniques and on KDEF-dyn were found to be comparable with human cognition.

The presented framework was evaluated on datasets which had images with expressions enacted in a controlled environment. For real-world images with spontaneous expressions the expressions may not be classified with such high accuracy. Future work is expected to focus on testing the efficiency of model in real-time facial expression decoding and developing it to be robust to pose variation, partial or full occlusion.

ACKNOWLEDGMENT

The authors wish to express most sincere gratitude to the anonymous reviewers and the editor for conscientiously going through the original manuscript and providing meaningful suggestions for enhancing its quality.

REFERENCES

[1] Ekman, P., Friesen, W.V. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2): 124-129. https://doi.org/10.1037/h0030377

[2] Matsumoto, D., Willingham, B. (2009). Spontaneous facial expressions of emotion of congenitally and noncongenitally blind individuals. Journal of Personality and Social Psychology, 96(1): 1-10. https://doi.org/10.1037/a0014037

[3] Matsumoto, D. (1992). More evidence for the universality of a contempt expression. Motivation and Emotion, 16: 363-368. https://doi.org/10.1007/BF00992972

[4] Lyons, M.J., Akamatsu, S., Kamachi, M., Gyoba, J. (1998). Coding facial expressions with Gabor wavelets. Proc. 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200-205. https://doi.org/10.1109/AFGR.1998.670949

[5] Singhal, A., Kumar, P., Saini. R., Roy, P.P., Dogra, D.P., Kim, B.G. (2018). Summarization of videos by analyzing affective state of the user through crowdsource. Cognitive Systems Research, 52: 917-930. https://doi.org/10.1016/j.cogsys.2018.09.019

[6] Caballero, A.F., Martinez-Rodrigo, A., Manuel Pastor, J., Castillo, J.C., Lozano-Monasor, E., Lopez, M.T., Zangroniz, R., Latorre, J.M., Fernandez-Sotos, A. (2016). Smart environment architecture for emotion detection and regulation. Journal of Biomedical Informatics, 64: 55-73. https://doi.org/10.1016/j.jbi.2016.09.015

[7] Huang, D., De la Torre, F. (2010). Bilinear kernel reduced rank regression for facial expression synthesis. In European Conference on Computer Vision, pp. 364-377. https://doi.org/10.1007/978-3-642-15552-9_27

[8] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I. (2010). The extended Cohn-Kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 94-101. https://doi.org/10.1109/CVPRW.2010.5543262

[9] Calvo M.G., Fernández-Martín, A., Recio, G., Lundqvist. D. (2018). Human observers and automated assessment of dynamic emotional facial expressions: KDEF-dyn database validation. Frontiers in Psychology, 9: 2052. https://doi.org/10.3389/fpsyg.2018.02052

[10] Lundqvist, D., Flykt, A., Öhman, A. (1998). The Karolinska directed emotional faces (KDEF). CD ROM from Department of Clinical Neuroscience, Psychology Section, Karolinska Institute, 91(630): 2-2.

[11] Ekman, P., Friesen, W.V. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto.

[12] Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J. (1999). Classifying facial actions. IEEE

Transactions on Pattern Analysis and Machine Intelligence, 21: 974-989. https://doi.org/10.1109/34.799905
[13] Li, Y., Wang, S., Zhao, Y., Ji, Q. (2013). Simultaneous facial feature tracking and facial expression recognition. IEEE Transactions on Image Processing, 22: 2559-2573. https://doi.org/10.1109/TIP.2013.2253477
[14] Yurtkan, K., Demirel, H. (2014). Entropy-based feature selection for improved 3D facial expression recognition. Signal Image Video Processing, 8(2): 267-277. https://doi.org/10.1007/s11760-013-0543-1
[15] Mohammadian, A., Aghaeinia, H., Towhidkhah, F. (2016). Incorporating prior knowledge from the new person into recognition of facial expression. Signal Image Video Processing, 10(2): 235-242. https://doi.org/10.1007/s11760-014-0732-6
[16] Ashir, A.M., Eleyan, A. (2017). Facial expression recognition based on image pyramid and single-branch decision tree. Signal Image Video Processing, 11(6): 1017-1024. https://doi.org/10.1007/s11760-016-1052-9
[17] Agarwal, S., Santra, B., Mukherjee, D.P. (2018). Anubhav: Recognizing emotions through facial expression. The Visual Computer, 34(2): 177-191. https://doi.org/10.1007/s00371-016-1323-z
[18] Happy, S.L., Routray, A. (2015). Robust facial expression classification using shape and appearance features. 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR), Kolkata, pp. 1-5. https://doi.org/10.1109/ICAPR.2015.7050661
[19] Salmam, F.Z., Madani, A., Kissi, M. (2019). Fusing multi-stream deep neural networks for facial expression recognition. Signal Image and Video Processing, 13(3): 609-616. https://doi.org/10.1007/s11760-018-1388-4
[20] Zhang, Y., Ji, Q. (2005). Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(5): 699-714. https://doi.org/10.1109/TPAMI.2005.93
[21] Aleksic, P.S., Katsaggelos A.K. (2006). Automatic facial expression recognition using facial animation parameters and multistream HMMs. IEEE Transactions on Inf. Forensics Security, 1(1): 3-11. https://doi.org/10.1109/TIFS.2005.863510
[22] Mlakar, U., Potocnik, B. (2015). Automated facial expression recognition based on histograms of oriented gradient feature vector differences. Signal Image and Video Processing. 9(1): 245. https://doi.org/10.1007/s11760-015-0810-4
[23] Wehrle, T., Kaiser, S., Schmidt, S., Scherer, K.R. (2000). Studying the dynamics of emotional expression using synthesized facial muscle movements. Journal of Personality and Social Psychology, 78(1): 105-119. https://doi.org/10.1037//0022-3514.78.1.105
[24] Sun, N., Li, Q., Huan, R., Liu, J., Han, G. (2019). Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recognition Letters, 119: 49-61. https://doi.org/10.1016/j.patrec.2017.10.022.
[25] Mase, K. (1991). Recognition of facial expression from optical flow. IEICE Transactions on Information and Systems, 74(10): 3474-3483.
[26] Essa, I.A., Pentland, A. (1994). A vision system for observing and extracting facial action parameters. In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 76-83. https://doi.org/10.1109/CVPR.1994.323813
[27] Yacoob, Y., Davis, L.S. (1996). Recognizing human facial expressions from long image sequences using optical flow. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(6): 636-642. https://doi.org/10.1109/34.506414
[28] Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J. (1999). Measuring facial expressions by computer image analysis Psychophysiology, 36: 253-263. https://doi.org/10.1017/S0048577299971664
[29] Niese, R., Al Hamadi, A., Farag, A., Neumann, H., Michaelis, B. (2012). Facial expression recognition based on geometric and optical flow features in colour image sequences. IET Computer Vision, 6(2): 79-89. https://doi.org 10.1049/iet-cvi.2011.0064
[30] Pu, X., Fan, K., Chen, X., Ji, L., Zhou, Z. (2015). Facial expression recognition from image sequences using twofold random forest classifier. Neurocomputing, 168: 1173-1180. https://doi.org/10.1016/j.neucom.2015.05.005
[31] Shojaeilangari, S., Yau, W.Y., Nandakumar, K., Li, J., Teoh, E.K. (2015). Robust representation and recognition of facial emotions using extreme sparse learning. IEEE Transactions on Image Processing, 24(7): 2140-2152. https://doi.org/10.1109/TIP.2015.2416634
[32] Zhao, J., Mao, X., Zhang, J. (2018). Learning deep facial expression features from image and optical flow sequences using 3D CNN. The Visual Computer, 34(10): 1461-1475. https://doi.org/10.1007/s00371-018-1477-y
[33] Dalal, N., Triggs, B. (2005). Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, pp. 886-893. https://doi.org/10.1109/CVPR.2005.177
[34] Kumar, P., Happy, S.L., Routray, A. (2016). A real-time robust facial expression recognition system using HOG features. In 2016 International Conference on Computing, Analytics and Security Trends (CAST), pp. 289-293. https://doi.org/10.1109/CAST.2016.7914982
[35] Wang, X., Jin, C., Liu, W., Hu, M., Xu, L., Ren, F. (2013). Feature fusion of HOG and WLD for facial expression recognition. In Proceedings of the 2013 IEEE/SICE International Symposium on System Integration, pp. 227-232. https://doi.org/10.1109/SII.2013.6776664
[36] Nigam, S., Singh, R., Misra, A.K. (2018). Efficient facial expression recognition using histogram of oriented gradients in wavelet domain. Multimedia Tools and Applications, 77: 28725-28747. https://doi.org/10.1007/s11042-018-6040-3
[37] Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., Dobaie, A.M. (2018). Facial expression recognition via learning deep sparse autoencoders. Neurocomputing, 237: 643-649. https://doi.org/10.1016/j.neucom.2017.08.043
[38] Gan, Y., Chen, J., Yang, Z., Xu, L. (2020). Multiple attention network for facial expression recognition. IEEE Access, 8: 7383-7393. https://doi.org/10.1109/ACCESS.2020.2963913
[39] Wei, W., Jia, Q., Feng, Y., Chen. G., Chu, M. (2020). Multi-modal facial expression feature based on deep-neural networks. Journal of Multimodal User Interfaces 14: 17-23. https://doi.org/10.1007/s12193-019-00308-9
[40] Qin, S., Zhu, Z., Zou, Y., Wang, X. (2020). Facial expression recognition based on Gabor wavelet transform and 2-channel CNN. International Journal of

Wavelets, Multiresolution and Information Processing, 18(02): 2050003. https://doi.org/10.1142/S0219691320500034

[41] Viola, P., Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, pp. I-I. https://doi.org/10.1109/CVPR.2001.990517

[42] Brox, T., Bruhn, A., Papenberg, N., Weickert, J. (2004). High accuracy optical flow estimation based on a theory for warping. In: Pajdla T., Matas J. (eds) Computer Vision - ECCV 2004. ECCV 2004. Lecture Notes in Computer Science, vol 3024. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-24673-2_3

[43] Horn, B.K.P., Schunck, B.G. (1981). Determining Optical Flow. Artificial Intelligence, 17: 185-203. https://doi.org/10.1016/0004-3702(81)90024-2

[44] Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black M.J., Szeliski R. (2011). A database and evaluation methodology for optical flow. Intl J. of Computer Vision, 92: 1-31. https://doi.org/10.1007/s11263-010-0390-2

[45] Allwein, E.L., Schapire, R.E., Singer, Y. (2000). Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research, 1: 113-141.

[46] Fan, X., Tjahjadi, T. (2019). Fusing dynamic deep learned features and handcrafted features for facial expression recognition. Journal of Visual Communication and Image Representation. 65: 102659. https://doi.org/10.1016/j.jvcir.2019.102659

[47] Meena H.K., Sharma K.K., Joshi, S.D. (2020). Effective curvelet-based facial expression recognition using graph signal processing. Signal, Image and Video Processing, 14: 241-247. https://doi.org/10.1007/s11760-019-01547-9

[48] Makhmudkjujaev, F., Abdullah-Al-Wadud, M., Iqbal, M.T.B., Ryu, B., Chae, O. (2019). Facial expression recognition with local prominent directional pattern. Signal Processing: Image Communication, 74: 1-12. https://doi.org/10.1016/j.image.2019.01.002

[49] Cheng, S., Zhou, G. (2020). Facial expression recognition method based on improved VGG convolutional neural network. International Journal of Pattern Recognition and Artificial Intelligence, 34: 2056003. https://doi.org/10.1142/S0218001420560030

[50] Chen, W., Hu, H. (2020). Joint prominent expression feature regions in auxiliary task learning network for facial expression recognition. Electronics Letters, 55: 22-24. https://doi.org/10.1049/el.2018.7235

[51] De la Torre, F., Chu, W.S., Xiong, X., Vicente, F., Ding, X., Cohn, J. (2015). IntraFace. 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, pp. 1-8. https://doi.org/10.1109/FG.2015.7163082

[52] Li, K., Jin, Y, Akram, M.W., Han, R., Chen, J. (2020). Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. The Visual Computer, 36: 391-404. https://doi.org/10.1007/s00371-019-01627-4

[53] Sadeghi, H., Raie, A.A. (2019). Histogram distance metric learning for facial expression recognition. Journal of Visual Communication and Image Representation, 62: 152-165. https://doi.org/10.1016/j.jvcir.2019.05.004