Electronic Letters on Computer Vision and Image Analysis 20(2):1-21; 2021

Modeling and Analysis of Facial Expressions Using Optical Flow-Derived Divergence and Curl Templates

Shivangi Anthwal* and Dinesh Ganotra*

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

Received 29th of July, 2020; accepted 28th of April, 2021

Abstract

Facial expressions (FEs) are integral part of non-verbal paralinguistic communication as they provide cues vital for perceiving one's emotional state. Assessment of emotions through FEs is an active research domain in computer vision due to its potential applications in multifaceted domains. In this work, an approach is presented wherein FEs are modeled and analyzed with dense optical flow-derived divergence and curl templates that embody the ideal motion pattern of facial features pertaining to the unfolding of an expression on the face. Two types of classification schemes based on multi-class support vector machine and k-nearest neighbor have been employed for evaluation. The efficacy of the approach has been validated with promising results obtained from a comparative analysis of the proposed approach with the state-of-the-art FE recognition techniques on CK+ and JAFFE datasets and with human cognition and pre-trained Microsoft face application programming interface on KDEF dataset.

Key Words: Facial Expression Recognition, Emotion Analysis, Optical Flow, Multi-Class Support Vector Classification, k-Nearest Neighbor Classification, Human Cognition versus Machine Analysis.

1 Introduction

The face houses the apparatus for producing both verbal and non-verbal cues essential for interpersonal communication. To appraise the effect of lexical content, prosodic cues, and the facial expressions (FEs) in a conversation, Mehrabian [1] conducted an empirical cognitive investigation and concluded that for a particular message conveyed, the spoken words would contribute to just 7% of the overall impact of the message, while the voice intonation of the speaker contributes to 38%. FEs of the speaker play the most integral role in communication by contributing for a substantial 55% to the overall impact of the spoken message. Therefore, a paradigm shift of incorporating FE as a communication channel in human-machine interaction is expected to render the interactive process more effectual, thereby optimizing the user experience. With the ubiquity of smart devices and environments, assessment of human affective behavior through facial FEs has engendered considerable interest. A recent noteworthy instance is multi-national mass media and entertainment conglomerate the Walt Disney attempting to automatically gauge the audience response to its movies by capturing their faces with infrared cameras during its movie screenings [2]. They used a novel algorithm that tracked facial behavior and could even predict when the audience would smile or

Correspondence to: shivangianthwal@hotmail.com
Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.1275
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

laugh at specific moments in the movies. This provided a more accurate and reliable insight into how its audience actually felt about its movies rather than typical reviews and surveys that could be prejudiced and may have people suppressing their genuine opinions.

After extensive cross-cultural investigations, Ekman and Friesen [3] proffered the discrete emotion theory that asserts the existence of certain fundamental emotions, namely anger, disgust, fear, happiness, sadness, and surprise, that have a prototypical expression unfolding pattern associated with each of them (Figure 1). Subsequently, the expression of contempt was also added to the list of these emotions expressed universally in similar fashion [4]. Automated recognition of these discrete emotions has been demonstrated to find utility in an expansive range of domains such as affective video summarization [5] and recommendation [6], ambient assisted living [7], providing interactive aid to kids with autism spectrum disorder [8], and interactive video gaming [9]. In pursuit of gauging performance of different methods for recognition of these discrete emotions conveyed through single image or image sequences, diverse benchmark databases [10–14] having subjects from different cultures, ethnicities, and belonging to different age groups have been proposed in the last two decades. Typically, methods aiming to attain facial expression recognition (FER) begin by taking images from a dataset and subsequently locating and cropping facial region in those images. This is followed by extracting relevant features that facilitate the characterization and subsequent analysis of the features with final stage of categorization of the FE portrayed by the subject in the images/videos by an adequate classifier.

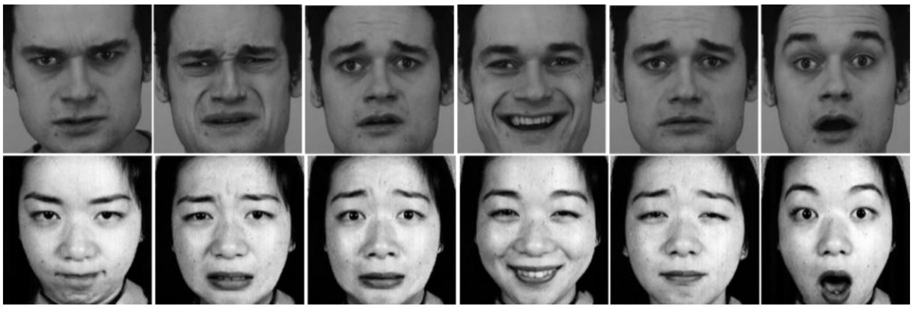


Figure 1: Six universal expressions. (Left to Right): *anger, disgust, fear, happiness, sadness, surprise* [11,12]

Motivated by the growing need of inducing emotional intelligence in machines, in this work an approach to recognize the seven principal emotions, namely *anger*, *contempt*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*, with their corresponding FEs has been introduced. The key features of the work can be summarized as follows:

- This work presents holistic features to model and quantify discrete emotions with pattern of facial motion arising during unfolding of their corresponding FE depicted in progressive image sequences. Divergence and curl templates derived from the dense optical flow field corresponding to facial motion are utilized as descriptors characterizing the motion. To the best knowledge of authors, this is the first work that uses divergence- and curl-based features as global or holistic descriptors, and thereby differs from previous similar approaches, which were atomistic, i.e., region-based. Furthermore this is the first time that for training purpose, divergence and curl templates are used instead of using the spatial distribution of divergence and curl of individual sequences.
- To validate the features, a cross-database evaluation was performed on Karolinska Directed Emotional Faces database and the results were compared with human cognition and pre-trained Microsoft face application programming interface (API). The results obtained are in consonance with the human cognition, thus validating the features as most affect-sensitive artificial vision frameworks intend to emulate the innate cognitive capacity of humans. Thereby, an agreement of the results obtained with human perception of FEs substantiates the usefulness of the presented features in proactive vision-based affect-sensitive systems.

The rest of the paper is structured as follows. Section 2 presents an overview of the relevant literature in the field of FER using visual information. Section 3 elucidates the concept of optical flow highlighting its significance and contribution in emotion recognition. The proposed method is described in detail in Section 4, whereas the summary of results obtained is discussed in Section 5. Conclusion and future directions are succinctly outlined in Section 6.

2 Related work

The pioneer model that combined physical cues with the anatomical knowledge of facial behavior was presented by Ekman and Friesen [15], who developed facial action coding system (FACS) that describes facial movements by mapping them onto a facial action unit (AU) space. FACS manual provides a concise linguistic elucidation of subtle and profound changes in facial configurations in terms of AUs for an objective measurement of facial activity. Presence of a single AU or combination of multiple AUs can represent a wide spectrum of FEs. For example, for conveying disgust, either the "nose wrinkle" or "upper lip raiser" must be present; for conveying happiness, "lip corner puller" must be there [10]. Inspired by FACS, Tong et al. [16] used dynamic Bayesian network to characterize probabilistic relationships among different AUs. They demonstrated with their experiments that systematically integrating AU measurement with temporal dynamics of AUs and their probabilistic relationship with other AUs yielded higher recognition rate of different AUs.

Techniques presented hitherto for automatic FER from visual information can be broadly categorized into static or dynamic, depending on whether temporal information is utilized or not. Static- or frame-based methods categorize the emotion from still images embodying the momentary appearance of the FE, generally in its peak form. Silva et al. [17] presented a compact and effective description of face depicting an FE based on horizontal and vertical distance between distinct fiducial points whose locations were known a priori. In one of the experiments, they integrated this geometric feature vector with Gabor filters to extract appearance features and demonstrated that complementing the two features enabled a better discrimination between different emotions. Lopes et al. [18] presented a system that used convolutional neural networks (CNNs) for extracting visual features such as shapes, edges, corners, and end-points of eyes, eyebrows, and lips. They also used pre-processing to eliminate effects of pose, brightness change, improper lighting, etc. The method designed by Ashir et al. [19] integrated multiresolution pre-processing for feature extraction with compressive sensing theory for dimensionality reduction. Each input facial image was fed to a pyramid level wherein features based on image gray levels were extracted and subsequently concatenated with corresponding features from other levels to form a feature vector. A random variable Gaussian matrix was employed for collecting compressed measurements from different pyramid levels and were fed to a multi-class support vector machine (SVM) classifier. Ding et al. [20] employed the illumination invariant logarithmic Laplace domain and extracted double local binary pattern (LBP)-based features from raw images described with Taylor series expansion. Similarity between extracted features was estimated with nearest neighbor classifier using chi-square distance. Bougourzi et al. [21] introduced a novel the pyramid multi-level (PML) face representation and integrated transformed handcrafted features with deep features for static FER. The appraised the optimal level of PML features of the handcrafted descriptors and combined them with the transformed face layers to obtain a compact image descriptor. They obtained accuracies competent with the state-of-the-art FER approaches in both within-database and cross-database experiments.

Dynamic methods generally model the temporal development of facial features and the correlation among them across different frames. Zhang et al. [22] attempted to capture temporal evolution of facial physical structure for expressions portrayed in a given video. They designed a hierarchical recurrent neural network for extracting dynamic features based on facial landmarks. The landmarks were decomposed into four different parts, with each part being fed as input to a separate subnet. Their architecture also comprised multi-signal CNN with one signal to enhance the variation among separate expression classes and the other to alleviate disparities among same expressions. Agarwal et al. [24] modeled a real time FER framework Anubhav which extracted features only from salient parts of the face carrying expression-related information. They exploited both spatial and temporal dimensions to achieve competitive recognition accuracies on benchmark datasets. Siddiqi et al. [25] presented an offline FER architecture that used stepwise linear discriminant analysis complemented with hidden conditional random fields model. The former extracted relevant features from input expression images with the help of partial F-test values to curtail intra-class differences and inflate inter-class variation. The latter, adept to approximate complex distributions with Gaussian density functions, was used for classification of the extracted features. Salmam et al. [25] presented a hybrid model that coupled a CNN representing appearance-based features such as wrinkles and skin folds and a deep neural network based on geometric features characterizing salient facial parts such as eyes, nose, and mouth. They demonstrated the increase in efficiency of FER by integrating both types of features. Danelakis et al. [26] accomplished dynamic FE retrieval for 3D face scans with the aid of spatial information of facial landmarks and their subsequent wavelet transformation.

In a video with subjects involved in a spontaneous conversation, the speech articulation process conspicuously influences facial configuration and has been observed to reduce the FER accuracy as compared to the case where the subjects are not talking. Bursic et al. [27] noted that while examining FEs of subjects involved in such conversations, the speaking effect needs to be regarded as a crucial factor. They developed a deep neural network-based model that analyzed cues related to facial features and speech articulation extracted from a model trained for lipreading. Their experiments on RAVDESS dataset validated their conjecture that the incorporation of features associated with speech articulation process increased the FER accuracy. Wehrle et al. [28] remarked that employing temporal information led to a better understanding of FEs as a neutral face image could be used as a reference state. However, for the situations with unavailability of a neutral face, static techniques for FER have an edge over dynamic techniques. To circumvent this issue, "average human face" as an alternative to neutral face could also be employed in a dynamic model [29]. The approach presented in this work is dynamic and uses optical flow-based descriptive features to encode visual motion appearance. A brief sketch on flow-based approaches employed for FER is provided in the next section.

3 Optical flow

Estimating optical flow is a vital step for dynamic scene interpretation as it gives the relative displacement of image gray values in a time-varying image sequence. Assessing flow from image sequences is an active research area and has witnessed overwhelming progress in the last two decades. FEs may be regarded as dynamic variation of facial components such that muscular movement of ocular region, forehead, lips, cheeks, nose, etc. result in large scale variations, whereas fine scale variations arise because of subtle skin deformations [30]. The difference in appearance of facial features can be represented by the optical flow between emotional and neutral faces. Two images, one with a neutral face and one depicting an expression (Figures 2a and 2b) will have an optical flow field associated with them. This 2D vector field (Figure 2c) indicates apparent velocity of each pixel quantifying the apparent facial motion that arises when the face advances from neutral state to emotional state. At discrete spatial locations, optical flow is denoted by a vector, whose orientation gives flow direction and length characterizes flow magnitude. There are discernible changes around the forehead and in the mouth region when a face turns from neutral to "angry." This explains maximum flow component in the corresponding region in the optical flow field diagram.

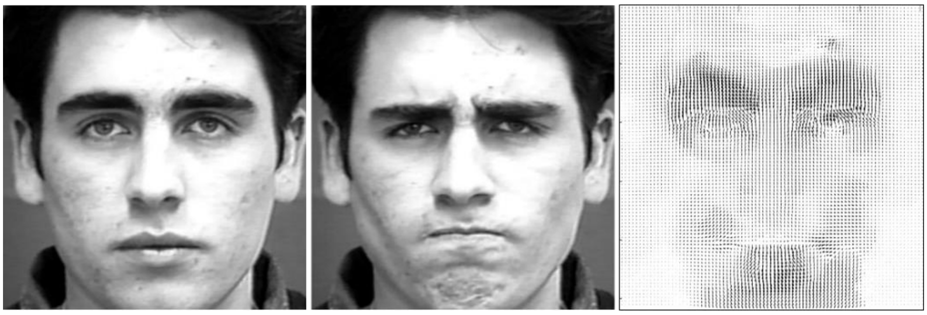


Figure 2a: Neutral face; 2b: Face conveying *anger*; 2c: Corresponding optical flow field

3.1 Optical flow computation techniques and applications

Classical flow computation techniques typically function by finding the spatial and temporal derivatives of pixel gray levels and optimizing a local or global objective function. Notwithstanding their simplicity, such methods are unsuitable for real-world scenarios with motion discontinuities and large displacements. Modern approaches employ supervised or unsupervised learning-based architectures to attain reliable estimates of optical flow. However, such approaches suffer from computational complexities.

Optical flow finds application in a wide array of domains such as cell deformation analysis [31], medical image registration [32], blood flow estimation [33], video indexing and retrieval [34,35], and obstacle detection and avoidance in real and virtual environments [36].

3.2 Optical flow and FER

About two decades ago, Mase [37] pioneered the analysis of FEs using optical flow. He utilized top-down and bottom-up approach to study movement of facial muscles and classified expressions into *anger, disgust, happiness,* and *surprise*. To study expressions with spatiotemporal models, Essa and Pentland [38] included temporal dimension in the FACS framework and modified it to FACS+. They tracked the facial changes occurring during the portrayal of an expression by superimposing a mesh on the face images and tracking the mesh corners using optical flow. Pu et al. [39] presented a novel framework for analysis of FEs by recognizing AUs from input image sequences with a two-fold random forest (RF) classifier. Facial motion was quantified by tracking active appearance model-based feature points with optical flow-based tracker giving the displacements of the feature points between the neutral and peak expression frames. The resultant displacement vectors were fed to the first level of RF to detect the AUs present in the corresponding input expression sequences. The detected AUs were fed as input to the second level of RF for categorization of FEs.

Zhao et al. [40] introduced accumulated optical flow between non-consecutive input facial frames as a feature descriptive of the global motion. Using both static and dynamic features as input to 3D CNNs, they achieved high recognition accuracies. Multi-channel deep spatial temporal feature fusion neural network presented by Sun et al. [29] fused spatial and temporal features for analyzing expressions from image pairs. The gray levels of input emotional face images acted as spatial features, whereas the optical flow field corresponding to the changes between peak expression face and neutral face was utilized as the temporal feature. They also proposed the use of average human face as a substitute to the neutral face for the cases where neutral face was not available for reference. Pan et al. [41] utilized the magnitude of the optical flow between successive frames in a video to characterize their relative motion as a form of a temporal channel in their spatiotemporal video-based FER model.

4 Evaluation methodology

In this work, a series of experiments were conducted, and for evaluation, images were taken from the Extended Cohn-Kanade (CK+) [10], the Karolinska Directed Emotional Faces (KDEF) [11], and the Japanese Female Facial Expression (JAFFE) [12] datasets. For CK+, the sequences investigated had either of the seven FE labels viz. *anger, contempt, disgust, fear, happiness, sadness,* and *surprise*. For JAFFE and KDEF, the labels were anger, disgust, fear, happiness, sadness, surprise. During evaluation, the first step was to extract the facial region in the images and crop it from the rest of the image. During the training stage, the optical flow-derived motion templates were used to find the feature descriptors that describe the spatial distribution of divergence and curl across the entire facial region and were fed to either multi-class SVM or k-nearest neighbor (k-NN)-based classifiers. For testing, the optical flow field associated with image pair comprising an emotional face and the corresponding neutral face was used to determine the spatial distribution of divergence and curl of the flow field that were used as input for recognition with the classifiers. The following sub-sections explicate the steps used for evaluation.

4.1 Image pre-processing: face localization and resizing

The foremost step for expression analysis from an image with a subject displaying an emotion is to extract the region containing salient information, i.e., the facial region. This was done by employing the

Viola-Jones algorithm [42], a technique suitable for an expeditious and reliable detection of frontal upright faces in input images. It utilizes histogram of oriented gradients features, Haar-like features, and LBPs complemented with cascaded classifiers trained by boosting. The extracted facial region was eventually cropped from the image and resized to a 256 × 256 size grayscale image (Figure 3).

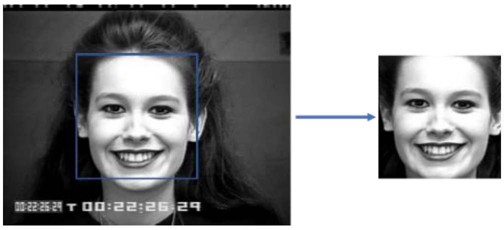


Figure 3: Locating face in the given image using Viola-Jones algorithm and cropping the same from the background

4.2 Optical flow computation

For determination of optical flow, global variational mechanism devised by Brox et al. [43] was utilized due to its validated robustness to noise and illumination changes. Global models find a dense field by operating over the entire image domain. To determine a reliable flow field, the model uses the following assumptions.

4.2.1 Gray value constancy assumption or brightness constancy assumption (BCA)

Since the seminal work by Horn and Schunck [44], it has been presumed that gray value of a pixel does not vary despite the change in its position in successive frames. Mathematically,

$$I(x, y, t) = I(x + u, y + v, t + 1) \tag{1}$$

Taylor expansion of (1) gives the optical flow constraint:

$$I_x u + I_y v + I_t = 0 \tag{2}$$

4.2.2 Gradient constancy assumption (GCA)

Brox et al. [47] coupled BCA with gradient constancy to develop a technique robust to illumination changes.

$$\nabla I(x, y, t) = \nabla I(x + u, y + v, t + 1) \tag{3}$$

The global deviations over the entire image domain can be measured by the data term given as:

$$E_{data}(u, v) = \int (|I(\mathbf{x} + \mathbf{w}) - I(\mathbf{x})|^2 + \gamma |\nabla I(\mathbf{x} + \mathbf{w}) - \nabla I(\mathbf{x})|^2) \mathbf{dx} \tag{4}$$

Where **x**=(x,y,t)T and **w**=(u,v,1)T and γ is the weight between the two assumptions. The integral $\int$ covers the entire spatial domain of the image.

4.2.3 Smoothness assumption

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

$$E_{smooth}(u,v) = \int (|\nabla u|^2 + |\nabla v|^2)\mathbf{dx} \qquad (5)$$

The total function supposed to be minimized is given as:

$$E(u,v) = \int [(|I(\mathbf{x}+\mathbf{w}) - I(\mathbf{x})|^2 + \gamma|\nabla I(\mathbf{x}+\mathbf{w}) - \nabla I(\mathbf{x})|^2) \; + \; ((|\nabla u|^2 + |\nabla v|^2)]\mathbf{dx} \qquad (6)$$

Where, the integral $\int$ covers the entire spatial domain of the image. The values of u and v that minimize E(u,v) should satisfy Euler-Lagrange equations. The system of equations resulting from the discretization for derivatives, is solved with successive over relaxation iterations. In this work, in an emotional-neutral image pair, neutral facial image is given by I(x,y,t) and the emotional face is represented by I(x+u,y+v,t+1). (u,v) gives the optical flow for this image pair. The color or the gray value for any point on the face represented by a pixel or a group of pixels should remain unaltered in different frames even when there is motion. In other words, these values would remain constant regardless of the motion. Any plausible changes in different frames due to unwanted illumination variations or noise have been validated as being well-handled by the method proposed by Brox et al. [43]. Thereby, the method is suitable for the computation of the associated optical flow field.

4.3 Designing motion templates associated with each expression using optical flow

A principal step in this work was the computation of motion templates corresponding to each FE that depict the ideal motion pattern arising when a face advances from neutral to emotional. These templates were subsequently used for deriving divergence and curl templates/descriptors used for training the classifiers. To derive a template that can embody the motion pattern of unfolding of an FE, distinct image pairs of neutral and emotional faces associated with that FE portrayed by different subjects were taken and the flow fields (size: 256 × 256) corresponding to each pair were computed. Motion template (size: 256 × 256) associated with the FE was computed as the mean of all those optical flow fields. In this fashion, for each FE, three such templates were computed. The reason for using a template-based representation derived from mean flow fields to represent the ideal motion pattern is two-fold. First, to avoid irregularities and discontinuities that may occur in flow field for some subjects due to head pose variation, inconsistent illumination, eye blink, etc., as the effects get minimized when an average of different subjects is taken. Moreover, despite the similarities in their pattern, the intensity with which emotions are manifested, i.e., the level of expressiveness varies for different individuals (Figure 4). Effects of extremely low or high emotion intensities also get alleviated in the mean field. Thus, mean flow field-based template representation is a more adequate and effective representation for the ideal motion pattern corresponding to different FEs. The template-derived descriptors used in this work are outlined in the next section.

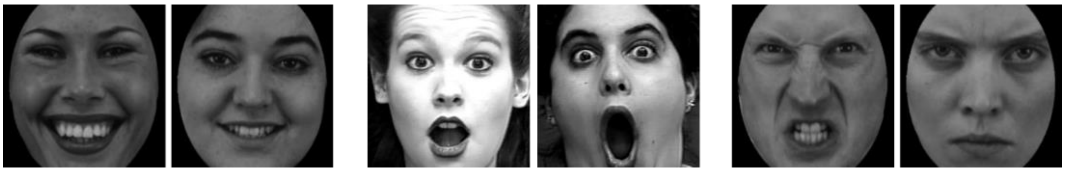


Figure 4: Variation in expression intensity of (a) *happiness* (b) *surprise* (c) *anger* for different individuals

4.4 Flow-based descriptors

The fundamental notion for the approach presented is to categorize the FEs based on the motion pattern of facial components that arises when a face advances from neutral to emotional state. As remarked by Black and Yacoob [45] and Shojaeilangari et al. [46], divergence and curl of flow field representing the motion may give pertinent information related to the expansion/contraction and circular motion of facial components. The spread of the optical flow field at any point may be quantified by the value of its divergence at that point. The regions with a greater spread have higher divergence value as opposed to the regions with zero or no spread. Likewise, curl of the optical flow field is anticipated to be greater at regions with a significant circular motion in comparison to the regions with absence of circular motion. Thus, divergence and curl of the flow field computed at each spatial location can yield a distinct pattern that can effectually characterize the facial motion and form adequate global descriptors for effective representation of facial motion. To the best knowledge of authors, this is the first work that uses divergence- and curl-based features as global or holistic descriptors in contrast to previous approaches which were atomistic, i.e., region-based. Furthermore, this is the first time that divergence and curl are computed from flow-derived motion template rather than the optical flow field. Consider a sample motion template $\mathbf{w} = (\mathrm{u}, \mathrm{v})$ with u and v the horizontal and vertical flow, respectively. The three types of features/descriptors derived from the motion templates are described below:

4.4.1 Divergence-based descriptor

The first type of descriptor *FlowD* (size: 256 × 256) is based on the divergence of motion field. It quantifies the expansion or contraction of facial components occurring while a face goes from neutral to emotional. As an illustrative example, Figures 5a and 5b depict a neutral face and a face portraying expression of happiness. Figure 5c depicts the optical flow field representing the motion pattern that arises during the portrayal of happiness. With horizontal widening of the mouth and cheeks, the motion pattern can be described by similar values of divergence of the flow field on either side in the lower half of the face. Also, the two eyes slightly widen symmetrically giving identical divergence values for the spatial regions corresponding to the two eyes. Thus the high/low values of divergence create a distinct pattern that can adequately characterize the facial motion.

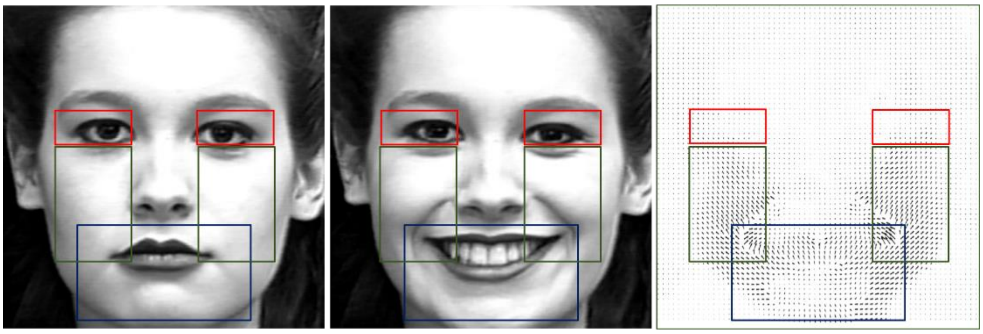


Figure 5: (a) Neutral face (b) Face conveying *happiness* (c) Corresponding flow field

For the three templates representing the motion field, the divergence value was determined at each spatial location giving three corresponding divergence-based templates used as descriptors for training. The divergence value at each spatial location in the image domain is computed as the sum of partial derivatives of horizontal flow along x axis and vertical flow along y axis at that point, i.e.:

$$\mathrm{divergence}\,\mathbf{w} = \frac{\partial \mathrm{u}}{\partial \mathrm{x}} + \frac{\partial \mathrm{v}}{\partial \mathrm{y}} \tag{7}$$

4.4.2 Curl-based descriptor

The second descriptor *FlowC* (size: 256 × 256) is derived from the component of curl of motion field perpendicular to itself. It gauges the circular motion of facial constituents. Figure 6 depicts the transformation of a neutral face to a face with the expression of anger and the associated flow field. Circular motion arises in the highlighted areas due to pursing of lips and with the nasal edges of the eyebrows drooping downward. Both the regions have a symmetric rotation of the two halves occurring about the center of that region but in opposite direction, giving curl values similar in magnitude but opposite in sign. Thus, the circular motion arising during portrayal of anger can be quantified in terms of its curl.

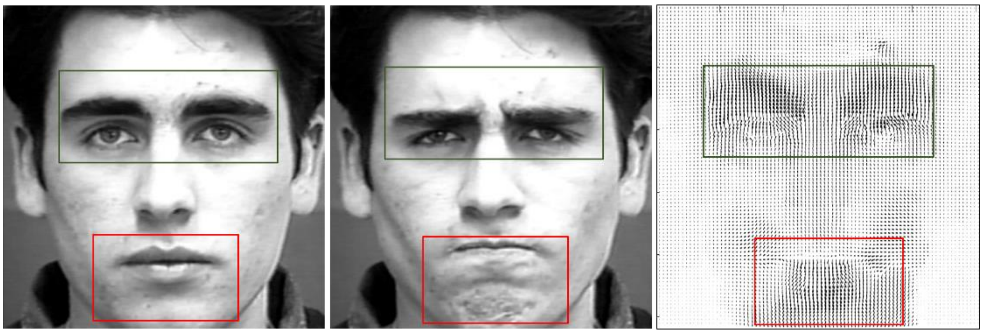


Figure 6: (a) Neutral face (b) Face conveying *anger* (c) Corresponding flow field

For the three motion templates, the curl value can be determined at each spatial location to obtain three corresponding curl-based templates used as descriptors for training. The curl value at each spatial location in the image domain can be derived by subtracting partial derivative of horizontal flow along y axis from partial derivative of vertical flow along x axis at that point, i.e.:

$$\mathrm{curl}\,\mathbf{w} = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \tag{8}$$

4.4.3 Divergence- and curl-based descriptor

The third form of descriptor *FlowDC* (size: 256 × 256) is derived simply and straightforwardly with the concatenation of the divergence and curl descriptors associated with each motion template. It furnishes information regarding both the motion of facial components in terms of expansion/contraction and spin at each spatial location in the image domain.

4.5 Classification techniques

In an FER scheme, after the computation of characteristic features for describing the FE, they are fed to a robust classifier for categorizing the FE. Two different types of classification schemes were used for recognition:

4.5.1 Multi-class SVM classification [47]

A linear SVM aims at determining a suitable hyperplane or a decision boundary that divides the given data into two distinct classes. Distance between nearest data point from a particular class and the hyperplane is known as margin. There are multiple hyperplanes to categorize the data, but the one that has the largest margin between the two classes is the most appropriate as its leads to higher probability of classifying test data with precision. Error correcting output codes (ECOC) method segregates multi-class classification problem into multiple binary classification problems. Two different ECOC coding schemes namely "one-versus-one" and "one-versus-all" were adopted for different experiments in this work for a comprehensive evaluation with variation in parameters. Under the first setting, to train a model for p different labels, ECOC uses m(m-1)/2 linear SVM models wherein for every binary learner first class is considered positive and the other negative, ignoring the rest. The second coding scheme one-versus-all entails for each binary learner first class to be positive and all others negative. For both the schemes, ultimately, the class with maximum positive votes is assigned to the test data. SVMs are used in this work due to their capability in efficient training even with a small set of samples.

4.5.2 k-NN classification [48]

It is a non-parametric classification scheme based on lazy, i.e., instance-based learning. The input data is assigned the class which is most common amongst its k nearest neighbors. For k = 1, the class assigned is that of the single nearest neighbor. The scheme is chosen to ease in its implementation and robustness to linearly inseparable data. The two parameters required for tuning the classifier are distance metric and the value of k. In this work, to measure the distance between different data points, the Euclidean distance metric is utilized, and the number of nearest neighbors is 1.

5 Experiments and results

This section outlines the details of the evaluation scheme and the results obtained on the databases employed for training and testing.

5.1 Evaluation on CK+ and JAFFE datasets

The CK+ dataset consists of 327 image sequences labeled with one of the seven discrete expressions. The sequences start from a face with a neutral state and progress toward a face with peak expression. The JAFFE dataset comprises grayscale images with Japanese female subjects portraying fundamental emotions. The total number of sequences for each emotion in CK+ and JAFFE are encapsulated in Table 1.

	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
CK+	45	18	59	25	69	29	82
JAFFE	30	-	30	30	30	30	30

Table 1: Sequences for each emotion in CK+ and JAFFE

In this work, a subject-independent protocol was deployed to validate the generalization capability of the proposed features. For experimental evaluation, the sequences were divided into two equal sets of training and test samples. The training set was further divided into three equal subsets each having sequences associated with the fundamental FEs. All the sequences in each subset corresponding to the same emotion formed a group. The motion template corresponding to each FE was derived by computing mean of the flow fields associated with emotional and corresponding neutral face image pairs in that group. Thus, three motion templates associated with each expression were derived from those sequences as discussed previously. Using each template, corresponding divergence and curl templates were derived that represented the global distribution of their values across the facial image domain. The computed divergence and curl templates that embody the facial motion pattern were fed to the classifiers as descriptive feature vectors individually (FlowD and FlowC) and in concatenation with each other (FlowDC) along with the associated expression label for training. In this manner, for each emotion category, there were three training templates each describing divergence, curl, and both.

From the testing sequences, first frame from each sequence was used as neutral frame and the last three peak expression frames were used as emotional frames, thereby giving three emotional-neutral faces image pairs from one sequence. The values of divergence and curl of the flow field associated with the emotional-neutral faces image pair were computed at each spatial location and fed to the classifiers, individually and in concatenation as the test data. Thus the classifiers were trained with divergence and curl templates whereas testing was done using divergence and curl of optical flow field corresponding to the neutral and emotional faces of test samples. The output of the classifier was the one of the expression labels that were used during training. The experiment was conducted twice by changing the training and test sequences and the average recognition accuracies in percentage are encapsulated in Table 2. The descriptor FlowDC that exploits information related to both circular and expansive/contractive motion was found to have the most superior performance.

	Multi-class SVM	k-NN
FlowD	85.45	87.15
FlowC	91.50	91.16
FlowDC	97.52	97.80

Table 2: Recognition rates (in percentage) for different classifiers and descriptors for CK+

Normalized confusion matrices for the two types of classifiers are shown in Tables 3–5.

	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**91.11**	0.00	0.00	0.00	0.00	8.89	0.00
Contempt	0.00	**100**	0.00	0.00	0.00	0.00	0.00
Disgust	0.00	0.00	**100**	0.00	0.00	0.00	0.00
Fear	0.00	0.00	0.00	**100**	0.00	0.00	0.00
Happiness	0.00	4.35	0.00	2.90	**92.75**	0.00	0.00
Sadness	0.00	0.00	0.00	0.00	0.00	**100**	0.00
Surprise	1.22	0.00	0.00	0.00	0.00	0.00	**98.78**

Table 3: Normalized confusion matrix for multi-class SVM classifier (one-versus-one) with FlowDC

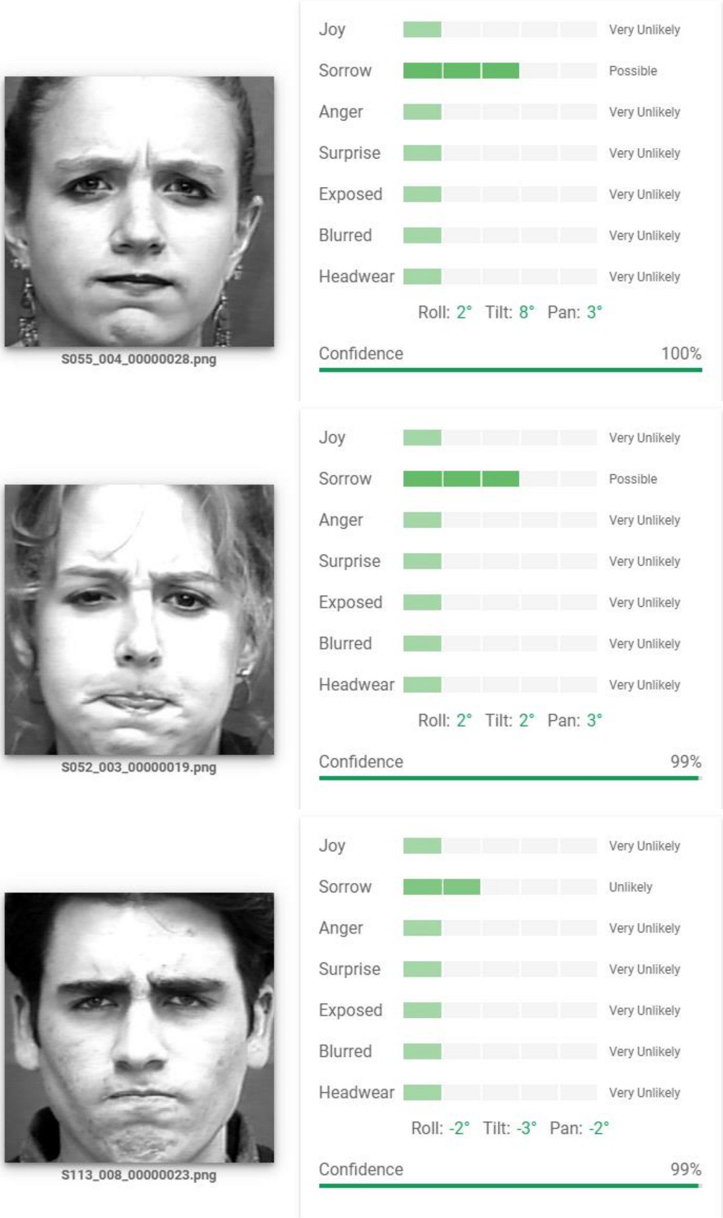
	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**84.44**	4.44	0.00	0.00	0.00	11.11	0.00
Contempt	0.00	**100**	0.00	0.00	0.00	0.00	0.00
Disgust	0.00	0.00	**100**	0.00	0.00	0.00	0.00
Fear	0.00	0.00	0.00	**88.00**	4.00	0.00	8.00
Happiness	0.00	7.25	0.00	2.90	**89.86**	0.00	0.00
Sadness	0.00	0.00	0.00	0.00	0.00	**100**	0.00
Surprise	1.22	0.00	0.00	0.00	0.00	0.00	**98.78**

Table 4: Normalized confusion matrix for multi-class SVM classifier (one-versus-all) with FlowDC

	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**93.33**	0.00	0.00	0.00	0.00	6.67	0.00
Contempt	0.00	**100**	0.00	0.00	0.00	0.00	0.00
Disgust	1.69	0.00	**98.31**	0.00	0.00	0.00	0.00
Fear	0.00	0.00	0.00	**100**	0.00	0.00	0.00
Happiness	0.00	2.90	0.00	2.90	**94.20**	0.00	0.00
Sadness	0.00	0.00	0.00	0.00	0.00	**100**	0.00
Surprise	1.22	0.00	0.00	0.00	0.00	0.00	**98.78**

Table 5: Normalized confusion matrix for k-NN classifier with proposed FlowDC

The overall accuracy attained with k-NN classifier is slightly higher than multi-class SVM. The expressions of *contempt*, *fear,* and *sadness* were identified with full accuracy by both the classifiers. There were a few cases of misclassification of *happiness* as *contempt* and *fear*. This could be due to the slight structural similarity that *happiness* bears with the aforementioned expressions involving horizontal widening of the lower half of the face. For both the classifiers the recognition rate was the lowest for anger. In all the cases of misclassification, anger was inaccurately predicted as sadness. The emotions anger and sadness are negative emotions associated with displeasure of some sort. The confusion of anger as sadness may be attributed due to similarity in the expression portrayal as for both, the most noticeable changes occur around mouth and ocular regions and slight changes occur around the nose. An emotional expression that the humans may perceive unerringly can confuse a machine due to such similarities. Figure 7 further illustrates this point where a few images from CK+ and KDEF datasets labeled as “anger” were erroneously classified as “sorrow” by the powerful pre-trained Google cloud vision API [49]. When an image is fed to Google cloud vision API for face analysis, it predicts the likelihood for the subject in the image to be conveying “anger,” “joy,” “sorrow,” and “surprise.” For the input images displaying *anger* the highest possibility or likeliness was given to *sorrow.*



Joy	Very Unlikely
Sorrow	Possible
Anger	Very Unlikely
Surprise	Very Unlikely
Exposed	Very Unlikely
Blurred	Very Unlikely
Headwear	Very Unlikely

Roll: 2° Tilt: 8° Pan: 3°

Confidence 100%

Joy	Very Unlikely
Sorrow	Possible
Anger	Very Unlikely
Surprise	Very Unlikely
Exposed	Very Unlikely
Blurred	Very Unlikely
Headwear	Very Unlikely

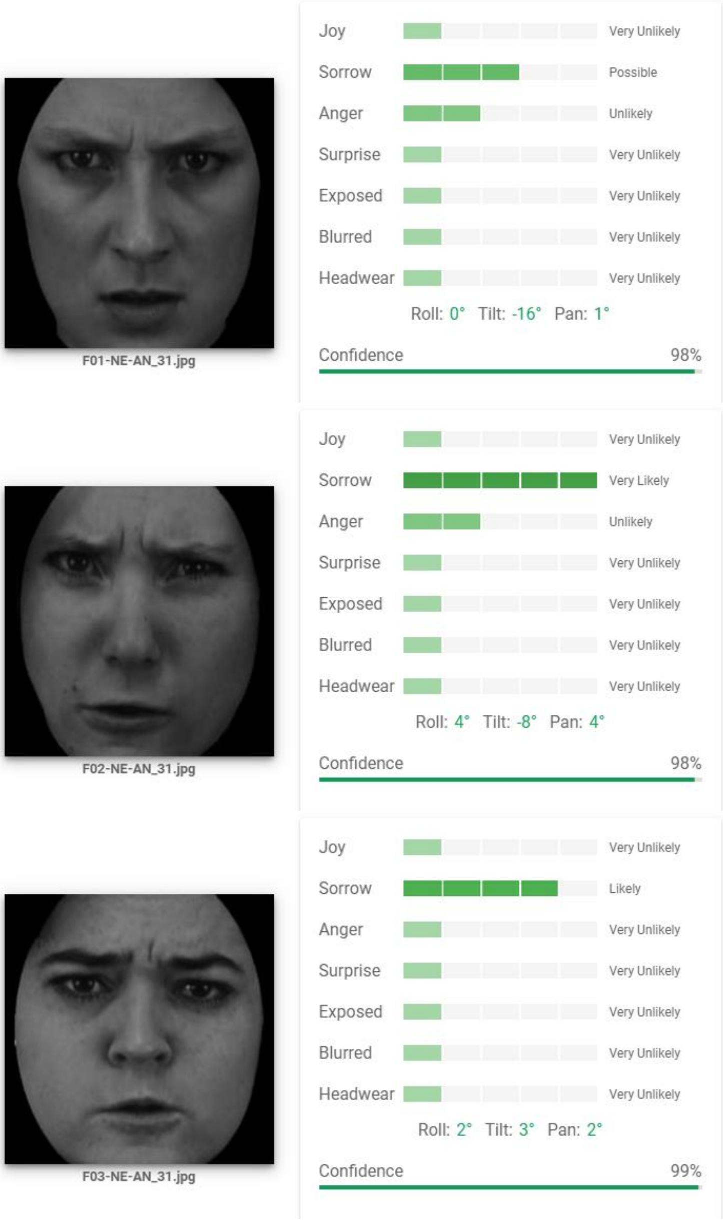
Roll: 2° Tilt: 2° Pan: 3°

Confidence 99%

Joy	Very Unlikely
Sorrow	Unlikely
Anger	Very Unlikely
Surprise	Very Unlikely
Exposed	Very Unlikely
Blurred	Very Unlikely
Headwear	Very Unlikely

Roll: -2° Tilt: -3° Pan: -2°

Confidence 99%



Joy	Very Unlikely
Sorrow	Possible
Anger	Unlikely
Surprise	Very Unlikely
Exposed	Very Unlikely
Blurred	Very Unlikely
Headwear	Very Unlikely

Roll: 0° Tilt: -16° Pan: 1°

Confidence 98%

Joy	Very Unlikely
Sorrow	Very Likely
Anger	Unlikely
Surprise	Very Unlikely
Exposed	Very Unlikely
Blurred	Very Unlikely
Headwear	Very Unlikely

Roll: 4° Tilt: -8° Pan: 4°

Confidence 98%

Joy	Very Unlikely
Sorrow	Likely
Anger	Very Unlikely
Surprise	Very Unlikely
Exposed	Very Unlikely
Blurred	Very Unlikely
Headwear	Very Unlikely

Roll: 2° Tilt: 3° Pan: 2°

Confidence 99%

Figure 7: Erroneous classification of ***anger*** as ***sorrow*** for different cases by Google cloud vision API

The recognition accuracies attained on CK+ were compared with those obtained by the state-of-the-art methods and were observed to be on par with them under both six-emotion class (contempt excluded) and seven-emotion class (contempt included) categorization paradigms adopted by different models (Table 6).

No. of classes	Research work	Accuracy (%)
Seven	Fan and Tjahjadi (2019) [50]	92.50
Seven	Maheswari et al. (2020) [51]	93.89
Seven	Hu et al. (2019) [52]	94.00
Seven	Wei et al. (2020) [53]	94.41
Seven	Makhmujaedaev et al. (2019) [54]	94.50
Seven	Bin Iqbal et al. (2020) [55]	95.13

Seven	Cheng and Zhou (2020) [56]	96.00
Seven	Gan et al. (2020) [57]	96.28
Seven	Qin et al. (2020) [58]	96.81
Seven	Salmam et al. (2019) [25]	96.92
Seven	Allaert et al. (2020) [59]	97.25
Seven	**FlowDC (multi-class SVM, one-versus-one)**	**97.52**
Seven	**FlowDC (k-NN)**	**97.80**
Seven	Zhu et al. (2021) [60]	98.46
Six	Meena et al. (2020) [61]	92.85
Six	Maheswari et al. (2020) [51]	94.85
Six	Khan et al. (2019) [62]	94.90
Six	Xie et al. (2019) [63]	95.88
Six	Kim et al. (2019) [64]	96.46
Six	Bin Iqbal et al. (2020) [55]	96.77
Six	Salmam et al. (2019) [25]	96.83
Six	Pan et al. (2020) [65]	97.01
Six	**FlowDC (multi-class SVM, one-versus-one)**	**97.46**
Six	Meena et al. (2019) [66]	97.61
Six	**FlowDC (k-NN)**	**97.92**

Table 6: Comparison between accuracies of FER approaches

As evident from the tabulated comparison, recognition accuracy of proposed descriptor FlowDC is on par with the state-of-the-art FER techniques. It is to be noted that the competence with state-of-the-art techniques is validated in terms of recognition accuracy. However, the computational time is not considered as a parameter in this work and will be considered as a direction to explore in future.

For JAFFE dataset, images for the FE corresponding to *contempt* were not available, thus the other six emotions were studied, and the resultant confusion matrices are presented in Tables 7–9. The overall highest accuracy was attained with multi-class SVM under one-versus-all coding and the lowest with k-NN classifier. The confusions/misclassifications of *sadness* as *fear* and *disgust* as *sadness* or *anger* were the most prevalent in the experiments with JAFFE.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**100**	0.00	0.00	0.00	0.00	0.00
Disgust	10.00	**73.33**	0.00	0.00	16.67	0.00
Fear	0.00	0.00	**100**	0.00	0.00	0.00
Happiness	0.00	0.00	0.00	**100**	0.00	0.00
Sadness	0.00	0.00	3.33	0.00	**96.67**	0.00
Surprise	0.00	0.00	0.00	0.00	0.00	**100**

Table 7: Normalized confusion matrix for multi-class SVM classifier (one-versus-one) with FlowDC for JAFFE

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**100**	0.00	0.00	0.00	0.00	0.00
Disgust	23.33	**76.67**	0.00	0.00	0.00	0.00
Fear	0.00	0.00	**100**	0.00	0.00	0.00
Happiness	0.00	0.00	0.00	**100**	0.00	0.00
Sadness	0.00	0.00	3.33	0.00	**96.67**	0.00
Surprise	0.00	0.00	0.00	0.00	0.00	**100**

Table 8: Normalized confusion matrix for multi-class SVM classifier (one-versus-all) with FlowDC for JAFFE

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**90.00**	0.00	0.00	3.33	6.67	0.00
Disgust	6.67	**66.67**	6.67	6.67	13.33	0.00
Fear	0.00	0.00	**100**	0.00	0.00	0.00
Happiness	0.00	0.00	0.00	**100**	0.00	0.00
Sadness	0.00	0.00	10.00	0.00	**90.00**	0.00
Surprise	0.00	0.00	10.00	0.00	0.00	**90.00**

Table 9: Normalized confusion matrix for k-NN classifier with FlowDC for JAFFE

To validate the descriptor further, a cross-database analysis was performed on the KDEF dataset. Classifiers trained with FlowDC derived from CK+ facial images were utilized for generating output expression labels corresponding to input facial images from KDEF dataset.

5.2 Evaluation on KDEF dataset

Forty front facing images for each expression and a neutral face corresponding to that subject were taken from KDEF set A for evaluation. Figure 1 top row shows a KDEF subject displaying six basic emotions. Flow was estimated between the image depicting an expression and corresponding neutral face. To interpret the results of evaluation on KDEF, a comparative analysis (Figure 8) was performed with the following:

Human judgment: The results were derived from a perceptual study on KDEF conducted by Calvo and Lundqvist [67]. The study of human judgment on this dataset used 40 different facial images corresponding to each emotion. The participants of the study were asked to identify the emotion portrayed by the images displayed to them for different durations. For the purpose of comparison in this work the two cases considered are: (a) when the participants were given 25 msec to judge expression portrayed in each image and (b) When the participants were given free-viewing time i.e. no time bound was there to judge the expression.

Microsoft face API [68]: When an input image is fed to pre-trained Microsoft face API, it analyzes information related to facial features, gender, age and attempts to detect and classify emotions such as *anger, contempt, disgust, fear, happiness, sadness, and surprise* portrayed by the subjects in the images. For evaluation in this study, test images were fed to the Microsoft face API. For each class, the accuracy was determined by dividing correctly identified expression with the total number of observations for that class.

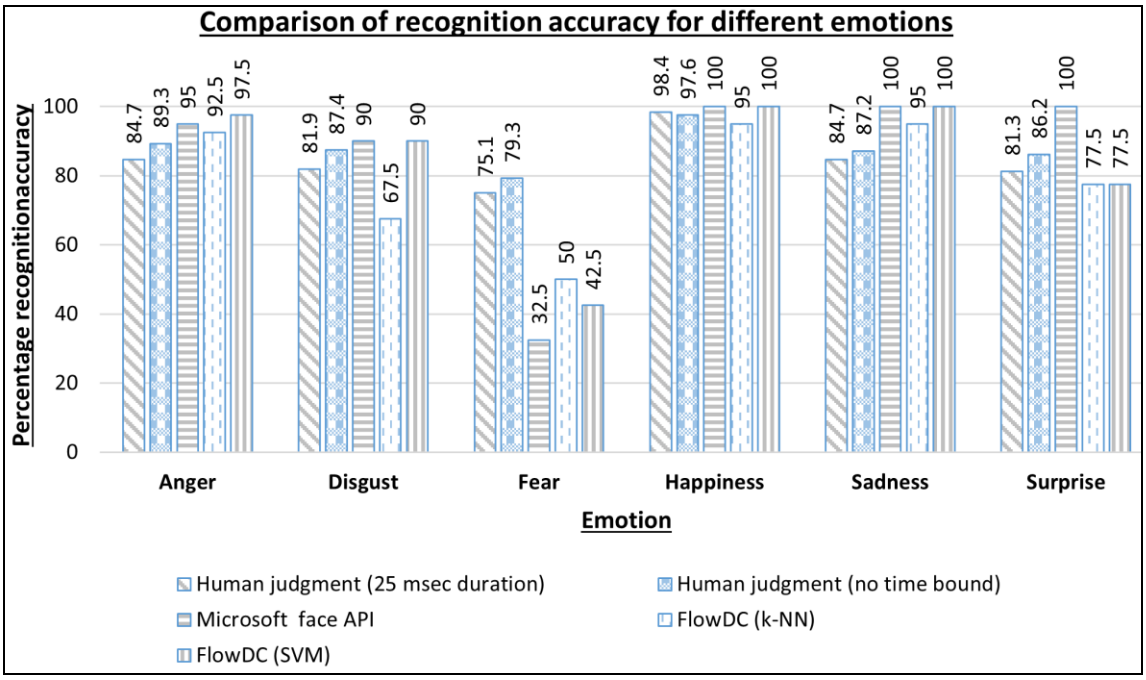


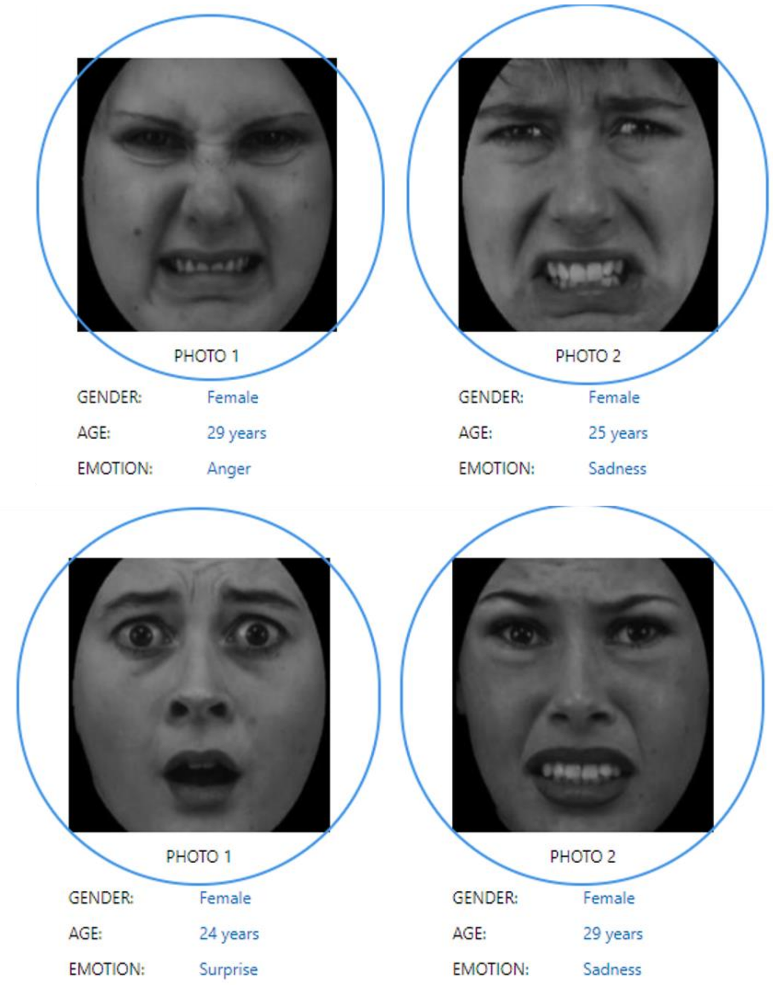
Figure 8: Illustrative comparison of different techniques on KDEF dataset

The expression identified with the least accuracy for all the cases was “fear.” The best identified expression was happiness, for all the methods. The multi-class SVM and Microsoft face API attained 100% accuracy and surpassed human judgment accuracy for happiness. Both anger and sadness were better recognized by the proposed descriptor FlowDC than human participants in the study by Calvo and Lundqvist [67]. Inevitably, human judgment with no time bound yielded the highest overall recognition accuracy of 87.8%. followed by Microsoft face API of 86.3%. For such a challenging dataset. even human observers made lapse of judgment in many cases. The recognition rate of proposed descriptor FlowDC with multi-class SVM classifier of 84.6% surpassed human judgment for 25 msec duration with accuracy 84.4%. Some cases of misclassification by Microsoft API are illustrated in Figure 9.

To further assess the results, the Pearson correlation coefficient was computed for recognition rates corresponding to different FEs attained by human judgment, Microsoft face API and the presented descriptors with the two classifiers. The correlation values are depicted in Table 10. It can be seen that results obtained by the proposed scheme are more in agreement with the human perception results Calvo and Lundqvist [67] presented. All the values are positive, implying a positive correlation. The highest correlations are of no time bound human judgment with FlowDC with multi-class SVM and k-NN, respectively. An agreement of the results obtained with human perception of FEs validates the usefulness of the divergence and curl-based feature descriptors for modeling FEs.

	Microsoft face API	FlowDC (k-NN)	FlowDC (SVM)
Human judgment (25msec)	0.63	0.76	0.72
Human judgment (no time bound)	0.73	0.78	0.80

Table 10. Correlation between human perception and different approaches



GENDER: Female
AGE: 29 years
EMOTION: Anger

GENDER: Female
AGE: 25 years
EMOTION: Sadness

GENDER: Female
AGE: 24 years
EMOTION: Surprise

GENDER: Female
AGE: 29 years
EMOTION: Sadness

Figure 9: Instances of misclassification by Microsoft face API (Top) ***disgust*** misclassified as a***nger*** and ***sadness***; (Bottom) ***fear*** misclassified as ***surprise*** and ***sadness*** by Microsoft face API

6 Conclusion

In this work an approach to determine FEs associated with discrete emotions has been presented. To attain successful classification, divergence and curl templates derived from motion templates and flow fields corresponding to each fundamental FE were used for the training and testing, respectively. High recognition rates were achieved on CK+, JAFFE, and KDEF datasets. Moreover the recognition rates on KDEF were found to be in agreement with human perception.

The futuristic vision-based architectures in smart devices and environments are primarily driven by the objective of emulating human vision and cognition for comprehending their surroundings. Thereby, an agreement of the results obtained with human perception of FEs substantiates the usefulness of the divergence- and curl-based features in vision-based affect analysis systems.

It is to be noted that the model is considered competent with state-of-the-art techniques in terms of recognition accuracy. However, the computational time is not considered as a parameter in this work. The future work will focus on comparison with other techniques based on time and other factors. Also, there will be focus on developing the model to be robust to unconstrained environments such as with high intensity fluctuations in image pairs, partial or full occlusion, and testing its applicability and performance in real-time. In addition, it will be attempted to use average face instead of neutral face for situations where the neutral face is unavailable for the subjects in the input images.

ACKNOWLEDGMENT

The authors wish to express their deepest gratitude to the editor and reviewers whose insightful remarks and suggestions helped in enhancing the quality of the originally submitted manuscript.

References

1. A. Mehrabian, *Silent Messages*, First Edition, Wadsworth publishing Belmont, California, p. 44, 1971.
2. Z. Deng et al. “Factorized variational autoencoders for modeling audience reactions to movies,” International Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA., 2017. 10.1109/CVPR.2017.637
3. P. Ekman, W.V. Friesen, “Constants across cultures in the face and emotion,” *Journal of Personality and Social Psychology*, 17(2), 124-129, 1971. https://doi.org/10.1037/h0030377
4. D. Matsumoto, “More evidence for the universality of a contempt expression,” *Motivation and Emotion*, 16, 363-368, 1992. https://doi.org/10.1007/BF00992972
5. A. Singhal et al., “Summarization of videos by analyzing affective state of the user through crowdsource,” Cognitive Systems Research, 52, 917-930, 2018. doi.org/10.1016/j.cogsys.2018.09.019.
6. S. Mo, J. Niu, Y. Su, S.K. Das, “A novel feature set for video emotion recognition,” Neurocomputing 291, 11-20, 2018. doi.org/10.1016/j.neucom.2018.02.052
7. A. Fernandez Caballero et al., Smart environment architecture for emotion detection and regulation. Journal of Biomedical Informatics, 64, 55-73, 2016. doi.org/10.1016/j.jbi.2016.09.015
8. N. Harrold, C.T. Tan, D. Rosser, T.W. Leong, “CopyMe: an emotional development game for children,” Proc. Conference on Human Factors in Computing Systems, pp. 503-506, 2014 doi.org/10.1145/2559206.2574785
9. D. Huang, F. De la Torre F, “Bilinear kernel reduced rank regression for facial expression synthesis,” In: Daniilidis K., Maragos P., Paragios N. (eds) Computer Vision –Lecture Notes in Computer Science, vol 6312. Springer, Berlin, Heidelberg, 2010. doi.org/10.1007/978-3-642-15552-9_27
10. P. Lucey et al, “The Extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression,” Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Computer Society pp. 94–101, 2010. http://doi.org/10.1109/CVPRW.2010.5543262
11. D. Lundqvist, A. Flykt, A. Öhman, “The Karolinska directed emotional faces,” - KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9, 1998.
12. M.J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, “Coding facial expressions with Gabor wavelets,” Proc. 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200-205, 1998. http://doi.org/10.1109/AFGR.1998.670949
13. L. Yin et al., “A 3d facial expression database for facial behaviour research,” Proc. IEEE Automatic face and gesture recognition, FGR, pp. 211–216, 2006. http://doi.org/10.1109/FGR.2006.6
14. G. Zhao et al., “Facial expression recognition from near-infrared videos,” Image and Vision Computing, **29**(9) 607–619, 2011. https://doi.org/10.1016/j.imavis.2011.07.002
15. P. Ekman, W.V. Friesen, “Facial Action Coding System: A technique for the measurement of Facial Movement,” Consulting Psychologists Press, Palo Alto, 1978
16. Y. Tong, W. Liao, Q. Ji, “Facial action unit recognition by exploiting their dynamic and semantic relationships,” IEEE Transactions on Pattern Analysis and Machine Intelligence, **29**, 1683-1699, 2007. https://doi.org/10.1109/TPAMI.2007.1094

17. da Silva F.A.M., H. Pedrini H, "geometrical features and active appearance model applied to facial expression recognition," International Journal of Image and Graphics **16**(4) 1650019, 2016. https://doi.org/10.1142/S0219467816500194
18. A.T. Lopes, E. de Aguiar, A.F. De Souza, T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," Pattern Recognition. **61**, 610-628, 2017. https://doi.org/10.1016/j.patcog.2016.07.026
19. A.M. Ashir, A. Eleyan, (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
20. Y. Ding, Q. Zhao, X. Yuan, "Facial Expression recognition from image sequence based on LBP and Taylor expansion," IEEE Access, **5**, 19409–19419, 2017. https://doi.org/10.1109/ACCESS.2017.2737821
21. F. Bougourzi, F. Dornaika, K. Mokrani, A. Taleb-Ahmed, Y. Ruichek, "Fusion transformed deep and shallow features (FTDS) for image-based facial expression recognition," Expert Systems with Applications, 156, 113459, 2020.
22. K. Zhang, Y. Huang, Y. Du, L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," IEEE Transactions on Image Processing 26(9) 4193-4203, 2017. https://doi.org 10.1109/TIP.2017.2689999
23. S. Agarwal, B. Santra, D.P. Mukherjee, "Anubhav: recognizing emotions through facial expression." The Visual Computer, 34(2), 177-191, 2018. https://doi.org/10.1007/s00371-016-1323-z
24. M.H. Siddiqi et al., "Human Facial Expression Recognition using stepwise linear discriminant analysis and hidden conditional random fields," IEEE Transactions on Image Processing, 24(4), 1386-1398, 2015. https://doi.org /10.1109/TIP.2015.2405346
25. F.Z. Salmam, A. Madani, and M. Kissi, "Fusing multi-stream deep neural networks for facial expression recognition," Signal Image and Video Process., vol. 13 no. 3, pp. 609–616, 2019.
26. A. Danelakis, T. Theoharis, I. Pratikakis, "A spatio-temporal wavelet-based descriptor for dynamic 3d facial expression retrieval and recognition," The Visual Computer 32(6-8) 1001–1011, 2016. https://doi.org/10.1007/s00371-016-1243-y
27. S. Bursic, G. Boccignone, A. Ferrara, A. D'Amelio, R. Lanzarotti, "Improving the accuracy of automatic facial expression recognition in speaking subjects with deep learning," Appl. Sci., 10, 4002, 2020.
28. T. Wehrle et al., "Studying the dynamics of emotional expression using synthesized facial muscle movements," Journal of Personality and Social Psychology, 78(1), 105-119, 2000. https://doi.org/10.1037/0022-3514.78.1.105
29. N Sun et al., "Deep spatial-temporal feature fusion for facial expression recognition in static images," Pattern Recognition Letters, 119, 49-61, 2019. https://doi.org/10.1016/j.patrec.2017.10.022
30. J. Chi, C. Tu, C. Zhang, "Dynamic 3D facial expression modeling using Laplacian smooth and multi-scale mesh matching," The Visual Computer 30(6-8), 649-659, 2014. https://doi.org/10.1007/s00371-014-0960-3
31. Kim, I-H, Chen, Y-C.M, Spector, D.L., Eils, R., Rohr, K. (2011): Nonrigid registration of 2D and 3D dynamic cell nuclei images for improved classification of subcellular particle motion. IEEE Transactions on Image Processing, 20(4), 1011–1022. https://doi.org/10.1109/TIP.2010.2076377
32. D. Rueckert et al., "Nonrigid registration using free-form deformations: application to breast MR images," IEEE Transactions on Medical Imaging, 18(8), 712–721, 1999. https://doi.org/10.1109/42.796284
33. T.-C. Huang et al., "Quantification of blood flow in internal cerebral artery by optical flow method on digital subtraction angiography in comparison with time-of-flight magnetic resonance angiography," PloS ONE, 8(1), e54678, 2013 https://doi.org/10.1371/journal.pone.0054678

34. G. Piriou et al. “Recognition of dynamic video contents with global probabilistic models of visual motion,” IEEE Transactions on Image Processing, 15(11), 3418–3431, 2006 https://doi.org/10.1109/TIP.2006.881963
35. C.-W. et al., “Motion flow-based video retrieval,” IEEE Transactions on Multimedia. 9 (6), 1193–1201, 2007. 10.1109/TMM.2007.902875.
36. H. Chao, Y. Gu, M. Napolitano, “A survey of optical flow techniques for robotics navigation applications,” Journal of Intelligent & Robotic Systems, 73(1-4), 361–372, 2014. https://doi.org/10.1007/s10846-013-9923-6
37. K. Mase, “Recognition of facial expression from optical flow,” IEICE Transactions on Information and Systems, E74, 3474-3483, 1991.
38. I.A. Essa, A. Pentland, “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, 1994. https://doi.org/10.1109/CVPR.1994.323813
39. X. Pu et al., “Facial expression recognition from image sequences using twofold random forest classifier,” Neurocomputing, 168, 1173-1180, 2015. https://doi.org/10.1016/j.neucom.2015.05.005
40. J.Zhao, X. Mao, J. Zhang, “Learning deep facial expression features from image and optical flow sequences using 3D CNN,” The Visual Computer, 34 (10),1461–1475, 2018. https://doi.org/10.1007/s00371-018-1477-y
41. X. Pan et al., “Video-based facial expression recognition using deep temporal–spatial networks,” IETE Technical Review. 2020. https://doi.org/10.1080/02564602.2019.1645620
42. P. Viola, M. Jones, “Rapid object detection using a boosted cascade of simple features.” In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001, vol. 1, pp. 1–511. https://doi.org/10.1109/CVPR.2001.990517
43. T. Brox, A. Bruhn, N. Papenberg, J. Weickert, “High accuracy optical flow estimation based on a theory for warping,” Proceedings of 8th European Conference on Computer Vision; Prague, Czech Republic Springer Lecture Notes in Computer Science 3024, T. Pajdla and J. Matas (Eds.), Vol. 4, pp. 25-36, 2004. https://doi.org/10.1007/978-3-540-24673-2_3
44. B.K.P. Horn, B.G. Schunck, “Determining optical flow. Artificial Intelligence,” 17, 185-203. 1981. https://doi.org/10.1016/0004-3702(81)90024-2
45. M.J. Black, Y. Yacoob, “Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion,” In Proceedings of IEEE International Conference on Computer Visio, 1995. https://doi.org/ 10.1109/ICCV.1995.466915
46. S. Shojaeilangari et al., “Robust representation and recognition of facial emotions using extreme sparse learning,” IEEE Transactions on Image Processing, 24(7), 2140-2152, 2015. https://doi.org /10.1109/TIP.2015.2416634
47. E.L. Allwein, R.E. Schapire, Y. Singer, “reducing multiclass to binary: a unifying approach for margin classifiers,” Journal of Machine Learning Research, 1, 113-141, 2000.
48. N.S. Altman, An introduction to kernel and nearest-neighbour non-parametric regression. The American Statistician. 46 (3): 175–185, 1992.
49. https://cloud.google.com/vision/ [Google vision API]
50. X. Fan and T. Tjahjadi, “Fusing dynamic deep learned features and handcrafted features for facial expression recognition,” J. Vis. Comm. and Image Representation, vol. 65, 102659, 2019.
51. V.U. Maheswari, G. Varaprasad, and S.V. Raju, “Local directional maximum edge patterns for facial expression recognition,” J. Ambient Intell. Human Comput., 2020. https://doi.org/10.1007/s12652-020-01886-3
52. K. Hu, G. Huang, Y. Yang, C.-M. Pun, W.-K. Ling, and L. Cheng, “Rapid facial expression recognition under part occlusion based on symmetric SURF and heterogeneous soft partition network,” Multimed Tools Appl, vol. 79, pp. 30861–30881, 2020.
53. W. Wei, Q. Jia, Y. Feng, G. Chen., and M. Chu, “Multi-modal facial expression feature based on deep-neural networks,” J. Multimodal User Interfaces, vol. 14, pp. 17–23, 2020.

54. F. Makhmudkjujaev, M. Abdullah-Al-Wadud, M.T.B. Iqbal, B. Ryu, and O. Chae, “Facial expression recognition with local prominent directional pattern,” Signal Process.: Image Commun., vol. 74, pp. 1–12, 2019.
55. M.T.B. Iqbal, B. Ryu, A.R. Rivera, F. Makhmudkhujaev, O. Chae, and S.-H. Bae, “Facial expression recognition with active local shape pattern and learned-size block representations,” IEEE Trans. Affect. Comput., 2020. doi.org/10.1109/TAFFC.2020.2995432
56. S. Cheng and G. Zhou, “Facial expression recognition method based on improved VGG convolutional neural network,” Int. J. Pattern Recognit. and Artific. Intell., vol. 34 no. 7, pp. 2056003, 2020.
57. Y. Gan, J. Chen, Z. Yang, and L. Xu, “Multiple attention network for facial expression recognition,” IEEE Access, vol. 8 pp. 7383–7393, 2020.
58. S. Qin, Z. Zhu, Y. Zou, and X. Wang, “Facial expression recognition based on Gabor wavelet transform and 2-channel CNN,” Int. J. Wavelets, Multiresolution and Inf. Process., vol. 18, no. 2, 2020.
59. B. Allaert, I.M. Bilasco, and C. Djeraba, “Micro and macro facial expression recognition using advanced local motion patterns,” IEEE Trans. Affect. Comput., 2020. https://doi.org/10.1109/TAFFC.2019.2949559.
60. X. Zhu, S. Ye, L. Zhao, Z. Dai, “Hybrid attention cascade network for facial expression recognition,” Sensors, 21(6), 2003, 2021.
61. H.K. Meena, K.K. Sharma, and S.D. Joshi, “Effective curvelet-based facial expression recognition using graph signal processing,” Signal, Image and Video Process., vol. 14, pp. 241–247, 2020.
62. R.A. Khan, A. Meyer, H. Konik, and S. Bouakaz, “Saliency-based framework for facial expression recognition,” Frontiers in Comput. Sci., vol. 13, pp. 183–198, 2019.
63. S. Xie, H. Hu, and Y. Wu, “Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition,” Pattern Recognit., vol. 92, pp. 177–191, 2019.
64. J.-H. Kim, B.-G. Kim, P.P. Roy, and D.M. Jeong, “Efficient facial expression recognition algorithm based on hierarchical deep neural network structure,” IEEE Access, vol. 7 pp. 41273–41285, 2019.
65. X. Pan, “Fusing HOG and convolutional neural network spatial–temporal features for video-based facial expression recognition,” IET Image Process., vol. 14, no. 1, pp. 176–182, 2020.
66. H.K. Meena, S.D. Joshi, and K.K. Sharma, “Facial expression recognition using graph signal processing on HOG,” IETE J. Res., 2019. https://doi.org/10.1080/03772063.2019.1565952
67. M.G. Calvo, D. Lundqvist, “Facial expressions of emotion (KDEF): Identification under different display-duration conditions.” Behavior Research Methods, 40(1) 109-115, 2008. doi.org/10.3758/BRM.40.1.109
68. https://aidemos.microsoft.com/face-recognition [Microsocft face API]