Computers, Materials & Continua
DOI:10.32604/cmc.2021.014782
Article

Tech Science Press

A Machine Learning Approach for Expression Detection in Healthcare Monitoring Systems

Muhammad Kashif[1], Ayyaz Hussain[2], Asim Munir[1], Abdul Basit Siddiqui[3], Aaqif Afzaal Abbasi[4], Muhammad Aakif[5], Arif Jamal Malik[4], Fayez Eid Alazemi[6] and Oh-Young Song[7,*]

[1]Department of Computer Science & Software Engineering, International Islamic University, Islamabad, 44000, Pakistan
[2]Department of Computer Science, Quaid-i-Azam University, Islamabad, 44000, Pakistan
[3]Department of Computer Science, Capital University of Science & Technology, Islamabad, 44000, Pakistan
[4]Department of Software Engineering, Foundation University, Islamabad, 44000, Pakistan
[5]Department of Computer Science, Abdul Wali Khan University, Mardan, 23200, Pakistan
[6]Department of Computer Science and Information Systems, College of Business Studies, PAAET, 12062, Kuwait
[7]Department of Software, Sejong University, Seoul, 05006, Korea
[*]Corresponding Author: Oh-Young Song. Email: oysong@sejong.edu

Received: 16 October 2020; Accepted: 13 December 2020

Abstract: Expression detection plays a vital role to determine the patient's condition in healthcare systems. It helps the monitoring teams to respond swiftly in case of emergency. Due to the lack of suitable methods, results are often compromised in an unconstrained environment because of pose, scale, occlusion and illumination variations in the image of the face of the patient. A novel patch-based multiple local binary patterns (LBP) feature extraction technique is proposed for analyzing human behavior using facial expression recognition. It consists of three-patch [TPLBP] and four-patch LBPs [FPLBP] based feature engineering respectively. Image representation is encoded from local patch statistics using these descriptors. TPLBP and FPLBP capture information that is encoded to find likenesses between adjacent patches of pixels by using short bit strings contrary to pixel-based methods. Coded images are transformed into the frequency domain using a discrete cosine transform (DCT). Most discriminant features extracted from coded DCT images are combined to generate a feature vector. Support vector machine (SVM), k-nearest neighbor (KNN), and Naïve Bayes (NB) are used for the classification of facial expressions using selected features. Extensive experimentation is performed to analyze human behavior by considering standard extended Cohn Kanade (CK+) and Oulu–CASIA datasets. Results demonstrate that the proposed methodology outperforms the other techniques used for comparison.

Keywords: Detection; expressions; gestures; analytics; pain; patch-based local binary descriptor; discrete cosine transform; healthcare

1 Introduction

In the last two decades, excessive research has been done in the field of facial expression recognition (FER). Facial expression is the most expressive way of communication among humans and is generally categorized into seven basic expression types named anger, disgust, fear, happiness, sad, surprise, and neutral [1]. Human behavior detection through recognition of expressions plays a very important role in human-computer interaction and has attracted much attention in the areas of surveillance, healthcare, forensics, missing individual identification, crime investigation, interactive games, intelligent transportation and many other applications [2]. Facial expression recognition is a problem related to pattern recognition and computer vision, where a two-dimensional image of the face is acquired to extract the features for classification. Thorough research has been carried out in recent years and several techniques have been proposed to achieve better performance. These techniques usually produce good results in a constrained environment. Though the task to identify expressions from images captured in an unconstrained environment is still challenging due to the presence of variations in resolution and illumination in certain datasets. The proposed method contains the ability to deal with such images that resemble real-world images in terms of these factors to estimate the expression adequately.

In facial expression recognition, the main step is the feature extraction from still images or video frames to obtain the appearance-based and geometric-based variations that map to a target facial expression [3]. This paper investigates the use of multiple LBPs [TPLBP, FPLBP] in facial expression recognition. Coded images are obtained after the extraction of features using TPLBP and FPLBP. In the next step, these features are converted into the frequency domain using DCT. The most discriminative features are computed to analyze emotions from both forms of coded DCT images to obtain a fused feature vector. SVM, K-NN and NB classifiers are used to classify the emotions from publicly available databases namely, CK+ [4–10] and Oulu–CASIA [11–13]. The proposed technique gives better performance in terms of accuracy and robustness than the techniques presented in the literature survey.

This research paper is categorized into five sections. Section 2 briefly describes the relevant literature review. The proposed technique is presented in Section 3. Section 4 describes the experimental results and discussion. Finally, Section 5 presents the conclusion and future work.

2 Related Work

Texture-based and appearance-based descriptors are popular and these are extensively used for multi-scale facial expressions such as principal component analysis (PCA) to reduce dimensionality, linear discriminate analysis (LDA) for feature selection, and local binary patterns (LBP) [14,15]. LBP [16–19] is a successful feature extraction technique for emotion recognition and other image processing applications. LBP is calculated for each pixel in an image by involving the neighbors around that pixel. It provides binary numbers using 8 neighbors and a threshold is applied on these eight values corresponding to the central pixel. The binary values are used to generate histograms representing the appearance-based regions. Many LBP variations have been investigated in the previous related work to resolve the issues of illumination, multi-scale, and high dimension variations in FER. Automatic emotion recognition is investigated in [20,21] using the weber local descriptor (WLD) technique for frontal and spontaneous images and implemented in the e-healthcare environment. The WLD histogram features are computed using the Fisher Discriminate Ratio (FDR). These features are classified through a support vector machine (SVM). The WLD-based system gives better performance using the JAFFE and Cohn Kanade Databases. Guo et al. [22] proposed an enhanced deep learning hybrid CNN-BiLSTM (EJH-CNN-BiLSTM) algorithm to detect pain intensity using facial expression. The fine-tuned VGG-Face pre-trainer is used as a feature extraction tool by considering the balanced UNBC-McMaster Shoulder Pain Archive Database principle. Principal Component Analysis was applied to reduce dimensionality and enhance efficiency. The algorithm is used to estimate four various stages of pain. The results explored that the algorithm is the potential tool in medical diagnostics for automatic pain detection.

A novel algorithm called online sequential extreme learning machine and spherical clustering (OSELM-SC) is proposed by Muhammad et al. [23]. In this approach, different techniques are applied to original face images. This includes the Voila–Jones detector for face detection and cropping and histogram equalization for illumination variations. Features are extracted by applying the curvelet transform to every region of the face image. Then the statistical features are extracted through mean, standard deviation, and entropy. The features are classified through the proposed algorithm OSELM. The best performance is achieved in the case of frontal and spontaneous images. A new facial decomposition technique named IntraFace (IF) algorithm is presented that uses landmarks to compute regions of interest (ROI). Texture, shape-based LTP, HOG, LBP, and CLBP features are extracted and classified through SVM. The better performance based on the recognition rate is achieved with this technique than other decomposition-based techniques. The appearance-based and geometric-based features are extracted by computing local face regions and then combined [24]. The LBP and normalized central moments (NCM) are used as features. The proposed technique compares the local region-based and grid-based holistic representation. The local region-based method performs better than grid-based representation after applying the SVM classifier on the features. The histogram of oriented gradients (HOG) and the most discriminant discrete cosine transform (DCT) features are extracted in [25]. The proposed system is accurate and reliable in handling illumination variation and multi-scale (resolution variance) problems. The system achieved better performance in the case of MMI and CK+ datasets images feature classified with KNN, Sequential Minimal Optimization (SMO), and Random Forest (RF). Donia et al. [26] proposed a new framework DSAE , which is based on Deep Sparse Network that automatically recognizes the expressions. This technique is only feature-based in which the Active Appearance Model (AAM), Principle Component Analysis (PCA), and Histogram of Oriented Gradients (HOG) features are extracted. The output features are used as input to Deep Sparse Coding Network, Deep Sparse Auto Encoder (DSAE) which gives the better performance considering the accuracy. Weber's local binary image cosine transform (WLBI-CT) has been proposed in [27]. This technique investigates the multi-orientation and multi-scale face images. The frequency components of images are computed through local binary descriptors (LBP) and Weber local descriptors (WLD). The WLD and LBP generate face image features that are obtained through DCT with orientation and without orientation, respectively. Then an evaluation is done using these feature vectors with different classifiers including Naïve Bayes classifier (NB), Sequential Minimal Optimization (SMO), Multilayer Perceptron (MLP), K-nearest neighbors (KNN), and Classification Tree. The main aim of the research is to recognize the basic emotions [28]. Microsoft Kinect is used for 3D face modeling, which models the face using 121 specific points and arranges them based on face points. And plot it in coordinates by the Kinect Device. The six Action Units are used to describe the emotions presented by the FACS System and are used as features that are classified through KNN and MLP. Min Guo et al. proposed an algorithm K-ELBP by integrating Extended Binary Local Pattern (ELBP) and Karhunen–Loeve Transform (KLT) [29]. The ELBP is used for uniform patterns and removed the others. KLT is used to reduce the dimensionality. The K-ELBP histograms are obtained from the segmented blocks. The multi SVM classifier is applied to a combined histogram to find accurate expressions. A salient geometric feature-based framework is presented by Ekman et al. [29] for the automatic FER system. The elastic bunch graph matching (EBGM) algorithm and Kanade Lucas Tomaci (KLT) tracker are used to create and track facial points and feature points initialization, respectively. Three different geometric-based points, line, and triangle features are extracted from generated tracked facial point results. The line and triangle discriminant features are extracted through the Extreme Learning Machine (ELM) and AdaBoost. SVM is used to classify the selected features. The line and triangle-based features are computed when the features are selected while the point-based features are computed directly. The best performance is obtained using multiple datasets. To improve the power of learning deep features, a novel island loss technique is implemented for convolutional network CNN [30]. The island loss technique with CNN (IL-CNN) outperforms the baseline CNN. It is used to reduce the intraclass variations that happen due to head position changes, occlusions and illumination variations. The IL-CNN outperforms the other techniques while using the CK+ dataset for a class of seven expressions and the Oulu–CASIA dataset. For the enhancement of the services of healthcare in smart cities, the FER technique [30] is proposed to extract the sub-bands by applying the "band let" transform to the face image. The weighted center-symmetric local binary pattern (CS-LBP) is implemented for every sub-band in the image in a block-wise manner. The feature vector is formed by combining CS-LBP histograms.

The most dominant features are extracted and classified by Support Vector Machine (SVM) and Gaussian Mixture Model. The performance of the technique is better in the case of JAFFE and CK datasets [30]. A novel FER system is proposed with a support vector machine (FERS) by Bargshady et al. [31]. The faces are detected from the image by combining self-quotient image (SQI) filter and Haar-like features. The SQI filter is used to overcome the light variations. The features are computed using the angular radial transforms (ART), discrete cosine transform (DCT), and the Gabor filter (GF) from the faces. The support vector machine (SVM) classifies the features and gives the best performance in terms of recognition rate for training and testing the patterns. "Simultaneous feature and dictionary learning" (SFDL) technique is proposed for sets of face images. In each training and testing set, the images were captured with different illumination and pose variations. SFDL method is implemented for the raw face pixels that learned the features and dictionaries. In stage one of the learning procedure, the facial image sets are manipulated together. The deep SFDL (D-SFDL) method is proposed for non-linear face samples of image-sets, by learning both class-specific dictionaries and hierarchical non-linear transformations. A shallow module is executed to extract the most discriminant information from the global and local regions to learn the low-level features. Then a part-based module is constructed to extract and learn dynamic local region information related to facial expressions. The long-short term memory (LSTM) and gated recurrent unit (GRU) layers are used to learn long-term dependencies. The extensive experiments show that the proposed technique gives better performance in the case of CK+ and Oulu–CASIA datasets.

3 Proposed Patch Based Multiple Descriptors Technique

The proposed technique describes novel patch-based multiple LBP descriptors using TPLBP and FPLBP. Image representation is computed from local patch values through these descriptors and encodes the properties of the local micro-texture around every pixel using short binary strings. Three patch LBP and four patches LBP [31] are implemented that encode the most discriminate types of local texture-based information. This technique consists of four steps. In the first step, faces are detected and cropped to overcome the multi-scale variations in the preprocessing step using the Viola–Jones algorithm described in Section 3.1. The feature engineering and fusion step introduces texture-based features. It helps in face detection through multiple LBP-based techniques. The discriminative features are extracted through DCT and fusion is performed. The feature vector is formed by fusing the selected features in Section 3.2. Finally, the classification step identifies the expression type. The experimental results for each data set are discussed in Section 3.3. The architecture of the proposed patch-based multiple descriptors technique is shown in Fig. 1.

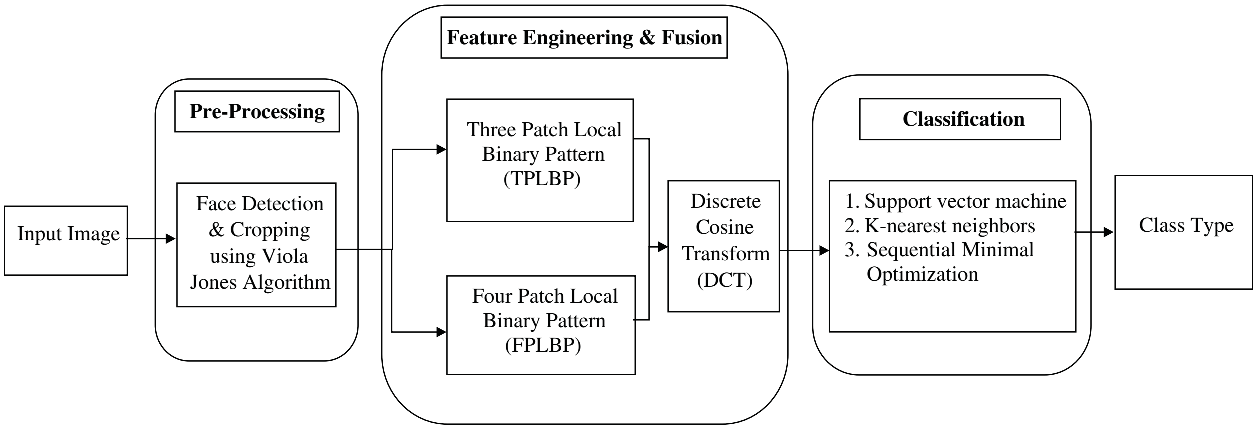

Input Image
Pre-Processing
Face Detection & Cropping using Viola Jones Algorithm
Feature Engineering & Fusion
Three Patch Local Binary Pattern (TPLBP)
Four Patch Local Binary Pattern (FPLBP)
Discrete Cosine Transform (DCT)
Classification
1. Support vector machine
2. K-nearest neighbors
3. Sequential Minimal Optimization
Class Type

Figure 1: Architecture of the proposed patch-based multiple descriptors technique

3.1 Preprocessing

The input images are preprocessed using the Viola Jones algorithm to detect and crop the face from the entire image as shown in Fig. 2. The faces are converted to grayscale if the images are already in RGB form. The pre-processing is performed due to the variance in the resolution of the images that uniformly maps all the data into 384 × 288 for CK+ and 65 × 65 for Oulu–CASIA datasets. The features are extracted and encoded in the feature engineering and fusion step.

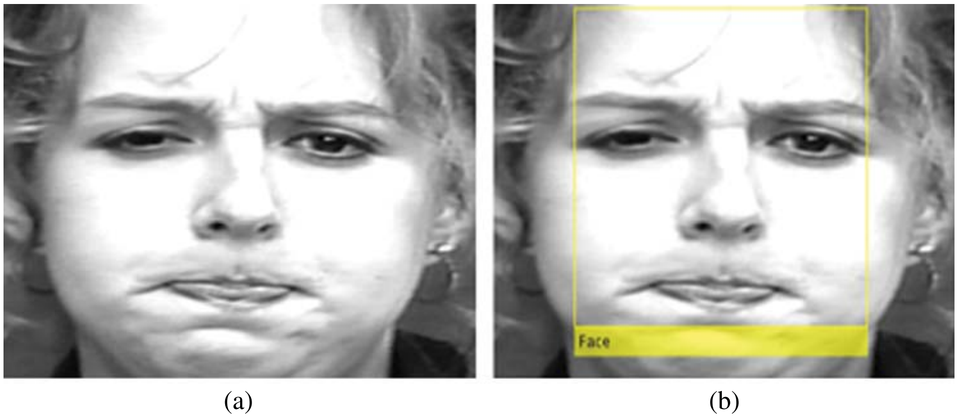


Figure 2: (a) Input image (b) face detection using Viola Jones algorithm

3.2 Feature Engineering and Fusion

This section includes a description of three patch LBP and four patch LBP (TPLBP, FPLBP) feature extraction mechanism, and encoding process that are performed on pre-processed images.

3.2.1 Three Patch LBP (TPLBP) Coding

TPLBP and FPLBP work like the simple local binary pattern (LBP) [19] technique but are extended and developed to introduce a patch-based version. Three patch pixel values are compared with each other to produce a single value and assign it to every pixel in the image to form a TPLBP coded image. A w × w patch positioned at the central pixel is computed for every pixel in the input face image. The S extra patches are allocated consistently around a central patch in a circle denoted by radius r and the central patch is compared with each z pair of patches values. The single bit value is defined, based on the two patches values that is closer to the middle patch thus resulting in S bits per pixel code. The following formula is executed for every pixel in the image to produce three-patch LBP coding.

$$TPLBP_{r,S,w,z}(\mathrm{p}) = \sum_{i=1}^{S} f\left(d\left(Y_{i,} Y_{p}\right) - d\left(Y_{i+Zmod\,S}, Y_{p}\right)\right) 2^{i} \tag{1}$$

In Eq. (1), Y_p denotes the central patch and the two patches are denoted by Y_i and $Y_{i+Zmod\,S}$ along the ring. To calculate the distance between any two patches, the d(p1, p2) is used (e.g., the gray level differences between p1 and p2 is L2 norm). The function f is formulated as:

$$f(y) = \begin{Bmatrix} 1 & if \;\; y \geq t \\ 1 & if \;\; y < t \end{Bmatrix} \tag{2}$$

For uniform regions some stability is provided using t value in Eq. (2) (e.g., t = 0.01). The processing speed is increased by obtaining the patches through nearest-neighbor sampling instead of interpolating their values.

A coded image is produced by encoding the input image similar to the CSLBP descriptor [24]. The coded image is split into a grid of non-overlapping regions and for every region, the histogram is computed that measures the frequency of every binary value. Unit length is produced by normalizing each histogram region; their values are truncated at 0.2 and normalized to unit length again. A single vector is formed by concatenating these histograms generated for an image.

3.2.2 Four Patch LBP Coding

In four patched LBP, the two circles of the radii r_1 and r_2 are positioned on the central pixel for every pixel in the input image. S extra patches of size w × w are split out around each circle consistently. In the internal circle, two central patches are compared with the two center patches in the external circle that is located z patches apart from each other in the circle. By comparing the two pairs, the one with higher similarity is used to define one bit in every pixel's value. S/2 center symmetric pairs for S extra patches along every circle are used for the computing the binary coded length [14]. By executing the two-step process, the coded image is computed. The following equation computes FPLBP coded image.

$$FBLBP_{r_1 r_2 s,w,z}(p) = \sum_{i=1}^{s/2} f(d\left(Y_{1i,} Y_{2,i+z} mod\, S\right) - d\left(Y_{1,i+s/2,} Y_{2,i+\frac{s}{2}+z} mod\, S\right))2^{i} \tag{3}$$

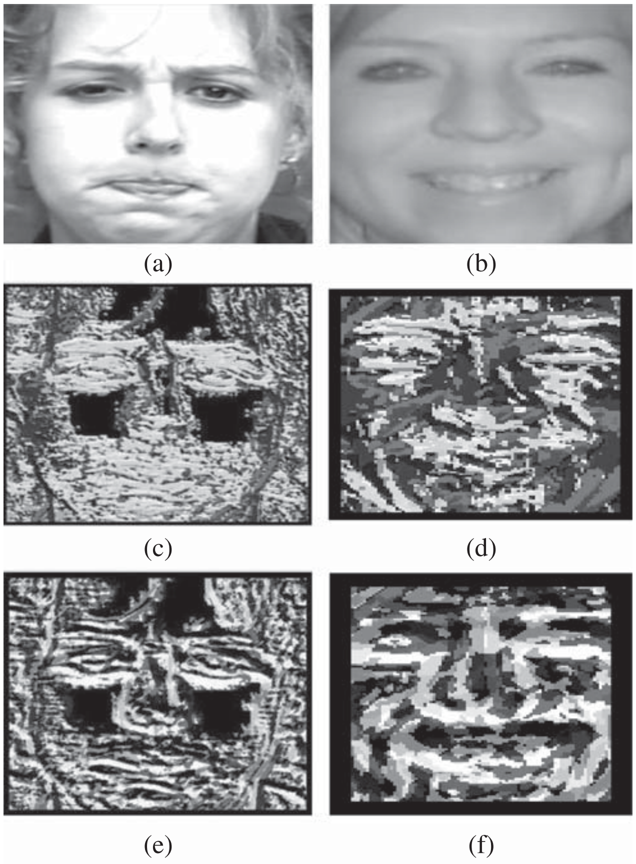


Figure 3: (a, b) Original images of the CK+ and Oulu–CASIA dataset (c, d) TPLBP coded images of the CK+ and Oulu–CASIA dataset (e, f) FPLBP coded image of the CK+ and Oulu–CASIA dataset

3.2.3 Discrete Cosine Transform (DCT)

DCT is a technique which describes an image as a sum of sinusoids or just like sinusoidal waves of varying magnitudes and frequencies. For feature extraction, DCT-2 technique is implemented in the proposed approach. Two-dimensional (2-D) DCT is applied on an input image that converts it into DCT coefficients of the same matrix as the input image. The most significant information is stored in just a few coefficients on the top left corner of the transform output called low frequencies. These are extracted in a zigzag manner and high frequencies are discarded as shown in Figs. 3c, 3d. Due to this reason, the DCT is often used in image compression applications. For example, in JPEG, for an X × Y input image, the DCT is computed by the following equation:

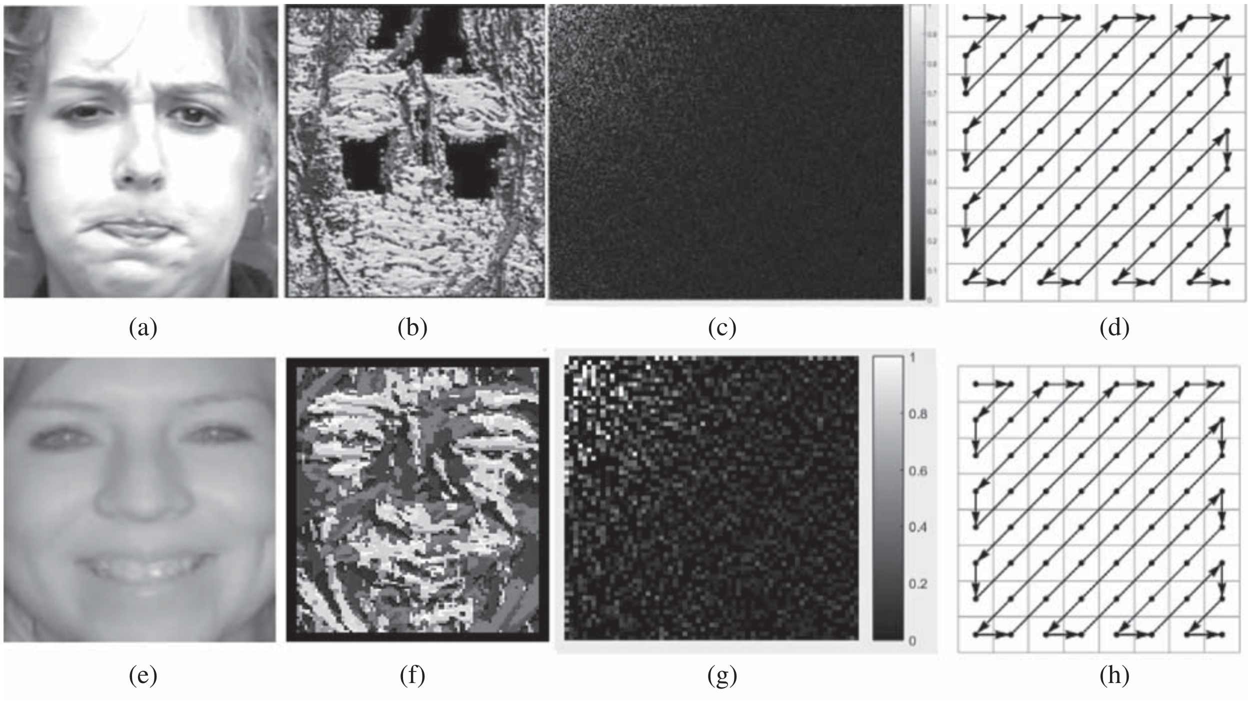
$$F(x,y) = \frac{1}{\sqrt{AB}}\alpha(x)\alpha(y)\sum_{u=0}^{A-1}\sum_{v=0}^{B-1}f(u,v)\cos\left(\frac{(2u+1)x\pi}{2A}\right)\cos\left(\frac{(2v+1)y\pi}{2B}\right) \tag{4}$$

$x = 0, 1, \ldots, A$ and $y = 0, 1, \ldots, B$

where f(u, v) function denotes the image intensity while F(x, y) function denotes the computed DCT coefficients in a 2D matrix form.

3.2.4 Feature Fusion

The TPLBP and FPLBP codes are generated and features are extracted through DCT, separately from each of these methods in the feature engineering phase. The features are also fused to construct a feature vector in a zigzag manner. The zigzag function takes a matrix and a certain number of features such as 64 (8×8) as an input and returns a one-dimensional array consisting of the results of zigzag scans. For example, it stores, the value of the first pixel and flows in the right and down direction until the 8 x 8 matrix is complete as shown in Figs. 4d, 4h. The same process is repeated for all the images that are coded through TPLBP and FPLBP respectively to form a concatenated feature vector of 128 values (64 and 64 for each image) as shown in Tab. 1.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4: (a, e) Original images of CK+ and Oulu–CASIA dataset (b, f) TPLBP coded images of original image (c, g) DCT output Image of coded images (d, h) Feature extraction in zigzag manner from DCT output image

3.3 Classification

The features are encoded and extracted in the feature engineering section followed by the feature fusion section, separately to form the feature vector of each dataset. SVM, KNN and SMO classifiers [31] are used to classify facial expressions using selected features.

4 Experimental Results and Discussion

Experiments have been performed on two different facial expression datasets the extended Cohn–Kanade (CK+) and partial Oulu–CASIA dataset. The detail of datasets is given in Tab. 2.

Execution of the proposed technique has been assessed using performance measures such as precision, recall, accuracy, specificity, sensitivity, and F-score. To determine the efficiency of the proposed technique, the comparison has been performed with the existing methods based on the included datasets.

Table 1: Concatenated feature vector record

Images	TPLBP + DCT	FBLBP + DCT	Expression labels
Image 1	64 (8 × 8) features	64 (8 × 8) features	1
Image 2	64 (8 × 8) features	64 (8 × 8) features	2
Image 3	64 (8 × 8) features	64 (8 × 8) features	3
Image 4	64 (8 × 8) features	64 (8 × 8) features	4
Image 5	64 (8 × 8) features	64 (8 × 8) features	5
Image 6	64 (8 × 8) features	64 (8 × 8) features	6

Table 2: Specifications of facial expression datasets

Database	Expressions	No. of subjects	No. of sequences	Gray/color	Resolution	Type
CK+	Six basic expressions	18	540 image sequences	Mostly grey	384 × 288	Posed, spontaneous smiles
Oulu–CASIA	Six basic expressions	20	7200 image sequences	Mostly grey	65 × 65	Posed, illumination variant

4.1 Experiments on Extended Cohn Kanade (CK+) Dataset

Initially, experiments are performed on the CK+ dataset according to the descriptions mentioned in Tab. 3 and the distribution of the images from dataset is shown in Tab. 4. Fig. 5 shows the performance of multiple classifiers such as SVM, KNN, and SMO using CK+ dataset.

Table 3: Contents of facial expression datasets

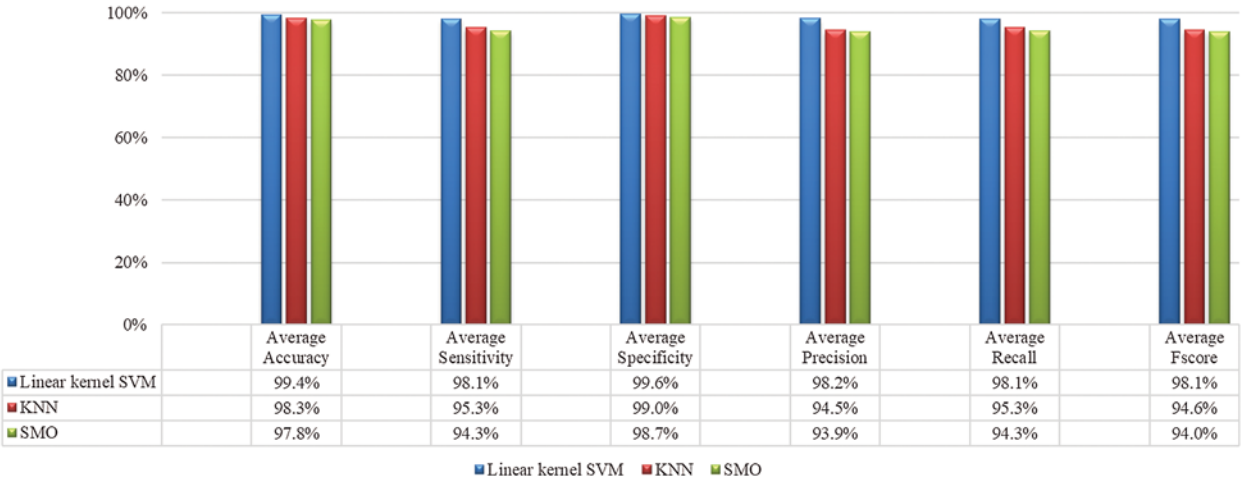
Feature set	Three patch and four patch LBP
Classifiers	LBP features extraction and encoding using discrete cosine transform 1. Support vector machine 2. K-nearest neighbor 3. Sequential minimal optimization

Tab. 5 shows the recognition rate of linear kernel SVM classifier. Values in bold indicate the best recognition cases for each class of the CK+ dataset. The recognition rate is high but anger is misclassified as disgust and happy.

Tab. 6 shows the recognition rate when KNN classifier is used. The recognition rate is good but anger and disgust are confused with fear and disgust is misclassified as happy.

Table 4: Number of CK+ images per expression

Expressions	Number of Images
Anger	90
Disgust	90
Fear	90
Happy	90
Sad	90
Surprise	90



	Average Accuracy	Average Sensitivity	Average Specificity	Average Precision	Average Recall	Average Fscore
■Linear kernel SVM	99.4%	98.1%	99.6%	98.2%	98.1%	98.1%
■KNN	98.3%	95.3%	99.0%	94.5%	95.3%	94.6%
■SMO	97.8%	94.3%	98.7%	93.9%	94.3%	94.0%

Figure 5: Performance metric values using CK+ dataset for multiple classifiers

Table 5: Confusion matrix for linear kernel SVM classifier using CK+ database

Expressions	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**38**	1	0	1	0	0
Disgust	0	**47**	0	0	0	0
Fear	0	0	**32**	0	0	0
Happy	0	0	1	**35**	0	0
Sad	0	0	0	0	**32**	0
Surprise	0	0	1	0	0	**28**

Tab. 7 shows the recognition rate in the case of SMO classifier. The recognition rate is high, but anger is confused with sad while disgust and sad are misclassified as anger.

A comparison of the results of different techniques presented in literature work along with the proposed technique using the CK+ dataset is presented in Tab. 8. A histogram of oriented features is generated along the discrete cosine transform. The hybrid technique has performed better when images with multi-scale and illumination variations are used. The face images are used to obtain local binary pattern (LBP) and weber local descriptor (WLD) features along with feature extraction mechanism based on DCT to perform classification. It is observed that the technique is not robust to noise. Tab. 8 shows the performance of the proposed system is comparable with other relevant techniques and the results describe the effectiveness of the proposed technique.

Table 6: Confusion matrix for KNN classifier using CK+ database

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**43**	1	0	0	0	0
Disgust	0	**23**	0	0	0	0
Fear	2	1	**39**	0	0	2
Happy	0	4	0	**30**	0	0
Sad	0	0	1	0	**35**	0
Surprise	0	0	0	0	0	**35**

Table 7: Confusion matrix for SMO classifier using CK+ database

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**32**	3	1	1	5	0
Disgust	0	**48**	0	0	0	0
Fear	0	0	**29**	0	0	0
Happy	0	0	0	**33**	0	0
Sad	3	0	1	0	**35**	0
Surprise	0	0	0	0	0	**25**

Table 8: Comparison of proposed technique with existing techniques using CK+ dataset

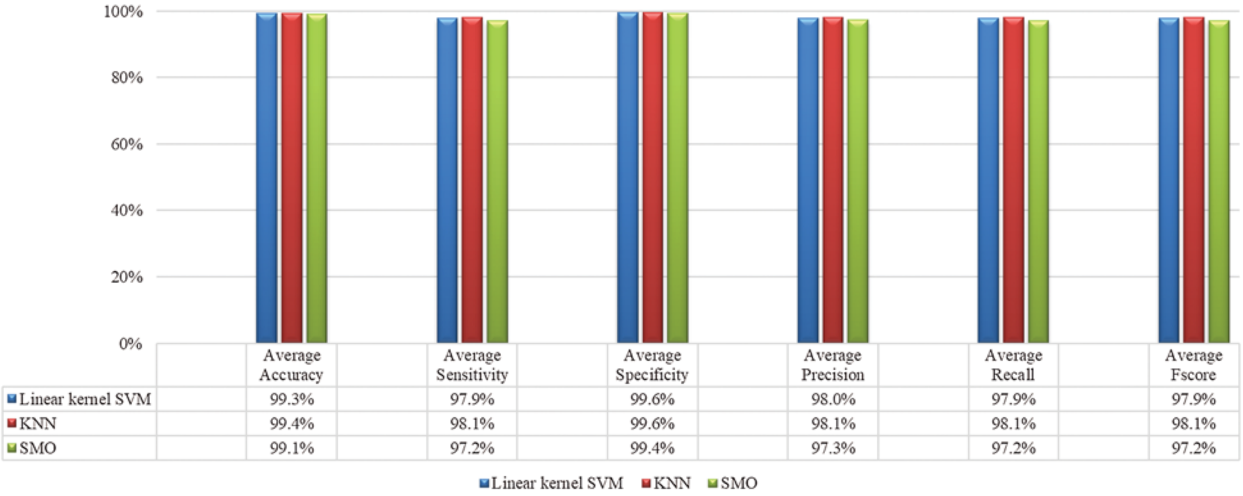
Database	Technique	No. of classes	Accuracy (%)
CK+	**WLD [SVM]** [1]	7	99.28
	Hybrid (HOG & DCT) [KNN, SMO, MLP] [5]	6	**99.60**
	AAM, HOG, PCA [DSAE] [6]	6	95.79
	WLBI-CT		99.30
	[SMO, NB, KNN, MLP, classification trees] [7]	6	99.30
	EBGM and KLT [SVM] [9]	6	97.80
	Proposed technique	6	99.40

4.2 Experiments on Oulu–CASIA Dataset

The performance of the proposed technique is also evaluated on the subset of Oulu–CASIA dataset. The dataset consists of two camera types named NIR (near to infrared) and VL (visible light) to produce image sequences. The images are also captured in different illumination situations including dark, strong, and weak light with these cameras. The main dataset contains the image sequences of 80 different persons. Data of 20 persons is obtained from VL camera with occlusion (with glasses) and without occlusion (without glasses), separately. Image sequences having dark, strong and weak light for each of 20 persons are combined in both cases, with occlusion and without occlusion for each expression, separately.

4.2.1 Experiments on Oulu–CASIA with Occlusion

The experiments are evaluated on a combination of dark, strong, and weak illumination having frontal and spontaneous images with glasses. Fig. 6 shows the average accuracy, sensitivity, specificity, precision, recall and F-score, using multiple classifiers such as SVM, KNN and SMO.



	Average Accuracy	Average Sensitivity	Average Specificity	Average Precision	Average Recall	Average Fscore
🟦Linear kernel SVM	99.3%	97.9%	99.6%	98.0%	97.9%	97.9%
🟥KNN	99.4%	98.1%	99.6%	98.1%	98.1%	98.1%
🟩SMO	99.1%	97.2%	99.4%	97.3%	97.2%	97.2%

Figure 6: Performance of multiple classifiers using Oulu–CASIA dataset with occlusion

Tab. 9 shows the recognition rate of linear kernel SVM classifier using Oulu–CASIA dataset with occlusion. The recognition rate is good but disgust is misclassified as anger while happiness and surprise are confused with fear.

Table 9: Confusion matrix for linear kernel SVM using Oulu–CASIA with occlusion dataset

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**115**	2	1	0	1	0
Disgust	0	**119**	0	0	0	0
Fear	0	0	**110**	4	0	2
Happy	0	0	1	**132**	0	0
Sad	1	1	0	2	**120**	0
Surprise	0	0	0	0	0	**109**

Tab. 10 shows the recognition rate in the case of the KNN classifier. Bold values indicate the best recognition rate of Oulu–CASIA with occlusion dataset. The recognition rate is accurate, but disgust misclassifies with fear and happiness while fear is confused with the surprise factor.

Table 10: Confusion matrix for KNN classifier using Oulu–CASIA with occlusion dataset

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**113**	0	0	0	0	0
Disgust	0	**115**	0	0	0	0
Fear	0	3	**123**	0	0	0
Happy	0	2	0	**126**	0	1
Sad	0	0	1	0	**127**	0
Surprise	0	0	7	0	0	**118**

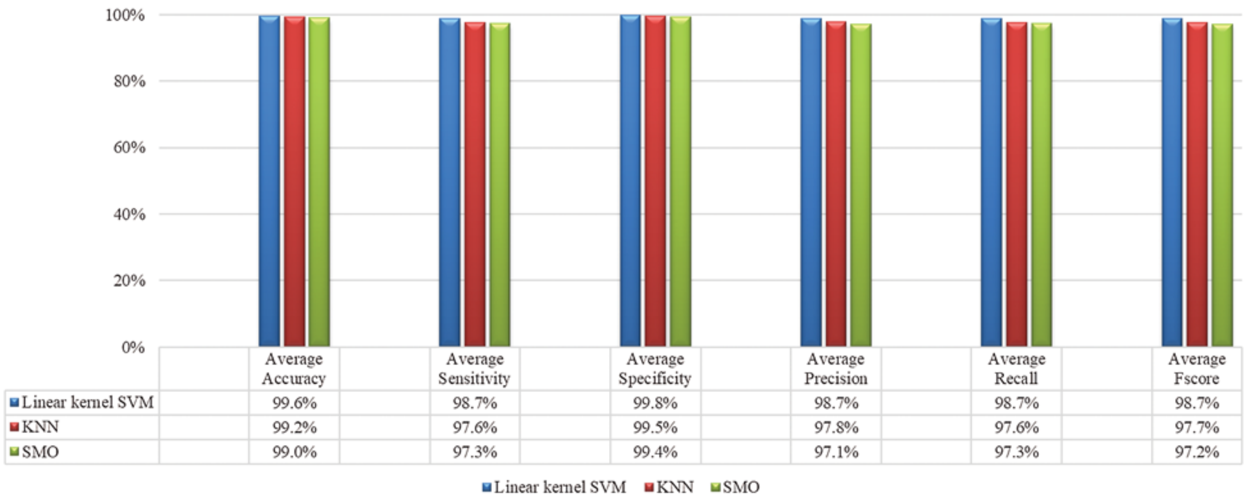
Tab. 11 shows the recognition results of SMO classifier using Oulu–CASIA with occlusion dataset. The confusion matrix indicated that disgust is confused with anger and sad while fear is classified as surprise. Sad is also confused with anger, disgust and fear.

Table 11: Confusion matrix for SMO classifier using Oulu–CASIA with occlusion dataset

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**110**	2	0	0	3	0
Disgust	1	**119**	2	0	3	0
Fear	0	0	**121**	0	2	4
Happy	0	0	0	**126**	0	0
Sad	0	2	1	0	**132**	0
Surprise	0	0	0	0	0	**103**

4.2.2 Experiments on Oulu–CASIA Without Occlusion

The experiments are performed on the dark, strong, and weak illumination frontal and spontaneous images having faces without glasses and combined as one dataset. Fig. 7 shows the average accuracy, sensitivity, specificity, precision, recall, and F-score using multiple classifiers.



	Average Accuracy	Average Sensitivity	Average Specificity	Average Precision	Average Recall	Average Fscore
■Linear kernel SVM	99.6%	98.7%	99.8%	98.7%	98.7%	98.7%
■KNN	99.2%	97.6%	99.5%	97.8%	97.6%	97.7%
■SMO	99.0%	97.3%	99.4%	97.1%	97.3%	97.2%

Figure 7: Performance of multiple classifiers using Oulu–CASIA without occlusion dataset

Tab. 12 shows the results of linear kernel SVM classifier using Oulu–CASIA without occlusion dataset. The recognition rate shows that anger is confused with disgust.

Table 12: Confusion matrix for linear kernel SVM using Oulu–CASIA without occlusion dataset

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**126**	0	0	0	0	0
Disgust	6	**113**	0	0	0	0
Fear	0	0	**121**	0	0	5
Happy	0	0	0	**117**	0	0
Sad	2	1	0	0	**114**	0
Surprise	3	0	0	0	0	**112**

Tab. 13 describes the results obtained by employing the KNN classifier using Oulu–CASIA dataset without occlusion.

Table 13: Confusion matrix for KNN classifier using Oulu–CASIA without occlusion dataset

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**117**	1	0	0	1	0
Disgust	4	**115**	0	0	0	0
Fear	0	0	**114**	0	1	1
Happy	0	0	0	**133**	0	0
Sad	0	0	1	0	**123**	0
Surprise	0	0	0	0	0	**109**

The results obtained by employing SMO classifier using Oulu–CASIA without occlusion dataset are listed in Tab. 14. The values reveal that disgust is confused with fear and happiness.

Table 14: Confusion matrix for SMO classifier using Oulu–CASIA without occlusion dataset

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**113**	0	0	0	0	0
Disgust	0	**115**	0	0	0	0
Fear	0	3	**123**	0	0	0
Happy	0	2	0	**110**	0	1
Sad	0	0	1	0	**127**	0
Surprise	0	0	7	0	0	**118**

The average recognition rate of the proposed technique is compared with existing methods using the same Oulu–CASIA dataset. IL-CNN technique is often used to reduce the intra-class variations. The performance and accuracy are low with the IL-CNN technique as compared to the proposed technique. The long-short term memory (LSTM) and gated recurrent unit (GRU) layers are used to learn long-term dependencies. Tab. 15 shows that the performance of the proposed system is better than other techniques.

Table 15: Comparison of proposed technique with existing using Oulu–CASIA dataset

Database	Technique	No. of classes	Accuracy (%)

Oulu–CASIA	LBP-TOP [SVM] [22]	6	NIR-STRONG = 79.40
			NIR-WEAK = 73.03
			NIR-DARK = 76.03
			VL-STRONG = 79.40
			VL-WEAK = 74.53
			VL-DARK = 58.80

	Multi-task global-local network (MGLN) [30]	6	MGLN-LSTM = 90.4
			MGLN-GRU = **90.4**

	Proposed technique	6	NIR—with Occlusion = **99.4** NIR—Without Occlusion = **99.2**

5 Conclusion and Future Work

A novel patch-based multiple LBP descriptors techniques namely three patch local binary patterns (TPLBP) and four patch local binary patterns (FPLBP) have been proposed. The proposed system exploits the feature extraction ability of TPLBP and FPLBP along with DCT to overcome the issues of illumination, scale variations, high dimensions, noisy images, and higher computational complexity of texture-based features. Multiple classifiers are used to classify standard CK+ and Oulu–CASIA datasets with posed, spontaneous emotions, illumination variant and multi-scale face images. The proposed technique can obtain a high-performance rate, which is relatively tough in situations with variations in angles and noise. The performance can be further improved to manage these factors using some pre-processing techniques along with TPLBP and FPLBP.

Funding Statement: This research was supported in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-2016-0-00312) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation) and in part by the Faculty Research Fund of Sejong University in 2019.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Alhussein, “Automatic facial emotion recognition using Weber local descriptorfor e-Healthcare system,” *Cluster Computing*, vol. 19, no. 1, pp. 99–108, 2016.

[2] A. Uçar, Y. Demir and C. Güzeliş, “A new facial expression recognition based on curvelet transformand online sequential extreme learning machine initialized with spherical clustering,” *Neural Computing and Applications*, vol. 27, no. 1, pp. 131–142, 2016.

[3] K. Lekdioui, R. Messoussi, Y. Ruichek, Y. Chaabi and R. Touahni, “Facial decomposition for expression recognition using texture/shape descriptors and SVM classifier,” *Signal Processing Image Communication*, vol. 58, pp. 300–312, 2018.

[4] D. Ghimire, S. Jeong, J. Lee and S. H. Park, "Facial expression recognition based on local region specific features and support vector machines," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 7803–7821, 2017.

[5] M. Nazir, Z. Jan and M. Sajjad, "Facial expression recognition using histogram of oriented gradients based transformed features," *Cluster Computing*, vol. 21, no. 1, pp. 1–10, 2017.

[6] N. Zeng, H. Zhang, B. Song, W. Liu and Y. Li, "Facial expression recognition via learning deep sparse autoencoders," *Neurocomputing*, vol. 273, pp. 643–649, 2018.

[7] S. A. Khan, A. Hussain and M. Usman, "Reliable facial expression recognition for multi-scale images using weber local binary image based cosine transform features," *Multimedia Tools and Applications*, vol. 77, no. 1, pp. 1–33, 2019.

[8] M. Guo, X. Hou, Y. Ma and X. Wu, "Facial expression recognition using ELBP based on covariance matrixtransform in KLT," *Multimedia Tools and Applications*, vol. 76, no. 2, pp. 2995–3010, 2017.

[9] D. Ghimire, J. Lee, Z. Li and S. Jeong, "Recognition of facial expressions based on salient geometric features and support vector machines," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 7921–7946, 2017.

[10] T. L. New, S. W. Foo and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2017.

[11] A. S. Al-Waisy, R. Qahwaji, S. Ipson and S. Al-Fahdawi, "A multimodal deep learning framework using local feature representations for face recognition," *Machine Vision and Applications*, vol. 29, no. 1, pp. 35–54, 2018.

[12] A. Pillai, R. Soundrapandiyan, S. Satapathy, S. C. Satapathy, K. Jung *et al.,* "Local diagonal extrema number pattern: A new feature descriptor for face recognition," *Future Generation Computer Systems*, vol. 81, pp. 297–306, 2018.

[13] U. Mlakar, I. Fister, J. Brest and B. U. Potočnik, "Multi-objective differential evolution for feature selection in facial expression recognition systems," *Expert Systems with Applications*, vol. 89, pp. 129–137, 2017.

[14] L. Wolf, T. Hassner and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1978–1990, 2011.

[15] T. Ojala, M. Pietkainen and T. Maenpaa, "Multi resolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[16] M. J. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba, "Coding facial expressions with Gabor Wavelets," in *Proc. AFGR*, Nara Japan, pp. 200–205, 1998.

[17] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar *et al.,* "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. CVPRW*, San Francisco, CA, United States, pp. 94–101, 2010.

[18] G. Zhao, X. Huang, M. Taini, S. Z. Li and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[19] G. Huang, V. Jain and E. Learned-Miller, "Unsupervised joint alignment of complex images," in *IEEE Int. Conf. on Computer Vision*, Rio de Janeiro, Brazil, 2007.

[20] M. Heikkilä, M. Pietikäinen and C. Schmid, "Description of interest regions with Center-symmetric local binary patterns," in *Proc. ICCVGIP*, Madurai, India, pp. 20–32, 2006.

[21] D. D. Lewis, *Representation and Learning in Information Retrieval.* University of Massachusetts at Amherst, Computer and Information Science Dept. Graduate Research Center Amherst, MA, USA, 1992.

[22] Y. Guo, G. Zhao and M. Pi Pietikäinen, "Discriminative features for texture description," *Pattern Recognition*, vol. 45, no. 10, pp. 3834–3843, 2012.

[23] G. Muhammad, M. Alsulaiman, S. U. Amin, A. Ghoneim, M. F. Alhamidet *et al.,* "A facial-expression monitoring system for improved healthcare in smart cities," *IEEE Access*, vol. 5, no. 1, pp. 10871–10881, 2017.

[24] H. H. Tsai and Y. Chang, "Facial expression recognition using a combination of multiple facial features and support vector machine," *Soft Computing*, vol. 22, no. 13, pp. 4389–4405, 2018.

[25] J. Lu, G. Wang and J. Zhou, "Simultaneous feature and dictionary learning for image set based face recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 4042–4054, 2017.

[26] M. M. F. Donia, A. A. A. Youssif and A. Hashad, "Spontaneous facial expression recognition based on Histogram of oriented gradients descriptor," *Computer Information Science*, vol. 7, no. 3, pp. 31–37, 2014.

[27] C. Hamburger, "Quasimonotonicity, regularity and duality for nonlinear systems of partial differential equations,"*AnnalidiMatematicaPuraedApplicata*, vol. 169, no. 1, pp. 321–354, 1995.

[28] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly *et al.,* "Island loss for learning discriminative features in facial expression recognition," in *Proc. FG 2018*, Xian, Shaanxi, China, pp. 302–309, 2018.

[29] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3, pp. 169–200, 1992.

[30] M. Yu, H. Zheng, P. Zhifeng, J. Dong and H. Du, "Facial expression recognition based on a multi-task global-local network," *Pattern Recognition Letters*, vol. 131, pp. 166–171, 2020.

[31] G. Bargshady, X. Zhou, R. C. Deo, J. Soar, F. Whittaker *et al.,* "Enhanced deep learning algorithm development to detect pain intensity from facial expression images," *Expert Systems with Applications*, vol. 149, pp. 113–121, 2020.