International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I11P230
Volume 69 Issue 11, 236-247, November, 2021

An Aggregated Optical Flow Vectors for Micro Expression Recognition Using Spatio-Temporal Binary Pattern Coding

Sammaiah Seelothu[1], Dr. K. Venugopal Rao[2]

[1]*Research Scholar, Department of Computer Science Engineering, Jawaharlal Nehru Technological University, Hyderabad, Telangana, India*
[2]*Professor, Department of Computer Science Engineering, G Narayanamma Institute of Technology and Science, Hyderabad, Telangana, India.*

sammaiahmanuu@gmail.com

Abstract **-** *Micro Expressions (MEs) are unique facial expressions when individual experiences an emotion but intentionally tries to hide their genuine emotion. MEs are involuntary and spontaneous, and their recognition has gained a significant research interest due to their potential applications. However, Micro Expression Recognition (MER) is an arduous task due to its short duration, subtle and local movements of faces. This paper proposes an effective descriptor called Composite Local Binary Pattern on Three Orthogonal Planes (CLBP-TOP) for Micro Expressions Recognition to address these problems. We also propose a novel Aggregated Optical Flow Vectors (AOFVs) Computation mechanism where the neighbour optical flows in a particular period are aggregated to measure motion intensities. Based on these motion intensities, we compute a weight matrix, and it is multiplied with CLBP-TOP histogram features to get weighted histogram features. For classification purposes, we employ the Support Vector Machine (SVM) algorithm. Extensive experimental evaluation of the CASME II dataset shows that our proposed approach significantly improves recognition accuracy and shows superior performance than the state-of-art methods.*

Keywords - *MER, feature extraction, aggregated optical flow vectors, Composite local binary pattern, accuracy.*

I. INTRODUCTION

Expressions are the one kind of information that reveals the information about the human mind at a time. These are often expressed through different models like facial expressions, body gestures and speech signals. Among these models, the facial expression is declared as the most significant model through which the emotion of human beings can be determined [1]. The vital role of facial expressions has made the researchers develop many methods for automatically recognising facial expressions. For instance, Vural et al. [2] developed a system for identifying driver's drowsiness and the method proposed by White-hill et al. [3] aimed at the procurement of feedback or responses of students while they are being taught. Hence, facial expression recognition has been applied in several fields like computer vision, fatigue detection, online feedback collection, physiological counselling [4] and criminal detection [5].

In general, facial expressions are divided into two categories; they are macro expressions and micro-expressions. Macro expressions are the typical expressions present in our daily lives when people interact with each other. The average period of macro expression varies between ½ seconds and 4 seconds. On the other hand, micro-expressions are incurred in high-stakes situations where people try to hide or suppress their genuine emotions [6]. Such kinds of concealed emotions have an approximate period of 1/5 second to 1/25 second, and hence they are called Micro-Expressions (MEs). Haggard and Isaacs [7] initially identified the Micro expression and then Ekman and Friesen [8].

Along with the short duration of MEs, they have loess intensities. Due to these reasons, the recognition of MEs is much difficult for a human to execute. Moreover, Ekman [9] suggested that Micro Expression Recognition (MER) task results in only an average performance without proper training. Hence there is a need for an automatic and reliable MER system to help people in recognition of MEs more accurately, especially in the fields like interrogations [10, 11], emotional interfaces [12], lie detection, sentimental analysis and clinical diagnosis [13].

Recently, the MER has become a hot research topic, and so many researchers have tried to develop so many methods for the task of automatic MER [14]. Typically, the MER is composed of two components: feature extraction from video clips and classification. In the context of feature extraction, for a given video clip, the features are extracted, which help describe the micro expression. Once the features are extracted, they are fed to a classifier that tries to find the concealed emotion. Along with these stages, there may exist one more stage called preprocessing. In the micro expression

related video clips, the emotion-related information is present at only a few frames. So the frames' determination is required at the preprocessing stage such that the processing burden is reduced over the system. The determination of that frames is called ME spotting. In almost all micro expression related databases, the database creators already mentioned the period of ME spotting. Some authors tried to spot the MEs through new methods, and then they were processed for recognition.

On the other hand, some authors utilised the specified ME spotting and processed the frames in that particular span for recognition. In both methods, the frames are subjected to feature extraction, and among the earlier feature extraction methods, Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) is widely employed in video-based MER and in some other computer vision-based tasks [15]. However, the LBP-TOP is susceptible to information loss, and there is no matter of encoding the neighbour pixels relation at different scales. The centre pixel is encoded only concerning the differences between itself and neighbour pixels. However, there is no consideration for second-order information.

In this paper, to address these problems, we propose a new feature extraction method by considering the LBP-TOP as a base reference method. Two variants of LBP-TOP are developed here; one is based on the radial difference, and another is based on angular differences. A centre pixel is encoded based on differences between itself and neighbour pixels, the difference between neighbour pixels at different scales, and the difference between neighbour pixels. After the encoding, the features are transformed into histograms. Simultaneously the ME frames are processed to compute cumulated optical flows followed by weights computation for each local block. The weight is multiplied with histogram features to find the final features, and they are fed to the SVM classifier for emotion recognition.

The remaining paper is organised as follows; Section II explores the information related to Related Work. The details of the proposed methodology are explored in section III. Section IV explores the details of simulation experiments, and finally, the concluding remarks are provided in section V.

II. RELATED WORK

For the recognition of ME, many approaches use the Spatio-temporal features, which can represent the small and subtle movements of MEs. Most methods extracted the Spatio-temporal features by dividing the image or frame into several non-overlapping blocks of size $N \times N$ and fetched the features from every block. Next, they are concatenated and finally generate a global feature vector. Yan et al. [16] divided the frame into 5×5 and 8×8 blocks and extracted LBP-TOP features for grouping the MEs of the CASME II dataset. The accuracy obtained at this method is chosen as baseline accuracy for MER.

Wang et al. [17] proposed a new Tensor Independent Color Space (TICS)" model for the recognition of MEs. This approach was adopted for CIELab [19] and CIELuv colour spaces and proved they also help recognise MEs. For a given ME video clip, they extracted a fourth-order tensor, i.e., a four-dimensional array. Among these four values, the first three belong to LBP-TOP, and the last one belongs to colour features. They also applied a magnification method on block-based "Histogram of Image Gradient Orientation (HIGO)" features.

Hang et al. [18] developed a "Spatio-temporal LBP with integral projection (LBP-IP)" for MER. First, to preserve the shape of the facial image, it was subjected to integral projection (both horizontal and vertical) based on different images. Following LBP-TOP, features are extracted from both projections. For classification, they employed an SVM algorithm with Chi-Square Kernel.

I. P. Adegun and V. Hima Bindu [20] applied LBP-TOP for feature extraction and "Extreme Learning Machine (ELM)" for classification. In this work, the feature extraction is applied only on the apex frame through LBP, and the micro expression videos are segmented into different image sequences through Spatio-temporal LBP-TOP. Here Support Vector Machine (SVM) is adopted as a base reference, and its training time is compared with the training time of ELM.

Y. Zhang et al. [21] proposed a new MER framework by combining the feature selection with local region division. The local region division is done based on Facial Action Coding System (FACS). Over every action unit, they employed three types of feature extraction methods such as LBP-TOP, "Histograms of Oriented Gradients on Three Orthogonal Planes (HOG-TOP)" [22, 23] and "Histograms of ImageOriented Gradients on Three Orthogonal Plans (HIGO-TOP)" [24]. Moreover, they applied a new ReliefF[25] algorithm for dimensionality reduction, and the simulations were done over CASME II and SMIC databases.

"Centralised Binary Patterns on Three Orthogonal Planes (CBP-TOP)" [26] is one more feature extraction method that can extract adequate information from spatial and temporal domains. In this method, the micro expression video clip is initially preprocessed, including face detection, interception, normalisation of size and ME spotting. Next, they applied CBP-TOP for feature extraction over the spotted frames and finally fed it to the ELM algorithm for MER.

Yandan Wang et al. [27] proposed a volumetric descriptor called LBPs with six intersection points (LBP-SIP) for MER. LBP-SIP is derived based on three intersection lines crossing over the centre point. LBPSIP lessens the redundancy in LBP-TOP and ensures a compact representation which leads to ales computational complexity. For classification, they employed an SVM classifier with leaving one sample out of cross-validation (LOCV).

Lu G. et al. [28] applied the RliefF algorithm with manifold learning based on Locally Linear Embedding (LLE) for the dimensionality reduction of LBP-TOP features after extracting them from a microexpression video clip. After reducing the dimensions, they applied the SVM algorithm

with "Radial Basis Function (RBF)" to classify MEs. They considered five emotion categories to recognise and used the CASME II dataset through "Leave one subject out cross-validation (LOSO-CV)" for simulation.

Even though LBP-TOP is widely employed, it is still not compact enough for feature extraction. Based on this inspiration, Yandan Wang e al. [29] proposed two descriptors such as LBPSIP and Super Compact LBP-Three mean Orthogonal Planes (MOP). These two methods preserve the crucial patterns and lessen the redundancy that significantly affects the discriminability of encoded features.

Even though LBP-TOP was implemented in a widespread fashion, there are two critical problems in LBP-TOP. The first problem is its limited performance at micro-expression analysis. LBP-TOP extracts the motion and appearance information based on the difference on sign basis between two pixels but not considered the other information. The LBP-TOP generally uses the conventional pattern models, which are not suitable for some applications related to local structure analysis. X. Huang et al. [30] proposed a new method called "Spatio-Temporal Completed Local Quantization Pattern (STCLQP)" for ME analysis. Initially, STCLQP fetches three essential components such as orientation, magnitude and sign. Next, an effective quantisation of vectors and codebook selection method is designed for every component in the temporal domain to study the discriminative and compact codebooks to generalise conventional pattern models.

Optical flow components [34] are one more features that can alleviate the motion information, and some of the researchers applied them for MER. Y. J. Liu et al. [31] proposed a simple method called "Main Directional Mean Optical Flow (MDMO)" for MER. The MDMO is a Region of Interest (ROI) based normalised statistical feature that considers local spatial location and static motion information. The most exciting fact of MDMO is its smallest size. Finally, they employed an SVM classifier for the recognition of micro-expressions.

S. L. Happy and A. Routray [32] explored the temporal features linked with facial MEs and proposed a "Fuzzy Histogram of Optical Flow Orientation (FHOFO)" feature for MER. The FHOFO is constructed with appropriate histograms from optical flow vector orientations with the help of histogram fuzzification to encode the temporal pattern. This approach also exposed the effect of exclusion and inclusion of motion magnitudes at the time of FHOFO features extraction. Next, Lu et al. [41] proposed a new method called "Fusion of Motion Boundary Histograms (FMBH)" that integrates the vertical and horizontal displacements of differential OF vectors. Next, a new proposition method is proposed by S. T. Liong et al. [33] in which only two frames are utilised for every video, are the onset frame and the apex frame. A simple emotion descriptor called "Bi-Weighted Oriented Optical Flow (Bi-WOOF)" is proposed to encode the expression of the required features of the apex frame.

Recently, J. Wu et al. [35] proposed two optical flow filtering methods; are "Optical Flow Filtering based on Two Branches Decisions (OFF2BD)" and "Optical Flow Filtering based on Three-Way decision (OFF3WD)". These two methods can filter the low quality ME video clips. OFF2BD used the conventional binary logic for the image classification and divided the images into positive and negative for further filtering. On the other hand, OFF3WD includes the boundary domain to delay to judge the quality of motion in images.

III. PROPOSED APPROACH

A. Overview

In this section, we discuss the complete particulars of the developed MER system. Here we mainly focused on extracting features from micro expression video clips, and towards such an objective, we employ two different kinds of features extraction methods: Aggregative Optical Flow (AOF) vectors and Composite Local Binary Pattern (CLBP) Coding. Here we consider a microexpression video with several frames as input and encode it through cumulative optical flow vectors followed by composite local binary pattern coding. Initially, we compute a cumulated optical flow vector for an input video and measure a weight matrix for every block in each frame. Simultaneously we compute Composite Local Binary Patterns followed by Histograms. Under the composite LBP, we consider the LBP-TOP instead of uncomplicated LBP.

Moreover, the LBP-TOP is computed under three phases: normal LBP-TOP, Scalable LBP-TOP (SLBP-TOP) and Orientational LBP-TOP (OLBP-TOP). The proposed new LBP-TOP encodes the first-order information and encodes the second-order discriminative information in two ways: scaling and orientation. The proposed LBP-TOP is more effective, and it can capture local, subtle changes of pixel intensities and ensure a robust discriminative power for the recognition system. Next, the obtained composite LBP-TOP histograms are multiplied with the weight matrix to get the final future set and it is fed to SVM classifier for micro-expression recognition. The overall block diagram of the proposed method is shown in figure.1.

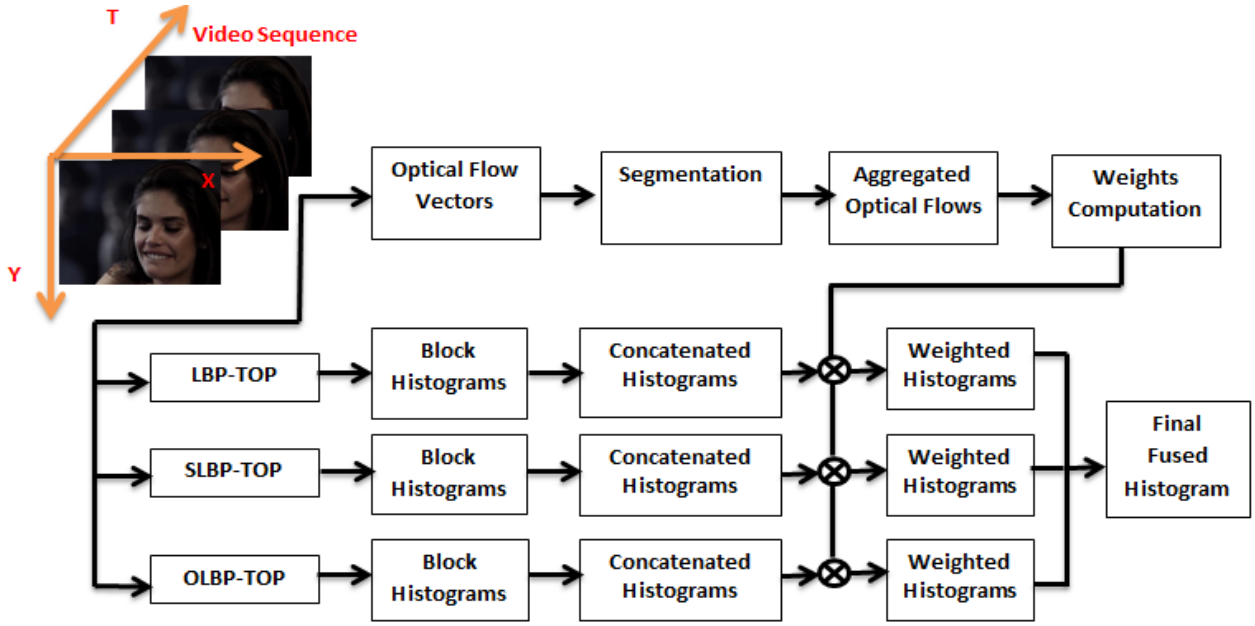

Figure.1 Block diagram of the proposed methodology

B. Aggregative optical flows

Optical Flow (OF) methods [36] have been gained tremendous research interest in current applications related to computer vision. Optical flow vectors can measure the spatial and temporal changes in pixel intensity. In the contest of micro expressions, the video clips contain many frames with only minor motions. The initial frames won't carry any motion information, and the real motion starts from the onset frame, and the peak of the motion reaches the apex frame. Moreover, the movements that occur from frame to frame are minor. A low-motion capturing method is required to acquire this kind of minor movement, and official flow is one possible solution. The optical flow is derived by computing the difference and velocity and direction of facial movements between successive frames in an input video. Mainly OF vectors are computed based on three assumptions (1) Constant brightness (2) Similar velocity (3) Objects change gradually over time but not drastically. In general, these assumptions are also possible because the pixels in a small block are of the same characteristics; the changes in face are gradually varying in nature with time, brightness, particularly in the face expression data sets, is reserved remain constant. In the real world, the illumination is adequately constant for the period of ME.

Here the optical flow vectors are measured over the frames of the video clip. The frames are considered only from onset to offset. As the databases have specified the frame numbers of onset and offset, we consider only those sets of frames for optical flow vectors measurement instead of entire frames. For example, consider subject 11's first video clip in the REVIEW dataset, the total number of frames is 64, and the keyframes (onset to offset) are only 31. The onset frame is the 8th frame, and the offset frame is the 39th frame. These details are specified in the REVIEW dataset. These sets of frames are considered keyframes since they carry most of the information required for micro-expression recognition. Hence we consider the only keyframes, and they are fed directly as an input for our optical flow vectors computation.

Consider a video clip, a pixel at location(x, y, t) with the intensity $I(x, y, t)$ have moved Δx, Δy and Δt between two frames. The optical flows are measured between two successive frames. Hence for a video clip having N frames, the total number of optical flows obtained is N-1. For a given video clip, the optical flow is represented with two displacements. They are horizontal displacement and vertical displacement. These two displacements are divided based on the assumption of constant brightness. For a constant brightness, the pixel intensity at t is equal to the pixel intensity at $t + \Delta t$, i.e.,

$$I(x, y, t) = I(x+, y + \Delta y, t + \Delta t) \quad (1)$$

The horizontal and vertical displacements are obtained by subjecting the Eq.(1) to partial differentiation with x, y and t. Consider $d_h^t(x, y)$ and $d_v^t(x, y)$ are the horizontal and vertical displacement, then the corresponding optical flow $O_F(x, y)$ is represented as

$$O_F(x, y) = [d_n^t(x, y) d_v^t(x, y)] \quad (2)$$

After computing optical flow vectors, we perform a segmentation operation and divide the entire optical flows into several groups. An aggregative optical flow is computed by integrating up all the OFs in that particular group. Such kind of accumulation is called temporal accumulation because the optical flows in each segment are differentiated with time. For segmentation, the size of the group needs to be

defined. Consider D to be the size of the segment, N be the total number of frames in the video clip. The total number of segments for aggregating optical flows is measured as

$$s = \frac{N-1}{D} \quad (3)$$

According to the Eq.(3), consider N = 10, the total number of optical flows obtained are N-1 = 10-1 = 9. These 9 optical flows are segmented into several groups. Consider the size of segment D = 3, then the total number of segments obtained are S = 9/3 = 3 means each segment have three optical flows, and the total number of such segments is 3. Here we compute aggregative optical flows for every group by integrating up the OFs in each group. The aggregative optical flows are of two types. They are horizontal and vertical. The horizontal aggregative optical flows are obtained by the accumulation of horizontal displacements, while the accumulation of vertical displacements obtains the vertical aggregative optical flows. In both cases, the aggregation is done as;

$$C_H^{s\in S} = \sum_{t=1}^{p} d_h^t(x,y) \quad (4)$$

And

$$C_V^{s\in S} = \sum_{t=1}^{p} d_v^t(x,y) \quad (5)$$

Where A_h^s and A_v^s are the Aggregative horizontal optical flows and Aggregative vertical OFs of segments *s* where *s* varies from 1 to S. D is the total count of OFs in every segment. Due to this kind of accumulation, the displacement is produced by noise shrinkages because they are inconsistent and random. On the other hand, due to this aggregation, the displacement caused by ME improve because they are consistent in direction between successive frames. During the aggregation process, the selection of group size is significant. A smaller number of optical flows in every group produces aggregated optical flows that have the displacements caused due to noise. On the other hand, a more significant number of optical flows in each group results in aggregated optical flows, including the movements of Heads, and produce a weight matrix with larger values.

C. Weights computation

Once AOFs are computed for each group, then they are again aggregated, and we consider the magnitudes of the final result for Weights Computation. Here the computation of the weights is done based on the block-based features. The AOFs are segmented into $W \times H$ non-overlapping blocks. Then we compute the motion intensity for each block by the summation of optical flows in the corresponding block. Consider the motion intensity of a block located at the i^{th} row and j^{th} column as $M_{i,j}$, it is calculated as

$$M_{i,j} = \sum_{s=1}^{S} \sum_{x=1}^{W} \sum_{y=1}^{H} \sqrt{C_H^s(x,y)^2 + C_V^s(x,y)^2} \quad (6)$$

Where *i* and *j* are the indices of blocks, *H* and *W* are the height and width of blocks, respectively, *s* is the index of AOF, and it varies from 1 to S, where S is the total number of groups. Finally, the weight matrix $W_{i,j}$ is computed by dividing the motion intensity of each block $W_{i,j}$ with the maximum motion intensity of all AOFs. The $W_{i,j}$ is calculated as follows;

$$W_{i,j} = \frac{M_{i,j}}{\max(M_{i,j})} \quad (7)$$

Where max(.)determines the maximum motion intensity in the entire set. The weights obtained here are used for the determination of optimal Histogram features of LAPTOP.

D. LBP-TOP

LBP is a fine-scale descriptor that can capture small texture details [37, 38]. LBP is a good feature descriptor since it encodes the fine details of facial appearance information over a range of coarse scales. Moreover, the LBP is much resistant to changes in intensity. LBP was discovered in the mid-1990s, and it was successfully implemented in different fields like face expression analysis, Dynamic texture analysis, human action recognition. LBP explores the superior construction of p pixels that are distributed consistently on a ring of radius *r* and have a centre pixel q_c at the centre position. Specifically for a centre pixel q_c having p neighbouring pixels that were correspondingly spread out on the ring of radius *r*, the LBP is computed as

$$LBP_{r,p}(q_c) = \sum_{n=0}^{p-1} s(q_{r,p,n} - q_c) 2^n \quad (8)$$

Where

$$s(x) = \begin{cases} 1, if\ x \geq 0 \\ 0, if\ x < 0 \end{cases} \quad (9)$$

Where $s(.)$ denotes a significant fraction. LBP is a grey scale-invariant, and it can encode significant local patterns like edges, blobs, and lines because it computes the difference between the centre pixel and the surrounding neighbour pixels. Even though LBP is an effective method for texture encoding, it is allowed only in the spatial domain and cannot be applied in the temporal domain. The temporal domain needs to be applied in every frame that is signified with time *t*. This process constitutes a substantial computational burden. To address these problems, an extended variant of LBP called LBP-TOP [15] is introduced in which the texture descriptor is formed by the concatenation of LBPs on three orthogonal planes such as XY, XT, and YT. LBP-TOP encodes the video by considering only co-occurrence statistics in three directions. Generally, a video sequence is a stack of XY planes in *T*axis, but it was ignored that the video sequence can be viewed as a stack of XT planes on the Y-axis and YT planes on the X-axis. The main advantage of this kind of representation is its ability to explore the space-time transition. LBP-TOP considers the distribution of features from every separate plane and then concatenates them together such that the obtained feature vector is very short, even for an increased number of neighbour points.Figure.2 shows the process of LBP-TOP computation.

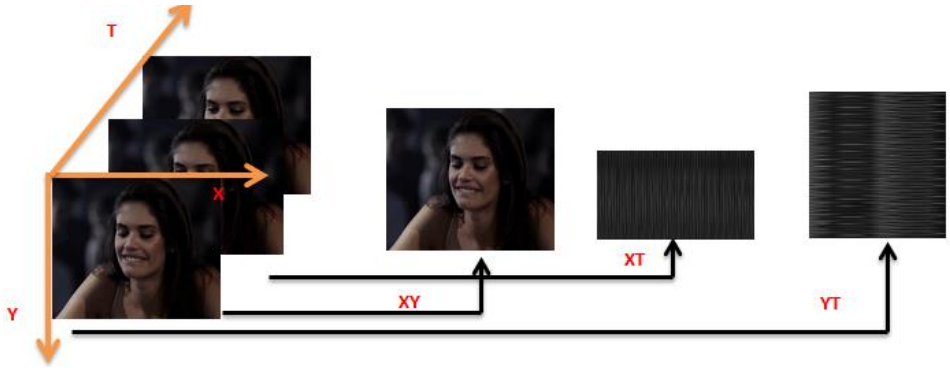

Figure.2 LBP-TOP: XY Plane, XT plane and YT plane

As shown in figure 2(a) the image is shown in the XY plane, (b) the image is shown in the XT plane, which shows a visual impression of one row changing in time and (c) the image is shown in YT plane which gives a visual impression of one column changing in time. FOR ALL PIXELS, the LBP code derived from XY, XT, and YT planes is denoted as LBP-XY, LBP-XT and LBP-YT. These three codes result in the statistics of three different planes, and they are concatenated into a single histogram. In such kind of representation, appearance and motions are encoded by LBP-XY, LBP-XT and LBP-YT. Among these three codes, the LBP-XY explores spatial domain information, and the remaining two (LBP-XT and LBP-YT) explore the temporal domain information. The figure shows the LBP Histogram from each plane and the corresponding concatenated histogram .

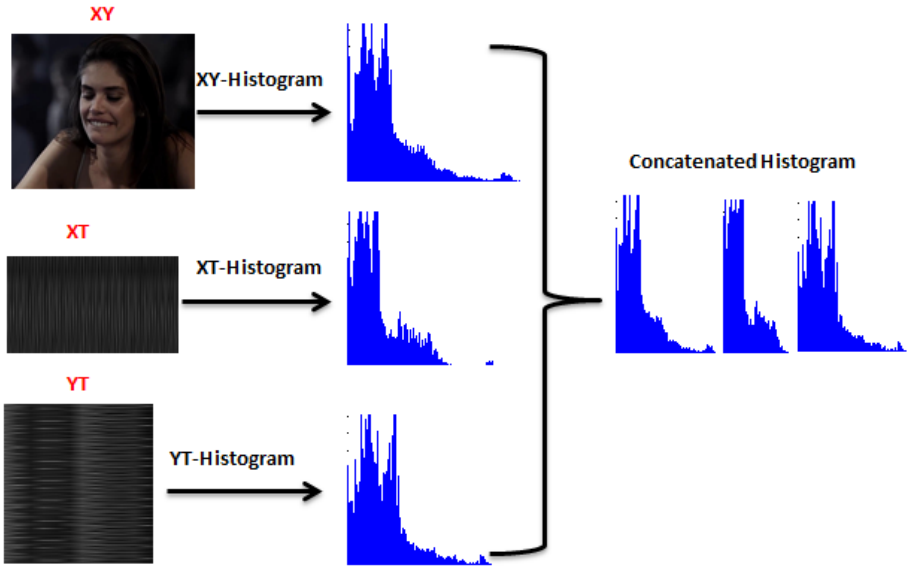

Figure.3 LBP-TOP Histogram

E. Composite LBP-TOP

Even though the LBP-TOP has gained much significance in providing sufficient Spatio-temporal variances, the information loss observed is more because the proportion of uniform patterns may be too small to acquire the variations. Next, a more extensive sampling size is much effective in providing more local information and better representation. However, as the sampling points increase, the LBP-TOP consequences in a huge dimensionality of a feature vector. These two problems lead us to derive a new variant of LBP-TOP. In this work, we explore the second-order information of a local patch in two aggregations: Scalable and orientational differences. Here we propose two compliments for LBP-TOP, and they are named SLBP-TOP and OLBP-TOP.

a) SLBPTOP

Here, the SLBP-TOP encodes the relation between first and second-order pixels of the centre pixel. In the conventional LBP-TOP, the centre pixel is encoded against the neighbour pixels only, but they don't consider the second-

order pixels, which means the pixels next to the neighbour pixel of the centre pixel are not considered. Unlike the conventional LBP-TOP, which encodes the information between the centre pixel and its neighbour pixels on the same circle (single scale), the SLBP considers the pixels on different circles (different scales) for encoding. In the proposed SLBP-TOP, we consider two radii, one is at 'r', and another is at' $r-\delta$'. For both of these two rings, the centre pixel is typical. In both rings, the SLBP-TOP considers only p number of pixels distributed exactly on two different rings.

To get the SLBP-TOP code, the difference between the pixels in the first and second rings is computed. Then the obtained differences are thresholded against 0. Based on the formal definition, the computation of SLBP is done as follows [39];

$$SLBP_{r,p,\delta}(q_c) = \sum_{n=0}^{P-1} s\left(q_{r,p,n} - q_{r-\delta,p,n}\right)2^n (10)$$

Where r and $r-\delta$ are the radii of outer and inner circles, respectively, Figure. 5 shows the simple process of RLBP computation.

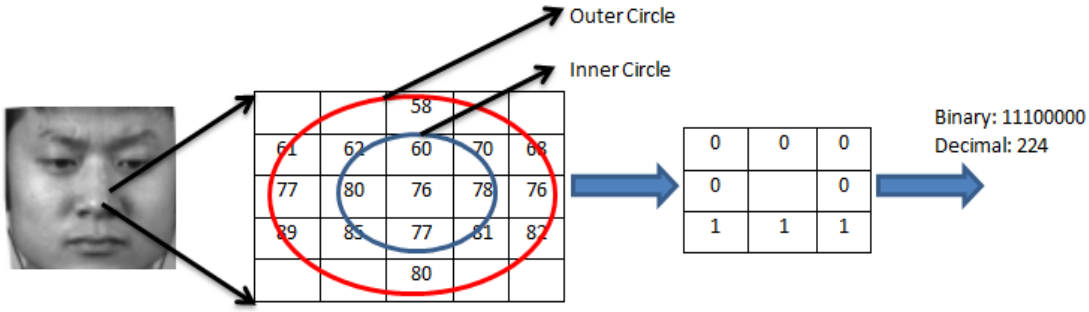

Figure.4 SLBP computation

b) OLBP-TOP

One drawback of LBP is its inability to encode the second-order information between pixels on the same ring. The conventional LBP only encodes the relation between the centre pixel and its neighbour pixels but not between the neighbour pixels themselves. If we consider such kind of second-order information, then it can reduce the ambiguity of noise addition. In general, the neighbour pixels are almost in correlation except when the noise is added. In the case of external disturbances like noises, colour artefacts, the pixel intensities of neighbour pixels have much deviated, and if we can analyse such discrimination, it provide more information about expression. Based on this inspiration, we propose a new variant of LBP called Orientational LBP-TOP, which encodes the relation between neighbour pixels at different orientations. The OLBP-TOP encodes the difference between neighbour pixels or compares them in a clockwise direction. The OLBPTOP computes the difference between neighbour pixels, and the obtained difference is thresholded against 0. This process is confined to only one ring. Here the difference between neighbour pixels is considered as second-order information. According to the theory of OLBP, the mathematical definition is given as [39]

$$OLBP_{r,p}(q_c) = \sum_{n=0}^{P-1} s\left(q_{r,p,n+1} - q_{r,p,n}\right)2^n (11)$$

Where *n* and *n+1* are the indices of two successive neighbour pixels, OLBP is more compact and provides more helpful information. Figure.5 shows an example of OLBP computation.

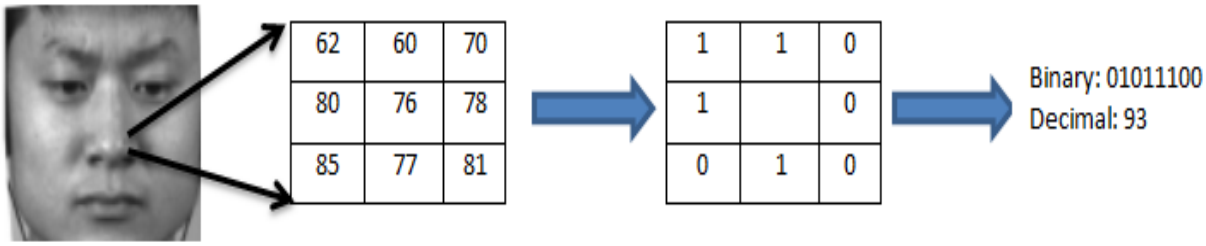

Figure.5 OLBP computation

Once the LBP, SLBP and OLBP are computed for XY planes, they are computed for the remaining two planes, such as XT-plane and YT-plane. After this process, they were subjected to histogram evaluation, and all three histograms were concatenated to form a final histogram of each block. Then every block histogram feature is multiplied with weight matrix (Eq.7) to obtain weighted histograms. After the feature extraction completion, we fed the obtained optimal histograms to the SVM algorithm for classification.

IV. SIMULATION RESULTS

In this section, we explore the details of simulation experiments conducted over the proposed model with the help of the CASME II micro Expression dataset. For simulation purposes, we used the MATLAB 2015 software with an image processing toolbox. Initially, we explain the details of the dataset, and then we explore the details of simulation results obtained. Finally, we also explore the comparison between the proposed and several existing methods; thereby, the performance effects can be proved.

A. Datasets

For the simulation purpose, here we used the most standard CASME II dataset. CASME II [16] is one of the most widely used databases, consisting of 247 ME video clips and they are acquired with the help of 26 subjects. The frames rate of each video clip is 200 frames per second. The total number of classes into which the database is divided is five; they are Repression (27 samples), Disgust (64 samples), Surprise (25 samples), Happiness (32 samples) and other (99 samples). The frame resolution is 640×480, while the resolution of each frame after cropping is reduced to 340×280. The CASME II has Action unit's labels following the facial action coding system.

B. Results

To explore the proposed method's effectiveness in recognising micro expression, we conducted a vast set of experiments by varying different parameters. First, we applied the three descriptors such as LBP-TOP, SLBP-TOP and OLBP-TOP as feature extraction methods. For all these methods, initially, the weights are computed based AOFs, and the obtained histograms after LBP descriptors are multiplied with weights to find the final histogram features. Once the final histograms are extracted, then they are fed to SVM classifier for expression recognition. Here for experimental validation, we employed Leave One Subject Out stagey in which among the available subjects, the video clips of one subject are used for testing and the remaining is used for training. In this way, we conduct five-fold cross-validation by interchanging the subjects used for training and testing. At every validation, we compute Recall, Precision, F1-Score and False Negative Rate. Based on these values, we further measure mean accuracy and weighted F1-Score. Since the F1-Score is the Harmonic mean of recall and precision, we compute those two metrics. The mean accuracy is measured by averaging the accuracies of all subjects.

Table.1 Results of CLBP-TOP with AOF over CASME II dataset

Emotion/Metric	Recall (%)	Precision (%)	F1-Score (%)	FNR (%)
Happy	76.9230	62.5230	68.9795	23.0769
Disgust	77.2385	90.9000	83.5142	22.7615
Surprise	86.3648	81.8181	84.0300	13.6352
Repression	53.7541	85.7142	66.0722	46.2459
Others	85.6647	81.3953	83.4754	14.3353

Table.2 Results of LBP-TOP with AOF over CASME II dataset

Emotion/Metric	Recall (%)	Precision (%)	F1-Score (%)	FNR (%)
Happy	73.6635	56.2530	63.7916	26.3365
Disgust	72.6258	85.7145	78.6292	27.3742
Surprise	82.4874	64.5441	72.4209	17.5126
Repression	49.5471	71.4352	58.5112	50.4529
Others	82.3397	73.8112	77.8425	17.6603

Table.3 Results of SLBP-TOP with AOF over CASME II dataset

Emotion/Metric	Recall (%)	Precision (%)	F1-Score (%)	FNR (%)
Happy	65.5421	47.0623	54.7858	34.4579
Disgust	65.2333	76.1998	70.2914	34.7667
Surprise	70.8964	53.8547	61.2116	29.1036
Repression	39.3658	66.6774	49.5045	60.6342
Others	74.3636	69.0547	71.6109	25.6364

Table.4 Results of OLBP-TOP with AOF over CASME II dataset

Emotion/Metric	Recall (%)	Precision (%)	F1-Score (%)	FNR (%)
Happy	69.4586	64.2935	66.7763	30.5414
Disgust	68.6639	80.9541	74.3042	31.3361
Surprise	77.8947	61.5447	68.7611	22.1053
Repression	45.6571	55.5696	50.1280	54.3429
Others	78.6638	71.4385	74.8773	21.3362

Table.1 shows the performance results computed after validating the proposed CLBP-TOP over the CASME II dataset. At this validation, we applied the complete feature extraction process to extract the features from micro expression video clips. The CLBP-TOP specifies composite LBP-TOP means the accumulation of all the features obtained through LBP-TOP, SLBP-TOP and OLBP-TOP. At this simulation, we found better results at the values of *r=2, p=8* and $\delta = 2$. For LBP-TOP and OLBP-TOP, the better performance is observed at *r=2, p=8*while for SLBP-TOP, the optimal performance is found at *r=2, p=8* and $\delta = 2$. The performance results at *r=2, p=8*for LBP-TOP and OLBP-TOP are demonstrated in Table.2 and Table.4, respectively, while the results of SLBP-TOP are demonstrated in Table.3. From these four tables, the performance weightage can be assigned as CLBP-TOP, LBP-TOP, OLBP-TOP and SLBP-TOP, in which the high weightage is assigned for the composite feature extraction method while the low priory is assigned for the Scalable feature extraction method.

Since the composite feature extraction method covers all kinds of features, it can provide more knowledge to recognise the variations in MEs. As the system can acquire more knowledge about MEs, it can recognise them more accurately. Due to this reason, the composite method is observed to have better recall, precision, F1-Score and less FNR. Next, the most minor performance is observed at the Scalable feature extraction method because, in SLBP-TOP, the encoding is done for the difference between pixels on multiple scales, but the centre pixel has less impact with the distant pixels when the block size is large. Next, the OLBP-TOP is also shown better performance but not the traditional LBP-TOP. Here the LBP-TOP has improved its performance due to the involvement of AOFs. Next, regarding expression, the higher performance (with respect to recall) is gained at two expressions, others and Surprise. The main reason is that the surprise emotion has an apparent muscle movement at the mouth because even n micro-expression, the subjects opened their mouth for a surprise emotion. Hence it is recognised more accurately than the other emotions. Next, the most minor performance is attained at repression emotion because the repression emotion is a relatively more minor range of muscle movements and hence more difficult to detect and classify.

The best accuracy and weighted F1-score obtained after the simulation of proposed descriptors over CASME II are demonstrated in Table.5. In the CASME II dataset, the maximum accuracy is observed at r=2 and p=8 for both LBP-TOP and OLBP-TOP, whereas the optimal performance for SLBP-TOP is observed at r=3, p=8 and $\delta = 2$. Since the SLBP is a scaling-based encoded, which needs a minimum of three scales, it has observed a superior performance at those parameters. In the accomplishment of CLBP-TOP, we have used these values only and gained maximum accuracy compared to all three individual methods. The maximum accuracy attained through the proposed method is observed as 75.9890, while for traditional LBP-TOP, it is only 72.1327, which means the proposed method have gained approximately 3% improvement in accuracy.

Table.5. Accuracy and Weighed F-Score for different methods over CASME II dataset

Method/Metric	Accuracy	Weighted F1-Score (%)	(r, p)
CLBP-TOP	75.9890	67.2143	-
LBP-TOP	72.1327	60.2391	**(2,8)**
SLBP-TOP	63.0802	51.4808	**(3,8,2)**
OLBP-TOP	68.0676	56.9694	**(2,8)**

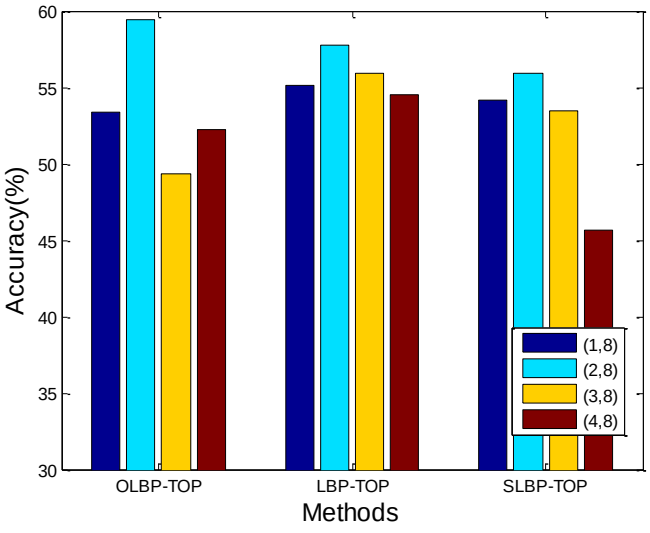

Figure.6 Accuracy attained through different methods at different parameters

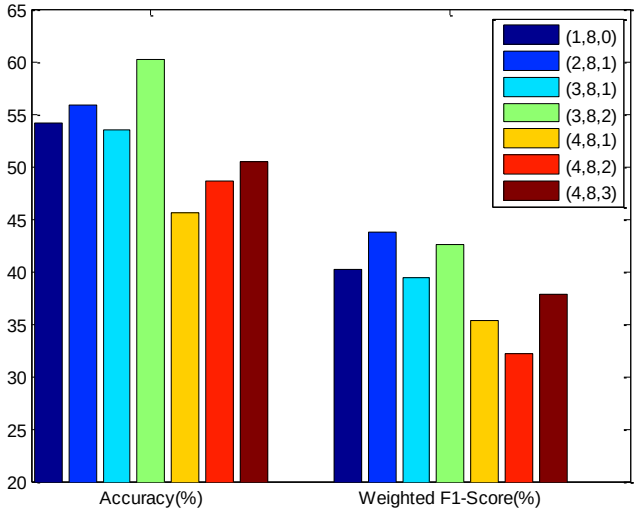

Figure.7 Performance of SLBP-TOP at different parameters

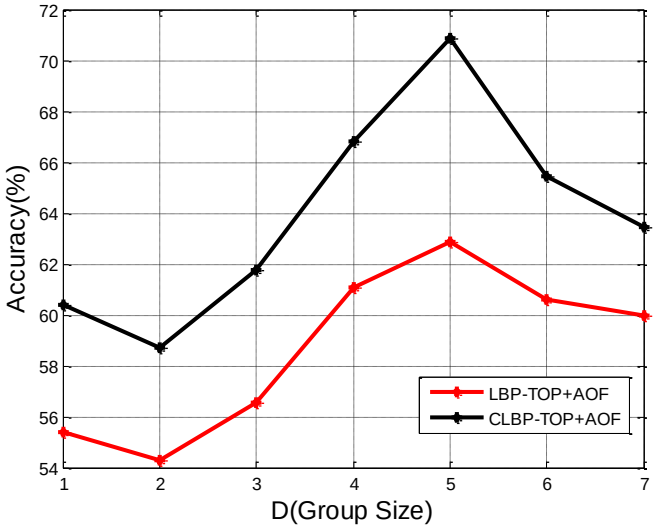

Figure.8 Accuracy at different group sizes

Figure.6 and Figure.7 demonstrates the performance of proposed descriptors at different parameters of LBP. Here the performance analysis of the proposed approach is done by varying *r* and δ values. Here the r-value is varied from 1 to 4, and δ is varied from 1 to 3. In all the simulation studies, we fixed the p values to 8. To encode the centre pixel for every simulation, we consider only 8 neighbour pixels even though the scale is increased. After this kind of simulation, we observed that the higher accuracy measured through OLBP-TOP and LBP-TOP is observed as *r=2* and *p=8*. Next, the maximum accuracy attained through SLBP-TOP is observed at *r=2*, *p=8*and $\delta = 2$. But the maximum weighted F1-Score of SLBP-TOP is observed at *r=2*, *p=8*and $\delta = 1$. From these two figures, we can see that the SLBP-TOP has gained less recognition accuracy when compared with other proposed methods.

Next, to alleviate the performance of the proposed AOF methodology, we conducted one more simulation study by varying the group size. Here *D* represents the group size, and it is varied from 1 to 7, and the observed accuracy results are shown in Figure.8. From this figure, we can see that the maximum accuracy is observed at group size 5 while the minimum accuracy is0bserved at group size 2. For group size, the accuracy observed is approximately 63.5523% which is more than the accuracy at group size 2. As the group size is 1, state that the methodology considers an entire set of frames for recognition and hence it has gained better recognition accuracy. As the group size increases, the accuracy also increases but up to a particular instance only. After group size 5, we can see that the accuracy is decreasing because the AOF may include the movements of the head and some other parts into the optical flow vector. These extra movements result in more false positives, and hence, the accuracy is less at more significant group sizes.

C. Comparison

Table.6 shows the comparison of the proposed approach with several existing methods. Since our main contributions are made over the optical flow and LBPs, we referred only to those methods which are considered these methods as base references. Most of the earlier methods employed LBP-TOP and Optical flows in the MER research. However, the maximum accuracy attained by them is noticed as 69.11% (attained at FMBH [41]). This approach adopted different Motion boundary histograms and fused them. However, the MBH is optimal for macro movements like human actions, and it has very much limited performance in MER. Histograms won't affect the analysis of textures which is the central aspect of the expression analysis. LBPs are a more efficient and optimal method for texture analysis, while the OFs are optimal methods for micro movement's analysis. In the context of LBP, the LBP-TOP based methods had shown a better performance. As Hang et al. [18] and Yandan Wang et al. [27, 29] applied only the extension of LBP, but the methods developed by A. K. Davisonet al. [42]applied LBP-TOP and achieved better accuracy. The one more variant of LBP, STLQP, considered a sign, magnitude and orientation based difference between pixels in the neighbourhood. However, they were not focused on the extraction of second-order information and the detection of local movements.

On the other hand, the optical flow method was not concentrated on texture analysis, and they only tried to extract the local movements between frames. For instance, the FHOFO encoded the temporal pattern with the help of the fuzzification of histograms of optical flow vectors. However, the analysis at spatial domain is missed.

Table.6 Accuracy Comparison

Author(s)	Features	Classifier	Dataset	Accuracy (%)
A. K. Davisonet al. [42]	LBP-TOP	SVM	CASME II	67.80
Hang et al. [18]	LBP-IP	SVM with Chi-Square kernel	CSAME II	59.51
Yandan Wang et al. [27]	LBP-SIP	SVM	CASME II	46.56
X. Huang et al. [30]	STLQP	SVM with Linear Kernel	CASE II	58.39
S. L. Happy and A. Routray [32]	PHOTO	KNN, SVM and LDA	CASE II	56.64
Lu et al. [41]	FMBH	SVM	CASME II	69.11
S. T. Liong et al. [33]	Bi-WOOF	SVM	CASE II	62.20
Yandan Wang et al. [29]	LBPMOP	SVM with RBF kernel	CASE II	44.13
Y. J. Liu et al. [31]	MDMO	SVM	CASE II	67.37
J. Wu et al. [35]	OFF2BD	SVM, LDA and 3-point KNN	CASME II	50.46
	OFF3WD	SVM, LDA and 3-point KNN	CASE II	51.68
Proposed	LAPTOP + AOF	SVM	CASE II	**72.66**
Proposed	SLBPTOP + AOF	SVM	CASE II	**68.45**
Proposed	LAPTOP + AOF	SVM	CASE II	**69.34**
Proposed	LAPTOP + AOF	SVM	CASE II	**74.68**

Next, S. T. Liong et al. [33] developed Bi-WOOF in which the weighted oriented optical flows are measured for the apex frame after its spotting from a microexpression video clip. However, they missed temporal information accumulation. Y. J. Liu et al. [31] proposed an extension to the OF called MDMO, which considers the mean and directionality of Optical Flow Vectors. This method attained a better performance compared to the all-optical flow methods. However, it lacks the spatial analysis of micro-expressions. J. Wu et al. [35] combined OFVs with LBP variants and tested the performance at several instances. Even though they computed weight matrix based on the accumulation of Optical flow vectors, they didn't contribute towards the LBP, and simply they used existing versions for validation. Compared to all the existing methods, the proposed composite method has gained a superior performance due to the following reasons; (1) considered the OFVs for the analysis of micro-movements and assigned a priority index (weight) for the blocks that have sufficient motion intensity, (2) considered the second-order information at pixel encoding through LBP.

V. CONCLUSION

In this paper, we proposed a simple and efficient micro expression descriptor called CLBP-TOP for MEs recognition. The proposed descriptor is the combined form of LBP-TOP, SLBP-TOP and OLBP-TOP. Along with CLBP-Top, we also proposed an AOF vector to assess micro-movements in a microexpression video clip. Based on the magnitudes of AOF vectors, a weight matrix is generated that explores the weightage of motion intensities in the ME video clip. Once the histograms are derived for CLBP-TOP, they are multiplied with a weight matrix, and weighted histograms are calculated, called final descriptors of a ME. Next, they are processed through the SVM algorithm for expression recognition. Simulation experiments are conducted over the CASME II dataset, and the performance is measured through accuracy and F1-Score. The obtained performance results have shown that pour proposed approach surpasses the state-of-art methods.

REFERENCES

[1] I. Cohen , N. Sebe , A. Garg , L.S. Chen , T.S. Huang , Facial expression recognition from video sequences: temporal and static modeling, Comput. Vision ImageUnderst. 91 (1) (2003) 160–187.

[2] E. Vural , M. Bartlett , G. Littleworth , M. Cetin , A. Cecil , J. Movellan , Discrimination of moderate and acute drowsiness based on spontaneous facial expressions, in: 20th International Conference on Pattern Recognition (ICPR), IEEE, (2010) 3874–3877 .

[3] J. Whitehill, M. Bartlett, J. Movellan, Automatic facial expression recognition for intelligent tutoring systems, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'08), (2008) 1–6 .

[4] T.A. Russell, E. Chu, M.L. Phillips, A pilot study to investigate the effectivenessof emotion recognition remediation in schizophrenia using the micro-expressiontraining tool, Br. J. Clin. Psychol. 45 (4) (2006) 579–583.

[5] X. Ben, P. Zhang, R. Yan, M. Yang, G. Ge, Gait recognition and micro-expressionrecognition based on maximum margin projection with tensor representation,Neural Comput. Appl. 27 (8) (2016) 2629–2646.

[6] L. Su, MD. Levine,High-stakes deception detection based on facial expressions, in: 22nd IEEE International Conference on Pattern Recognition (ICPR), (2014) 2519–2524.

[7] E.A. Haggard, K.S. Isaacs, Micro-momentary facial expressions as indicators of ego-mechanisms in psychotherapy, in: Methods of Research in Psychotherapy, Springer, (1966) 154–165.
[8] P. Ekman, W.V. Friesen, Non-verbal leakage and clues to deception, Psychiatry 32 (1) (1969) 88–106.
[9] P. Ekman, M. O'Sullivan, and M. G. Frank, A few can catch a liar,"Psychol. Sci., 10(3) (1999) 263_266.
[10] M. G. Frank, C. J. Maccario, and V. Govindaraju, Behavior and security, in Protecting Airline Passengers in the Age of Terrorism. Santa Barbara,CA, USA: ABC-CLIO, LLC, (2009) 86_106.
[11] M. O'Sullivan, M. G. Frank, C. M. Hurley, and J. Tiwana, Police liedetection accuracy: The effect of lie scenario, Law Hum. Behav., 33(6) (2009) 530.
[12] X. Li, X. Hong, A. Moilanen, X. Huang, T. P_ster, G. Zhao, andM. Pietikäinen, Towards reading hidden emotions: A comparative studyof spontaneous micro-expression spotting and recognition methods, IEEETrans. Affect. Comput., 9(4) 563_577, Oct./Dec. 2017.
[13] M. Frank, M. Herbasz, K. Sink, A. Keller, and C. Nolan, ``I see how youfeel: Training laypeople and professionals to recognise meeting emotions,"in Proc. Annu. Meeting Int. Commun. Assoc. Sheraton, New York, NY,USA, (2009).
[14] Yee-Hui Oh, John See, Anh Cat Le Ngo, Raphael C. W. Phan and Vishnu M. Baskaran, A Survey of Automatic Facial Micro-Expression Analysis: Databases, Methods, and Challenges, Frontiers in Psychology, 9 (2018) 11-21.
[15] G. Zhao and M. Pietikäinen, Dynamic texture recognition using localbinary patterns with an application to facial expressions, IEEE Trans.Pattern Anal. Mach. Intell., 29(6) (2007) 915_928.
[16] W. J. Yan, X. Li, S. J. Wang, G. Zhao, Y. J. Liu, Y. H. Chen, X. Fu, CASME II: An improved spontaneous micro-expression database and the baseline evaluation, PLoS One 9 (1) (2014) e86041.
[17] S. J. Wang, W. J. Yan, X. Li, G. Zhao, C. G. Zhou, X. Fu, M. Yang, J. Tao, Micro expression recognition using color spaces, IEEE Trans. Image Process. 24 (12) (2015) 6034–6047.
[18] X. Huang, S. J. Wang, G. Zhao, M. Piteikainen, Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, (2015) 1–9.
[19] Zeileis, Achim; Hornik, Kurt; Murrell, Paul (2009). Escaping RGBland: Selecting Colors for Statistical Graphics, Computational Statistics & Data Analysis. 53 (9) (2009)3259–3270.
[20] I. P. Adegun and V. Hima Bindu, Facial micro-expression recognition: A machine learning approach, Scientific African, 8 (2020) e00465
[21] Yanliang Zhang , Hanxiao Jiang, Xingwang Li, Bing Lu, Khaled M. Rabie, and Ateeq Ur Rehman, A New Framework Combining Local-RegionDivision and Feature Selection forMicro-Expressions Recognition", IEEE Access, 8 (2020) 94499-94509.
[22] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. (2005) 886-893.
[23] M. R. Guidera, A. E. Qadi, M. R. Lrit, and M. E. Hassouni, A novel method for image categorisation based on histogram oriented gradient and support vector machine, in Proc. Int. Conf. Electr. Inf. Technol. (ICEIT) (2017) 1-5.
[24] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietik¨ainen, Reading Hidden Emotions: Spontaneous Micro-expression Spotting and Recognition, arXiv:1511.00423v1 [cs.CV] 2 Nov (2015).
[25] K. Kira and L. A. Rendell, A Practical Approach to Feature Selection, in Proc. 9th Int. Workshop Mach. Learn. (ML), Dec. (1992) 249-256.
[26] Guo, Y.; Xue, C.; Wang, Y.; Yu, M., "Micro-expression recognition based on CBP-TOP feature with ELM, Optik2015, 126, 4446–4451.
[27] Wang, Y.; See, J.; Phan, W.; Oh, Y.H. LBP with Six Intersection Points: Reducing Redundant Information in LBP-TOP for Micro-expression Recognition. In Asian Conference on Computer Visio; Springer: Cham, Switzerland, (2014) 525–537.
[28] Lu G., Yang C., Yang W., Yan J., Micro-expression recognition based on LBP-TOP features, Journal of Nanjing Institute of Posts and Telecommunications, 37(6) (2017) 1-7
[29] Yandan Wang, John See, Raphael C. W. Phan, Yee-Hui Oh, Efficient Spatio-Temporal Local Binary Patterns for Spontaneous Facial Micro-Expression Recognition, PLoS ONE 10(5): e0124674.
[30] X. Huang, G. Zhao, X. Hong, W. Zheng, M. Pietikäinen, Spontaneous facial micro-expressionanalysis using spatiotemporal completed local quantised patterns,Neuro-computing 175 (2016) 564–578.
[31] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, X. Fu, A main directionalmean optical flow feature for spontaneous micro-expression recognition, IEEETrans. Affect. Comput. 7 (4) (2016) 299–310.
[32] S. Happy, A. Routray, Fuzzy histogram of optical flow orientations formicro-expression recognition, IEEE Trans. Affect. Comput. (2017).
[33] S.-T. Liong, J. See, K. Wong, R.C.-W. Phan, Less is more: Micro-expression recognition from video using apex frame, Signal Process., Image Commun. 62 (2018) 82–92.
[34] Benjamin Allaert, Isaac Ronald Ward, Ioan Marius Bilasco, ChaabaneDjerba and Mohammed Bennamoun, Optical Flow Techniques for Facial Expression analysis - a Practical Evaluation Study, arXiv:1904.11592v2 [cs.CV] 4 Jan 2021.
[35] Junjie Wu, Jianfeng Xu, Deyu Lin and Min Tu, Optical Flow Filtering-Based Micro-Expression Recognition Method, Electronics (2020), 9, 2056.
[36] M.J. Black, P. Anandan, The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields, Computer Vision and Image Understanding, 63 (1) (1996) 75–104.
[37] T. Ojala, M. Pietikäinen, and T. Mäenpää, Multi-resolution gray-scale and rotation invariant texture classi_cation with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell., 24(7) (2002) 971_987.
[38] T. Ojala, M. Pietikäinen, and D. Harwood, A comparative study of texture measures with classification based on featured distributions," Pattern Recognit., 29(1) (1996) 51_59.
[39] Chengyu Guo, Jingyun Liang, Geng Zhan, Zhong Liu, Matti Pietikäinen, and Li Liu, Extended Local Binary Patterns for Efficient and Robust Spontaneous Facial Micro-Expression Recognition, IEEE Access, 7(2019) 174517-174530.
[40] Petr Husa'k, Jan C˘ each, Jir˘´ı Matas, Spotting Facial Micro-Expressions In the Wild, 22nd Computer Vision Winter Workshop Nicole M. Artner, Ines Janusch, Walter G. Kropatsch (eds.) Retz, Austria, February (2017) 6–8.
[41] H. Lu, K. Kpalma, J. Ronsin, Motion descriptors for micro-expression recognition, Signal Process., Image Commun. 67 (2018) 108–117.
[42] A. K. Davison, W. Merghani, and M. H. Yap, Objective classes for micro-facial expression recognition," J. Imag., 4(10) (2018) 119.