VIDEO SEQUENCING BASED FACIAL EXPRESSION DETECTION WITH 3D LOCAL BINARY PATTERN VARIANTS

Kennedy Chengeta and Serestina Viriri

[1]University of KwaZulu Natal
[2]School of Computer Science and Mathematics,
Westville Campus, Durban, South Africa

ABSTRACT

Facial expression recognition in the field of computer vision and texture synthesis is in two forms namely static image analysis and dynamic video textures. The former involves 2D image texture synthesis and the latter dynamic textures where video sequences are extended into the temporal domain taking into account motion. The spatial domain texture involves image textures comparable to the actual texture and the dynamic texture synthesis involves videos which are given dynamic textures extended in a spatial or temporal domain. Facial actions cause local appearance changes over time, and thus dynamic texture descriptors should inherently be more suitable for facial action detection than their static variants. A video sequence is defined as a spatial temporal collection of texture in the temporal domain where dynamic features are extracted. The paper uses LBP-TOP which is a Local Binary Pattern variant to extract facial expression features from a sequence of video datasets. Gabor Filters are also applied to the feature extraction method. Volume Local Binary Patterns are then used to combine the texture, motion and appearance. A tracker was used to locate the facial image as a point in the deformation space. VLBP and LBP-TOP clearly outperformed the earlier approaches due to inclusion of local processing, robustness to monotonic gray-scale changes, and simple computation. The study used Facial Expressions and Emotions Database(FEED) and CK+ databases. The study for the LBP -TOP and LGBP-TOP achieved bettered percentage recognition rate compared to the static image local binary pattern with a set of 333 sequences from the Cohn–Kanade database.

KEYWORDS

Local binary patterns on Three Orthogonal Planes (LBPTOP) · Volume Local Binary Patterns(VLBP)

1. INTRODUCTION

Video based facial expression analysis has received prominent roles of late, in crowd analysis, security and border control among others[13]. Its also used in image retrieval, clinical research centers and social communication. Facial expressions remain the most effective way of emotion display. Previous work on 2D facial expression recognition has focused more into single image frame based analysis than video or image sequence analysis. The former assumes one image is

Dhinaharan Nagamalai et al. (Eds) : CSEIT, NCS, SPM, NeTCoM - 2018
pp. 137–149, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81811

representing the facial expression where as for the video image sequence, each facial image is a temporal dynamic process[14, 15, 11].

The study focuses on facial motions and locating key facial components namely, nose, eyes, face and mouth. The dynamic features are then extracted using key algorithms like local Gabor binary patterns from three orthogonal planes (LGBP-TOP) which is a LBP variant with Gabor filtering as well [14, 15, 11]. This feature descriptor (LBP from Three Orthogonal Planes LBP-TOP) is proposed to extract dynamic textures from video sequences to characterize facial appearance changes [14, 15, 11]. The facial expression video image sequences are then modeled as a histogram sequence which is a sum of concatenated local facial regions[1]. Support Vector Machines and KNN algorithms are used to classify the datasets. The experiments used the extended Cohn-Kanade (CK+) database and the FEED database. The facial expression sequence is modeled as a histogram sequence by concatenating the histogram pieces of all the local regions of all the LGBP-TOP maps. For recognition, support vector machine (SVM) is exploited. The experimental results on the extended Cohn-Kanade database (CK+) demonstrate that the proposed method has achieved the best results compared to other methods in recent years.

Video-based face recognition system typically consists of face detection, tracking and recognition[16, 10, 6, 17]. The video sequences picked up depict key universal expressions (surprise, sadness, joy, disgust and anger). Each signal expression is performed by 7 different subjects beginning from the neutral expression. This paper mainly focuses on the integration of spatial-temporal motion LBP with Gabor multi-orientation fusion and compares the 3 LBP histograms on three orthogonal planes to accuracy of face expression recognition[1, 14]. An ensemble voting classifier is used for each plane and the overall LBP-TOP algorithm. The LBP-TOP is also compared against its variants like LBP-MOP [1, 14]. Experiments conducted on the extended Cohn-Kanade (CK+) database and FEED database show that our approach is robust in dealing with video-based facial expression recognition problems compared better than the 2D image texture local binary pattern variants.

2. LITERATURE REVIEW

In recent researches use of spatio-temporal representations has been successfully used to address limitations of static image analysis[8, 3–5, 2]. Successful research has been done by fusing PCA, Gabor Wavelets, local binary patterns permanent features like eyes ,moth or lips' feature vectors were generated from the facial appearances in the spatial and the frequency domains. And local directional patterns. Classification has included support vector machines(SVM), Adaboost, k-nearest neighbor and neural networks[7, 8, 14]. The permanent features like eyes ,moth or lips' feature vectors were generated from the facial appearances in the spatial and the frequency domains.

2.1. Static Facial Expression Analysis Background

Clinical research has been widely studied with 2D images either as a combination of facial expressions in 2D images or universal global facial [7, 8, 3, 5] The Facial Action Coding System (FACS) has been developed to describe facial expressions using a combination of action units (AU)[7, 8, 3, 5]. Each action unit corresponds to a specific muscular activity that produces momentary changes in facial appearance. The global facial expression handles the expressions as a whole without breaking up into AUs. The most commonly studied universal expressions include

happiness, sadness, anger and fear, which are referred to as universal emotions. While most of the work has been on static 2D images, the Facial Expression Coding System (FACES) has been designed to analyze videos of facial expressions, in terms of the duration, content and valence of universal expressions [7, 8, 3, 5]. However, these methods need intensive human intervention to rate the images and videos of facial expressions. Such rating methods are prone to subjective errors, and have difficulties in providing unified quantitative measurements. There is need for automated, objective and quantitative measurements of facial expressions.

Challenges with static 2D images Based Methods Major clinical research in facial expression analysis includes subjective and qualitative scenarios in the 2D image family[8, 3]. The 2D static images lack temporary dynamics[14, 4, 13, 15, 16, 18]. They also are prone to subjectivity and poor qualitative features. The 2D static images do not capture temporary dynamics and expression changes [14, 4, 13, 15, 16, 18]. Therefore, there was need for automated, objective and quantitative measurements of facial expressions captured using videos. In this paper, we present a computational framework that uses videos for the analysis of facial expression changes. This framework explores the dynamic information that is not captured by static images during emotion processing, and provides computationally robust results [14, 4, 13, 15, 16, 18]. The study's chosen framework includes the video face detection and tracking incorporating shape variability. Based on tracking results, features are extracted from faces and then weighted facial expression classifiers applied on the given histograms[18, 1].

2.2. Facial Recognition With Video Image Sequences

Temporal information has capability to improve static image classification. In the work of Yacoob et al. , each facial expression is divided into three segments: the beginning, the apex and the ending [18, 14, 15, 2, 1]. Rules are defined to determine the temporal model of facial expressions. Such rules are ad-hoc and cannot be generalized to complex environments. In the work of Cohen et al., facial expressions are represented in terms of magnitudes of predefined facial motions, termed Motion-Units (MU) [18, 1]. A Tree-Augmented-Naive Bayes classifier is successfully applied to recognize facial expressions on static images, and then a multi-level Hidden Markov Model (HMM) structure is applied to recognize video sequences based facial expressions [14, 16]. Yeasin et al. applied a two stage approach to classify images in 3D by measuring the video intensity as we; using optical flow [1, ?,16]. Several probabilistic methods like particle filtering and condensation can also track facial expression in video sequences [18, 1]. Separate manifold substances have also been applied in video based facial expression analysis. To track video sequences models like 3d wireframe models, facial mesh models, net models and ASM models were successfully used. Videos subtle changes of facial expression can be measured on video facial expression recognition than on static image analysis[14, 16, 1, 10, 12].

3. LOCAL BASED FACIAL EXPRESSION FEATURE EXTRACTION

Facial expression analysis influences wide areas in human computer interaction. Local binary patterns and their wide 2D and 3D variants have been used in this field. Holistic and local based feature extractors have been used successfully. PCA are prominent holistic algorithms and local binary patterns, Gabor filters and Gabor wavelets and local directional patterns have been successfully applied as local feature extractors.

3.1. Local Binary Patterns (LBP) For Static Image Feature Extraction

Local binary patterns are based on facial images being split into local sub regions. The challenges of facial occlusion and rigidness are faced though grey scale image conversion is used to reduce illumination[8, 7, 3]. Local binary patterns are invariant to grey level images. Localized feature vectors derived are then used to form the histogram which is used by machine learning classifiers or deep learning methods. The local features are position dependent [8, 7, 3]. For local binary patterns, the facial region is divided into small blockers like mouth, eyes, ears, nose and forehead[4]. The basic local binary pattern non center pixels use the central pixel as the threshold

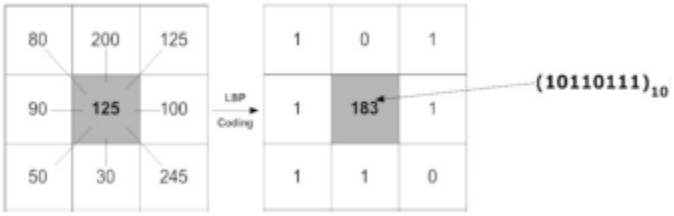

Fig. 1. Local Binary Patterns (LBP)

value taking binary values [8, 7, 3]. Uniform binary patterns are characterized by a uniformity measure corresponding to the bitwise transition changes. The local binary pattern has 256 texture patterns. The local binary LBP r,n operator is represented mathematically as follows The LBP feature for a local neighborhood of radius r, with n number of neighbor pixels is defined as:

$$LBP_{(n,r)} = \sum_{n-1}^{n=0} s(p_n - p_c)2^n. \qquad (1)$$

The neighborhood is depicted as an m-bit binary string leading to n unique values for the local binary pattern code. The grey level is represented by 2n-bin distinct codes. The value pc is the grayscale value of the center pixel, pn is the gray scale value of a neighbor pixel

LBP Variants Various LBP variants were successfully proposed and used. These include TLBP for Ternary Local Binary Pattern as well as Central Symmetric Local Binary Patterns [8, 9, 4]. Over-Complete Local Binary Patterns (OCLBP) is another key variant that takes into overlapping into adjacent image blocks. The rotation invariant LBP is designed to remove the effect of rotation by shifting the binary structure [8, 4]. Other variants include the monogenic and central symmetric (MCS-LBP).

Local directional patterns For local directional patterns or LDP a key edge detection local feature extractor, the images were divided into LDPx histograms, retrieved and then combined into one descriptor[9, 3, 5, 1, 11].

$$LDP_x(\sigma) = \sum_{K}^{r=0} \sum_{L}^{r=0} f(LDP_q(o,u),\sigma). \qquad (2)$$

The local directionary pattern, includes edge detection using the kirsch algorithm.

For video sequencing facial image analysis Volume Local Directional Binary Pattern (VLDBP) and Local Gabor Binary Patterns from Three Orthogonal Planes have been successfully been used.

$$\begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix} \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix} \begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} \begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix}$$

$$\begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix} \begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix} \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix} \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix}$$

Fig. 2. Local Directional Patterns (LDP)

Volume Local Directional Binary pattern (VLDBP) Volume Local Directional Binary pattern (VLDBP) is used as an extension of LBP in the dynamic texture field. Dynamic texture extends the temporal domain and is used in video image analysis. The face regions of the video sequence images are modeled with VLDBP which incorporates movement and appearance together. It uses three parallel planes and the middle plan contains the center pixel to derive the localized binary patterns. VLBP will also consider co-occurrence of all neighboring points from the 3 planes and generate binary representative codes. The extraction considers a local volumetric neighborhood against each pixel. The grey levels of the central pixels and the surrounding pixels are then compared against each other.

$$VLBP_{LPR} = \sum_{q=0}^{3P+1} v_q 2^q \qquad (3)$$

3.2. Local Gabor Binary Patterns from Three Orthogonal Planes

3D dynamic texture recognition which concatenates three histograms from LBP on three orthogonal planes was proposed. The three orthogonal planes namely XY, XT, and YT have been widely [1, 14, 17]. LBP-TOP extracts features from the local neighborhoods over the 3 planes. The spatial-temporal information can be regarded as a set of volumes in the (X, Y, T) space, where X and Y represent the spatial coordinates, while T denotes the frame index (time) in temporal domain [1, 14, 15]. The neighborhood of each pixel no longer falls in a two dimensional space, where LBP operation can be used to extract features into histograms. Instead, we need to compute feature descriptor in the three dimensional space (X, Y, T). LBP-TOP was proposed to describe the spatial-temporal information in the three dimensional space. LBP-TOP computes the local binary patterns of a center pixel through thresh holding the neighboring pixels [1, 14, 17]. The algorithm decomposes the 3 dimensional volume into 3 orthogonal planes namely XY, XT and YT[1, 2]. The XY plane indicates appearances features in the spatial domain. The XT is a visual representation of a row with respect to time. The YT represents the features of motion for a column in the temporal space domain.

The spatial plane XY is similar to the regular LBP in static image analysis. The vertical spatio-temporal YT plane and horizontal XT plane are the other 2 planes in the 3 dimensional space. The resulting descriptor enables encoding of spatio-temporal information in video images. The

performance and accuracy of the latter was also comparable to the LBP-TOP. The LBP STCLQP or spatio-temporal completed local quantized patterns (STCLQP) was also used to consider the pixel sign, orientation and size or magnitude. Local Gabor Binary Patterns from Three Orthogonal Planes (LGBPTOP) add Gabor Filtering to improve accuracy. With the added filtering algorithm rotational misalignment of consecutive facial images is mitigated[1, 16, 10]. To avoid LBP-TOP statistical instability, a re- parametrization technique whose foundation is second local Gaussian jet was proposed [1, 14, 17].

$$k = (H_L, X_Y, H_L, XT, H_L, Y_T, H_C, X_Y, HC, XT, HC, YT) \quad (4)$$

(LBP/C)T OP feature is denoted in vector form where Hv, m (v= LBP or C, and m = XY,XT,YT where m = XY,XT,YT) are the 6 LBP sub-histograms which contrast feature in the three orthogonal planes [1, 14, 17]. The LBP-TOP algorithm describes video sequence changes in both spatial and temporal domains hence captures structural information of the former domain and longitudinal data of the latter [16, 18, 1]. LBP histogram features encode spatial data in the XY plane and the histograms from the XT and YT planes include the temporal and spatial data. With facial actions causing local and expression changes over time, the dynamic descriptors have an edge in facial expression analysis over the static descriptors [16, 18, 1].

$$H_i, j = \sum x, y, t f_j(x, y, t) = i \quad (5)$$

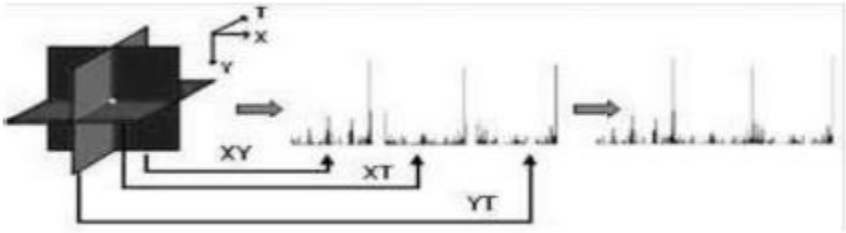

Fig. 3. LBP from three orthogonal planes. Three planes intersecting one pixel. LBP histogram of each plane and concatenating the histograms [1].

Likewise, the contrasts in the three orthogonal planes are also computed, which are denoted as Cm (m= XY, XT and YT) [1]. These contrast values are then represented as three sub-histograms Hx ,y (x= C and y= XY, XT , YT) [16, 18, 1]. Because of the contract, Cy is used to refer to the 3 features in all three orthogonal planes, and is called LBP CT-OP variant. Image device quality of the facial expression videos also impacts frame rates and spatial resolution quality as well [16, 18, 1].

Six Intersection Points (SIP) The LBP-SIP or Local Binary Pattern— Six Interception Points (LBP-SIP) considered 6 unique points along the intersecting lines of the 3 orthogonal planes to derive the binary pattern histograms [16, 18, 1].

$$AB, DF, EG = L_A \cap L_B \cap L_C \quad (6)$$

where AB,DF and EG are intersection points. 6 unique neighbor points carry sufficient information to describe the spatio-temporal textures centered upon point C[16, 18, 1]. LBP-SIP produces a compact set of features in high dimensional features spaces where there is sparse data [16, 18, 1].

LBP-Three Mean Orthogonal Planes (MOP) LBP-MOP or mean orthogonal plane is another variant to have been successfully used by concatenating mean images from image stacks derived along the 3 orthogonal planes[16, 18, 1]. It also preserves essential image patterns and reduces redundancy which affects encoded features

4. FACIAL EXPRESSION IMPLEMENTATION

The implementation involves analyzing video streams to track facial feature points over time. The feature vectors are then calculated and emotions detected from the trained models. Training and classification of the models is done using the popular algorithms namely support vector machines, k-nearest neighbor and neural network machine learning classifiers. Recognition of the model and new expressions on new images is then done on the selected annotated databases which includes CK+ database and FEED database. The section describes the approach, databases selected and then classification algorithm chosen and implemented.

4.1. Approach

The study's objective was to recognize facial expressions from video sequences. The approach involved locating and tracking the faces and expressions during the video segmentation and sequential modeling phase. The video sequence detection involved landmark detection and tracking, which define the facial shapes[15]. Viola Jones openCV detection tools are used. The features were then extracted using various 3D video feature extraction variants of the LBP-TOP algorithm. Gabor filters [16, 18, 1, 12, 5] were then applied during preprocessing. Geometric features were normalized and they were immune from skin color and illumination changes. Several machine learning classifiers namely support vector machines, k-nearest neighbor and neural networks were used for the classification. The algorithm used is shown in Algorithm 1.The study analyzed a sequence of frames that change from one form to another to detect faces from a live video based on the CK+ dataset and the FEED dataset [16, 18, 1].

Data: Copy and preprocess video image datasets
Result: Facial expression classification results for the image datasets
while For each image I inside the CK+ and FEED database do
1. divide the database into training and test sets;
2. for each image inside the given datasets;
3. apply Viola Jones algorithm for extraction and preprocess the image
4. using Principal Component Analysis;
5. extract the LBP-TOP, LBP-XY,LBP-XT and LBP-YT features;
6. extract the features using the LBP-MOP, LBP-SIP algorithm
7. apply Gabor Filters to get the LGBP-TOP and LGBP-MOP features
8. calculate the Euclidian distance matrix;
9. apply the classification on each with different classifiers;
10. the best classification results is then labeled the best algorithm;

End For'
end

Algorithm 1: Local Gabor Binary Patterns from Three Orthogonal Planes to analyse video sequences[20, 1]

4.2. Facial Expression Preprocessing

The first step of establishing the PCA classifier was to determine parameters such as the number of principal components to consider (PCs) and the number of training images [20, 1]. Gabor filters (a linear filter) were then used to detect edges in texture analysis. In the spatial domain. A given Gabor filter acts like a Gaussian kernel function modulated by a sinusoidal plane wave as shown in the equation below [20, 16, 18, 1] . The Gabor filters extract expression-invariant features.

$$G\_c[i,j] = Be^{-\frac{(i^2+j^2)}{2\sigma^2}} \cos(2\pi f(i\cos\theta + j\sin\theta)); \quad (7)$$

$$G\_s[i,j] = Ce^{-\frac{(i^2+j^2)}{2\sigma^2}} \sin(2\pi f(i\cos\theta + j\sin\theta)); \quad (8)$$

where B and C are normalizing factors that will be derived[20].

4.3. Facial Expression Databases

The study used video sequences lasting around 10 seconds with 15 second frames per second. The facial expressions are dynamic and evolve over time from the start, when reaching the apex and offsets. The video sequence datasets that could have been used included the CK+, YouTube Faces Database, Acted Facial Expressions in the Wild (AFEW) as well as the BU-3DFE, MMI+ dataset and the Facial Expressions and Emotions Database (FEED) [16, 18, 1]. The FEED dataset included 400 webcam video extracts from 18 voluntary participants in mpg format of sizes 480 times 640. There were labeled as 6 facial expression classes [16, 18, 1]. YouTube Faces Database data set contains over 3 000 videos from about a thousand and five hundred video sequence images. The videos average around 2 to 3 seconds and clips frame sizes from 48 to 6000 frames with a mean of 180. The BU-4DFE database has 101 subjects for identifying the emotion and it also has 83 feature points to recognize the emotion. In the total 101 subjects, 58 subjects are female and remaining 43 subjects are male. The study chose the FEED and CK+ dataset for implementation [16, 18, 1]. For static image analysis the study used the CK+ dataset and Google set dataset. The static image analysis was then compared to the video sequence databases.

CK+ dataset The CK+ dataset includes 593 video sequences and 7 expression types from 123 participants. The participants included African-Americans and euro- Americans and other races accounted for 6 percent [16, 18, 1]. The video sequences were 640 by 490 by 640 by 480 pixels. The grey images made with 8-bit precision made up the frames dataset[13]. The study used 90 participants and considered the 6 expressions namely anger, disgust, fear, happiness, sadness, and surprise[9].

4.4. Facial Expression Video Sequences Classification

The study uses the k-nearest neighbor, random forest, neural networks and support vector machines[19, 8, 5]. For a KNN machine learning classifier *k* NN, the nearest neighbor, given x_q, with *k* nearest discreet neighbors, will take a mean of *f* values of *k* nearest neighbors[19, 9, 12, 15].

$$kNN = f^{(}x_q)\frac{\sum_{i=1}^{k} f(x_i)}{k} \quad (9)$$

Support Vector Machine Support vector machines consider the that points close the given class boundaries[10]. A hyperplane is chosen to separate 2 classes which are initially given as linearly separable. The hyperplane separating the two classes is represented by the given equation[19, 9, 12, 15]:

$$w^T x_n + b = 0, \quad (10)$$

such that:

$$w^T, x_n + b1 \qquad y_n = +1, \quad (11)$$

5. EXPERIMENT AND RESULTS

Static image analysis experimental results For the 2D experiments the CK+ dataset was tested against an ensemble of classifiers as well as major local binary and directional patterns. The highest results were experienced when local binary patterns and binary patterns were applied with an ensemble of classifiers. Whilst the 2D classification results showed greater accuracy they lacked the 3D and dynamic spatial properties. The best classification was found on a combined LBP+ELBP and Gabor Filters combination with a 16, 2 radius combination that resulted in a classification rate of 99.15 percent for the ensemble voting classifier. The voting classifier had support vector machines, random forests and k-nearest neighbour with a ratio of 2:4:2 respectively.

GoogleSet Data	kNN+	Support Vector Machine	RF	Voting Classi-fier	Ave Time(s)	CK+ Data	kNN+	Support Vector Machine	RF	Voting-Classifier
$LGBP_{8,2}$	92.29%	96.23%	95.32%	97.36%	53.41s	$LGBP_{8,2}$	94.73%	96.24%	94.96%	95.67%
$LGBP_{16,2}$	91.45%	96.64%	97.13%	94.11%	45.43s	$LGBP_{16,2}$	92.67%	94.45%	93.31%	95.13%
$CS\text{-}LGBP_{8,2}$	89.31%	97.48%	97.13%	98.26%	52.97s	$CS\text{-}LGBP_{8,2}$	91.22%	95.32%	96.09%	96.96%
$CS\text{-}LGBP_{16,2}$	91.45%	94.92%	93.08%	97.31%	54.89s	$CS\text{-}LGBP_{16,2}$	88.45%	95.52%	95.88%	98.21
$ELGBP_{16,2}$	91.56%	93.42%	98.09%	97.19%	54.99s	$ELGBP_{8,2}$	88.65%	84.41%	96.06%	94.27%
$ELGBP_{16,2}$	87.89%	88.12%	94.84%	96.09%	51.09s	$ELGBP_{16,2}$	86.01%	85.65%	96.5%	96.1%
$LGTP_{16,2}$	87.93%	96.74 %	97.34%	97.24%	53.12s	$LGTP_{16,2}$	85.97%	95.44 %	97.43%	96.13%
$RLGBP_{8,2}$	86.01%	96.91 %	95.98%	96.98%	53.22s	$RLGBP_{8,2}$	85.21%	94.96 %	94.68%	97.81%
$RLGBP_{16,2}$	89.93%	96.21 %	93.09%	92.64%	52.11s	$RLGBP_{16,2}$	88.83%	96.81 %	95.67%	98.26%
$LDP{+}ELGBP_{8,2}$	94.21%	96.61%	97.12%	98.13%	52.33s	$LDP{+}ELGBP_{8,2}$	93.61%	95.62%	94.88%	98.03%
$LDP{+}ELGBP_{16,2}$	94.85%	97.97%	96.32%	99.26%	54.16s	LDP+ $ELGBP_{16,2}$	94.21%	97.85%	97.61%	99.15%

Fig. 4. 2D static image classifier for CK+ and GoogleSet combined Dataset with Gabon Filters applied

5.1 Experimental Results on CK+ and FEED 3D Datasets

Three experiment types were executed on the CK+ dataset's facial motions. Recognition rates were executed for the LBP-XY plane, LBP-XT plane as well as the LBP-YT planes. The combined recognition rate for the LBP-TOP was also calculated.

The 4 scenarios used an ensemble classifier of support vector machines, k-nearest neighbor as well as random forest classifiers with different weighted ratios. The following tables indicate classification of the CK+ and FEED datasets based on the XY, YT and XT plane dimensions with an average length of 0.9 seconds. The minimum length was 0.78 seconds and the highest length was 0.934 seconds.

%	XY Plane	XT Plane	YT Plane	Weight	Accuracy	length
$LBP\text{-}TOP_{8,8,8,1,1,1}$	0.969	0.972	0.966	4;3;1	0.975	0.932
$LBP\text{-}MOP_{8,8,8,1,1,1}$	0.974	0.969	0.965	4;2;3	0.977	0.933
$LBP\text{-}SIP_{8,8,8,1,1,1}$	0.973	0.976	0.971	3:2:5	0.976	0.912
$LBP\text{-}TOP_{4,4,4,1,1,1}$	0.976	0.977	0.966	6:2:3	0.978	0.9310
$LBP\text{-}TOP_{2,2,2,1,1,1}$	0.978	0.981	0.972	3:2:1	0.983	0.885
$LBP\text{-}TOP_{8,8,8,3,3,3}$	0.976	0.984	0.976	4:2:1	0.982	0.897
$LGBP\text{-}TOP_{8,8,8,1,1,1}$	0.973	0.976	0.969	2:4:2	0.979	0.930
$LGBP\text{-}MOP_{8,8,8,1,1,1}$	0.977	0.976	0.970	3:4:1	0.982	0.915
$LGBP\text{-}SIP_{8,8,8,1,1,1}$	0.979	0.979	0.976	2:3:5	0.979	0.930
$LGBP\text{-}TOP_{4,4,4,1,1,1}$	0.979	0.978	0.869	3:6:2	0.983	0.932
$LGBP\text{-}TOP_{2,2,2,1,1,1}$	0.981	0.988	0.977	2:3:1	0.984	0.933
$LGBP\text{-}TOP_{8,8,8,3,3,3}$	0.993	0.994	0.984	3:6:1	0.989	0.934

Table. 1. Video Sequence Classification on CK+ Dataset with 593 video image sequences

For the CK+ dataset the combined LGBP-TOP with Gabor Filtering and an ensemble of voting classifier combination of support vector machines, k-nearest neighbor and random forest achieved a higher accuracy of 98.9 percent from a sequence of 593 video sequences. For the FEED database with 400 video image sequences the corresponding accuracy was 99.5 percent.

	XY Plane	XT Plane	YT Plane	Weight	Accuracy	length
$LBP\text{-}TOP_{8,8,8,1,1,1}$	0.973	0.962	0.969	4;2;4	0.977	0.90
$LBP\text{-}MOP_{8,8,8,1,1,1}$	0.967	0.974	0.979	4;3;3	0.975	0.90
$LBP\text{-}SIP_{8,8,8,1,1,1}$	0.963	0.973	0.976	5;4;1	0.979	0.90
$LBP\text{-}TOP_{4,4,4,1,1,1}$	0.975	0.979	0.982	4;2;4	0.985	0.78
$LBP\text{-}TOP_{2,2,2,1,1,1}$	0.978	0.982	0.985	3;5;2	0.989	0.89
$LBP\text{-}TOP_{8,8,8,3,3,3}$	0.981	0.983	0.986	4;4;1	0.991	0.90
$LGBP\text{-}TOP_{8,8,8,1,1,1}$	0.977	0.966	0.971	2;5;3	0.983	0.90
$LGBP\text{-}MOP_{8,8,8,1,1,1}$	0.972	0.976	0.983	5;1;4	0.984	0.90
$LGBP\text{-}SIP_{8,8,8,1,1,1}$	0.968	0.975	0.979	6;2;2	0.982	0.90
$LGBP\text{-}TOP_{4,4,4,1,1,1}$	0.979	0.985	0.983	3;5;2	0.989	0.90
$LGBP\text{-}TOP_{2,2,2,1,1,1}$	0.981	0.986	0.986	6;1;3	0.993	0.90
$LGBP\text{-}TOP_{8,8,8,3,3,3}$	0.986	0.987	0.991	4;3;3	0.995	0.90

Table. 2. Video Sequence Classification on Facial Expressions and Emotions Database(FEED) Dataset

The combined feature extractor LBP-TOP achieved higher classification rates as compared to the specific dimension LBP-XT, LBP-XT and LBP-YT accuracy rates. Better variation was experienced for the LBP-XT based plane. The second experiments evaluated the efficiency of using Gabor Filters to enable multi-orientation fusion to the spatial temporal advantages of the LBP-TOP algorithm. Support vector machines, k– nearest neighbor and random forest ensemble classifier was also used in this scenario The combined classifier with Gabor-Filters and LBP-TOP feature extractor showed greater accuracy to the normal LBP-TOP algorithm. The other LBP-TOP variants like SIP and MOP also achieved greated accuracy but the LGBP-TOP with parameters of 8,3 on each dimension achieved better accuracy to all the LGBP-TOP variants.

5.2. Video Sequence Confusion Matrices

The confusion matrix obtained from the video datasets for showed an overall success of 99.51 percent and 99.1% when Gabor Filtered on the CK+ and FEED database respectively. The following two confusion matrices give detail of the precision recall accuracy for the CK+ dataset which included 593 video datasets with video lengths of less than 1 second.

	precision	recall	f1-score	support
anger	0.978	1	0.989	115
disgust	0.976	0.985	0.993	62
fear	0.998	0.983	0.984	124
happy	0.983	0.978	0.982	109
neutral	1	0.992	1	93
sadness	0.991	0.983	0.99	72
avg/total	0.992	0.995	0.995	**593**

	Confusion Matrix						
anger	[[**111,**	0,	0,	2,	1,	1],	
disgust	[	0,	**60,**	1,	1,	0,	0],
fear	[	3,	0,	**120,**	1,	1,	0],
happy	[	2,	2,	2,	**103,**	0,	0],
neutral	[	1,	0,	1,	1,	**90,**	0],
sadness	[	0,	0,	0,	1,	1,	**70**]]

Fig. 5. LBP-TOP , CK+ Dataset Facial Expression Recognition dataset from 593 video sequences

The FEED Dataset had 400 video sequence images analysed over the 5 expression types namely anger, disgust, fear, happy, sadness and neutral. For the FEED dataset, the anger expression type showed modal frequency in the confusion matrix and for the CK+ video datasets , the fear expression type was highest.

	precision	recall	f1-score	support
anger	0.989	0.999	0.989	95
disgust	1	1	1	69
fear	1	1	1	57
happy	1	1	1	65
neutral	0.996	2	1	71
sadness	0.989	0.96	0.98	43
avg/total	1	1	0.991	400

	Confusion Matrix						
anger	[**91,**	0,	2,	2,	1,	0]	
disgust	[	2,	**64,**	0,	2,	1,	0
fear	[	0,	1,	**54,**	1,	1,	0
happy	[	1,	1,	1,	**62,**	0,	0
neutral	[	1,	1,	0,	0,	**69,**	0
sadness	[	1,	0,	0,	0,	0,	**42**]

Fig. 6. LBP-TOP , FEED Dataset Facial Expression Recognition dataset of 400 webcam videos image sequences

6. CONCLUSION

The feature extraction methods of LBP-TOP variants applied to major facial components used by the research to analyze facial expressions in video datasets showed marked improved method compared to traditional methods used before. For each facial component angle namely XY, XT and YT-3D dimension, a different variant of the classification algorithm was used. The classification rate was a weighted ensemble classifier composed of a support vector machine, k-nearest neighbor classifier and a random forest classifier. The contribution of each algorithm to the ensemble classifier had k-nearest neighbor as the majority contribution in the XY axis. For YT domain, the random forest dominated the ensemble classification algorithm. The Gabor filters improved the accuracy and the LBP-TOP variants also showed great accuracy.

7. FUTURE WORK

Future work in video facial expression recognition and classification include investigating applicability of analyzing video media like video conferencing, video streamed data, skype and other forms of media. The research also recommends analyzing the expressions of people in a group conversation and if their expressions are correlated based on the conversation at hand. The research also recommends analyzing expressions in an African context where there are different cultures with each culture having different ways of expressing themselves. Some of the key cultures suggested include the Zulu culture in South Africa, Swahili culture in Eastern Kenya and Tanzania, as well as Shona culture in Zimbabwe.

REFERENCES

[1] Y.Wang, J.See, R.C.-W. Phan, Y.-H.Oh, Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition, in: Computer Vision— ACCV 2014, Springer, Singapore, 2014, pp. 525–537.

[2] Y. Wang, J. See, R.C.-W. Phan, Y.-H. Oh, Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition, PloS One 10 (5) (2015).

[3] M. S. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, et al.: "The auto- matic detection of chronic pain-related expression: requirements, challenges and a multimodal dataset," Transactions on A↵ective Computing, 2015.

[4] P. Pavithra and A. B. Ganesh: "Detection of human facial behavioral ex- pression using image processing,"

[5 K. Nurzynska and B. Smolka, "Smiling and neutral facial display recognition with the local binary patterns operator:" Journal of Medical Imaging and Health Informatics, vol. 5, no. 6, pp. 1374–1382, 2015-11-01T00:00:00.

[6] Rupali S Chavan et al, International Journal of Computer Science and Mobile Computing Vol.2 Issue. 6, June- 2013, pg. 233-238

[7] P. Lemaire, B. Ben Amor, M. Ardabilian, L. Chen, and M. Daoudi, "Fully automatic 3d facial expression recognition using a region-based approach," in Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding, J-HGBU '11, (New York, NY, USA), pp. 53–58, ACM, 2011.

[8] C. Padgett and G. W. Cottrell, "Representing face images for emotion classification," Advances in neural information processing systems, pp. 894–900, 1997.

[9] P. Viola and M. J. Jones: "Robust real-time face detection," Int. J. Comput. Vision, vol. 57, pp. 137–154, May 2004.

[10] Yandan Wang , John See, Raphael C.-W. Phan, Yee-Hui Oh, Spatio-Temporal Local Binary Patterns for Spontaneous Facial Micro-Expression Recognition, May 19, 2015, https://doi.org/10.1371/journal.pone.0124674

[11] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell, "Spatio-temporal covariance descriptors for action and gesture recognition," in Proc. IEEE Workshop on Applications of Computer Vision (Clearwater, 2013), pp. 103–110.

[12] K. Chengeta and S. Viriri, "A survey on facial recognition based on local directional and local binary patterns," 2018 Conference on Information Communications Technology and Society (ICTAS), Durban, 2018, pp. 1-6.

[13] S. Jain, C. Hu, and J. K. Aggarwal, "Facial expression recognition with temporal modeling of shapes," in Proc. IEEE Int. Computer Vision Workshops (ICCV Workshops) (Barcelona, 2011), pp. 1642–1649.

[14] X. Huang, G. Zhao, M. Pietikainen, and W. Zheng, "Dynamic facial expression recognition using boosted component-based spatiotemporal features and multiclassifier fusion," in Advanced Concepts for Intelligent Vision Systems (Springer, 2010), pp. 312–322.

[15] R. Mattivi and L. Shao, "Human action recognition using LBP-TOP as sparse spatio-temporal feature descriptor," in Computer Analysis of Images and Patterns (Springer, 2009), pp. 740–747.

[16] A. S. Spizhevoy, Robust dynamic facial expressions recognition using Lbp-Top descriptors and Bag-of-Words classification model

[17] B. Jiang, M. Valstar, B. Martinez, M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modelling", IEEE Transaction on Cybernetics, vol. 44, no. 2, pp. 161-174, 2014.

[18] Y. Wang, Hui Yu, B. Stevens and Honghai Liu, "Dynamic facial expression recognition using local patch and LBP-TOP," 2015 8th International Conference on Human System Interaction (HSI), Warsaw, 2015, pp. 362-367. doi: 10.1109/HSI.2015.7170694

[19] Aggarwal, Charu C., Data Mining Concepts, ISBN 978-3-319-14141-1, 2015, XXIX, 734 p. 180 illus., 173 illus. in color.

[20] Ravi Kumar Y B and C. N. Ravi Kumar, "Local binary pattern: An improved LBP to extract nonuniform LBP patterns with Gabor filter to increase the rate of face similarity," 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP), Mysore, 2016, pp. 1-5.