A Review of Local, Holistic and Deep Learning Approaches in Facial Expressions Recognition

Kennedy Chengeta, Serestina Viriri
School of Mathematics, Statistics and Computer Science
University of KwaZulu-Natal, Durban, South Africa
216073241@ukzn.ac.za, viriris@ukzn.ac.za

***Abstract*—In facial expression identification, algorithms with higher classification rates and lower computational costs are preferred. To achieve that, feature extraction and classification should be accurate and efficient. Feature extraction optimization involves selecting the optimal feature descriptor. Various algorithms in computer vision involve holistic, local and deep learning algorithms. Holistic algorithms analyze the whole facial image and includes algorithms like Linear Discriminant Analysis or fisherfaces, eigenfaces (PCA), Histograms of Oriented Gradients and Gray Level Co-occurrence Matrix (GLCM). Local feature descriptors involve using local facial components separately then aggregating them into a combined histogram. Local binary patterns (LBP), local directional patterns (LDP) and scale-invariant feature transform (SIFT) feature extraction algorithms have been successfully used in local feature extraction. Deep learning involves using convolutional neural networks for image analysis. The most popular models are AlexNet, VGG-Face and GoogleNet. The study evaluates computational accuracy and efficiency of the three forms of facial expression recognition namely holistic, local and deep learning algorithms. The JAFFE and CK+ datasets are used for analysis. Gabor Filters are used for preprocessing filtering of the images whilst Viola Jones OpenCV toolset is used for image visualization. The study concludes that local algorithms compete very well with deep learning algorithms in terms of accuracy but use less processing power than convolutional networks. For real time facial expression analysis with minimal processing power and need for quick response times, LBP algorithms are recommended.**

1. Introduction

Automated facial expression recognition has many applications in computer vision application in advanced robot systems, social and sentimental analysis, medical diagnosis centers, manufacturing and security systems [10], [17]. Facial expression applications normally represent the 7 major expressions namely happiness, sadness, anger, neutral, fear, contempt and surprise [10], [24], [31]. Research in facial expression recognition is categorized into 4 main steps namely facial detection, facial alignment and preprocessing, feature extraction and lastly classification [19], [21], [22], [23]. Facial detection involves use of software and hardware systems to identify facial images from a human image be it from a video or static picture.

Major feature extraction approaches in facial recognition include local directional patterns (LDP), principal component analysis (PCA), linear discriminant-analysis (LDA), local binary patterns (LBP) and their variants, convolutional neural networks and Gabor Filters [19], [21], [22], [23]. Local binary patterns have been widely used as texture classifiers in facial expression using an effective arithmetic operator that possesses rotational and gray scale invariance properties. LDA or linear discriminant analysis is a holistic algorithm that finds linear transformations through maximizing between class variance as well as reducing inter class variance [19], [21], [22], [23]. Classification has largely used major machine learning supervised algorithms like different flavours of neural networks, random forests, bagging algorithms, support vector machines, k-Nearest Neighbour, C4.5 decision trees, extreme gradient boosting and ensemble bagging classifiers. The study compares accuracy and performance of the three major facial recognition algorithm classes, namely local, holistic and deep learning algorithms. The first section reviews the literature of the 3 approaches, followed by the approach, implementation and analysis of the comparison of the results. [1]

2. Literature Review

Facial expression research has been widely researched [19], [21], [22], [23]. The Facial Action Coding System (FACS) has been used to describe facial expressions using action units [19], [21], [22], [23]. Each action unit represents a change in facial expression or appearance. Common expressions include the 7 mostly used expressions namely happiness, sadness, anger and fear [19], [21], [22], [23]. Facial expression recognition follows 4 stages namely, facial detection using algorithms like Viola Jones detector. It also includes feature extraction steps where holistic algorithms like PCA and LDA or local algorithms like local binary patterns or local directional patterns are used. Other holistic algorithms like histogram of oriented gradients (HOG) have

ISBN 978-1-5386-7365-2/19/$31.00 ©2019 IEEE

also succeeded in facial expression and facial recognition [19], [21], [22], [23]. Local descriptors have a reputation of good performance under different illumination variations [19], [22], [23]. They are also easily compared using histogram. The many variations of local binary patterns for instance like central symmetric or multi scale variants allow the descriptor to gain on issues like feature reduction or improved accuracy [21], [22].

Facial expression recognition also involves classification of the features extracted in the form of histograms using traditional machine learning classifiers like support vector machines, k-nearest neighbour and random forest classifier. Alternatively, research has also seen a surge in the use of deep learning convolutional networks or CNNs which do dual feature extraction and classification. The 3 key approaches in facial expression recognition include holistic, local and deep learning algorithms.

2.1. Holistic Algorithm Analysis

Holistic face feature extraction algorithms use the whole facial image to perform facial expression recognition which is a combined pixel information set of a facial image. Key algorithms include Linear Discrimination Analysis, HOG or Histogram of Oriented Gradients as well as Gray Level Co-occurrence Matrix or GLCM [24].

2.1.1. Linear Discriminant Analysis. Linear Discriminant Analysis or Fisher's Linear discriminant uses a small set of features that distinguish an individual by maximizing the Fisher Discriminant Criterion (Fisher 1936) [34]. The Linear Discriminant Analysis is represented as a function of P [34]

$$P(x|k) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} \exp\left(-\frac{1}{2}(x' - \mu_k')^T \Sigma^{-1}(x' - \mu_k')\right) \quad (1)$$

2.1.2. Histogram of Oriented Gradients (HOG):. This algorithm has been extensively used in facial expression recognition as feature descriptors and uses image gradients to extract regions of interest) [34], [36]. HOG captures features through enumeration of occurrences of gradient orientation frequencies in localized image portions, detection windows, or region of interest (ROI) [36]. The image is subdivided into smaller cells and a histogram of gradient orientations are calculated over them) [34], [36]. Each cell is discretized into angular bins based on its orientation with a specific gradient weight. Cells adjacent to each other are grouped into normalized spatial regions or blocks) [36]. These blocks then combine to form the histogram) [36]. The Gradient Calculation for each pixel at each (a, b) coordinate, the magnitude m(a,b)) [36] is represented as

$$fx(a,b) = f(a+1,b) - f(a-1,b) \quad (2)$$

$$fy(a,b) = f(a+1,b) - f(a-1,b) \quad (3)$$

$$m(a,b) = \sqrt{(fa(a,b)^2 + fb(a,b)^2)]} \quad (4)$$

$$\theta(a,b) = \arctan(fa(a,b))/(fb(a,b)) \quad (5)$$

The gradient mathematics based on the a and b horizontal and vertical axis allows for showing the image pixel brightness levels with f(a,b)) [36]

2.1.3. Gray Level Co-occurrence Matrix (GLCM) . This algorithm was born out of Haralick's research of image classification. GLCM measures how various combinations of pixel brightness occur in a facial picture [33], [34]. Fourteen textural features were defined and measured from a probability matrix based on observed combinations of intensities of positions relative to the other in the image [33]. Key features included contrast, correlation, energy and homogeneity [33]. A matrix was used to represent the number of rows and columns equal to the gray level frequency where the matrix X is the relative frequency where the intensity of 2 pixels is represented by 'i' and 'j' on a distance d and angle θ. J, the GLCM value is represented as follows

$$J_{\Delta x, \Delta y_{i,j}} = \frac{\sum_{p=1}^{n-\Delta x} \sum_{q=1}^{m-\Delta y} \begin{cases} 1 \text{ if } (I_{p,q} = i,\, I_{p+\Delta x, q+\Delta y} = j) \\ I_{p,q}, I(p+\Delta x, q+\Delta y) \in r \end{cases}}{p_r} \quad (6)$$

2.2. Convolutional Neural Networks in Facial Expression Recognition

Deep learning convolutional neural networks(CNN) in computer vision involve training and testing input images where each input image undergoes series of convolution of layers with filters or kernels [14], [17]. With CNN one or more convolutional layers are used together with a pooling and fully connected layers. Deep learning allows for high level data representation and allows for dual feature extraction and classification in the form of a multilayer neural network [14], [16] as shown in the following equation.

$$S(a,b) = (V * Z)(a,b) = \sum_x \sum_y Z(a-x, b-y)V(x,y) \quad (7)$$

Convolutional networks are a form of neural network that relies on convolution as opposed to general matrix calculus in at least of the layers. The convolutional network is represented by function x which can then be mapped to a function w mapped to a given feature map [14], [35]. For an image in 2D, given as I with a 2 dimensional kernel K, the neural network is represented as in equation (1) [14], [35]. The time index given as t takes only integer values and given that x and w are integers the discrete convolution value s(t) is given in equation (8) [17].

$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a) \quad (8)$$

$$input \otimes kernel = \sum_{y=0}^{columns} (\sum_{x=0}^{rows} input(x-a, y-b)kernel(x,y)))) \quad (9)$$

Deep learning convolutional neural networks(CNN) in computer vision involve training and testing input images where each input image undergoes series of convolution of layers with filters or kernels [14]. Deep learning allows for high level data representation and allows for dual feature extraction and classification in the form of a multilayer neural network [14], [35]. The convolutional neural networks models have three layers namely [15], [35], the convolutional layer, pooling layer and the output layer [14], [17], [35]. Pooling layers which can be more than one, are then applied on the images to reduce image spatial sizes and also trainable parameters and max pooling is a popular pooling algorithm. Average pooling and L2 norm pooling are the other 2 pooling options [14]. The output layer takes the extracted features from reduced parameter images having passed several layers of convolution and padding and generates output in class form [14], [35]. The output contains a loss function for error prediction. A single forward and backward pass forms a successful training cycle [14], [35].

2.2.1. Deep Convolutional Neural Networks Models. Four deep convolutional neural networks models are discussed in this section. The most popular model is AlexNet followed by VGG-Face and GoogleNet. Alexnet showed robustness in classical image and facial recognition with better performance than the traditional feature extraction and classifier methods. The model is mainly characterized by down sampling of middle representations using stridden convolutions alongside the highest pooling layers [14], [15]. The map at the end is a reconfigured vector which is used to be input data into 2 fully connected layers that produces an AlexNet image descriptor. The VGG-Face model has very deep architecture and uses small filters based convolutional layers [14], [15] . The layers are followed with a max-pooling layer. The last 2 layers are then fed into a VGG image descriptor. VGGNet has 16 convolutional layers and uses 3 by 3 convolutions with lots of filters [15]. The training is on 4 GPUs with 138 million parameters [14], [15] .

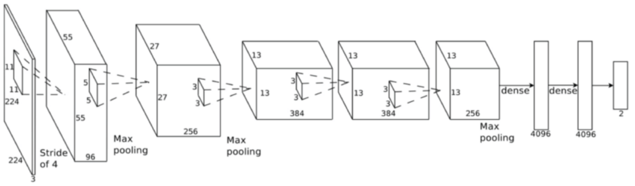

Figure 1. Alex Net with 5 Convolutional Layers and 3 Fully Connected Layers. [14], [15]

GoogleNet builds on an inception architecture. The architecture has blocks that bring together channel projections and spatial convolutions over various scales [14], [35]. Parameter spaces are reduced through decomposing of spatial convolutions with huge filters. A deeper network model results with fewer parameters and reduced complexity [14]. The model does not use any connected layers and the last convolutional map is an outcome of global average pooling and average activation values of the channels. The architecture is made up of 22 CNN layers with parameters reduced from 60 million to 4 million [14], [35].

2.3. Local Based Facial Expression Feature Extraction

Local feature extraction algorithms have been widely used in facial expression feature extraction both in 2D and 3D space [24]. The algorithms have shown robustness to illumination variation because of their gray scale capabilities as in the case of the local binary patterns. Scale-invariant feature transform(SIFT), Gabor wavelets, local binary patterns and local directional patterns have been successfully applied as local feature extractors [24].

2.3.1. Scale-invariant feature transform (SIFT). The algorithm transforms an image into a collection of local feature vectors [24], [32]. These feature vectors are made distinct and invariant to key transformations like rotation, translation and scaling of the picture. The algorithm is used in computer vision to detect and describe local salient and stable image features [24], [32]. It also describes characteristics of the small image regions around specific points both quantitatively and qualitatively. Key properties include locality which describes the property that features are local and robust to occlusion and clutter [24]. The features are also distinct and can be matched to an object repository. The quantity property means the objects can result in many features generated from few small objects. The efficiency and extensibility properties indicate the close to real time performance SIFT has as well as the extensibility capability to adding robust features [24], [32].

2.3.2. Local Binary Pattern (LBP). The algorithm is based on facial images being subdivided into smaller facial sub regions that include the nose, mouse, forehead and mouth for instance [20]. The localized feature vectors obtained are then used to form an aggregate histogram which is fed into machine learning classifiers [30]. The algorithm is not immune to occlusion, and rigidness though several variants have been invented after to address some of these issues [19], [22], [23], [30]. The local binary pattern features are also position dependent [19], [22], [23], [30].

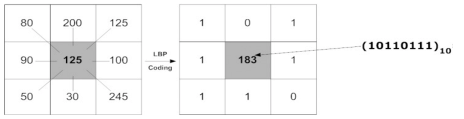

Figure 2. Local Binary Patterns (LBP)

The basic local binary pattern non center pixels are centered around a central pixel value in binary form taking only values zero or one [19], [22], [23]. Uniform binary patterns have 256 texture patterns and have uniformity in the bitwise transformations. The local binary LBP C, D operator is represented mathematically as follows

$$LBP_(C, D)(c_y, d_y) = \sum_{K-1}^{V=0} q(p_y - p_c)2^V . \tag{10}$$

The neighborhood is depicted as an m-bit binary string leading to V unique values for the local binary pattern code. The grey level is represented by 2V-bin distinct codes.

2.3.3. LBP Variants. Various LBP variants were successfully applied in computer vision. These include TLBP for Ternary Local Binary Pattern and the Central Symmetric Local Binary Patterns [20], [23], [24]. Over-Complete Local Binary Patterns (OCLBP) accounts for adjacent image block overlapping's. The rotation invariant LBP eliminates rotation effects by shifting the binary setup [20], [23]. Other variants include the monogenic local binary patterns(CS-LBP) and local binary pattern with TOP for dynamic expression recognition.

2.3.4. 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 [19], [21], [24].

$$LDP_x(\sigma) = \sum_{K}^{r=0} \sum_{L}^{r=0} f(LDP_q(o,u),\sigma). \tag{11}$$

The local directional pattern, includes edge detection using the kirsch algorithm and is represented by the following convolution matrix.

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

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

Figure 3. Local Directional Patterns (LDP)

3. Facial Expression Implementation

The implementation involves analyzing facial expression approaches based on local, holistic and deep learning algorithm approaches. Local binary pattern feature extractors are used for local algorithms whilst GLCM, LDA and PCA algorithms are used as holistic algorithms [33]. 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 JAFFE databases [26]. The section describes the approach, databases selected and the classification algorithm chosen and implemented.

3.1. Approach

The study compares the HOG, GLCM, LBP, LDP, CS-LBP, LDA as well as the deep learning CNN algorithm in terms of accuracy of classification of the CK+ and JAFFE databases. Facial detection is done using the Viola Jones Open CV detection algorithm. Feature extraction is done using the algorithms above. Classification is through machine learning classifiers namely support vector machines (SVM), random forest, weighted classifiers and k-nearest neighbor algorithm.

```
Data:  Copy and preprocess facial image datasets
Result: Facial expression classification results for
        the image datasets
while For each image I inside the CK+ and JAFFE
 dataset 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;
    4. apply Gabor Filters and PCA for
     dimensional reduction of features
    5. apply HOG, GLCM, LBP, LDP, CS-LBP,
     LDA and CNN;
    6. apply the classification on each with different
     classifiers;
    End While'
end
```

Algorithm 1: Local, Holistic and Deep Learning Algorithm approaches [29]

3.2. Facial Expression Preprocessing

The first step of establishing the PCA classifier was to determine parameters such as the number of principal components to consider and the number of training images [29]. 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 [29] . The Gabor filters extract expression-invariant features [29].

$$G\_c[i,j] = Be^{-\frac{(i^2+j^2)}{2\sigma^2}} \cos(2\pi f(i\cos\theta + j\sin\theta)); \tag{12}$$

$$G\_s[i,j] = Ce^{-\frac{(i^2+j^2)}{2\sigma^2}} \sin(2\pi f(i\cos\theta + j\sin\theta)); \tag{13}$$

where B and C are normalizing factors that will be derived [29].

3.2.1. Dimensional reduction with Principal Component Analysis. Noise was eliminated by remodeling the data to ensure the images have uniform dimensions. Dimensional reduction was achieved by using principal component analysis (PCA) holistic algorithm [28], [34]. For PCA an optimal hyperplane is calculated where all projected points spread with highest spreads. The principal components of the face. For a training sample set [28], [34] m, the mean m of the training sample is calculated as

$$\bar{\boldsymbol{m}} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{m}_k \tag{14}$$

The covariance matrix X based on given training samples is

$$X = \sum_{k=1}^{n} (\boldsymbol{x}_k - \boldsymbol{m})\ (\boldsymbol{x}_k - \boldsymbol{m})^T \tag{15}$$

3.3. Facial Expression Databases

The study used the JAFFE or Japanese Female Expression database, Labeled Faces in the Wild (LFW) dataset as well as the CK+ database. The given emotions ranged from fear, sadness, happiness or joy, disgust to neutral [15], [30]. The study participants performed several action units and facial displays and the images changed from neutral to peak through seven categories

[18], [33]. The JAFFE database has 213 images from 10 different female subjects and 7 different expressions namely anger, disgust, fear, happy, sadness, surprise and neutral. The images use 256 by 256 pixel images. The JAFFE facial dataset was chosen due to its small dataset structure and the CK+ due to its large dataset and different races. The CK+ dataset has around hundred individuals of American, Asian and Latin origin [18]. The Cohn-Kanade (CK) AU-coded expression dataset included over 100 students aged in their teenage years and early adult hood [24]. Labeled Faces in the Wild (LFW) a public available dataset with 5,749 unique identities and 13,233 face photographs was also used in the algorithm but its results were not considered due to its very small pictures which proved difficult to analyze.

3.4. Facial Expression Classification

The study uses template matching, k-nearest neighbor, random forest, neural networks and support vector machines [21], [23], [28]. For a kNN machine learning classifier kNN, the nearest neighbor, given x_q, with k nearest discreet neighbors, will take a mean of f values of k nearest neighbors [28] [24], [27].

$$kNN = \; f^{(}x_q)\frac{\sum_{i=1}^{k} f(x_i)}{k} \tag{16}$$

$$\sum_{i}^{k} 1/(i+1) = 1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{k} \tag{17}$$

3.4.1. Support Vector Machine. Support vector machines consider points close the given class boundaries [28]. 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 [24] [26] [27]:

$$w^T x_n + b = 0, \tag{18}$$

such that:

$$w^T, x_n + b1 \qquad y_n = +1, \tag{19}$$

3.4.2. Template Matching. The classification also uses the Kullback's test which is based on Chi Squared or χ^2 statistic distribution to check intra-image class covariance [28]. The Mahalanobis distance measures the distance between classes considering the covariance structure [28]. The distance between a pair of N dimensional points is extrapolated by the statistical differences. For two points $\vec{x}$ and $\vec{y}$ derived from similar distributions and covariance matrix $\mathbf{D}$, the Mahalanobis distance is calculated by the following equation as

$$((\vec{x} - \vec{y})'\mathbf{D}^{-1}(\vec{x} - \vec{y}))^{\frac{1}{2}}$$

The Kullback or χ^2 distance in computer vision is given as $\chi^2_{ij} = \frac{1}{2}\sum_{k=1}^{d}(x_k^i - x_k^j)^2/(x_k^i + x_k^j)$ where the $\boldsymbol{x}$'s are normalized histogram feature vectors for the given images [28].

4. Facial expression analysis results

The CK+ and JAFFE datasets were tested against support vector machines (SVM), k nearest neighbour(kNN) as well as random forest which is a bagging algorithm. The highest local binary pattern classification results showed when ELBP which is a local binary pattern variant was pre-filtered with Gabor Filters and PCA for dimensionality reduction to give a 98.92 percent accuracy. Convolutional neural networks showed much greater accuracy of 99.32 percent in the large datasets. The highest holistic algorithm LDA accuracy showed was 92.98 percent. The execution time between the 3 algorithms included 61 seconds for local binary patterns, 59 seconds for LDA and 3 hours for the convolutional neural network. The CNN implementation was executed on a 16GB and 2CPUs hardware iOS machine. The dataset used included 5000 images from the CK+ dataset and also 2000 images from the JAFFE dataset.

Figure 4. Image classifier for CK+ and GoogleSet combined Dataset with Gabon Filters applied

		CK+ Dataset(5000 images)				˜JAFFE(2000 images)˜			
Feature Extraction Models		**Classification Algorithms**			**Exec Time**	**Classification Algorithms**			**Exec Time**
	L(r,n)	**kNN**	**SVM**	**RF**		**kNN**	**SVM**	**RF**	**sec**
LBP	8,2	97.59	97.35	97.49	51	97.85	97.32	97.96	58
SIFT		94.32	94.11	93.87	59	94.51	94.65	94.73	61
GLCM		95.12	95.66	95.41	53	95.15	95.11	95.43	59
LDP	8,2	96.98	97.42	97.52	54	97.31	97.09	97.89	56
LBP+PCA	8,2	98.23	97.33	98.12	54	97.88	97.87	98.07	65
LDA		93.21	94.31	94.56	59	93.61	93.41	92.98	58
HOG		91.67	91.92	93.32	61	92.77	93.21	91.78	61
CSLBP	8,2	97.99	97.62	97.74	52	97.45	98.32	98.21	57
CSLBP	16,2	94.05	97.92	97.08	53	97.52	97.24	96.08	55
RLBP	8,2	96.81	97.04	97.83	61	97.68	97.76	98.43	51
LTP	8,2	98.32	97.44	98.42	58	97.97	98.21	98.33	49
LDP+ELBP	8,2	98.68	98.81	98.92	61	98.78	98.69	98.76	59
CNN(epochs)	500	99.21			183	99.11			191
CNN(epochs)	500	99.32			195	99.01			199
LBP	16,2	97.26	98.03	97.28	44s	97.33	98.11	98.61	55
RLBP	16,2	97.33	98.01	97.89	51	98.02	97.21	97.69	46

Figure 5. Classifier for CK+ and JAFFE Dataset with smaller Dataset-250

		CK+ Dataset(250 images)				˜JAFFE(100 images)˜			
Feature Extraction Models		**Classification Algorithms**			**Exec Time**	**Classification Algorithms**			**Exec Time**
	(r,n)	**kNN**	**SVM**	**RF**		**kNN**	**SVM**	**RF**	**sec**
LBP	8,2	96.32	97.35	97.49	62	97.42	97.22	97.12	53
SIFT		93.47	94.11	93.87	48	94.31	94.32	93.33	63
GLCM		94.84	95.66	95.41	49	94.43	94.31	94.09	55
LDP	8,2	96.33	97.42	97.52	51	96.94	96.99	97.57	58
LBP+PCA	8,2	97.13	97.33	98.12	58	96.97	97.27	99.12	66
LDA		92.14	94.31	94.56	66	93.41	92.67	93.45	59
HOG		90.41	91.92	92.82	51	92.21	92.11	90.98	64
CSLBP	8,2	97.57	97.62	97.74	46	97.11	98.09	98.01	55
CSLBP	16,2	93.81	97.92	97.08	47	96.99	97.12	96.08	58
RLBP	8,2	96.67	97.04	97.83	56	96.76	97.36	98.21	59
LTP	8,2	97.04	97.44	98.42	65	97.13	98.47	98.17	46
LDP+ELBP	8,2	98.18	98.25	98.71	58	98.01	98.13	98.32	62
CNN(epochs)	500	98.97			183	98.89			196
CNN(epochs)	500	99.01			191	99.06			187
LBP	16,2	97.23	98.14	97.18	46	97.33	98.11	98.61	52
RLBP	16,2	97.41	98.11	97.22	58	98.02	97.21	97.69	43

.

4.1. Facial Expression Results for Smaller Datasets-CK+ and JAFFE

For smaller datasets of 250 images for CK+ and 100 for JAFFE dataset the accuracy for the local algorithms, local binary patterns was equally impressive. The ELBP with radius 16,2 and Gabor Filter also gave a 97.67 percent accuracy. The CNN deep learning algorithm showed reduced accuracy over smaller datasets but due to number of epochs executed was able to achieve 98.32 percent accuracy. Holistic algorithm GLCM and LDA achieved 96.32 and 95.41 percent respectively. The execution time was averaging around three hours for the deep learning convolutional neural network algorithm compared to average 50 seconds for the local binary pattern algorithm and the LDA algorithm. Of the 6 emotions measured neutral had the best accuracy and happy showed the highest misclassified individuals.

The confusion matrixes for the 6 emotions for the CK+ dataset showed an overral success of 98.25 percent for ELBP with Gabor Filters algorithm compared to 99.23 for convolutional neural networks and 96.45, 93.65 for GLCM algorithm and LDA respectively on the larger datasets.

LDP+ELBP feature extraction classified by a random forest classifier gave accuracy of 98.76 percent and 99.92 on the JAFFE and CK+ databases. This was much improved accuracy compared

Figure 6. CK+ Dataset Facial Expression Recognition Large dataset (5000 images) and ELBP

	prec	**rec**	**f1**	**sup**	**Confusion Matrix**					
anger	0.968	1	0.981	917	**[[912**	0,	3,	0,	2],	
disgust	0.978	0.969	0.975	814	[0,	**811,**	1,	1,	1,	0],
fear	1	0.976	0.984	830	[2,	1,	**825,**	0,	1,	1],
happy	0.969	0.985	0.923	876	[1,	4,	1,	**870,**	0,	0],
neutral	1	0.992	1	814	[0,	1,	1,	1,	**811,**	0],
sad	0.993	0.989	0.967	749	[1,	1,	1,	5,	0,	**741]]**
avg/total	0.997	0.986	0.955	**5000**						

Figure 7. CK+ Dataset Facial Expression Recognition small dataset (250 images) and ELBP

	prec	**rec**	**f1**	**sup**	**Confusion Matrix**					
anger	0.977	1	0.949	22	**[[19**	0,	1,	1,	1],	
disgust	0.965	0.969	0.959	19	[1,	**16,**	0,	1,	1,	0],
fear	1	0.991	0.989	35	[2,	1,	**31,**	0,	1,	0],
happy	0.961	0.985	0.979	81	[0,	1,	0,	**79,**	0,	1],
neutral	1	0.986	1	44	[0,	0,	1,	1,	**42,**	0],
sad	1	0.982	0.991	49	[1,	0,	1,	1,	0,	**46]]**
avg/total	0.969	0.982	0.995	**250**						

Figure 8. JAFFE Dataset Facial Expression Recognition Large dataset (2000 images) and ELBP

	prec	**rec**	**f1**	**sup**	**Confusion Matrix**					
anger	0.956	1	0.956	326	**[[321**	1,	1,	2,	1],	
disgust	0.989	0.990	0.947	197	[2,	**190,**	0,	3,	1,	1],
fear	1	0.979	0.966	303	[0,	0,	**300,**	0,	1,	2],
happy	0.975	0.988	0.975	661	[1,	1,	4,	**654,**	0,	1],
neutral	1	0.969	1	164	[1,	0,	0,	1,	**162,**	0],
sad	1	0.984	0.989	349	[1,	1,	1,	1,	0,	**345]]**
avg/total	0.993	0.987	0.989	**2000**						

Figure 9. JAFFE Dataset Facial Expression Recognition small dataset (100 images) and ELBP

	prec	**rec**	**f1**	**sup**	**Confusion Matrix**					
anger	0.976	0.991	0.989	11	**[[8**	0,	1,	1,	1],	
disgust	0.958	0.984	0.986	21	[1,	**19,**	0,	0,	1,	0],
fear	1	0.985	0.991	10	[1,	0,	**8,**	0,	1,	0],
happy	0.976	0.991	0.983	15	[0,	1,	0,	**12,**	0,	2],
neutral	1	0.986	0.997	14	[0,	1,	1,	1,	**11,**	0],
sad	0.987	0.992	0.986	29	[0,	0,	1,	1,	1,	**26]]**
avg/total	0.981	0.994	0.993	**100**						

to other comparative holistic and slightly less than the deep learning algorithm. The datasets were also compared using the 2 template matching algorithms for expression identification of new faces. The local feature extraction(LBP) algorithms showed better performance than the holistic algorithm(LDA) on identification of facial expressions.

4.2. Conclusion

The study gave an insight into the different approaches to facial expression recognition namely local, holistic and deep learning approaches. The latter did not require a classifier. The holistic approaches were found to be able to capture distinct features and uniquely identify individuals using mathematical algebra. These algorithms were also found to include non-important features for facial expression. Feature based approaches showed robustness to facial changes to lighting, orientation and scaling. Deep learning approaches showed much better accuracy of the 3 and removed the need to have a classifier though they needed much more processing power and execution time was longer.

In terms of accuracy the LBP algorithm performed better than the holistic algorithms GLCM and LDA on the CK+ and JAFFE databases. This was in both scenarios where a huge and smaller dataset were considered. The deep learning CNN approach fared slightly better with small upward margins than the local feature based LBP algorithm on the CK+ and JAFFE larger datasets. This convolutional neural network approach was executed over 500 and 1000 epochs on the CK+ and JAFFE dataset. The upward margin over the smaller datasets showed a less upward margin percent advantage for the deep learning approach to the feature based approach. This showed the convolutional neural network in smaller datasets do lose accuracy.

The processing time in executing the convolutional time was more than 1000 percent time more than the holistic and feature based approaches. The hardware required for the CNN approach was also high. The study concludes that feature based approaches namely local binary patterns' accuracy was marginally less to deep learning approaches but extremely high albeit with less processing and execution time. The study recommends using different variants of local binary patterns to recognise facial expressions in real time.

4.3. Future Studies

Future studies involve not only comparing the three approaches but considering a fusion of all the 3 approaches to form a hybrid facial expression classification. There is also need to look at using trained models from local binary pattern to feed into a convolutional neural network. The study also wishes to expand the comparison to micro expression recognition over video sequences.

References

[1] HAN, J., and KAMBER, M. (2001). Data mining: concepts and techniques. San Francisco, Morgan Kaufmann Publishers.

[2] 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 Affective Computing, 2015.

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

[4] 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.

[5] Andrew J Calder and A.Mike Burton and Paul Miller and Andrew W Young and Shigeru Akamats: A principal component analysis of facial expressions, Vision Research, "41",number "9",pages:"1179 - 1208", "2001"

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

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

[8] M. S. Bartlett: Face image analysis by unsupervised learning, vol. 612.Springer Science and Business Media, 2012.

[9] P. M. Blom, S. Bakkes: C. T. Tan, S. Whiteson, D. Roijers, R. Valenti, and T. Gevers: "Towards personalised gaming via facial expression recognition,", AAAI Press (Palo, 2014), 2014.

[10] S. Biswas, K. Bowyer, and P. Flynn: "A study of face recognition of identical twins by humans," in Information Forensics and Security (WIFS), 2011 IEEE International Workshop on, pp. 1–6, Nov 2011.

[11] C. Longmore, C. Liu, and A. Young: "The importance of internal features in learning new faces," The Quarterly Journal of Experimental Psychology, vol. 68, no. 2, pp. 249–260, 2015.

[12] X. Zhao and S. Zhang: "Facial expression recognition based on local binary patterns and kernel discriminant isomap," Sensors, vol. 11, no. 10, pp. 9573– 9588, 2011.

[13] A. Ramirez Rivera, R. Castillo, and O. Chae: "Local directional number pattern for face analysis: Face and expression recognition," Image Process- ing, IEEE Transactions on, vol. 22, pp. 1740–1752, May 2013.

[14] Shima Alizadeh and Azar Fazel:Convolutional Neural Networks for Facial Expression Recognition, 1704.06756,Jun 2017

[15] Liang Chen and Meng Xi,Local binary pattern network: A deep learning approach for face recognition,2016 IEEE International Conference on Image Processing (ICIP)

[16] K. Chengeta and S. Viriri, 2018 Conference on Information Communications Technology and Society (ICTAS), A survey on facial recognition based on local directional and local binary patterns

[17] M. Z. Uddin, W. Khaksar and J. Torresen, "Facial Expression Recognition Using Salient Features and Convolutional Neural Network," in IEEE Access, vol. 5, pp. 26146-26161, 2017. doi: 10.1109/ACCESS.2017.2777003

[18] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. (CVPR4HB 2010), San Francisco, USA, 94-101.

[19] 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 Affective Computing, 2015.

[20] P. Pavithra and A. Ganesh:"Detection of human facial behavioral expression using image processing"

[21] 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.

[22] 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, USA), pp. 53–58, ACM, 2011.

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

[24] P. Viola and M. J. Jones: "Robust real-time face detection," Int. J. Comput.Vision, vol. 57, 2004.

[25] R. Valenti, N. Sebe and T. Gevers, "Facial Expression Recognition: A Fully Integrated Approach," 14th International Conference of Image Analysis and Processing - Workshops (ICIAPW 2007), Modena, 2007, pp. 125-130. doi: 10.1109/ICIAPW.2007.25

[26] K. Chengeta and S. Viriri, "A survey on facial recognition based on local directional and local binary patterns,",Conference on Information Communications Technology and Society, Durban, 2018

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

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

[29] 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.

[30] Pietikinen, M., Hadid, A., Zhao, G., Ahonen, T.,Computer Vision Using Local Binary Patterns,2011

[31] Ekman, P., and Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. Semiotica, 1, 49-98. Navneet Dalal and Bill Triggs. 2005. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - , Washington, DC, USA, 886-893.

[32] Y. Nakashima and Y. Kuroki, "Sift feature point selection by using image segmentation," 2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Xiamen, 2017, pp. 275-280. doi: 10.1109/ISPACS.2017.8266488

[33] Vishal S. Thakare and Nitin N. Patil. 2014. Classification of Texture Using Gray Level Co-occurrence Matrix and Self-Organizing Map. In Proceedings of the 2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies (ICESC '14). IEEE Computer Society, Washington, DC, USA, 350-355. DOI: https://doi.org/10.1109/ICESC.2014.66

[34] Sergios Theodoridis Sergios Theodoridis Konstantinos Koutroumbas; Pattern Recognition,2008 - 4th Edition - ISBN: 9781597492720, 9780080949123.

[35] H. Dikkers, M. Spaans, D. Datcu, M. Novak, and L. Rothkrantz: "Facial recognition system for driver vigilance monitoring," in Systems, Man and Cybernetics, 2004 IEEE International Conference on, vol. 4, pp. 3787–3792 vol.4, Oct 2004.

[36] Navneet Dalal and Bill Triggs. 2005. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - , Washington, DC, USA, 886-893