Selective local binary pattern with convolutional neural network for facial expression recognition

Syavira Tiara Zulkarnain, Nanik Suciati
Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Article Info

Article history:

Received Jul 11, 2021
Revised Jun 10, 2022
Accepted Jul 5, 2022

Keywords:

Convolutional neural network
Facial expression recognition
Feature selection
Local binary pattern

ABSTRACT

Variation in images in terms of head pose and illumination is a challenge in facial expression recognition. This research presents a hybrid approach that combines the conventional and deep learning, to improve facial expression recognition performance and aims to solve the challenge. We propose a selective local binary pattern (SLBP) method to obtain a more stable image representation fed to the learning process in convolutional neural network (CNN). In the preprocessing stage, we use adaptive gamma transformation to reduce illumination variability. The proposed SLBP selects the discriminant features in facial images with head pose variation using the median-based standard deviation of local binary pattern images. We experimented on the Karolinska directed emotional faces (KDEF) dataset containing thousands of images with variations in head pose and illumination and Japanese female facial expression (JAFFE) dataset containing seven facial expressions of Japanese females' frontal faces. The experiments show that the proposed method is superior compared to the other related approaches with an accuracy of 92.21% on KDEF dataset and 94.28% on JAFFE dataset.

Corresponding Author:

Nanik Suciati
Department of Informatics, Institut Teknologi Sepuluh Nopember
Teknik Kimia street, Surabaya, East Java 60117, Indonesia
Email: nanik@if.its.ac.id

1. INTRODUCTION

Humans communicate in verbal and non-verbal ways. Speech and writing are typical of verbal communication, while body language, gestures, and facial expressions are non-verbal. As one of the non-verbal communication modes, the prototypic of facial expressions depict human emotions of discrete expression such as fear, anger, disgust, pleasure, neutral, sadness, and surprise [1] or detect human emotions of encoded the basic anatomically change in facial muscles [2]. Facial expression recognition (FER) can be used to support human-computer interactions, robotics, learning process supervision, and mental health diagnosis [3], [4].

The conventional and deep learning approaches have been used to develop FER. The conventional approach generally consists of three stages, namely preprocessing, feature extraction, and classification [5], [2]. Face detection and cropping, face alignment, facial landmark identification, image normalization are common processes in the preprocessing stage. Preprocessing mainly aims to detect, localize the region of interest (face area) and, at the same time, provide a more stable face image, which is an essential stage in FER. The preprocessing generally consists of face acquisition and face normalization. Face acquisition is applied to localize face area and subtract irrelevant information in an image, such as complex background. This localization can reduce cost computation in further stages. Papageorgiou *et al.* [6] introduce a framework for object detection based on wavelet representation and pixel statistical analysis of class instances. The framework is tested for facial recognition, and the image representation is known as Haar features. Viola and Jones propose an object detection approach known as Viola-Jones using image representation based on integral calculation [7]. The Viola-Jones face detection can detect frontal images in almost real-time and achieves high detection rates [8]. Yan *et al.* [8] introduce the locally assembled binary (LAB) feature, which merges the co-occurrence binary Haar feature based on the existence of the local binary pattern (LBP). King introduces an open-source cross-platform known as Dlib-ml, which consists of two main components, an extensible linear algebra and a machine learning toolkit [9], [10]. The Dlib detector has a low mean absolute error (MAE) for face detection at night and daytime conditions [11].

Meanwhile, face normalization is used to normalize different light intensities to provide a more stable face image [5]. Research of Chen *et al.* [12] propose an image normalization based on discrete cosine transform (DCT). Due to the change of illumination that is mainly located in the low-frequency band, Chen's idea is discarding the low frequency of DCT coefficients to minimize light conditions' variability. Munir *et al.* [13] introduce merged binary pattern coding (MBPC), which merges horizontal, vertical, and diagonal direction bits (HVD code) in a histogram. HVD codes are obtained from a higher and lower bit magnitude of a discrete Fourier transform (DFT). Lee *et al.* [14] conclude that gamma distortion takes effect on different contrast of luminance value (V) in HSV color space. An image with good contrast, focus, and a few blurring has relatively high entropy. Lee *et al.* [14] reconsiders the non-linear response of human eyes toward brightness as a transformation function of adaptive gamma (AGT-Me). This method's theoretical basis is maximizing differential entropy by minimizing gamma distortion and applying automatic gamma adjustment for contrast enhancement.

The feature extraction process that takes important characteristics of the training data is generally performed separately from the classification process. This stage represents the important feature of an input image in a vector of some values. In any approach, feature extraction is essential because the high recognition rate depends on how reliable the manually defined feature extraction method is. There are two categories of feature extraction algorithms. One is an algorithm that extracts geometric features, and another is an algorithm that extracts appearance (texture) features [2]. The first category is based on the feature vector that corresponds to geometric information, such as shapes and locations, of the pair of facial components. Some algorithms that fall in this category are active shape model (ASM) that extract key points at the global shape of the human face; optical flow that is commonly used to track motion change and extract key points with the dense flow; haar-like feature extraction; and feature point tracking [5].

The second category extracts the features in whole or some regions of the input image based on texture appearance. Extracting the texture information can be performed in the frequency or spatial domain. The algorithms that belong to the frequency domain are DCT [1], weber local descriptor (WLD) [15], discrete wavelet transform (DWT) [16], and Gabor wavelet [5]. The well-known and widely-used algorithm that works on the spatial domain is LBP, proposed by Ojala *et al.* [17]. LBP has advantages in uncomplicated computation and tolerance to illumination changes. LBP has been successfully applied in various applications and is widely exploited to improve the robustness in extracting information [18]. Some research to enhance the LBP by exploring pixel's neighbor relationships are complete local binary pattern (CLBP), local directional pattern (LDP), and local phase quantization (LPQ) [15].

Another stage that affects FER performance is classification. There are two categories of classifiers, namely conventional and deep learning. Some classifiers that belong to the first category, such as k-nearest neighbors (KNN) [19], support vector machine (SVM) [20], and Adaboost (adaptive boosting) [21], have been applied in FER. The conventional approach results in high accuracy on datasets with less variability, such as a collection of frontal face images with the same illumination. The performance of conventional approaches usually decreases when applied to datasets with large variability. The robustness of the hand-crafted feature extraction method and the conformity of each stage bind the conventional approach performance [5], [19].

The deep learning approach aka convolutional neural network (CNN) unifies the processes of feature extraction and classification in one learning framework. The deep CNN consists of convolutional in first layer which are responsible for generating a feature map from the input image and fully connected in the second layers which function as classifier. Fei *et al.* [3] proposed a FER using features maps extracted by AlexNet (one of the CNN architecture) combined with linear discriminant analysis (LDA) for classification. Vedantham and Reddy [15] introduced a robust feature extraction using local descriptors with optimized deep belief network (DBN)-spider monkey optimization (SMO) for FER. The training for deep learning is performed to determine a large number of weights of these layers. The deep CNN models often produce comprehensive feature maps and high classification accuracy when trained on a huge number of images. The performance of this approach decreases if the number of training images is small.

Developing a FER system on datasets with a small number of images containing a variety of illuminations and head poses is a challenge. Variation in illumination and head pose can lead to misclassification if not handled properly. Several conventional approaches have been proposed to develop FER. Research by Ekweariri and Yurtkan [19] presents feature selection based on the variance of LBP images and classification based on distance measurement, which obtains an accuracy of 66.67%. Research by Islam *et al.* [22] extracts features in the segmented parts of the expression area (right and left eye, nose, mouth) using Gabor filters before being applied in the classification stage using the extreme learning machine (ELM). The proposed FER achieves an accuracy of 86.84%.

There are researches in a hybrid approach, which combines the conventional and the deep learning approaches to develop FER. The hybrid approach proposed by Levi and Hanssner [23] uses LBP, a feature extraction method often utilized in the conventional approaches, to handle variation in illumination and to produce a stable image representation. The LBP images are then used to train a CNN, as is usually performed in the deep learning approach. The FER proposed by Levi and Hanssner [23] achieves an accuracy of 61.29% Meanwhile, a hybrid approach by Fei *et al.* [3] develops FER using a combination of deep features extracted by AlexNet and LDA classifier. This work utilizes a pre-trained AlexNet, one of the CNN architectures that has been trained on a huge common image dataset, as a feature extractor. The generated features are then used to train LDA, one of the classical classifiers. They experiment with five datasets containing front-facing images. The performance of AlexNet as a classifier that represents the deep learning approach and the combination of AlexNet as a feature extractor with an LDA classifier that represents the hybrid approach is compared in the experiment. The AlexNet results in 84.7% accuracy, while AlexNet+LDA results in 87.8% accuracy.

In this research, we present an improvement of the hybrid approach for FER on datasets with a small number of images containing a variety of illuminations and head poses. We propose a new method for selecting the most discriminant features on face images using a median-based standard deviation of LBP images, to generate more stable image representation. The selective local binary pattern (SLBP) image representation is used for adjusting the weights of CNN layers in the training stage. We use the Karolinska directed emotional faces (KDEF) dataset in the experiment [24].

2. RESEARCH METHOD

2.1. Design of proposed method

The design of our FER system is shown in Figure 1. The preprocessing stage aims to subtract irrelevant information, normalize the illumination, and augment the number of images. The feature extraction intends to extract LBP features. Our proposed method, the selective LBP, aims to select the most informative patterns that distinguish expression classes. In the classification stage, a training process is conducted to build a CNN model that is further used in the testing process. The model is trained to obtain the CNN architecture weights suitable for both producing feature maps and classifying expression classes at the same time. The following subchapter discusses each stage in detail.

The preprocessing stage-apply face detection, face normalization. All image in the dataset contains a face with a background. We use face detector Dlib to subtract irrelevant information and provide only face area. Dlib is an open-source library of machine learning toolkit using the CNN approach called max-margin object detection (MMOD) [25]. CNN-based Dlib is more accurate and works well for non-frontal images than the original Dlib that is based on histogram of oriented gradients (HoG). All face area with various sizes produced by the face detector is trimmed from the images. Then all is resized to 128×128 and converted into grayscale. For face normalization, we apply illumination normalization using adaptive gamma transformation method (AGT-Me). The AGT-Me performs blind inverse gamma correction based on image maximum entropy. First, denote all pixel intensity (0 to 255) in the grayscale facial image as l_m, it is then transformed within the range [0,1] of u_m using (1). An optimal gamma parameter γ^* on a predefined masking area Ω is calculated using (2). The gamma parameter γ is then applied in (3) to perform gamma transformation g_m. The final step is transforming back the pixel value within its original range of the grayscale image using the inverse of (1).

$$u_m = \frac{l_m+0.5}{256} \tag{1}$$

$$\gamma^{\wedge} = \frac{1}{\frac{1}{N}\sum m \in \Omega l_n\,(u_m)} \tag{2}$$

$$g_m = f\,(u_m) = u_m^{\gamma} \tag{3}$$

$$t \times (gc - gp) = \begin{Bmatrix} 1, gp \geq gc \\ 0, otherwise \end{Bmatrix} \tag{4}$$

$$LBP(x,y) = \sum_{p=0}^{7} t \times (gc - gp) \times 2^p \quad (5)$$

The motivation of the proposed selective LBP is to extract the most informative LBP in facial images. The pattern is assumed to be located in the part of the face whose intensity often changes due to differences in facial expressions. The LBP texture descriptors are used because of their simplicity in representing textures and their ability to handle a variety of poses and illuminations. The selective LBP consists of three processes, i.e., generating an LBP image of all facial images in the training data, constructing a binary reference image used to mark informative pixel positions, and selecting informative LBP for each image.

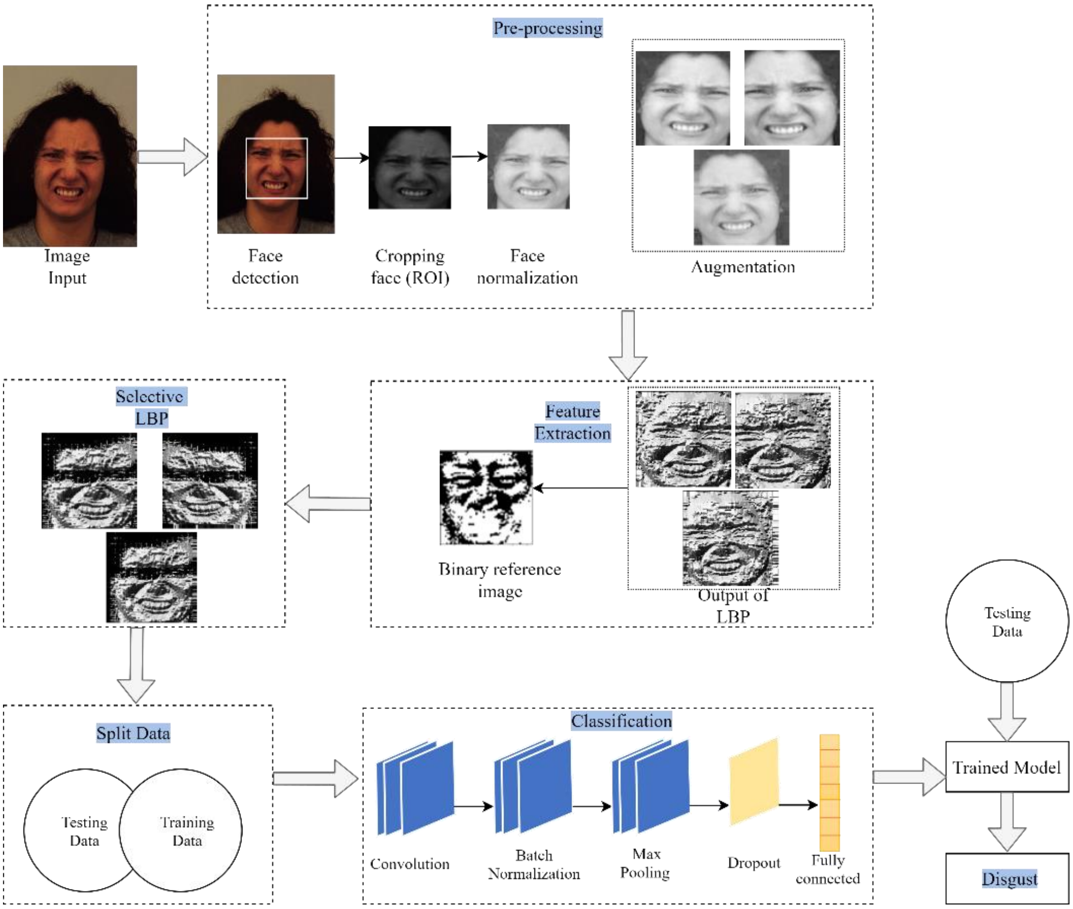
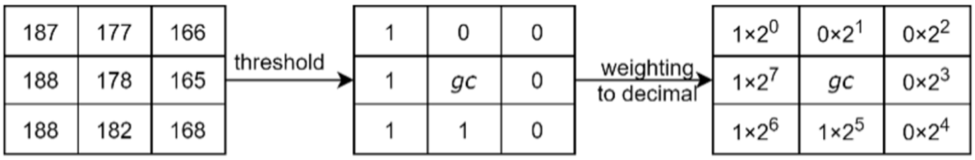


Figure 1. Design of the proposed system

Generating an LBP image-LBP, which was firstly proposed by Ojala *et al.* [17] captures the local texture patterns on an image, such as dot, edge, and line, by normalizing the intensity of each pixel with its neighbors. The utilization of the LBP texture feature can be in the form of a holistic histogram or a holistic image depending on the type of image used, whether it is a standard texture or a real-world image [26]. In this study, we apply the LBP using 3x3 neighbors. Since the facial images contain variations that naturally exist in the real world, so it is proper to use the LBP textures as holistic images for further processing. A 3x3 neighborhood window is run on all the pixels in the image. The LBP code of a pixel at the neighborhood center is computed by firstly comparing the intensity of the center pixel to its neighbors, as illustrated in Figure 2. The LBP code for the grayscale image in Figure 2(a) is calculated using a 3x3 neighbor thresholding process with the center pixel as the threshold to produce image in Figure 2(b), and is encoded using a weighted summation with the predefined weights in Figure 2(c). This comparison is like conducting a thresholding process using the center pixel as the threshold, as shown in (4). If the neighbor gp has an intensity lower than the center gc, then the pixel value is 0; otherwise, it is 1. The thresholding result is used to activate the predefined fixed order decimal weights. The LBP code of the center pixel is a summation of all weights. There are eight neighbor pixels gp and the LBP code is defined by (5).



187	177	166
188	178	165
188	182	168

(a)

1	0	0
1	*gc*	0
1	1	0

(b)

1×2^0	0×2^1	0×2^2
1×2^7	*gc*	0×2^3
1×2^6	1×2^5	0×2^4

(c)

Figure 2. The LBP code for (a) the grayscale image, (b) calculated using a 3x3 neighbor thresholding process with the center pixel as the threshold to produce the image, and (c) encoded using a weighted summation with the predefined weights

Construct a binary reference and selecting informative LBP-feature selection is a major proposal of this research, which selects discriminant features based on the standard deviation (square root of the variance) of LBP images. The important point that distinguishes the proposed feature selection method from others is the use of median-based standard deviation. The standard deviation provides more informative patterns, capturing less variability in intensity level [27], and presents a more precise calculation on each class [28]. Moreover, the calculation of median-based standard deviation provides a robust calculation estimator [29] applicable in any distribution. In symmetric distribution, the median is closer to modus than to mean. These reasons cause the median value is quite favored over the mean. The algorithm for selecting informative LBP:

- First, grouping all LBP images according to the expression classes.
- Calculate the standard deviation σ (square root of the variance σ^2) at each pixel position $f(x, y)$ in every image belongs to the same expression class, which is called a pixel deviation. Calculation of standard deviation σ using (9), and previously followed by calculating median-based variance σ^2 in (8).
- Calculate class deviation using the average of all pixel deviations shown in (6). Furthermore, the class deviation is used as a reference to determine whether a pixel location contains informative or non-informative features.
- Find pixel location that has an informative feature candidate. A pixel location is called an informative candidate if its pixel deviation higher than the class deviation.
- Store all of the pixel locations in a binary reference matrix. Stored the matrix conducted by replacing the LBP code at the corresponding pixel position in the LBP image contains an informative pattern with value 1. While the value 0 means the opposite. Every expression class has a separated binary reference matrix. Each is constructed using all images that belong to a certain class.
- Unification of the seven (according to amount classes of the dataset) matrices is conducted using logical operator AND.
- Selection of the informative LBP feature refers to the binary reference matrix. At a particular pixel location, the LBP value is considered informative and will be extracted if the binary reference matrix element at the corresponding location is 1. Otherwise, the LBP value is ignored. This method produces an informative holistic LBP image used in the classification process.

$$\mu = \frac{1}{LM}\sum_{x=0}^{L-1}\sum_{y=0}^{M-1} f(x,y) \quad x = 0,1,2,\dots,L-1; y = 0,1,2,\dots,M-1 \tag{6}$$

$$med = median[f(x,y)] \tag{7}$$

$$\sigma^2 = median[f(x,y) - med]^2 \tag{8}$$

$$\sigma = \sqrt{\sigma^2} \tag{9}$$

2.2. Experimental setting

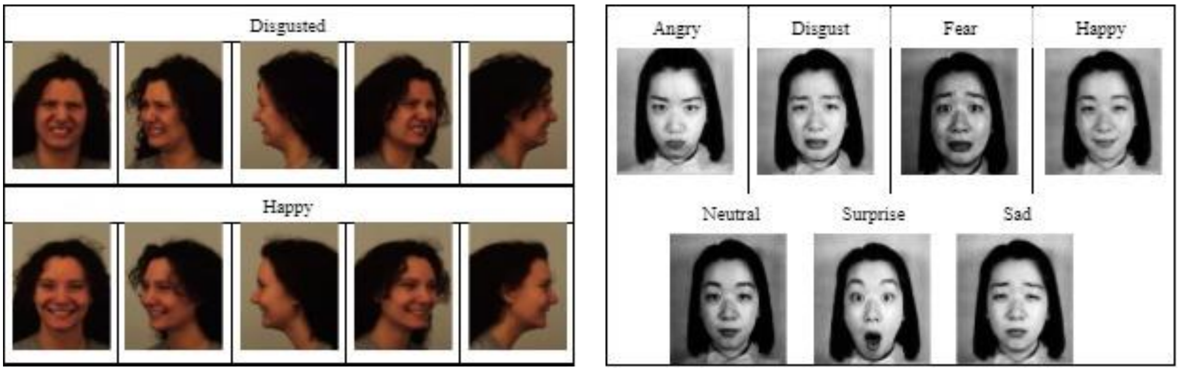
For model development aims to prepare layers, activation function, loss function, and other required parameters. We construct a CNN architecture shown in Table 1. which consists of five convolution layers, four batch normalizations used to normalize data, so it has a mean close to 0 and a standard deviation close to 1, four down-samplings with kernel size 2×2, two dropouts used to prevent overfitting, and a fully connected layer. Due to the experiment scenario, we set the CNN parameters with epoch 60, batch size 100, Adam optimizer with a learning rate of 0.01, and early stopping with loss monitor validation to avoid overfitting or under-fitting.

Table 1. Detail architecture

Layer	Input	Specification
Input layer	-	
Convolution Layer 1	128, 128, 1	Filter: 5×5, 32 Stride: 1x1 ReLU
Max Pooling 1	124, 124, 32	Kernel: 2×2
Convolution Layer 2	62, 62, 32	Filter: 5×5, 32 Stride: 1×1 ReLU
Max Pooling 2	58, 58, 32	Kernel: 2×2
Convolution Layer 3	29, 29, 32	Filter: 5×5, 32 Stride: 1×1 ReLU
Max Pooling 3	25, 25, 32	Kernel: 2×2
Dropout 1	12, 12, 32	Probability: 0.4
Convolution Layer 4	12, 12, 32	Filter: 5×5, 64 Stride: 1×1 ReLU
Max Pooling 4	8, 8, 64	Kernel: 2×2
Dropout 2	4, 4, 64	Probability: 0.1
Convolution Layer 5	4, 4, 64	Filter: 4×4, 7 Stride: 1×1 ReLU
Fully Connected	1, 1, 7	7

2.3. Dataset

We use Karolinska directed emotional faces (KDEF) and Japanese female facial expression (JAFFE) dataset [30] in experiment. The original of KDEF dataset consists of 4,900 human facial expression images. These images are made up of 35 male and 35 female participants, photographed in two different sessions. According to Figure 3(a), every participant has seven expressions in the first session, each expression consists of five images that are captured from different angles (frontal, half-left, left, half-right, right). The second session contains images taken with the same expressions and angles such as the first session but under different laboratory lighting conditions. Because of the augmentation process in preprocessing stage, the amount of dataset is 14,700. That dataset is divided into training and testing sets. The number of testing data is 2,675, consisting of randomly selected images from the second photo session. The number of training data is 12,025. All images from the first photo session and the rest of the second session are included in the training data. This division ensures that all individuals and all angle/head pose in the dataset already have representation in the training data. The JAFFE dataset consists of 213 human facial expression images. These datasets are made up of 10 Japanese female participants with a size image of 256×256. For deep learning purposes, the amounts of dataset are increased to 10,917 images after augmentation. Those images divided into 2,000 testing data and rest of it will be included in the training data. According to Figure 3(b) each participant has seven expressions such as happy, sad, surprise, angry, disgust, fear, and neutral.



(a) (b)

Figure 3. The facial expression image datasets used in the experiment are: (a) KDEF (several angel and expression of KDEF dataset) and (b) JAFFE (expression of JAFFE dataset)

3. RESULTS AND DISCUSSION

3.1. Result

For experiment purposes, we conduct two scenarios on two different datasets. The first scenario is an internal comparison on KDEF dataset and JAFFE dataset intended to obtain the optimum configuration of the proposed method, which are shown in Table 2. Tables 3 and 4 describes the performance in precision, recall and F1-score for each expression class of the best algorithm combination in the KDEF and JAFFE dataset. The first column in Table 2 presents information about the dataset and whether face detection is applied or not. The second column provides information about combination of the applied algorithms such as augmentation, normalization, feature extraction, and feature selection. The third column is the classifier that we used. The option of the normalization algorithms is histogram equalization and AGT-Me. The methods used in feature selection include median-based standard deviation and median-based variance. The same CNN architecture is used in all combinations of algorithms. The second scenario is an external comparison to measure how well the proposed method among the other existing methods. The result for KDEF dataset is shown in Table 5 and the JAFFE dataset are shown in Table 6. Table 5 describe all methods used on the KDEF dataset, except the first approach, performing on the BU-3DFE dataset. Figure 4 shows graphic of performance based on tuning parameter on KDEF and JAFFE dataset.

Table 2. Accuracy comparison among different algorithm combinations on KDEF and JAFFE dataset

Dataset	Combination	Classifier	Accuracy (%)
KDEF	Image: raw (without crop the images)	CNN	29.4
KDEF + Face detection	Image: cropped images	CNN	45
KDEF + Face detection	Image: cropped images Feature extraction: LBP	CNN	42.5
KDEF + Face detection	Image: cropped images Normalization: Histogram Equalization Feature extraction: LBP	CNN	17.9
KDEF + Face detection	Image: cropped images Normalization: AGT-Me Feature extraction: LBP	CNN	57.9
KDEF + Face detection	Image: cropped images Normalization: AGT-Me Feature extraction: LBP Feature selection: median-based standard deviation	CNN	73.06
KDEF + Face detection	Image: cropped images + Augmentation Normalization: AGT-Me Feature extraction: LBP	CNN	86.4
KDEF + Face detection	Image: cropped images + Augmentation Normalization: AGT-Me Feature extraction: LBP Feature selection: median-based variance	CNN	90.60
KDEF + Face detection	**Image: cropped images + Augmentation** **Normalization: AGT-Me** **Feature extraction: LBP** **Feature selection: median-based standard deviation**	CNN	**92.21**
JAFFE + Face detection	Image: cropped images + Augmentation	CNN	77.5
JAFFE + Face detection	Image: cropped images + Augmentation Feature extraction: LBP	CNN	80.50
JAFFE + Face detection	Image: cropped images + Augmentation Normalization: Histogram Equalization Feature extraction: LBP	CNN	81.87
JAFFE + Face detection	Image: cropped images + Augmentation Normalization: AGT-Me Feature extraction: LBP	CNN	88.25
JAFFE + Face detection	Image: cropped images + Augmentation Normalization: AGT-Me Feature extraction: LBP Feature selection: median-based variance	CNN	90.98
JAFFE + Face detection	**Image: cropped images + Augmentation** **Normalization: AGT-Me** **Feature extraction: LBP** **Feature selection: median-based standard deviation**	CNN	**94.28**

Table 3. Performance of each expression class in the KDEF dataset using the best algorithm's configuration

Feature	Precision (%)	Recall (%)	F1-score (%)
Afraid	92	80	85
Angry	91	89	90
Disgusted	92	93	92
Happy	98	99	99
Neutral	90	98	94
Sad	89	91	90
Surprised	93	96	94

Table 4. Performance of each expression class in the JAFFE dataset using the best algorithm's configuration

Feature	Precision (%)	Recall (%)	F1-score (%)
Angry	100	96	98
Disgust	100	100	100
Fear	96	85	90
Happy	81	100	89
Neutral	82	90	86
Sad	85	96	90
Surprised	100	75	86

Table 5. Performance of state-of-the-art comparison on KDEF dataset

References	Proposed technique	Accuracy (%)
A.N. Ekweariri and Yurtkan [19]	LBP + **Enhanced LBP** + KNN: **Feature selection with mean-based variance**, produce general binary images from all classes (from LBP's feature)	66.67
K. Rujirakul and So-In [31]	Histogram equalization + **Deep PCA**+ extreme learning machine	83.00
B. Islam *et al.* [22]	Viola-Jones face detection + **Facial region segmentation** + Gabor filter + extreme learning machine	86.84
Z. Fei *et al.* [3]	**Feature extraction: AlexNet and FC6** + LDA (classifier)	87.80
Z. Fei *et al.* [3]	**Feature extraction: AlexNet and FC6** + SVM (classifier)	86.40
Z. Fei *et al.* [3]	**Feature extraction: AlexNet and FC6** + KNN (classifier)	64.00
Our proposed method	Augmentation + AGT-Me **+ Selective LBP** + CNN: **Feature selection with median-based standard deviation**, produce references binary images from each class (from LBP's feature)	92.21

Table 6. Performance of state-of-the-art comparison on JAFFE dataset

References	Proposed technique	Accuracy (%)
Salmam *et al.* [32]	Viola-Jones face detection + Feature extraction: SDM method + **Euclidean distance** + **Distance ratio** + **Feature Selection** + NN	93.8
Our proposed method	Augmentation + AGT-Me **+ Selective LBP** + CNN: **Feature selection with median-based standard deviation**, produce references binary images from each class (from LBP's feature)	94.28

3.2. Discussion

From the first scenario is an internal comparison intended to obtain the optimum configuration of the proposed method, the best accuracy on the KDEF and JAFFE dataset is 92.21% and 94.28% obtained by applying selective LBP with median-based standard deviation. Hence, the best configuration is to incorporate pre-processing, consisting of Dlib face detector, AGT-Me normalization, data augmentation using horizontal flipping and zooming in; LBP feature extraction; selective LBP feature selection; and CNN classifier. It has listed the performance of precision, recall, and F1-score for each expression shows in Tables 3 and 4. Expression with the highest precision and recall in KDEF dataset is happy and in JAFFE dataset is disgust. Visual observation of the facial images mentions that the happy expression of KDEF dataset and disgust expression of JAFFE dataset does not resemble any other class expression; thus, it can be easily differentiated.

Another configuration is using the KDEF dataset that contains 4,900 raw images in the RGB channel without pre-processing, feature extraction, and feature selection as the input of the CNN classifier. This configuration obtains 29.4% accuracy. CNN with raw images as an input does not guarantee high performance, as observed by Levi and Hassner [23]. They proposed an ensemble CNN method using wild or real-world images as an input resulting in less than 55% accuracy. Even though CNN has a layer intended to do feature extraction, if the input image were not pre-processed or do not have a handcrafted feature extraction process beforehand, the CNN cannot handle the high variability of images accurately. The second experiment uses the face area of the 4,900 raw images produced by the face detection process without augmentation, normalization, feature extraction, and feature selection as the CNN classifier's input. This configuration achieves 45% accuracy and shows that providing a face area that contains the most important information for facial expression recognition as the input can improve CNN's performance. However, the use of LBP images of the face area directly as the input decreases the accuracy to 42.5%. Therefore, we apply two normalization algorithms, i.e., histogram equalization and AGT-Me, to normalize the face area before coded to the LBP image. We find that the AGT-Me was more suitable for expression recognition and increase the accuracy to 57.9% because the resulting image tends to have a similar contrast, by being able to adjust the contrast increase using maximizing differential entropy by minimizing gamma distortion and applying automatic gamma.

Therefore, in subsequent experiments, we choose AGT-Me to get better performance. The addition of a feature selection process using the selective LBP to the previous experiment configuration increases the expression recognition accuracy to 73.06%. This improvement suggests that the proposed selective LBP is suitable for selecting informative LBP patterns on the facial image. Meanwhile, the addition of the data augmentation process also increases the accuracy to 86.4%. This result implies that deep learning performs better when trained with more data. Another dataset that we used is JAFFE dataset. Because the amount of the data is small, we decided to perform augmentation process first in experiment and achieved an accuracy 77.5%. The next experiment of JAFFE dataset is using LBP as feature extraction, then adding normalization with AGT-me which achieved an accuracy 80.50% and 88.25% respectively. In the two last experiments for each dataset, we combine the strength of data augmentation and feature selection with two variants of the selective LBP method, namely median-based variance and median-based standard deviation. The use of median-based variance on the selective LBP increases the accuracy to 90.60% in KDEF dataset and 90.98% in JAFFE dataset.

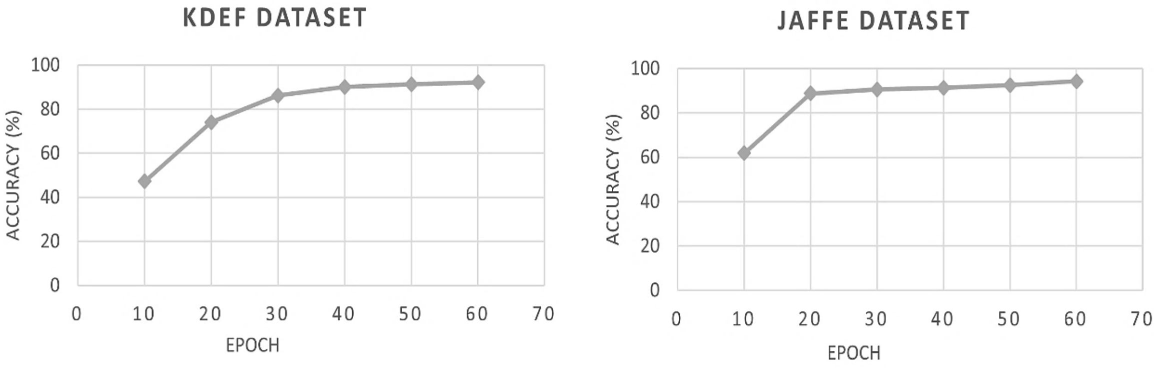
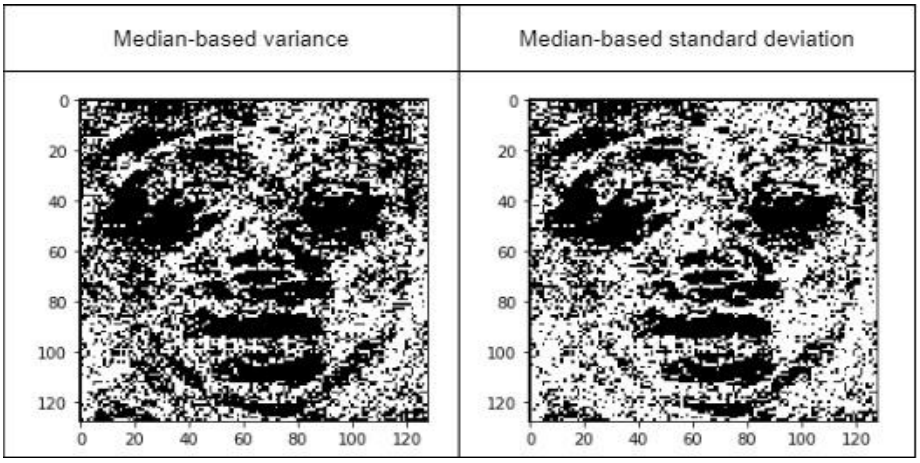


Figure 4. Performance on KDEF and JAFFE dataset by tuning the epoch parameter

The result of the second scenario that comparison of several methods in facial expression recognition which work on the KDEF, and JAFFE dataset proves the superiority of the proposed method with an accuracy of 92.21% and 94.28%. Other methods such as Ekweariri and Yurtkan [19] propose an enhanced LBP using mean-based variance calculation to construct a binary reference matrix for selecting features. They experiment on the BU-3DFE dataset (2D images) using a KNN classifier and achieve 66.67% of accuracy [31] propose Deep PCA for feature extraction, a deep learning approach to enhance the traditional PCA feature extraction. They perform on 140 images of the KDEF dataset using an extreme learning machine (ELM) classifier and obtain an accuracy of 83%. A method to handle Viola-Jones' error in detecting some parts of a face is proposed by Islam *et al.* [22]. They apply facial region segmentation to obtain coordinates, width, and height of frontal face images used as a reference to detect the face area properly. Their proposed facial region segmentation is tested on 980 frontal face images of the KDEF dataset using Gabor filter for extracting features and ELM for classification. Their experiment achieves 86.84% accuracy. The last comparison method is the work by Fei *et al.* [3]. They propose a combination of feature vectors produced by AlexNet, a deep learning architecture, with three classification methods: LDA, SVM, and KNN. They experiment on 980 frontal face images of the KDEF dataset. The best result is obtained using the LDA classifier with an accuracy of 87.80%. Salmam *et al.* [32] proposed dynamic feature extraction of facial expression recognition and feature selection used JAFFE dataset as an experiment achieved an accuracy 93.8%. Figure 4 shows the performance of tuning epoch parameter with Adam optimizer and learning rate of 0.01, both KDEF and JAFFE dataset shown an increase performance also followed by increasing epoch.

We tested our proposed method on the KDEF and JAFFE dataset, consisting of 4,900 face images with variations in head pose and illumination. The experiment showed the robustness of the proposed method with an accuracy of 92.21% and 94.28% on the KDEF and JAFFE dataset respectively. Feeding in the result of handcrafted feature calculation to the CNN architecture can give an accuracy improvement. This is based on theory of standard deviation that provides more informative patterns by capturing less variability in intensity level [27]. As proven in Figure 5, the standard deviation effectively captures more pixel (mark with white pixel) than the variance computation. Moreover, the calculation of median-based standard deviation provides a robust calculation estimator [29] applicable in any distribution. In symmetric distribution, the median is closer to modus than to mean. These reasons cause the median value is quite favored over the mean.


Median-based variance

Median-based standard deviation

Figure 5. Result of binary reference image with different calculation

4. CONCLUSION

This paper presents a facial expression recognition on the KDEF and JAFFE dataset consisting of facial images with illumination and head pose variations. We propose selective LBP (SLBP), a method to select the most informative features using median-based standard deviation on facial LBP images. Combination of the proposed feature selection method use of median-based standard deviation and data augmentation capable of handling head poses variation. Meanwhile, the use of AGT-Me for normalizing an image is suitable to manage illumination variation.

Our system consists of subsequent processes, i.e., Dlib face detection, AGT-Me normalization, data augmentation, LBP feature extraction, selective LBP feature selection, and CNN. Comparison with several related approaches shows that our system achieves the highest accuracy and has a high-performance recall, precision, and F1-score of each class expression. It is concluded that our proposed method (selective LBP), combining with a hybrid method (handcrafted feature extraction with CNN), improves facial expression recognition performance and solved the challenge with image containing a variety of illuminations and head poses. Representation of facial expression with head poses variations and illumination using handcrafted method based on local feature are still a problem especially for those with extreme variations. One way to solve the problem is by using feature representative enrichment, which are done by combining the proposed method with informative facial landmark or facial action units (AUs) which encodes the basic movement of the muscles that should be considered as the future work.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge financial support from the Institut Teknologi Sepuluh Nopember for this work, under project scheme of the Publication Writing and IPR Incentive Program (PPHKI).

REFERENCES

[1] W.-L. Chao, J.-J. Ding, and J.-Z. Liu, “Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection,” *Signal Processing*, vol. 117, pp. 1–10, Dec. 2015, doi: 10.1016/j.sigpro.2015.04.007.

[2] Y.-L. Tian, T. Kanade, and J. F. Cohn, “Facial expression analysis,” in *Handbook of Face Recognition*, New York: Springer-Verlag, 2005, pp. 247–275.

[3] Z. Fei *et al.*, “Deep convolution network based emotion analysis towards mental health care,” *Neurocomputing*, vol. 388, pp. 212–227, 2020, doi: 10.1016/j.neucom.2020.01.034.

[4] S. Li and W. Deng, “Deep facial expression recognition: A Survey,” *IEEE Transactions on Affective Computing*, 2020, doi: 10.1109/TAFFC.2020.2981446.

[5] Y. Huang, F. Chen, S. Lv, and X. Wang, “Facial expression recognition: A survey,” *Symmetry*, vol. 11, no. 10, Sep. 2019, doi: 10.3390/sym11101189.

[6] C. P. Papageorgiou, M. Oren, and T. Poggio, “A general framework for object detection,” in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, 1998, pp. 555–562, doi: 10.1109/ICCV.1998.710772.

[7] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 2001, vol. 1, pp. I--511--I--518, doi: 10.1109/CVPR.2001.990517.

[8] S. Yan, S. Shan, X. Chen, and W. Gao, “Locally assembled binary (LAB) feature with feature-centric cascade for fast and accurate face detection,” in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2008, pp. 1–7, doi: 10.1109/CVPR.2008.4587802.

[9] D. E. King, “Dlib-ml: A machine learning toolkit,” *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[10] S. Sharma, K. Shanmugasundaram, and S. K. Ramasamy, “FAREC-CNN based efficient face recognition technique using Dlib,” in *2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, May 2016, no. 978, pp. 192–195, doi: 10.1109/ICACCCT.2016.7831628.

[11] J. Paone, D. Bolme, R. Ferrell, D. Aykac, and T. Karnowski, “Baseline face detection, head pose estimation, and coarse direction detection for facial data in the SHRP2 naturalistic driving study,” in *2015 IEEE Intelligent Vehicles Symposium (IV)*, Jun. 2015, no. Iv, pp. 174–179, doi: 10.1109/IVS.2015.7225682.

[12] W. Chen, M. J. Er, and S. Wu, “Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain,” *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 458–466, Apr. 2006, doi: 10.1109/TSMCB.2005.857353.

[13] A. Munir, A. Hussain, S. A. Khan, M. Nadeem, and S. Arshid, “Illumination invariant facial expression recognition using selected merged binary patterns for real world images,” *Optik*, vol. 158, pp. 1016–1025, Apr. 2018, doi: 10.1016/j.ijleo.2018.01.003.

[14] Y.-H. Lee, S. Zhang, M. Li, and X. He, “Blind inverse gamma correction with maximized differential entropy,” *Signal Processing*, vol. 193, p. 108427, Apr. 2022, doi: 10.1016/j.sigpro.2021.108427.

[15] R. Vedantham and E. S. Reddy, “A robust feature extraction with optimized DBN-SMO for facial expression recognition,” *Multimedia Tools and Applications*, vol. 79, no. 29–30, pp. 21487–21512, Aug. 2020, doi: 10.1007/s11042-020-08901-x.

[16] M. A. Muqeet and R. S. Holambe, “Local appearance-based face recognition using adaptive directional wavelet transform,” *Journal of King Saud University-Computer and Information Sciences*, vol. 31, no. 2, pp. 161–174, Apr. 2019, doi: 10.1016/j.jksuci.2016.12.008.

[17] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, Jan. 1996, doi: 10.1016/0031-3203(95)00067-4.

[18] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, “Local binary patterns and its application to facial image analysis: A survey,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 6, pp. 765–781, Nov. 2011, doi: 10.1109/TSMCC.2011.2118750.

[19] A. N. Ekweariri and K. Yurtkan, “Facial expression recognition using enhanced local binary patterns,” in *9th International Conference on Computational Intelligence and Communication Networks (CICN)*, 2017, pp. 43–47, doi: 10.1109/CICN.2017.8319353.

[20] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on Local Binary Patterns: A comprehensive study,” *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, May 2009, doi: 10.1016/j.imavis.2008.08.005.

[21] V. K. Gudipati, O. R. Barman, M. Gaffoor, and A. Abuzneid, “Efficient facial expression recognition using adaboost and haar cascade classifiers,” in *2016 Annual Connecticut Conference on Industrial Electronics, Technology & Automation (CT-IETA)*, Oct. 2016, pp. 1–4, doi: 10.1109/CT-IETA.2016.7868250.

[22] B. Islam, F. Mahmud, and A. Hossain, “Facial region segmentation based emotion recognition using extreme learning machine,” in *2018 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, Nov. 2018, pp. 1–4, doi: 10.1109/ICAEEE.2018.8642990.

[23] G. Levi and T. Hassner, “Emotion recognition in the wild via convolutional neural networks and mapped binary patterns,” in *Proceeding 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 503–510, doi: 10.1145/2818346.2830587.

[24] D. Lundqvist, A. Flykt, and A. Ohman, “The Karolinska directed emotional faces (KDEF),” *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*. 1998.

[25] D. E. King, “Max-margin object detection,” *arXiv preprint arXiv:1502.00046*, Jan. 2015.

[26] B. Yang and S. Chen, “A comparative study on local binary pattern (LBP) based face recognition: LBP histogram versus LBP image,” *Neurocomputing*, vol. 120, pp. 365–379, Nov. 2013, doi: 10.1016/j.neucom.2012.10.032.

[27] R. C. Gonzalez and R. E. Woods, *Digital image processing*, 3rd Edition, Pearson Education, 2007.

[28] J.-M. Sung, D.-C. Kim, B.-Y. Choi, and Y.-H. Ha, “Image thresholding using standard deviation,” in *Image Processing: Machine Vision Applications VII*, Mar. 2014, vol. 9024, doi: 10.1117/12.2040990.

[29] A. Bovik, *Handbook of image and video processing*, Second Edi. Elsevier, 2005.

[30] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with Gabor wavelets,” in *Proceeding third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 200–205, doi: 10.1109/AFGR.1998.670949.

[31] K. Rujirakul and C. So-In, “Histogram equalized deep PCA with ELM classification for expressive face recognition,” in *2018 International Workshop on Advanced Image Technology (IWAIT)*, 2018, pp. 2018–2021.

[32] F. Z. Salmam, A. Madani, and M. Kissi, “Emotion recognition from facial expression based on fiducial points detection and using neural network,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 1, pp. 52–59, 2018, doi: 10.11591/ijece.v8i1.pp52-59.

BIOGRAPHIES OF AUTHORS

Syavira Tiara Zulkarnain is post graduate student at Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember (ITS), Indonesia. She received the bachelor's degree in Informatics in 2015 and master's degree in 2019 at ITS. Her researches interest are computer vision and digital image processing. She can be contacted at email: syavira.zulkarnain@gmail.com.

Nanik Suciati received the bachelor's degree in computer engineering from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia in 1994, the master's degree in computer science from University of Indonesia in 1998, and the doctoral degree in Information Engineering from University of Hisroshima in 2010. She is an Associate Professor at Department of Informatics, ITS. Her researches interest are computer vision, computer graphics, and artificial intelligence. She can be contacted at email: nanik@if.its.ac.id.