The Performance Analysis of Facial Expression Recognition System Using Local Regions and Features

Yining Yang*, Vuksanovic Branislav, and Hongjie Ma

School of Energy and Electronic Engineering, University of Portsmouth, Portsmouth, UK;
Email: branislav.vuksanovic@port.ac.uk (V.B.), hongjie.ma@port.ac.uk (H.M.)
*Correspondence: Yining.Yang1@myport.ac.uk (Y.Y)

***Abstract*—Different parts of our face contribute to overall facial expressions, such as anger, happiness and sadness in distinct ways. This paper investigates the degree of importance of different human face parts to the accuracy of Facial Expression Recognition (FER). In the context of machine learning, FER refers to a problem where a computer vision system is trained to automatically detect the facial expression from a presented facial image. This is a difficult image classification problem that is not yet fully solved and has received significant attention in recent years, mainly due to the increased number of possible applications in daily life. To establish the extent to which different human face parts contribute to overall facial expression, various sections have been extracted from a set of facial images and then used as inputs into three different FER systems. In terms of the recognition rates for each facial section, this result confirms that various regions of the face have different levels of importance regarding the accuracy rate achieved by an associated FER system.**

***Keywords*—facial expression recognition, facial features, principal component analysis, convolutional neural networks**

I. Introduction

The automatic detection of human emotions such as anger, happiness, and sadness from facial expressions has been actively investigated in recent decades. With the rapid development of technology and advances in Machine Learning (ML), the range of potential applications for Facial Expression Recognition (FER) systems has significantly increased, covering various aspects of human-computer interaction and emotion analysis as well as other more targeted applications such as monitoring driver fatigue by using simulators and computer games. Therefore, the number of FER-related studies and publications has grown rapidly over the last decade. A study on the published works in this area provides a good overview of the more traditional methods [1−4] and recent deep learning-based methods [5], along with the corresponding results. Over the past few years, the need to detect one's emotions has increased, and human emotion recognition has gained growing interest, including but not limited to human-machine interfaces, animation, medicine and security [6].

In most FER studies, FER has been applied to classify expression information into one of several basic emotions, facial expressions were grouped into seven categories: anger, disgust, fear, happiness, neutral, sadness and surprise [7]. In this research, neutral expression classifies faces that do not express human emotions.

Although difficult, the "FER via ML" problem has been addressed with increasing success in recent years. However, the underlying issue relating to the link between facial expression and the true emotional state of a person presents a different type of challenge to the machine-learning community. Another related problem is the question of which part of the human face contributes to facial expression to the greatest extent. This case perhaps reveals the most about the emotional state of a person.

Based on a previous study [8], the recognition process focuses more on the lower part of the face that performs sound communication and includes the mouth, nose, chin, and cheeks. Some researchers believe that recognition should exploit the intrinsic characteristics of how humans express emotions, that is, modifying their facial expressions by moving the landmark features of the face. The location of facial landmarks should be used, and this information should be incorporated into the classification process [9]. In researches [10, 11], eyebrows play an important part in face recognition [11], with a continuous and categorical model describing how humans understand emotions in facial expressions. The Viola-Jones algorithm is used to detect features of the human face and other developments. In Wegrzyn *et al*'s research [12], observers mostly relied on the eye and mouth regions when successfully recognising an emotion. Furthermore, the difference in the importance of the eyes and mouth allowed the expressions to be grouped in a continuous space, ranging from sadness and fear (reliance on the eyes) to disgust and happiness (the mouth). Therefore, based on previous research, the local facial parts that contribute to

Manuscript received July 12, 2022; revised September 20, 2022; accepted October 20, 2022.

doi: 10.18178/joig.11.2.104-114

facial expression to the greatest extent may clearly reveal most of the emotional state of a person.

An FER system usually consists of a number of steps common to many other machine vision systems, as depicted in Fig. 1. The first step in this system is the pre-processing of the captured images to enhanceimage quality. Pre-processing can include image adjustment, histogram normalisation, image scaling, and filtering. The next task is face detection, which can be combined or achieved by the detection of regions of interest (ROI) in the face, including the eyes, cheeks, mouth, and forehead. One of the most common algorithms employed to detect face and facial ROIs is the Viola-Jones method [13].

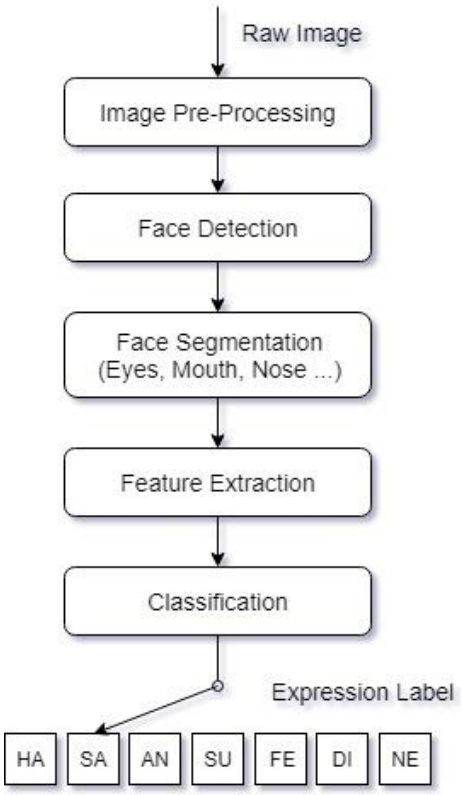


Figure 1. Main stages in the generic facial expression recognition system.

This stage represents a shift in the system from image to actual data processing. Some of the most popular features and tools for feature extraction include Local Binary Patterns (LBP) [14, 15], Gabor filters [16], Principal Component Analysis (PCA) [17], Independent Component Analysis [18], Linear Discriminant Analysis [19, 20], Local Directional Pattern [21, 22] among many other approaches.

Both the Gabor filter and the LBP are texture descriptors used to explain the organisation and texture of the face in different ways. In particular, the LBP approach assigns a label to each pixel in the P-neighbourhood (P equals spaced pixels within the same distance from the central observed pixel) by thresholding its values with the central pixel value and converting these threshold values into a decimal number via the LBP rule [14, 15]. Several different methods and variants of these approaches, such as local binary patterns from the three orthogonal planes (LBP-TOP) [23, 24] and local-oriented edges (LOE-LBP-TOP) [23], have been reported in the literature in recent years.

Another group of feature extractors is based on edge information in the images, which can be extracted in different ways. The Line Edge Map (LEM) and graphic processing unit-based Active Shape Model (GASM) are two prominent methods for edge feature extraction that can also be combined with other edge detection and image enhancement methods. The Histogram of Gradients (HOG) is another edge-related feature descriptor [24, 26].

In addition, some studies have used the attention neural network to improve the weight of the key local area of facial expression. In the case of facial expressions, many cues come from several areas of the face, such as the mouth and eyes, whereas other parts, such as ears and hair, play a small role in the output. This finding means that, ideally, the ML framework should focus on the more important parts of the face whilst being less sensitive to other facial regions [6].

Therefore, the aim of this study is to investigate the contributions and degree of importance of different sections of the human face to overall facial expression and automatic expression detection. To achieve this aim, three different sections of the faces available in the JAFFE database were extracted from each image and automatic classification, that is, the FER system, which were then applied to associate each section with a corresponding expression. These sections – (i) eyes, (ii) nose and mouth, and (iii) mouth and chin regions – are shown in Fig. 2 using an image from the JAFFE database. After comparing these to the "ground truth", the classification labels supplied with the JAFFE database were then made, and the accuracy of classification for each part of the face was calculated. Various sizes of each extracted facial region were considered to better assess the importance of accurate estimation and details of each section for classification accuracy.

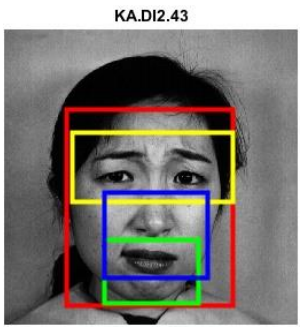


Figure 2. Sample facial image from the JAFFE database with its label and indicated segments used in this work.

Three different feature extraction and classification techniques were used in this study. Our aim is not to compare the performance of these methods but to discount the influence of the recognition algorithm and generalise the findings of this work. First, a "classical" PCA-based method [17, 20] was combined with a simple k-Nearest Neighbour (KNN) [27] classification algorithm and used on the standard set of JAFFE images. The second approach

combines an LBP [14, 15] method of feature extraction and a Support Vector Machine (SVM) [15, 26] to classify the extracted features and detect the expression. Finally, a more recent deep learning-based approach was applied where an artificial Convolutional Neural Network (CNN) was used to perform both the feature extraction and classification of presented facial and extracted sections of facial images. The same set of images and facial sections was used for each algorithm tested in this study. CNNs have been shown to outperform other techniques in most machine vision tasks, providing a large amount of data. In this case, facial images are available to train the network. However, the JAFFE database is relatively small, considering the training requirements of most CNNs. The issue of a small training set was somewhat overcome in this work by using one of the available pre-trained networks, Alexnet. This was slightly reconfigured and then trained in the final stage using a portion of the JAFFE images. This technique is typically referred to as transfer learning.

The remainder of this paper is organised as follows. The applied techniques and systems are briefly described in the Section II. The results are presented in Section III and conclusions are presented in the final section of this paper.

II. FER System and Techniques Used in This Work

A. LBP and SVM

The original LBP operator, introduced by Ojala *et al.* [14], proved to be a powerful means of texture description and has been successfully applied to FER [28, 29] and other image recognition tasks. The operator labels the pixels of an image by thresholding a 3×3 neighbourhood of each pixel with a centre value. The result of thresholding can be considered a binary number, as shown in Fig. 3. The 256-bin histogram of LBP labels can be computed over a region and used as a texture descriptor.

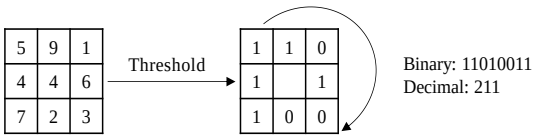


Figure 3. Basis LBP operator.

$$S(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \quad x = i_n - i_c \tag{1}$$

In the Eq. (1) above, i_n is the pixel value at coordinates in the neighbourhood of (x, y) and i_c is the pixel value at coordinate (x, y). For example, in Fig. 3, the i_c coordinate is (1,1) and the i_n coordinates are (0,0), (0,1), (0,2), (1,0), (1,2), (2,0), (2,1), and (2,2).

The binary number generated by the LBP operator compares the neighbouring pixel values with the central pixel value. The pattern with eight neighbours is expressed as follows:

$$LBP(x, y) = \sum_{n=0}^{n-1} 2^n \times s(i_n - i_c) \tag{2}$$

The limitation of the basic LBP operator is its small 3×3 neighbourhood that cannot capture dominant features with large-scale structures [30]. Therefore, the operator was later extended to use neighbourhoods of different sizes [31].

After labelling an image with the LBP operator, the histogram of this image contains information about the distribution of the local micropatterns over the entire image. Therefore, the histogram can be used to statistically describe the image characteristics. Facial images can be observed as a composition of micropatterns that can be effectively described by LBP histograms. Therefore, LBP features are intuitively used to represent facial images [30]. Notably, the LBP histogram computed over the entire face image encoded only the occurrences of the micropatterns without any indication of their respective locations.

An SVM classifier was selected because it is well founded in statistical learning theory and has been successfully applied to various tasks in computer vision [23]. SVM has been successfully used for facial expression classification. As a powerful ML technique for data classification, SVM performs an implicit mapping of data into a higher (perhaps infinite) dimensional feature space. It then finds a linear separating hyperplane with the maximal margin to separate data in this higher-dimensional space [30, 32]. A quadratic SVM was used for classification in this study. The multiclass SVM method is a one-against-one approach. When the model was trained, the SVM used five-fold cross-validation.

B. PCA and 2DPCA

One of the earliest successful approaches used for facial recognition and later extended to FER was the method proposed by Turk and Petland in 1991 [17]. Essentially, this method is based on (PCA, a statistical method that is widely researched and applied in many engineering and computing problems.

The purpose of PCA is to find an optimal set of orthogonal vector bases through a linear transformation of the observed dataset. The linear combination of the extracted basis vectors can be used to reconstruct the original data samples, thereby minimising the error between the reconstructed and original samples. When applied to facial images, each input image was transformed into a one-dimensional column vector and then used to form an input matrix. This data matrix is then analysed, after which a set of orthogonal basis vectors is extracted to represent a multidimensional face image as a linear combination of these basis vectors. Data dimensions can be reduced by linearly projecting data samples onto the subset of those vectors where the variances of all projected samples are maximised. These projections can be treated as features extracted from each data sample and then classified. Euclidian distance or other similar metrics, for example, Mahalanobis cosine distance, are typically used for this purpose.

The original one-dimensional PCA of high-dimensional data matrices was obtained by stacking column vectors into a large data matrix. This approach can be problematic because of the large size of the cumulative data matrix and, in some situations, a small sample number. In this case, PCA calculations are computationally demanding and less

accurate. To avoid this problem, a two-dimensional (2D) version of PCA, the 2DPCA, is proposed [33]. Here, each image matrix does not have to be transformed into a column vector or stacked into a large data matrix; instead, it is used without transformation. It has been established that this approach reduces the computational requirements and results in a more accurate decomposition of the original dataset.

Mathematically, if we use A to denote an $m \times n$ image matrix and take X to represent the m dimensional unitary column vector, the projection of matrix A onto X is a linear transformation:

$$Y = A^T X \tag{3}$$

The obtained column vector Y is called the projected feature vector of image A on vector X.

The optimal projection vector Xopt must be obtained to maximise the variance of the projected feature vector X_{opt} that must be obtained. The optimal projection vector X_{opt} maximises the scatter criterion function $J(X) = X^T G X$.

Here, G is the covariance matrix for the set of M images A_k, $k = 1, 2, \dots, M$ calculated using the following:

$$G = \frac{1}{M}\sum_{k=1}^{M}(A_k - \bar{A})^T(A_k - \bar{A}) \tag{4}$$

where $\bar{A}$ represents the average or mean image of this set, that is:

$$\bar{A} = \frac{1}{M}\sum_{k=1}^{M} A_k \tag{5}$$

G is a symmetric $M \times M$ matrix. M eigenvectors X_i and the associated eigenvalues λ_i of this matrix can be calculated using the eigenvalue decomposition method, that is,

$$[P, \Lambda] = eig(G) \tag{6}$$

where matrix P contains all extracted vectors as columns. Matrix Λ is a diagonal matrix containing the associated eigenvalues, sorted in descending order, as follows:

$$P = [X_1, X_2, \dots, X_M] \tag{7}$$

$$\Lambda = diag\{\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_M\}$$
$$\lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_M \tag{8}$$

The projection matrix P_{proj} can now be formed from the first d eigenvectors corresponding to d largest eigenvalues.

$$P_{proj} = [X_1, X_2, \dots, X_d] \tag{9}$$

Once the projection matrix is evaluated, each image collected from the facial expression gallery is projected into the 2DPCA subspace, after which the projections are classified using one of the numerous distance metrics (Table I).

TABLE I. SUMMARY OF 2DPCA APPLICABLE TO FER

Training Images	Test Image
Calculate the mean image for the whole training set Y_k	
$\bar{Y} = \frac{1}{M}\sum_{k=1}^{M} Y_k$	
Find the difference between the mean image and each image from the training set	Find the difference between the mean image and test image Z
$\hat{Y}_k = Y_k - \bar{Y}$	$\hat{Z} = Z - \bar{Y}$
Evaluate the covariance matrix	
$G = \frac{1}{M}\sum_{k=1}^{M} \hat{Y}\hat{Y}_k^T$	
Calculate the eigenvectors of the covariance matrix	
$[P, \Lambda] = eig(G)$	
Select the subset of extracted eigenvectors and form the projection matrix	
$P_{proj} = P(:, 1{:}\,d)$	
Project each training image using the projection matrix	Project the test image using the projection matrix
$\tilde{Y}_k = \hat{Y}_k^T P_{proj}$	$\tilde{Z} = \hat{Z}_k^T P_{proj}$

Once the above process has been completed, this is followed by the classification phase, which includes an evaluation of the distance between the projections of each centralised training image and the test image, that is, $d(\tilde{Y}_k, \tilde{Z})$. The search for the minimum distance would finally reveal the expression from the training set to be associated with the test image.

Different distance metrics can be used in this stage. Some of the most frequently used are the Euclidian distance, cosine angle distance, mean square error distance and Manhattan distance. In this work, a simple Euclidean distance is used, where for two matrices, X and Y of size $R \times C$ this distance is calculated as follows:

$$d(X, Y) = \sqrt{\sum_{i=1}^{R}\sum_{j=1}^{C}\left(x_{i,j} - y_{i,j}\right)^2} \tag{10}$$

C. Deep Learning

Deep-learning algorithms attempt to use unknown structures in the input distribution to find good representations, usually at multiple levels, using low-level features to define high-level learning features [34]. The deep learning method has been applied to the field of image recognition because the CNN has a strong recognition ability [35]. Compared with traditional ML methods, deep learning relies on massive training data because it requires a substantial amount of data to understand the potential patterns of data. Data dependence is one of the most serious problems in deep learning and, at the same time, in some areas, lack of training data is inevitable. Data collection is complex and expensive, making it extremely difficult to

build large-scale, high-quality annotation datasets. In this case, transfer learning is an important tool for solving the basic problem of insufficient training data in ML [36].

Transfer learning relaxes the assumption that the training data must be independent and identically distributed with the test data. For this reason, transfer learning has achieved great success in many areas where improvements have been difficult owing to insufficient training data [36]. Transfer learning can use a pre-trained network as a starting point for learning new tasks [37]. It attempts to transfer the learnt structure from the source domain to the target domain by relaxing the assumption that the training data must be independent and identically distributed to the test data.

Fig. 4 shows the process of transfer learning.

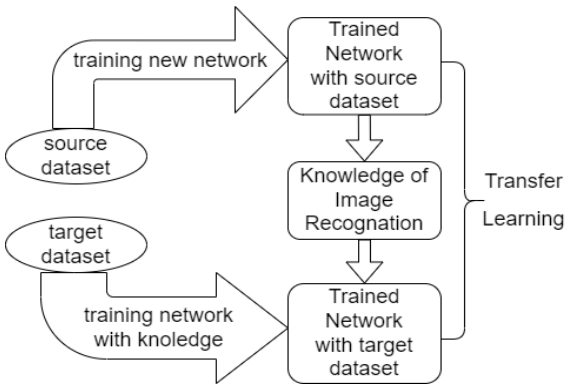


Figure 4. Transfer learning.

Transfer learning can be defined in the following way [36]:

Given a learning task T_t based on D_t, we can obtain help from D_s for the learning task T_s. Transfer learning aims to improve the performance of the predictive function $f_T(\cdot)$ for the learning task T_t by discovering and transferring latent knowledge from D_s and T_s, where $D_s \neq D_t$, and/or $T_s \neq T_t$. In addition, in most cases, the size of D_s is much larger than the size of D_t, $N_s \gg N_t$.

A domain can be represented by $D = \{\chi, P(X)\}$, which contains the feature space χ and edge probability distribution $P(X)$ where $X = \{x_1, \cdots, x_n\} \in \chi$.

A task can be represented as $T = \{y, f(x)\}$. It consists of a label space y and target prediction function $f(x)$, which can be regarded a conditional probability function $P(y|x)$.

This study used a pre-trained AlexNet CNN as the source domain, which was fine-tuned to classify the new set of images, trained on more than 1 million images, and classified images into 1000 object categories. The network learns the rich features of a large number of images. The labels of the objects in the image were taken as inputs and the probability of each object class as the outputs. Using a pre-trained network as a starting point for learning a new task, fine-tuning the network using transfer learning is usually much faster and easier than training the network from scratch using random initial weights. This also allows the experimenter to quickly transfer the learned features to a new task using a smaller number of training images [37].

AlexNet is a convolutional neural network that is eight layers deep. A total of 25 layers are available, as shown in Table II [38].

The first layer is the image input layer, which requires an input image size of 227×227×3, with a number of colour channels. Therefore, the input data must be processed accordingly during database preprocessing to match the AlexNet input data structure. The last three layers of the pretrained network were configured for 1000 classes. These three layers must be fine-tuned to solve a new classification problem. The new network extracts all but the last three layers from the pretrained network. Then, the last three layers were transferred to a new classification task by replacing the original AlexNet with the last three layers with fully connected layers, the SoftMax layer, and the classification output layer. According to the facial expression recognition classification problem, the new output should be an expression label for one of the seven basic expressions (‘anger’, ‘disgust’, ‘fear’, ‘happy’, ‘neutral’, ‘sad’ and ‘surprise’). The category data of the fully connected layer (fc8) must be reset to seven categories.

TABLE II. STRUCTURE OF THE ALEXNET

	Name	Type	Activations	Learnables	
1	data	Image Input	227×227×3	-	227×227×3 images with ‘zerocenter’ normalisation
2	conv1	Convolution	55×55×96	Weights 11×11×96 Bias 1×1×96	96 11×11×3 convolutions with stride [4 4] and padding [0 0 0 0]
3	relu1	ReLU	55×55×96	-	
4	norm1	Cross Channel Normalization	55×55×96	-	cross channel normalisation with 5 channels per element
5	pool1	Max Pooling	27×27×96	-	3×3 max pooling with stride [2 2] and padding [0 0 0 0]
6	conv2	Grouped Convolution	27×27×256	Weights 5×5×48×128×2 Bias 1×1×128×2	2 groups of 128 5×5×48 convolutions with stride [1 1] and padding [2 2 2 2]
7	relu2	ReLU	27×27×256	-	
8	norm2	Cross Channel Normalization	27×27×256	-	cross channel normalisation with 5 channels per element
9	pool2	Max Pooling	13×13×256	-	3×3 max pooling with stride [2 2] and padding [0 0 0 0]
10	conv3	Convolution	13×13×384	Weights 11×11×96 Bias 1×1×96	384 3×3×256 convolutions with stride [1 1] and padding [1 1 1 1]
11	relu3	ReLU	13×13×384	-	

12	conv4	Grouped Convolution	13×13×384	Weights 3×3×256×384 Bias 1×1×384	2 groups of 192 3×3×192 convolutions with stride [1 1] and padding [1 1 1 1]
13	relu4	ReLU	13×13×384	-	
14	conv5	Grouped Convolution	13×13×256	Weights 3×3×192×128×2 Bias 1×1×128×2	2 groups of 128 3×3×192 convolutions with stride [1 1] and padding [1 1 1 1]
15	relu5	ReLU	13×13×256	-	
16	pool5	Max Pooling	6×6×256	-	3×3 max pooling with stride [2 2] and padding [0 0 0 0]
17	fc6	Fully Connected	1×1×4096	Weights 4096×9216 Bias 4096×1	4096 fully connected layers
18	relu6	ReLU	1×1×4096	-	
19	drop6	Dropout	1×1×4096	-	50% dropout
20	fc7	Fully Connected	1×1×4096	Weights 4096×4096 Bias 4096×1	4096 fully connected layers
21	relu7	ReLU	1×1×4096	-	
22	drop7	Dropout	1×1×4096	-	50% dropout
23	fc8	Fully Connected	1×1×1000	Weights 1000×4096 Bias 1000×1	
24	prob	Softmax	1×1×1000	-	softmax
25	output	Classification Output	1×1×1000	-	crossentropyex with 'tench' and 999 other classes

III. Experiment Setup

Several datasets containing images that capture facial expressions have been made publicly available to test and compare the developed FER algorithms and systems. One commonly used dataset is the Japanese Female Facial Expression (JAFFE) dataset. JAFFE consists of 213 images containing facial expressions that can be classified as "anger," "disgust," "fear," "happy," "sad," "surprise" and "-neutral". Each female subject shown in these images had three or four expressions that were rated by a group of 60 female observers. The observers in this study were asked to grade each facial expression on the scale of "1 to 5" for each of the first six expressions listed, and then the scores were averaged. Subsequently, the face was labelled as belonging to one of the six groups based on the dominant score for that particular expression. In the absence of a dominant score for any expression, the facial expression in the image was classified as the seventh, "neutral" expression. A short excerpt of the data provided by the JAFFE dataset is provided in Table III. A sample image from this database is shown in Fig. 2.

TABLE III. JAFFE Data-Base Information

Image No.	HA	SA	SU	AN	DI	FE	Figure Label
7	5.00	1.13	1.26	1.10	1.10	1.06	HA
8	4.65	1.29	1.39	1.23	1.16	1.16	HA
9	1.42	4.00	1.55	2.39	3.26	3.03	SA
10	1.23	4.39	1.45	2.61	3.19	2.71	SA
11	1.32	4.00	1.87	2.60	3.77	3.19	SA
12	1.26	4.29	1.58	2.26	3.39	2.94	SA

This study aimed to investigate the contribution and importance of different sections of the human face to overall facial expression and automatic expression detection. Three different sections of the faces available in the JAFFE database were extracted from each image to achieve this aim. In addition, automatic classification, that is, the FER system, was applied to associate each section with a corresponding expression. Fig. 2 shows the following sections: (i) eyes, (ii) nose and mouth, and (iii) mouth and chin regions using an image from the JAFFE database. A comparison to the "ground truth" − classification labels supplied with the JAFFE database was performed, and the accuracy of classification for each part of the face was calculated. Various sizes of each extracted facial region were considered to better assess the importance of accurate estimation and details of each section for classification accuracy.

Three sections of each image from the database were considered to estimate the amount of information contained in different parts of the facial image related to facial expressions: (i) eyes, (ii) nose and mouth region, and (iii) mouth and chin region. Tests also included different combinations of these regions in determining the extent to which recognition accuracy can be improved by increasing the amount of information provided for recognition. Two different databases, JAFFE and CK+, were used in this experiment, but the images from these databases were not mixed. Cross-database FER is a separate and complex problem described in the literature. A recent study [39] has confirmed the significantly inferior performance of the FER system in this setup. Sample images from the two databases used in the experiments reported in this section are shown in Fig. 5.

a) JAFFE image b) CK+ image

Figure 5. Sample images from two databases.

The JAFFE [5] database consists of facial images of Asians, specifically Japanese females. The JAFFE database contains 10 subjects and three to five images for each of the seven expressions from each subject. The JAFFE database includes 213 static images.

The CK+ database contains facial images from western (Caucasian) populations. The CK+ database [40, 41] consists of 593 expression sequences from 123 subjects, where 327 sequences are labelled with one of the seven expressions (angry, disgust, fear, happy, sad, surprise, and contempt). The 123 subjects were from different regions with varying races, ages, and gender. Each image sequence contains a set of captured frames in which the subjects change their expressions from a neutral emotional state and end at the peak expression. The neutral frame and four peak frames of each sequence were selected from 327 labelled sequences. On the basis of the balance among the eight expressions, each expression included 100 images. Different from that in JAFFE database, where seven facial expressions were applied, the same facial expressions (except for contempt) were used for the CK+ database. Therefore, the CK+ database included 700 images. In addition to certain structural differences, cultural differences exist between the two databases. Thus, the two databases used vary significantly in type, including the number of available images and faces.

Images from these two databases were used to extract important facial sections for the FER. In addition, tests were performed using five different sizes of extracted facial regions. The extracted sections of different sizes are labelled as Masks 1−5 and are shown on a sample face image in Fig. 6. The entire face, shown at the top of this figure and labelled Mask 0, was also used in the FER system tests as an additional check for the achieved results. As an illustration of the images used for recognition, a set of Mask 1 — "eyes" regions from the JAFFE database is shown in Fig. 7; each row in this figure correspond to one of the standard facial expressions contained in this database.

The results in a pervious study [39] indicated that the eyes, nose, and mouth can have different influences on the results of the FER. For further research, three regions, namely, eyes and eyebrows, nose and mouth and mouth and chin, were specially selected. Different regions were intercepted in different sizes to observe the effect of region changes on FER results and to search for each region where important features are located.

The eye region is dominated by the eyes and contains eyebrows. Previous results showed that the feature information provided by the nose area is limited, so the nose and mouth areas are reduced and centred on the upper lip area. Lastly, the mouth and chin areas are reduced and centred on the lips. The influence on the FER results was explored through the extraction and analysis of facial features in different regions and areas.

The local region must be extracted according to the landmark coordinates to facilitate the comparison and quantitative analysis of the data.

The local region of the face was extracted based on the original image of the database and 68 landmarks in the Dlib library. Rectangular interception is performed; therefore, the coordinates of each of the four vertices are set as (x1, y1), (x2, y1), (x1, y2) and (x2, y2). (x1, y1) represents the coordinates of the upper left point, (x2, y1) indicates the coordinates of the upper right point, (x1, y2) represents the coordinates of the lower left point and (x2, y2) indicates the coordinates of the lower right point.

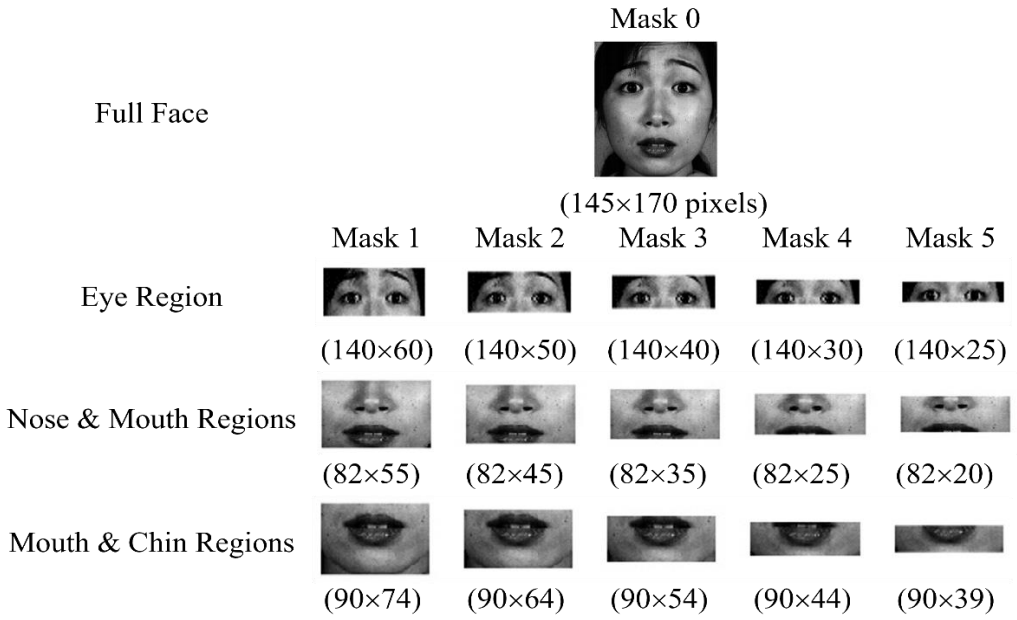


Figure 6. Facial sections used for tests.

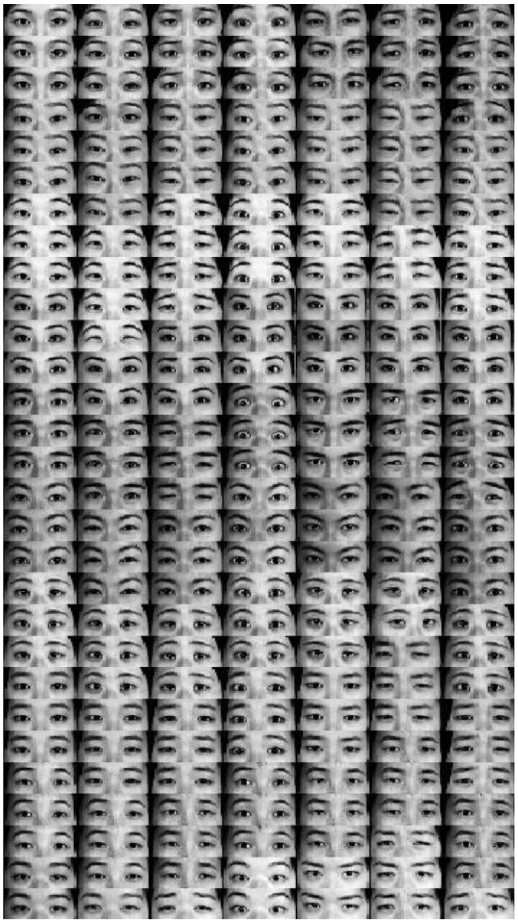


Figure 7. Eye region for all JAFFE images, sorted according to standard facial expression.

Local facial region extraction uses the coordinates (x1, y1) and (x2, y2). The width of the extracted region is defined as W (W = x2 − x1) and the height of the extracted region is defined as H (H =y2 − y1). The three local regions of the face were extracted and further processed according to five masks of different sizes in each local region, as shown in Fig. 6. The experimental input data were obtained, namely, five masks for the eyes, nose, mouth, mouth, and chin.

Following the eye region in Fig. 6, the extraction range was based on landmarks 1, 17, 20, and 30. Specifically, the abscissa of landmark 1 is x1 = 1x, the ordinate of landmark 20 is y1 = 20y, the abscissa of landmark17 is x2 = 17x and the ordinate of landmark 30 is y2 = 30y. Rectangle extraction was performed according to Table IV Eye region.

Regarding the nose and mouth regions in Fig. 6, the extraction range was based on landmarks 6, 29, 37, and 46. Specifically, the abscissa of landmark 37 is x1 = 37x, the ordinate of landmark 29 is y1 = 29y, the abscissa of landmark46 is x2 = 46x and the ordinate of landmark 6 is y2 = 6y. Rectangle extraction was performed according to Table IV Nose and Mouth Regions.

Following the mouth and chin regions shown in Fig. 6, the extraction range was based on landmarks 5, 9, 13, and 34. Specifically, the abscissa of landmark 5 is x1 = 5x, the ordinate of landmark 34 is y1 = 34y, and the abscissa of landmark 13 is x2 = 13x and the ordinate of landmark 9 is y2 = 9y. Rectangle extraction was performed according to Table IV: Mouth and Chin Regions.

TABLE IV. DETAILS OF THE LOCAL FACIAL REGIONS EXTRACTION

	Eye Region	Nose & Mouth Regions	Mouth & Chin Regions
(x1, y1)	(1x, 20y − 10)	(37x, 29y)	(5x, 34y)
(x2, y2)	(17x, 30y)	(46x, 6y)	(13x, 9y)
W1, W2, W3, W4, W5	17x − 1x	46x − 37x	13x − 5x
H1	30y − 20y + 10	6y − 29y	9y − 34y
H2	H1×50/60	H1×64/74	H1×45/55
H3	H1×40/60	H1×54/74	H1×35/55
H4	H1×30/60	H1×44/74	H1×25/55
H5	H1×25/60	H1×39/74	H1×20/55

IV. RESULTS

Fig. 8 illustrates the performance of the PCA and 2DPCA algorithms when applied to a full-face image (Mask 0 in Fig. 6) using the) JAFFE and CK+ databases. Although the accuracy of the PCA based algorithm increases significantly when the number of principal components (PCs) increases from 2 to 30, the 2DPCA-based algorithm is less dependent on the number of principal components and achieved an accuracy of approximately 92% when applied to the JAFFE and 98% for the CK+ database, regardless of the number of retained PCs. For this reason, the 2DPCA algorithm using only six PCs was selected and used to evaluate the performance of FER in this work when different sections of the facial image were used to recognise the emotional state of the person in the image.

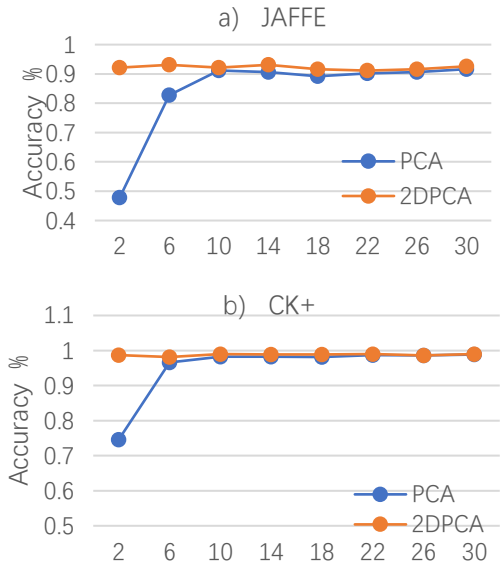


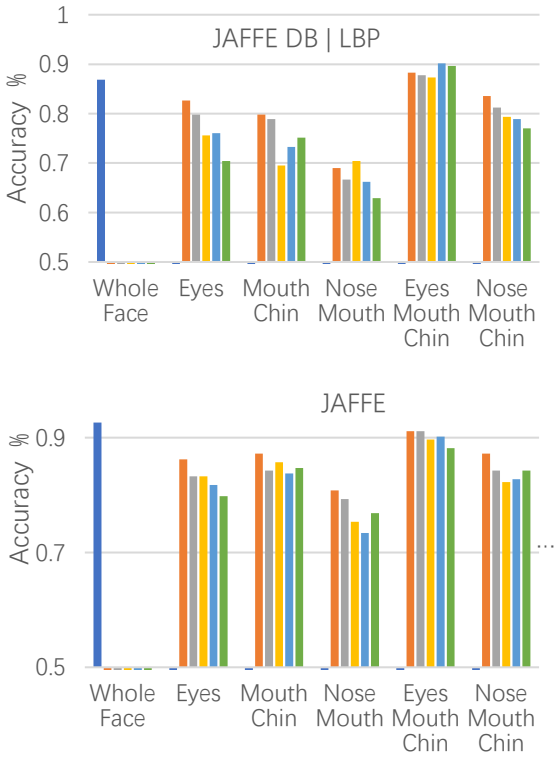
Figure 8. PCA versus 2DPCA − effect of the number of used PCs for a) JAFFE sections and b) CK+ sections (Mask 0).

Fig. 8 illustrates the results obtained using the facial sections extracted from the JAFFE images shown in Fig. 6. According to this set of results, the eye region and the mouth and chin regions seem to "hold" the most clues to a person's emotions. The accuracy achieved using only the nose and mouth region decreased by 3%-20% (in Mask 1) compared with that using the eye region alone. Compared with the eye region, the recognition accuracy did not change by more than 3% when using the mouth and chin regions in Mask1.

Notably all three algorithms (LBP, 2DPCA and CNN) achieved relatively similar recognition rates and behaved similarly for each facial section. The combined recognition accuracy of the eye region and mouth and chin regions was relatively high. Furthermore, the nose and mouth regions had the lowest recognition accuracy.

Combining the eye regions with the mouth and chin regions could achieve a significantly higher recognition rate. When using the LBP algorithm and AlexNet deep learning method, the combined result of the eye region and the mouth and chin regions in Mask 4 was significantly higher than the recognition accuracy of the whole face (i.e., Mask 0 in Fig. 6). For the PCA algorithm, the combined result of the eye region in Mask 4 and the mouth and chin regions in Mask 4 is 90%, which is only approximately 2% lower than the recognition accuracy of the entire face (i.e., Mask 0 in Fig. 6).

Simultaneously, the nose and mouth region results from the three algorithms (LBP, 2DPCA, and CNN) were analysed. The general trend is that as the facial expression information of most of the mouth region is lost, recognition accuracy also decreases. Similar results were obtained from the test when the nose and mouth regions and mouth and chin regions were combined.



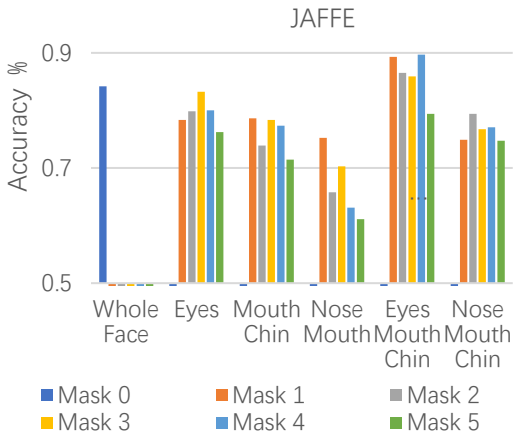
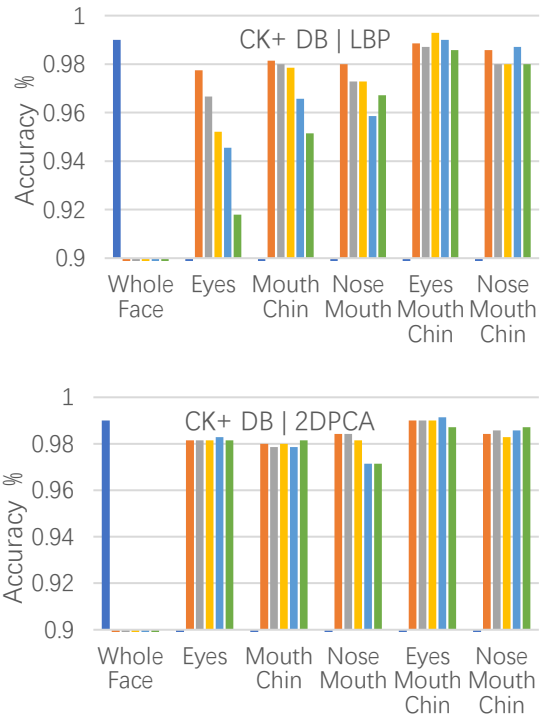


Figure 9. LBP, 2DPCA and CNN-based FER system accuracy applied to JAFFE sections.

The same set of results achieved using images, that is, sections of images from the CK+ database, is shown in Figure. These results relatively confirm the findings of the experiments conducted on the JAFFE database, although the FER accuracy is different for the same method. This result can be attributed to the significantly larger database, that is, a larger number of images available for training in the CK+ database.

Fig. 10 illustrates the results obtained using the face sections extracted from the CK+ image shown in Fig. 6. According to this set of results, the eye region and the mouth and chin regions "hold" most information about a person's emotion.



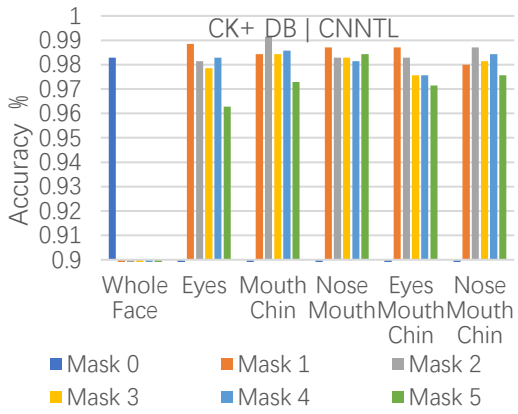


Figure 10. LBP, 2DPCA and CNN based FER systems accuracy applied to CK+ sections.

When combining the eye region with the mouth and chin regions, a significantly higher recognition rate was achieved. In many cases, however, this recognition rate was achieved when the entire face was analysed (i.e., Mask 0).

In the case of using the three regions independently, the mask is gradually reduced, and the main trend of the test results is to decline, particularly for the LBP algorithm. In the eye region, the eyes contain more information than the eyebrows. The accuracy dropped approximately 6% when only the eyebrows were extracted from the eye region. After combining the eyes, mouth, and chin regions and then the nose, mouth, and chin regions, the results of the corresponding masks have improved. The LBP algorithm has improved by approximately 3%, the 2DPCA algorithm has improved by less than 2%, and the CNN with transfer learning algorithm has changed by approximately 1%. Although the change in results is relatively smaller than that of JAFFE dataset, all three applied algorithms can still achieve significant improvements in results, with an overall recognition rate of more than 93%.

Notably, the recognition accuracy of the CNN-based FER system further increased, thereby confirming the notion that for larger training sets, CNN-based systems tend to outperform other techniques and approaches.

Therefore, this result indicates that for the machine-based recognition system, the region of the eyes carries most of the useful information regarding the emotional state of a person. The 'mouth and chin' regions also seem to be important for FER. However, the 'nose and mouth' regions are comparatively less important for FER tasks, particularly when the number of training images is limited.

V. Conclusion

According to previous studies, some parts of the human face may more clearly reveal a person's emotional state, and should, therefore, be considered more carefully in FER applications. The present study focuses on which part of the human face has the greatest influence on facial expression from the FER aspect. It aims to determine the contribution and importance of different parts of the face to the overall facial expression and automatic FER. Three main regions of the face were extracted from two different facial databases for each facial image to achieve this objective. In addition, automatic classification was used to associate each part with the corresponding expression. Three different well-established methods were used for these FER tasks: LBP combined with SVM, 2DPCA, and CNN.

The eye region, nose and mouth area, and mouth and chin area were selected for FER testing. The results obtained in these experiments show that although all three regions contain useful information, as shown in Figs. 9 and 10, the eye region contains relatively more expression information. However, as the range narrows, the amount of useful information in the eyebrow-only range decreases. Moreover, the mouth area contains relatively more expression information. Through analysis of the MC and NM areas, the upper lip area can be guided to contain a more effective expression information than other areas. This result shows that in FER, the eyes and mouth can be selected as the main features to replace all facial features for recognition to reduce the computational difficulty. Simultaneously, the accuracy of FER can be increased by combining different regional features to match all facial features. However, the 'eye' region seems to contain useful information for FER, whereas the accuracy decreases significantly with the reduction of the selected region and subsequent absence of the most important part. The second most important region for the FER task appears to be the 'mouth' region, specifically the 'upper lip' region. When combining this region with other less important parts of the face (i.e., the 'nose and mouth' region and the 'mouth and chin' region), only marginal improvements can be achieved.

Conflict of Interest

The authors declare no conflict of interest. Given his role as Editor-in-Chief, Branislav Vuksanovic had no involvement in the peer-review of this article and has no access to information regarding its peer-review.

Author Contributions

Yining Yang, Branislav Vuksanovic and Hongjie Ma conducted the research; Branislav Vuksanovic provided the paper structure. Yining Yang and Branislav Vuksanovic wrote the paper. Hongjie Ma reviewed the paper several times and suggested various modifications and additions. All authors had approved the final version.

References

[1] M. Pantic and L. Ü. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans Pattern Anal Mach Intell*, vol. 22, no. 12, pp. 1424–1445, 2000.

[2] J. Suthar and N. Limbad, "A literature survey on facial expression recognition techniques using appearance based features," *International Journal of Computer Trends and Technology*, vol. 17, no. 4, pp. 161–165, 2015. doi: 10.14445/22312803/ijctt-v17p131

[3] J. Kumari, R. Rajesh, and K. M. Pooja, "Facial expression recognition: A survey," *Procedia Comput. Sci.*, vol. 58, pp. 486–491, 2015. doi: 10.1016/j.procs.2015.08.011

[4] I. M. Revina and W. R. S. Emmanuel, "A survey on human face expression recognition techniques," *Journal of King Saud University - Computer and Information Sciences*, 2018. doi: 10.1016/j.jksuci.2018.09.002

[5] S. Li and W. Deng. (2018). Deep facial expression recognition: A survey. [Online]. pp. 1–25. Available: http://arxiv.org/abs/1804.08348

[6] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-Emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, pp. 1–16, 2021. doi: 10.3390/s21093046
[7] Z. Zhao, Q. Liu, and S. Wang, "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 6544–6556, 2021. doi: 10.1109/TIP.2021.3093397
[8] A. D. Sergeeva, A. V. Savin, V. A. Sablina, *et al.*, "Emotion recognition from micro-expressions: Search for the face and eyes," in *Proc. 2019 8th Mediterranean Conference on Embedded Computing (MECO)*, 2019, pp. 8–11.
[9] I. Tautkute, T. Trzcinski, and A. Bielski, "I know how you feel: Emotion recognition with facial landmarks," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1959–1961. doi: 10.1109/CVPRW.2018.00246
[10] A. P. Lim and A. Zahra, "Facial emotion recognition using computer vision," in *Proc. Indonesian Association for Pattern Recognition International Conference (INAPR)*, 2018, pp. 46–50.
[11] P. Sinha, B. Balas, Y. Ostrovsky, *et al.*, "Face recognition by humans: Nineteen results all computer vision researchers should know about," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1948–1962, 2006.
[12] M. Wegrzyn, M. Vogt, B. Kireclioglu, *et al.*, "Mapping the emotional face. How individual face parts contribute to successful emotion recognition," *PLoS One*, vol. 12, no. 5, pp. 1–15, 2017. doi: 10.1371/journal.pone.0177239
[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, vol. 1. doi: 10.1109/CVPR.2001.990517
[14] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996. doi: 10.1016/0031-3203(95)00067-4
[15] N. T. Cao, A. H. Ton-That, and H. Il Choi, "Facial expression recognition based on local binary pattern features and support vector machine," *Intern. J. Pattern Recognit. Artif. Intell.*, vol. 28, no. 6, 1456012, 2014. doi: 10.1142/S0218001414560126
[16] M. J. Lyons, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 12, pp. 1357–1362, 1999. doi: 10.1109/34.817413
[17] A. P. M. Turk, "Face recognition using eigenfaces," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991. doi: https://doi.org/10.1162/jocn.1991.3.1.71
[18] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *Factor Analysis*, vol. 13, no. 6, pp. 1450–1464, 2002. doi: 10.4324/9781315799476-12
[19] O. Barkan, J. Weill, L. Wolf, *et al.*, "Fast high dimensional vector multiplication face recognition," in *Proc. the IEEE International Conference on Computer Vision*, 2013, pp. 1960–1967. doi: 10.1109/ICCV.2013.246
[20] J. Yang and C. Liu, "Horizontal and vertical 2DPCA-based discriminant analysis for face verification on a large-scale database," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 4, pp. 781–792, 2007. doi: 10.1109/TIFS.2007.910239
[21] T. Jabid, M. H. Kabir, and O. Chae, "Facial expression recognition using Local Directional Pattern (LDP)," in *Proc. 17th IEEE International Conference on Image Processing*, 2010, pp. 1605–1608. doi: 10.1109/ICPR.2010.373
[22] T. Jabid, M. H. Kabir, and O. Chae, "Local Directional Pattern (LDP) – a robust image descriptor for object recognition," in *Proc. 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010, pp. 482–487. doi: 10.1109/AVSS.2010.17
[23] G. Zhao, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007. doi: 10.1109/TPAMI.2007.1110
[24] B. Sun, L. Li, G. Zhou, *et al.*, "Combining multimodal features within a fusion network for emotion recognition in the wild," in *Proc. the 2015 ACM on International Conference on Multimodal Interaction*, 2016, pp. 497–502. doi: 10.1145/2818346.2830586
[25] Y. Gizatdinova, V. Surakka, G. Zhao, *et al.*, "Facial expression classification based on local spatiotemporal edge and texture descriptors," in *Proc. the 7th International Conference on Methods and Techniques in Behavioral Research*, 2010, pp. 1–4. doi: 10.1145/1931344.1931365
[26] M. Dahmane and J. Meunier, "Prototype-Based modeling for facial expression analysis," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1574–1584, 2014.
[27] A. Poursaberi, H. A. Noubari, M. Gavrilova, *et al.*, "Gauss-Laguerre wavelet textural feature fusion with geometrical information for facial expression identification," *EURASIP J. Image Video Process*, vol. 2012, no. 17, pp. 1–13, 2012. doi: 10.1186/1687-5281-2012-17
[28] B. Sun, L. Li, G. Zhou, *et al.*, "Facial expression recognition in the wild based on multimodal texture features," *J. Electron. Imaging*, vol. 25, no. 6, 061407, 2016. doi: 10.1117/1.jei.25.6.061407
[29] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proc. the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 503–510. doi: 10.1145/2818346.2830587
[30] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009. doi: 10.1016/j.imavis.2008.08.005
[31] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002. doi: 10.1109/TPAMI.2002.1017623
[32] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000, pp. 46–53. doi: 10.1109/AFGR.2000.840611
[33] K. Young-gil and S. Young-jun, "Face recognition using wavelet transform and 2D PCA," in *Proc. the Korea Contents Association Conference*, 2004, pp. 348–351.
[34] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proc. ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 17–37.
[35] Y. Fan, J. C. K. Lam, and V. O. K. Li, "Multi-region ensemble convolutional neural network for facial expression recognition," in *Artificial Neural Networks and Machine Learning*, Springer International Publishing, 2018. doi: 10.1007/978-3-030-01418-6
[36] C. Tan, F. Sun, T. Kong, *et al.*, "A survey on deep transfer learning," in *Artificial Neural Networks and Machine Learning*, Springer International Publishing, 2018. doi: 10.1007/978-3-030-01424-7
[37] Transfer Learning Using AlexNet - MATLAB & Simulink - MathWorks United Kingdom. [Online]. Available: https://uk.mathworks.com/help/deeplearning/ug/transfer-learning-using-alexnet.html
[38] G. E. Hinton. A. Krizhevsky, and I. Sutskever, "Imagenet classification with deep convolutional neural networks," *Neural Inform.*, pp. 1097–1105, 2012. doi: 10.1201/9781420010749
[39] Y. Yang, B. Vuksanovic, and H. Ma, "Effects of region features on the accuracy of cross-database facial expression recognition," in *Proc. the 12th International Conference on Agents and Artificial Intelligence*, 2020, pp. 610–617. doi: 10.5220/0008966306100617
[40] P. Ekman, "Strong evidence for universals in facial expressions a reply to Russell's mistaken critique," *Psychol. Bull.*, vol. 115, no. 2, pp. 268–287, 1994. doi: 10.1037/0033-2909.115.2.268
[41] P. Lucey, J. F. Cohn, T. Kanade, *et al.*, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010, pp. 94–101. doi: 10.1109/CVPRW.2010.5543262