Received: October 9, 2023. Revised: November 12, 2023.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Geometrical Facial Expression Recognition Approach Based on Fusion CNN-SVM

Abbas Issa Jabbooree[1] **Leyli Mohammad Khanli**[1]* **Pedram Salehpour**[1]
Shahin Pourbahrami[2]

[1] *Department of Computer Engineering, Faculty of Electrical and Computer Engineering,*
University of Tabriz, Iran
[2] *Department of Computer Engineering, Technical and Vocational University (TVU), Tehran, Iran*
* Corresponding author's Email: psalehpoor@tabrizu.ac.ir

Abstract: Facial expression recognition (FER) effectively enhances human-computer interaction, robot interfaces, and emotion-aware smart agent systems. Most existing FER algorithms focus solely on extracting facial features from the pixels of the face, disregarding the relative geometric positions that depend on facial landmark points. However, these approaches need to be revised regarding accuracy and robustness. This paper introduces a novel approach to FER that combines convolutional neural network (CNN) and support vector machine (SVM) techniques to extract hybrid features, thereby enhancing discriminative power. The proposed system employs the β-skeleton undirected graph for improved geometric feature extraction through several pre-processing stages. A 1D-CNN is utilized for training the geometric features while simultaneously training the same image using a 2D-CNN. The resulting feature vectors from both sub-networks are merged for SVM classification. The performance of the proposed system is evaluated on two widely used datasets: the extended Cohn-Kanade (CK+) dataset and the Japanese female facial expression (JAFFE) dataset, which are commonly employed in face recognition research. Experimental results demonstrate the superiority of the proposed system over existing methods, achieving a recognition accuracy of 96.19% on the CK+ dataset and 89.23% on the JAFFE dataset.

Keywords: Fusion approach, β-Skeleton, Geometry features, CNN, SVM.

1. Introduction

Facial expression recognition plays a significant role in daily human communication by conveying nonverbal signals that reflect emotional states. The FER system has gained substantial attention in various research fields due to its wide range of applications. Some examples are the field of the field of computer vision, and the study of human behaviour in massive multimedia datasets [1, 2, 3]. However, when analyzing facial expressions, focusing on the most distinctive features associated with the expression is crucial while disregarding irrelevant facial information such as hair, glasses, and background clutter [4].

Accurate facial expression discrimination remains difficult for FER performance. FER allows intelligent machines to understand human behaviour, and a wide range of useful applications, such as psychological diagnostics, electronic entertainment, and mental healthcare, are achieved [5]. Traditional methods aim to enhance facial expression recognition features. For mobile internet applications, real-time performance and lightweight networking are becoming increasingly important [6]. Several approaches address the FER problem by focusing only on local features, such as Gabor filter with a genetic algorithm [7] and local binary pattern LBP with ORB features [8]. All these methods adopt SVM for the classification process and extract facial features from facial pixels that do not consider relative geometric position features [4].

Geometric information relative to position and self-deformation of facial components can also express emotional states from facial observations. Facial landmarks represent geometric facial knowledge [9], which inherently characterizes the geometric information of facial components in a graph structure. Moreover, the geometric facial description is preferable in practical FER applications because it is more resistant to variations in appearance caused by factors like lighting, spatial noise, and identity information [5]. This paper introduces a fusion approach based on two CNN to improve facial expressional recognition accuracy. Our approach considers two categories of facial features: appearance-based pixel features and geometric-based features. Using appearance-based techniques for feature extraction yields more reliable results, avoiding incorporating sensitive skin texture patterns [10]. Our method uses an input image, particularly one capturing a facial expression, to determine the region of interest (ROI). To enhance the process of feature extraction and facial landmark detection, multiple pre-processing steps are applied.

Furthermore, a novel grouping technique was devised to categorize the seven sets of facial landmarks according to their specific locations on the face [4]. These landmark groups are a basis for extracting geometric features using the β-skeleton undirected graph. The extracted geometric features are then trained using a 1D-CNN. Simultaneously, the same image undergoes training in a 2D-CNN. In the final step, the two sub-networks are integrated, with the input features from both networks concatenated along the axis to form a vector. This vector is subsequently fed into an SVM for the classification process. In this work, the SVM classification method replaces the top-level structure of the CNN classifier (fully connected feedforward neural network) to improve the expression classification accuracy further [11, 12]. The facial expression labels are; anger, happiness, fear, sadness, disgust, neutral, and surprise [13].

This paper employs a CNN as the baseline model. It conducts experiments to compare its performance with the Fusion-CNN-SVM model, specifically analyzing the impact of the β-skeleton method. The key contributions of the paper can be summarized as follows:

1) Introducing a novel fusion approach that combines 1D-CNN for training geometry features, 2D-CNN for appearance features, and SVM for classification.

2) Enhancing facial expression recognition by incorporating geometrical feature extraction based on the β-skeleton undirected graph.

The remaining content of the paper is organized as follows: Section 2 provides a concise review of related works in the field of FER. Section 3 presents a detailed explanation of the proposed system framework. Section 4 discusses the implementation of experiments, obtained results, and performance comparisons with the latest FER methods. The last section concludes the paper and outlines potential future work.

2. Related work

In this section, FER-related works based on facial feature extraction are presented. Facial feature extraction and recognition are two difficult areas of study in facial expression recognition. There are two types of facial feature extraction appearance-based and geometrical-based. The following is a brief description of each approach used in related works.

2.1 Appearance feature based FER approaches

This type of features refers to a statistical depiction of the pixel values present in a facial image. Adil Boughida Durga et al. [7] introduce a novel approach for FER that integrates Gabor filters with a genetic algorithm (GA). The method involves extracting Gabor features from a human face, specifically targeting the region of interest determined by landmarks. GA is employed to select and classify optimal features using SVM. The Gabor filter parameters were chosen manually, which may not guarantee optimal values. The authors in [14] propose a model based on a multilayer Maxout network linear activation function to initialize the CNN and LSTM techniques. The facial expression evaluation is based on a single-frame image and historical-related information. It was utilized to extract related information from the image frame. It uses the SVM classifier to improve the expression classification accuracy. The proposed method can accurately identify some expressions but struggles with neutral and angry expressions.

To enhance the FER functionality, the authors propose a local binary pattern (LBP) feature extraction [15]. In this method, the length of the feature vector is shortened, and the noise is removed. Then use an SVM for the classification process. Mobile internet applications have a growing demand for fast and lightweight networking. The authors of [6] introduce a lightweight A-MobileNet model to enhance local feature extraction of facial expressions. The center and softmax losses are then used to adjust the model parameters to reduce and maximize intra-class distances. This technique outperforms the original MobileNet series and other models due to its improved recognition accuracy without additional model parameters. The proposed method does not provide a detailed analysis of the misrecognized expression data and the reasons behind it. In [16], a novel FER technique using infrared technology is proposed to assess classroom non-verbal behaviours. This approach involves multi-label distribution learning, employing the Cauchy distribution-based label learning (CDLLNet) instead of conventional single-expression labels. These revised labels enable accurate facial expression recognition for natural and adjacent expressions using only one infrared image. These revised labels enable accurate facial expression recognition for natural and adjacent expressions using only one infrared image.

2.2 Geometrical features based FER approaches

These traits indicate the shape and spatial organization of the face and its elements, such as the lips and eyebrows. Based on this, the stability indicators provided by distance and shape signatures are important to the FER performance [10]. The pair of normalized distance and shape signature is then used to extract other features, such as statistical measurements like moment, range, and entropy. Specifically, this enhanced feature set is transmitted to the multilayer perceptron (MLP) so that it may have access to different classes of expressions. This study is the need for more extensive evaluation across a wider range of facial expression datasets. The authors in [17] proposed a FER system based on discrete wavelet transform with fuzzy combinations. Facial geometry was then determined using a new version of the Eyemap and Mouthmap algorithms. The area and angle of the triangle are then defined and classified by a neural network. The proposed method mentions that there were variations in the detection of facial geometry when a person closed their eyes, indicating a potential limitation in accurately FER in such scenarios. Quang Tran Ngoc et al. [18] introduced the FER system based on directed graph neural networks (DGNN), utilizing landmark features. In this approach, the landmarks are nodes within the graph structure, while the edges are constructed using the delaunay method in a directed graph. The performance of the proposed method depends on the accuracy of the landmarks, and it can be impossible to extract facial landmarks from certain images. Furthermore, noise can affect the performance of the proposed method, and as noise increases, the overall performance decreases.

2.3 Fusion-based FER approaches

Several methods used to construct a learner in conventional FER research to extract features from the image input to FER. This type of model can typically learn a large number of single-expression features from the facial image, but in rare circumstances, such as illumination changing and facial occlusion, its accuracy can even become unpredictable. The multimodal data fusion technique can make this scenario better. When the model cannot get all the facial features, it can still get them for prediction from other sources [19]. Authors in [12] use dynamic facial expression data with a hybrid model of CNN and SVM classifiers to create a novel FER framework. CNN and SVM classifiers used dense facial motion flows and geometry landmark flows of facial expression sequences to extract facial motion properties. The ideal hybrid classifier weighting combination outperforms individual classifiers in facial expression recognition.

Hao Wang et al. [20] present a FER technique that iteratively fuses multi-orientation gradient MO-HOG and deep-learned classifiers. It extracts MO-HOG features from whole and expression-rich local facial images. Deep-learned features are unreliable on small databases but include high-level semantic information. Therefore, the deep network extracts useful features. The proposed method adopts post-classifier fusion techniques by iterating a classifier fusion approach based on an optimization algorithm. Han Yi et al. [19] presented a multimodal system for FER problems. Their approach incorporates three distinct neural networks: CNN, LNN (neural network for landmarks), and HNN (neural network for histogram oriented gradient). These networks are employed to extract features from facial images. The proposed method offers improved accuracy and speed of detection compared to prior multimodal-based FER approaches. The FER method, which uses a pre-trained model for feature extraction and is based on CNN, was proposed by authors in [13]. It proposed three methods for creating a temporally aggregated feature vector: mean, standard deviation, and early fusion. According to the experiment results, the FER system successfully recognizes six universalist expressions and decreases slightly when dealing with micro-expression recognition (MER). Suparshya Babu Sukhavasi et al. [2] combined deep neural networks and SVM to create a unique hybrid network architecture to predict seven driver expressions in various poses, occlusions, and lighting situations. These systems help with good driving and estimate driver stability and road safety. Features have been discovered by fusing Gabor and

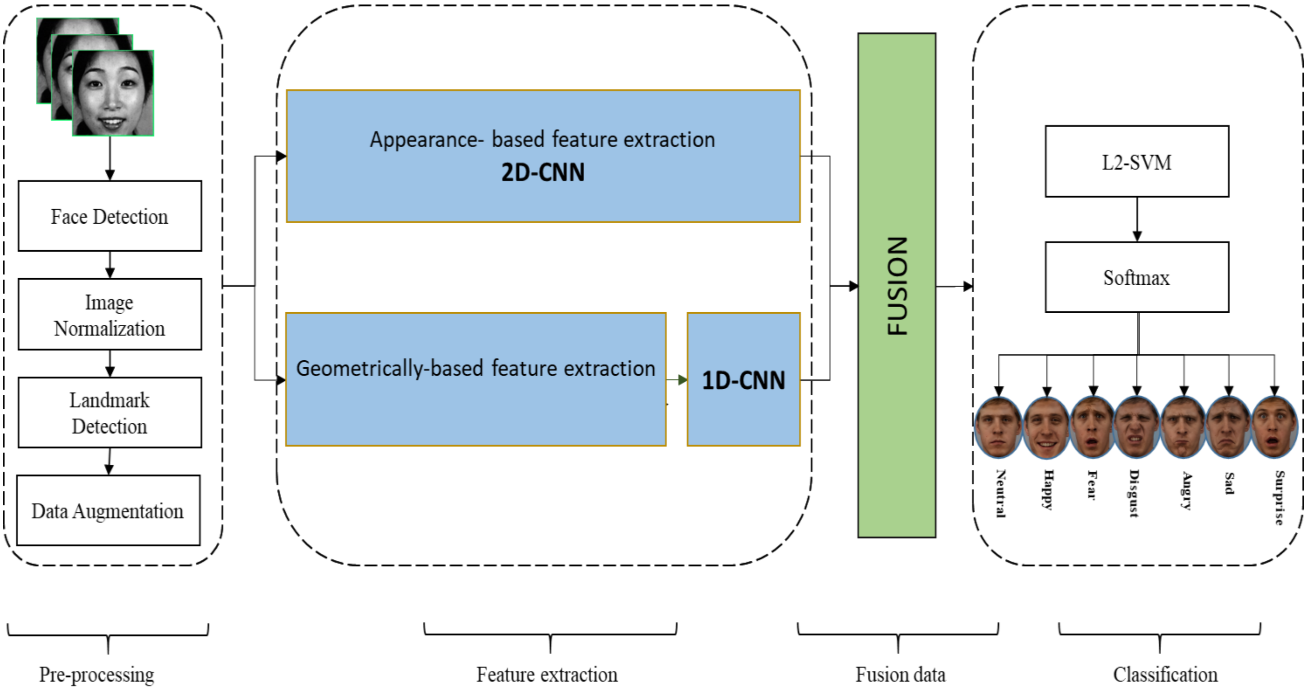


Figure. 1 The proposed Fusion-CNN-SVM FER architecture

LBP features and classified by combining an SVM classifier with CNN. The proposed approach is its inability to effectively detect high-intensity levels in the driver's emotions that may occur during challenging driving scenarios, such as mountain cliffs, deep valleys, and ridges. Additionally, the method faces limitations when drivers are wearing masks, as it was tested primarily on individuals without masks.

The above papers illustrate feature extraction using either pixel-based or geometrical-based methods. However, very few studies have examined the relationship between the geometric positions of facial features and pixels-based FER approaches. This paper presents a multimodal approach (geometrical and appearance) based on the geometrical β-skeleton using Fusion CNN-SVM to get a good FER performance.

3. Proposed system design

The proposed methodology is described in depth in this section. Pre-processing of input images is covered in detail in section 3.1. Section 3.2 gives details on geometric feature extraction. CNN, SVM, and the joint fusion classifier, as seen in Fig. 1, are discussed in detail in section 3.3.

3.1 Image pre-processing

Each image dataset can customize resolution, brightness, and pose. Fig. 1 displays all input processes in a sequence for data pre-processing, which must be general.

Face detection is an important step in facial expression recognition systems. After that, we removed unnecessary features, such as hair and other accessories, that could have hidden the subject face during cropping. Facial feature extraction is important in facial expression recognition. Unallocated faces might adversely affect recognition because more data will not be required when the image is fed into the model or feature extraction is applied [21]. Various face detection techniques are available; this paper employs the Viola-Joins Haar Cascade Algorithm [22]. The Viola-Jones detector is the best face-detection algorithm because of its high efficiency and fast detection [13].

There are two methods for normalizing an image; firstly, standardizing the image size. The input image or feature size must be constant for CNN to function properly. As a result, we adjusted the dimensions of every image to exactly 50×50 pixels. The convergence of the network is thus accelerated [23]. The second type is numerical image value adjustment. Normalized image pixels followed the Z-score standard normal distribution. Z-score normalization revealed a normal distribution with an average of 0 and a maximum deviation of 1 [24]. The 1-by-255 intensity normalization method can produce better results. Before feeding an image into a neural network, most deep-learning algorithms normalize it by 255 [25].

During image collection, variations in illumination, shadows, and other factors can result

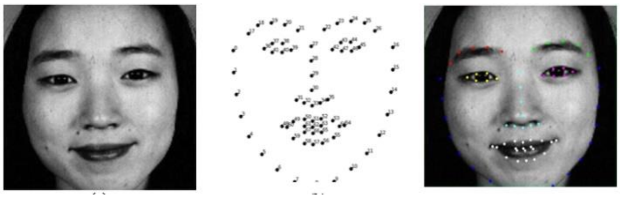


Figure. 2 Facial landmark detection process

in uneven light and shade distribution, itis a challenge in feature extraction. To address this issue, the gray image levels need to be equalized. Histogram equalization (HE) technique is employed in this paper, which utilizes the histogram graph to create an equal distribution [23]. In the facial landmark detection step, the Dlib model was adopted for automatic facial landmark detection [9]. The cascade suppressor estimates the positions of these landmarks using a sparse subset of image values. Through this process, 68 well-trained landmarks are generated from the recognized facial parts in the wild dataset. As shown in Fig. 2, the detected landmarks include the corners of the mouth, eyebrows, and eyes.

Deep learning techniques rely on significant training data to achieve effective results. To address this requirement, data augmentation techniques were employed to expand the FER dataset, thereby enabling the training of a CNN model without overfitting [3]. By leveraging the image data generator parameters, these techniques can generate multiple diverse images from a single input [17]. This study employed various image augmentation methods, such as rotating at different angles, flipping, Gaussian blur, and sharpening.

3.2 Using β-Skeleton for geometrical feature extraction

This method can be applied to facial expression analysis by representing facial landmarks as points in space, β-skeletons can capture the connectivity and shape of facial features [9]. Geometric data, encompassing the precise positioning and orientation of significant facial landmarks, can be extracted through this methodology. This data proves invaluable in characterizing a diverse spectrum of facial expressions. Upon identification of these facial landmarks, they are assigned Cartesian coordinates (x, y) within a two-dimensional framework. These landmarks, distributed across seven categories including the eyes, eyebrows, mouth, and nose, shed light on the intricate anatomy and physiology of the face [25]. Additionally, analyzing the lengths, branching points, and cycles within the facial expression β-skeleton makes it possible to extract meaningful information about facial expressions' intensity, complexity, and dynamics.

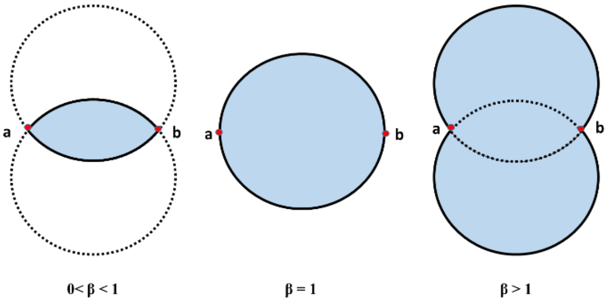


Figure. 3 β-skeleton approximation [28]

Existing graph-based neighbourhood algorithms, such as nearest neighbourhood and Gabriel graphs, often need explicit parameter definitions from graph theory. These algorithms describe neighbourhood areas primarily based on circular graphs, making optimization challenging [26, 27]. Extracting data point connections through these algorithms involves multiple phases and is time-consuming. In contrast, the β-skeleton graph improves graph accuracy and neighborhood grouping. It measures the empty neighborhoods around the edges to determine their structure [26]. The value of the β parameter plays a crucial role in defining the region where neighboring nodes are considered. A smaller β value results in more edges being plotted between nodes, while a larger value leads to the loss of edges in sub-graphs [26, 28]. Fig. 3 demonstrates the performance of this technique with different β values. In the β-skeleton plan, the neighboring graph area is determined, with the nearest neighbor being selected when β is greater than 1 [28].

The β-skeleton, based on Euclidean geometry, is a graph that does not have an un-direction [28]. Angles are crucial in defining the connections in geometric neighbourhood graphs, with β being a positive parameter, Eq. (1) using to calculate these angles [27]:

$$\theta = \begin{cases} \sin^{-1}\frac{1}{\beta} & \text{if } \beta \geq 1 \\ \pi - \sin^{-1} & \text{if } \beta < 1 \end{cases} \quad (1)$$

This method has some drawbacks, such as complex angle parameter calculations through trade-offs, but it also improves classification accuracy with adjustable neighbourhood area [26].

As shown in Fig. 4, the β-skeleton calculated using Cartesian coordinates for each pre-assembled face feature set. For instance, let's consider a feature vector v1 representing the left eyebrow, consisting of five facial characteristic points (x17leb, y17leb), (x18le, y18le),..., (x21leb, y21leb). Table 1 provides

Received: October 9, 2023. Revised: November 12, 2023.

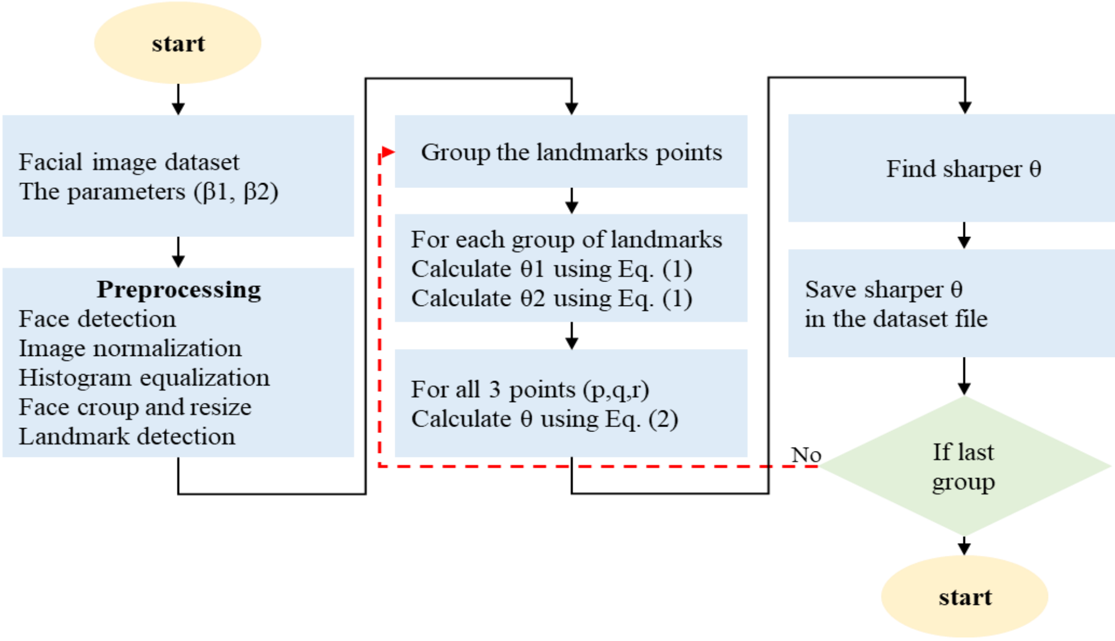
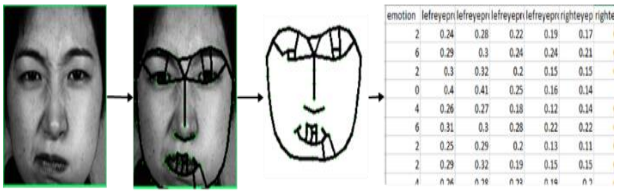


Figure. 4 Geometrical features extraction based on the β-skeleton flowchart

Table 1. Grouping strategy for facial landmark points

No.	Facial area	Group
1	Left-eyebrow (leb)	P17 to P21
2	Right-eyebrow (reb)	P22 to P26
3	Left-eye (le)	P36 to P41
4	Right-eye (re)	P42 to P47
5	Nose	P27 to P35
6	Mouth	P48 to P67
7	Jaw	P0 to P16



emotion	lefreyepr	lefreyepr	lefreyepr	lefreyepr	righteyep	righte
2	0.24	0.28	0.22	0.19	0.17	
6	0.29	0.3	0.24	0.24	0.21	
2	0.3	0.32	0.2	0.15	0.15	
0	0.4	0.41	0.25	0.16	0.14	
4	0.26	0.27	0.18	0.12	0.14	
6	0.31	0.3	0.28	0.22	0.22	
2	0.25	0.29	0.2	0.13	0.11	
2	0.29	0.32	0.19	0.15	0.15	
4	0.26	0.28	0.23	0.19	0.2	

Figure. 5 Geometrical feature extraction process based on β-skeleton

a comprehensive list of feature vectors representing these characteristics relative to the remaining face landmarks.

To determine the values of θ1 and θ2 based on the values of β1 and β2, we use Eq (1). While we use the cosines law [29] mentioned in Eq (2) to calculate the angle (θ) consisting of three points p, r, and q through the three sides a, b, and c formed between them.

$$\theta = \cos^{-1}(\frac{a^2+b^2-c^2}{2\times a\times b}) \tag{2}$$

We can identify whether the angle is sharper by comparing θ1 and θ2 with angle (θ), and the resulting information is recorded in a dataset file. The β-skeleton is computed for each cluster of facial features. Fig. 5 shows the calculation of the β-skeleton for each group, after which the groups are merged into a single vector.

3.3 The proposed CNN architecture

Unlike simple neurons, traditional CNNs in multi-layer neural N\networks (MLNNs) utilize convolutional kernels [30]. A CNN typically consists of multiple convolution layers, a max pooling layer, and a fully connected layer [30]. Each layer comprises multiple neuron-equipped 2D feature maps. The pooling layer, such as maximum, sum, or average pooling, is commonly employed to summarize and compress feature maps, thereby reducing network complexity [30]. In order to perform tasks like image classification, face alignment, and head posture prediction, CNNs apply pooling and convolution processes to the input layer, a 2D matrix of pixel values.

In our proposed architecture, a 2D-CNN is utilized for the appearance-based component. This design is specifically tailored to detect subtle changes in local facial characteristics, such as the positioning of the jaw and lips, which disproportionately impact conveyed expressions. The proposed 2D-CNN network design is described in Table 2 and shown graphically in Fig. 6.

1D-convolution kernels are used in a number of different domains, including speech recognition [31], human activities classification [32], and music genre classification [33]. Shorter 1D-CNN can effectively handle 1D data, but longer 2D-CNN typically

DOI: 10.22266/ijies2024.0229.40

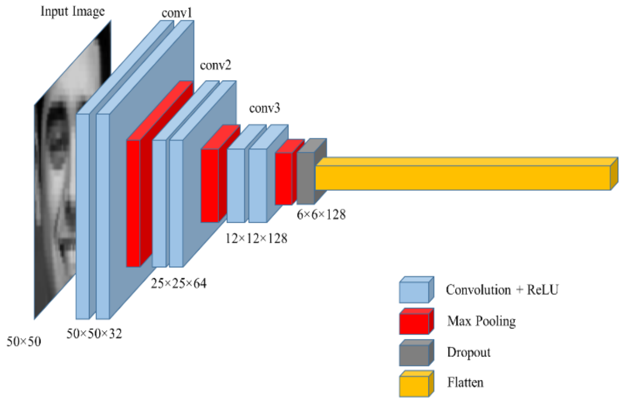


Figure. 6 The proposed 2D-CNN architectures

Table 2. The proposed 2D-CNN model architecture

2D-CNN model architecture	
Input	(50x50x1)
Conv1	(50x50x32)
Maxpool1	(25x25x32)
Conv2	(25x25x64)
Maxpool2	(12x12x64)
Conv3	(12x12x128)
Maxpool3	(6x6x128)
Dropout	(0.5)
Flatten layer	

require numerous layers to extract significant features from 2D inputs. Because of this, 1D-CNNs can be more easily implemented on CPUs and low-cost, real-time devices [34]. In this paper, the phase was determined by analyzing geometric data using a 1D-CNN. Three convolution layers, three pooling layers, a dropout layer with a rate of 0.25, and a flattening layer make up the 1D-CNN network's eight-layer structure. This layer's output is the addition of the first channel's output and the second channel's result. The combined output is then sent into a SVM for classification.

3.4 Data fusion architecture based on SVM

SVM is a classifier that uses statistical learning theory to minimize structural risk [9, 14]. It is very generalizable and highly sparse structurally. The following are some advantages of SVM over other machine learning techniques [14]: As a first benefit, SVM offers strong generalizability and solves issues with a small dataset. Second, the optimization target for the linear separable problem is one of two types of sample geometric interval maximizations. To make the nonlinearly separable problem linearly separable, the kernel function is included in the SVM. As a result, the complexity of the method remains unaffected by the size of the samples. Third, to further strengthen the robustness of the classification, the SVM employs soft interval technology to permit a small number of data to be misclassified while applying penalties on the misclassified samples. Finally, SVM is like three-layer feedforward neural network. The support vector determines the hidden layer nodes, while the global maximum point guarantees model convergence. It can choose the weight vector and hidden layer nodes.

Two improved algorithms are based on the Support Vector Machine model L1-SVM and L2-SVM. The difference from L1-SVM is that L2-SVM is differentiable and imposes greater quadratic loss and linear loss on points that violate the boundary [11].

Four multimodal data fusion methods have been investigated to identify the most effective combination. The first method involves fusing all the extracted and normalized multimodal feature vectors into a single feature vector, which is then used for classification. The second method averages the scores obtained from all the fused models. The third method selects the maximum scores from the fused models. The fourth method assigns appropriate weights to each model based on its performance [35]. This work uses the first method, which combines appearance-based with geometrical feature-based multimodal data sources (Fig. 1). The previous layer's output is merged into a vector for the L2-SVM classifier. Note that feature-level fusion generally yields better identification results than other approaches. Additionally, feature-level fusion incorporates additional input data [36]. The final step involves determining the facial expression label using each network's training dataset. Through these experiments, the effectiveness and reliability of the proposed method have been validated.

4. Experimental results and analysis

Experiments and data analysis are presented in this section. Specifically, a Core i5-10210U CPU at 2.11 GHz and 4GB of RAM were used in the testing scenario. The anaconda navigator IDE ran Python 3.6 and the deep learning frameworks TensorFlow and Keras. To find the optimal results for the experiments, we used Adam optimization technique and set the learning rate to 0.0001 and the decay rate to 0.05. The batch size during training was 32, and it lasted for 50 epochs in total. The CNN was designed with a dropout ratio of 0.5 to prevent overfitting. Both the first and second β values parameters were initially set to 1.1 and 0.9, respectively.

In section 3.3, we selected the standard 2D-CNN

Received: October 9, 2023. Revised: November 12, 2023.

with fully-connected layers as the baseline system for our experiments. The proposed method adopts facial landmarks for the β-skeleton technique to extract geometric features. The dataset was randomly split into 80% for training and 20% for testing [21]. Performance evaluation of the proposed method utilized various metrics, including precision, recall, F1-score, and validation accuracy [30].

4.1 CK+ Dataset result

The CK+ dataset is commonly used in FER setups. It comprises 593 videos featuring 123 unique individuals [9, 37]. The dataset includes image sequences with lengths ranging from 10 to 60 frames, capturing the transition from a neutral to an intense facial expression. Additionally, there are 981 still photos extracted from 327 video clips involving 118 individuals. It covers all seven major human emotions: anger, disgust, fear, happiness, sadness, surprise, and contempt. Fig. 7 visually represents a subset of images from the CK+ dataset [38].

According to Table 3, the experimental results demonstrate that the proposed baseline system achieves a recognition accuracy of 95.81%. However, the results of the experiments indicate that the proposed system outperforms the baseline system in terms of accuracy, with a 0.31% improvement. The inclusion of geometric fusion features obtained from facial landmarks enhances the precision of expression classification by capturing additional expression information beyond appearance-based features derived from raw face images. The proposed method's classification accuracy on the CK+ dataset is visualized in a normalized confusion matrix, as depicted in Fig. 9(a). The performance evaluation of the proposed system, precision, recall, and F1 scores generated from the confusion matrix are presented in Table 4. The confusion matrix highlights the model’s accuracy in classifying emotional expressions such as contempt and fear. The accuracy curves during training and validation are displayed in Fig. 10(a). illustrates that the proposed method converges after a certain number of epochs. The model starts to converge between 10 and 15 epochs. This demonstrates that the proposed method can achieve desirable results after fewer epochs. Table 3 compares the proposed method to state-of-the-art models on the CK+ dataset, employing different techniques.

Suparshya Babu Sukhavasi et al. [2] combined Deep neural networks and SVM to create a unique hybrid network architecture to predict seven driver expressions and achieve an accuracy of 95.05. Adil Boughida Durga et al. [7] introduce a novel approach for FER that integrates Gabor filters with a genetic algorithm (GA) and obtain an accuracy of 94.26%. The proposed system based on LBP feature extraction in [15] achieves an average accuracy of 93.90%. The novel FER technique using infrared technology is proposed in [16] to assess classroom non-verbal behaviors obtains a recognition result of 87.42%. Authors in [40] proposes a FER deep-CNN that can automatically discover deeper feature representations of facial expressions. The proposed system achieving accuracy rates of 80.303%. Quang Tran Ngoc et al. [18] introduced the FER system based on DGNN, utilizing landmark features and achieve the classification accuracy of 96.02%. Authors in [9] proposed GA-SVM system and

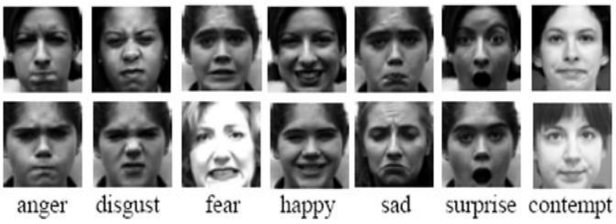


Figure. 7 Sample of the CK+ dataset [37]

Table 3. The CK+ dataset results compared with the existing approach

Year	Approaches	Accuracy
2022 [2]	GLFCNN+SVM	95.05
2022 [7]	Gabor Filter, GA+SVM	94.26
2021[15]	LBP+SVM	93.90
2021 [16]	CDLLNET	87.42
2021 [8]	LBP, ORB features +SVM	93.21
2017 [40]	CNN	80.30
2020 [18]	DGNN	96.02
2020 [9]	GA-SVM	95.58
2016 [4]	SACNN+ALSTM	95.15
	Proposed Baseline Method	95.81
	Proposed Fusion-CNN-SVM Method	96.19

Table 4. The CK+ dataset evaluation metrics for each expression

Expression	Precision	Recall	F1	Accuracy
Anger	93.07	94.95	94.00	94.00
Contempt	86.00	98.85	91.98	86.00
Disgust	97.98	94.17	96.04	97.00
Fear	88.00	100.00	93.62	88.00
Happy	100.00	99.87	99.94	100.00
Sad	93.00	84.55	88.57	93.00
Surprise	99.00	100.00	99.50	99.00

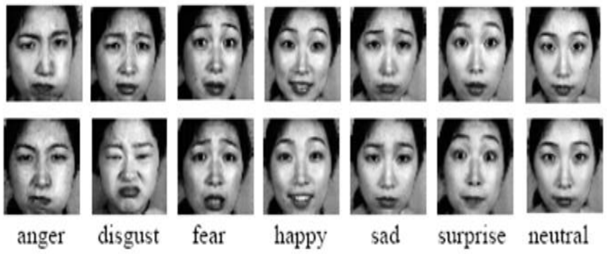


Figure. 8 Sample of the JAFFE dataset [39]

Table 5. The JAFFE dataset results compared with the existing approach

Year	Classifier	Accuracy
2021 [8]	LBP, ORB features +SVM	88.50
2017 [40]	CNN	76.74
2021 [15]	LBP+SVM	88.30
	Proposed Baseline Method	85.38
	Proposed Fusion-CNN-SVM Method	89.23

Table 6. The JAFFE dataset evaluation metrics for each expression

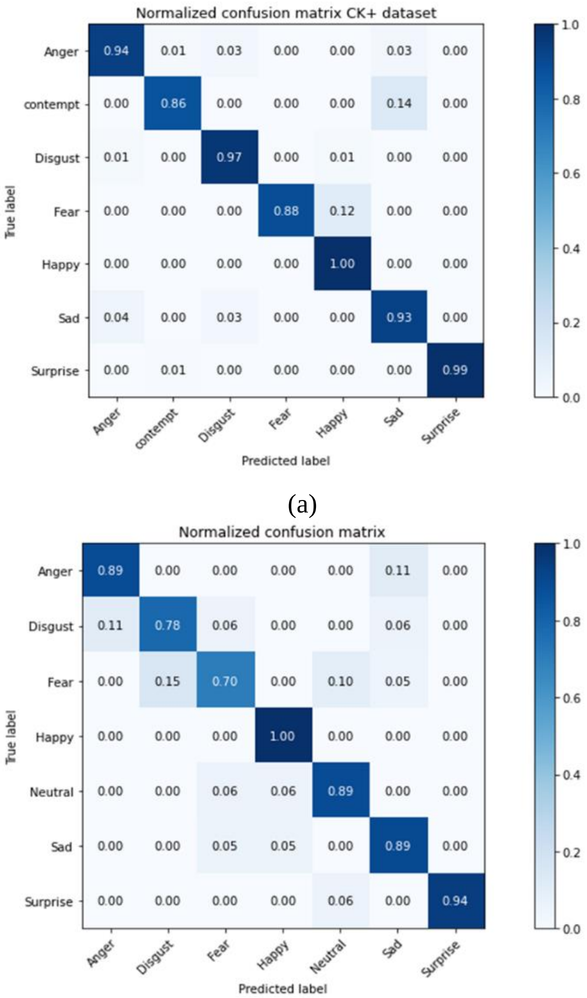
Expression	Precision	Recall	F1	Accuracy
Anger	89.00	89.00	89.00	89.00
Disgust	77.23	83.87	80.41	78.00
Fear	70.00	80.46	74.87	70.00
Happy	100.00	99.89	99.95	100.00
Natural	88.12	89.90	89.00	89.00
Sad	89.90	80.18	84.76	89.00
Surprise	94.00	100.00	96.91	94.00

obtain an average accuracy of 95.58% on CK+ database. The proposed model achieves an average recognition accuracy of 96.19% on the CK+ database.

4.2 JAFFE dataset result

The JAFFE dataset was created by training ten Japanese women to exhibit the main seven facial expressions [15, 39]. It consists of 213 grayscale still photographs, capturing emotions such as anger, happiness, neutral, surprise, sadness, fear, and disgust [38]. All images were taken in a controlled studio environment with stable lighting conditions and no distractions like hair or glasses [8]. The original image resolution is 256 × 256 pixels, and all facial expressions are captured in a frontal view. Fig. 8 provides the photographs in this dataset.

According to Table 5, the baseline system consistently achieves a recognition accuracy of 84.20 on the JAFFE dataset. However, empirical evidence demonstrates that the proposed system outperforms the reference model by 5.03% in accuracy. The classification accuracies of the propose system on the JAFFE dataset are visualized in the normalized confusion matrix presented in Fig. 9(b). Table 6 comprehensively evaluates the proposed method's performance, including precision, recall, and F1 score.

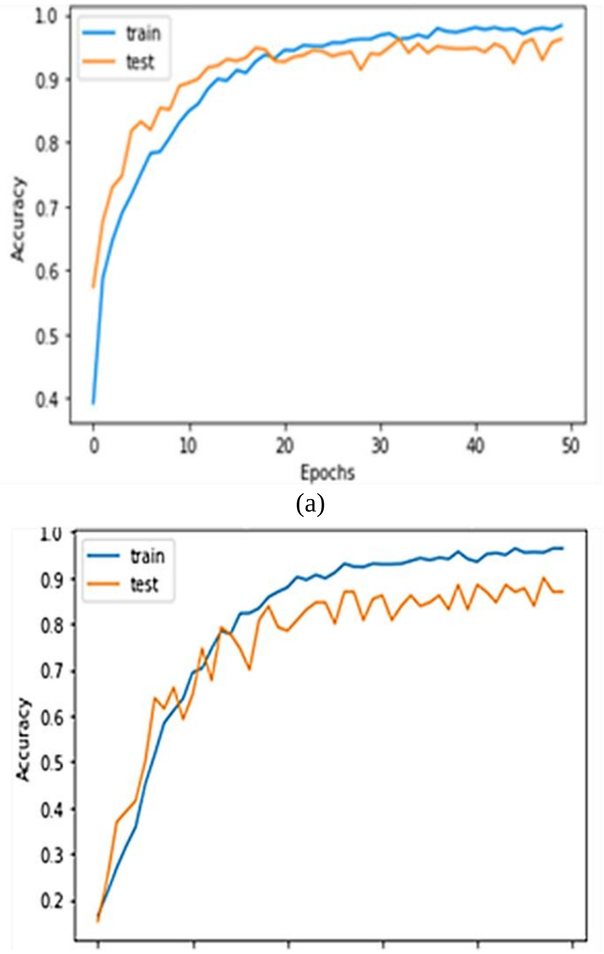


(a)

(b)

Figure. 9 The confusion matrices: (a) CK+ dataset and (b) JAFFE dataset

The confusion matrix reveals that the model can distinguish between happy and surprised expressions, two of the most challenging emotions for humans to categorize accurately. Differentiating between disgust and fear expressions can be more challenging due to potentially limited training data available. Nevertheless, when evaluated on the JAFFE dataset, the proposed system achieves an average accuracy of 89.23%. Fig. 10(b) illustrates the accuracy curves during training and validation. Table 5 indicates that the proposed approach outperforms state-of-the-art methods utilizing



(a)

(b)

Figure. 10 Training Vs validation accuracy curves: (a) CK+ dataset and (b) JAFFE dataset

various methodologies on the JAFFE dataset. The proposed method based on local binary pattern with ORB features and SVM classifier in [8] achieve accuracy of 88.50%. Authors in [40] proposes a FER deep-CNN that can automatically discover deeper feature representations of facial expressions. The proposed system achieving accuracy rates of 76.74%. The proposed system based on LBP feature extraction in [15] achieves an average accuracy of 88.30%.

5. Conclusions

This paper proposes a data fusion approach, Fusion-NN-SVM, which combines two types of CNNs and utilizes SVM as a classifier to extract diverse features for facial expression classification. By leveraging the geometric correlations between facial pixels and landmark points, this technique enables extracting more discriminative facial features for FER. The proposed system achieves impressive accuracy rates of 96.16% and 89.23% on the CK+ and JAFFE datasets, respectively, outperforming the performance of state-of-the-art methodologies. Facial landmarks are crucial in FER systems as they provide valuable geometry-level features. The results demonstrate higher recognition rates for surprise and happiness, while emotions like fear exhibit relatively lower rates. The proposed system shows convergence within a finite number of epochs, contributing to the efficiency of reaching the desired outcomes. Developing a Fusion-based FER system that incorporates modalities beyond facial features is envisioned. By integrating speech, text, and physical gestures, we can overcome the inherent limitations of relying on a single mode of communication. The incorporation of multiple modalities into the FER system holds the potential to enhance its precision and reliability.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

We certify that all authors contributed to this study. Paper conceptualization, methodology and investigation, Abbas Issa Jabbooree, Leyli Mohammad Khanli, Pedram Salehpour, and Shahin Pourbahrami; software, validation, formal analysis, resources, data curation, writing-original draft preparation and visualization, Abbas Issa Jabbooree; writing-review and editing, Leyli Mohammad Khanli, Pedram Salehpour, and Shahin Pourbahrami; supervision and project administration, Leyli Mohammad Khanli, Pedram Salehpour.

References

[1] S. Yasmin, R. K. Pathan, M. Biswas, M. Khandaker, and M. R. I. Faruque, “Development of a Robust Multi-Scale Featured Local Binary Pattern for Improved Facial Expression Recognition”, *Sensors*, Vol. 20, No. 5391, pp. 1–17, 2020.

[2] S. B. Sukhavasi, S. B. Sukhavasi, K. Elleithy, A. E. Sayed, and A. Elleithy, “A Hybrid Model for Driver Emotion Detection Using Feature Fusion Approach”, *Int. J. Environ. Res. Public Heal. MDPI*, Vol. 19, No. 3085, pp. 1–19, 2022.

[3] M. Sajjad, S. Zahir, A. Ullah, Z. Akhtar, and K. Muhammad, “Human Behavior Understanding in Big Multimedia Data Using CNN based Facial Expression Recognition”, *Mob. Networks Appl.*, Vol. 25, No. 4, pp. 1611–1621, 2020.

[4] C. Liu, K. Hirota, J. Ma, Z. Jia, and Y. Dai, “Facial Expression Recognition Using Hybrid

Features of Pixel and Geometry", *IEEE Access*, Vol. 9, pp. 18876–18889, 2021.

[5] R. Zhao, T. Liu, Z. Huang, D. P. K. Lun, and K. K. M. Lam, "Geometry-Aware Facial Expression Recognition via Attentive Graph Convolutional Networks", *IEEE Trans. Affect. Comput.*, Vol. 3045, No. c, 2021.

[6] Y. Nan, J. Ju, Q. Hua, H. Zhang, and B. Wang, "A-MobileNet: An approach of facial expression recognition", *Alexandria Eng. J.*, Vol. 61, No. 6, pp. 4435–4444, 2022.

[7] A. Boughida, M. N. Kouahla, and Y. Lafifi, "A novel approach for facial expression recognition based on Gabor filters and genetic algorithm", *Evol. Syst.*, Vol. 13, No. 2, pp. 331–345, 2022.

[8] B. Niu, Z. Gao, and B. Guo, "Facial Expression Recognition with LBP and ORB Features", *Comput. Intell. Neurosci. Hindawi*, Vol. 2021, pp. 1–10, 2021.

[9] X. Liu, X. Cheng, and K. Lee, "GA-SVM-Based Facial Emotion Recognition Using Facial Geometric Features", *IEEE Sens. J.*, Vol. 21, No. 10, pp. 11532–11542, 2021.

[10] A. Barman and P. Dutta, "Facial expression recognition using distance and shape signature features", *Pattern Recognit. Lett.*, Vol. 145, pp. 254–261, 2021.

[11] S. Liu, X. Tang, and D. Wang, "Facial expression recognition based on sobel operator and improved CNN-SVM", In: *Proc. of 2020 3rd IEEE Int. Conf. Inf. Commun. Signal Process*, pp. 236–240, 2020.

[12] J. C. Kim, M. H. Kim, H. E. Suh, M. T. Naseem, and C. S. Lee, "Hybrid Approach for Facial Expression Recognition Using Convolutional Neural Networks and SVM", *Appl. Sci.*, Vol. 12, No. 11, 2022.

[13] S. M. G. Lozoya, J. D. L. Calleja, L. Pellegrin, H. J. Escalante, M. A. Medina, and A. B. Ruiz, "Recognition of facial expressions based on CNN features", *Multimed. Tools Appl.*, Vol. 79, No. 19–20, pp. 13987–14007, 2020.

[14] F. An and Z. Liu, "Facial expression recognition algorithm based on parameter adaptive initialization of CNN and LSTM", *Vis. Comput.*, Vol. 36, No. 3, pp. 483–498, 2020.

[15] D. G. R. Kola and S. K. Samayamantula, "A novel approach for facial expression recognition using local binary pattern with adaptive window", *Multimed. Tools Appl.*, Vol. 80, No. 2, pp. 2243–2262, 2021.

[16] T. Liu, J. Wang, B. Yang, and X. Wang, "Facial expression recognition method with multi-label distribution learning for non-verbal behavior understanding in the classroom", *Infrared Phys. Technol.*, Vol. 112, No. December 2020, p. 103594, 2021.

[17] A. Joseph and P. Geetha, "Facial emotion detection using modified eyemap–mouthmap algorithm on an enhanced image and classification with tensorflow", *Vis. Comput.*, Vol. 36, No. 3, pp. 529–539, 2020.

[18] Q. T. Ngoc, S. Lee, and B. C. Song, "Facial landmark-based emotion recognition via directed graph neural network", *Electron.*, Vol. 9, No. 5, 2020.

[19] Y. Han, X. Wang, and Z. Lu, "Research on facial expression recognition based on Multimodal data fusion and neural network", *arXiv preprint arXiv:2109.12724*, 2021.

[20] H. Wang, S. Wei, and B. Fang, "Facial expression recognition using iterative fusion of MO-HOG and deep features", *J. Supercomput.*, Vol. 76, No. 5, pp. 3211–3221, 2020.

[21] M. Wang, P. Tan, X. Zhang, Y. Kang, C. Jin, and J. Cao, "Facial expression recognition based on CNN", *J. Phys. Conf. Ser.*, Vol. 1601, No. 5, 2020.

[22] J. Hyun, J. Kim, C. H. Choi, and B. Moon, "Hardware architecture of a haar classifier based face detection system using a Skip scheme", In: *Proc. of IEEE Int. Symp. Circuits Syst.*, Vol. 2021-May, 2021.

[23] H. Zhang, A. Jolfaei, and M. Alazab, "A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing", *IEEE Access*, Vol. 7, pp. 159081–159089, 2019.

[24] S. J. Park, B. G. Kim, and N. Chilamkurti, "A robust facial expression recognition algorithm based on multi-rate feature fusion scheme", *Sensors*, Vol. 21, No. 21, pp. 1–26, 2021.

[25] A. I. Jabbooree, L. M. Khanli, P. Salehpour, and S. Pourbahrami, "A novel facial expression recognition algorithm using geometry β – skeleton in fusion based on deep CNN", *Image Vis. Comput.*, Vol. 134, p. 104677, 2023.

[26]S. Pourbahrami, L. M. Khanli, and S. Azimpour, "An automatic clustering of data points with alpha and beta angles on apollonius and subtended arc circle based on computational geometry", In: *Proc. of 2020 28th Iran. Conf. Electr. Eng. ICEE 2020*, 2020.

[27] Y. Yang, D. Q. Han, and J. Dezert, "An angle-based neighborhood graph classifier with evidential reasoning", *Pattern Recognit. Lett.*, Vol. 71, pp. 78–85, 2016.

[28]S. Pourbahrami, L. M. Khanli, and S. Azimpour, "A novel and efficient data point neighborhood

construction algorithm based on Apollonius circle”, *Expert Syst. Appl.*, Vol. 115, pp. 57–67, 2019.

[29] A. E. Lagias, T. D. Lagkas, and J. Zhang, “New RSSI-Based Tracking for Following Mobile Targets Using the Law of Cosines”, *IEEE Wirel. Commun. Lett.*, Vol. 7, No. 3, pp. 392–395, 2018.

[30] K. Yan and X. Zhou, “Chiller faults detection and diagnosis with sensor network and adaptive 1D CNN”, *Digit. Commun. Networks*, No. June 2021, 2022.

[31] F. Li, M. Liu, Y. Zhao, L. Kong, L. Dong, and X. L. M. Hui, “Feature extraction and classification of heart sound using 1D convolutional neural networks”, *EURASIP J. Adv. Signal Process.*, Vol. 2019, No. 1, 2019.

[32] J. P. Zhu, H. Q. Chen, and W. B. Ye, “Classification of human activities based on radar signals using 1D-CNN and LSTM”, In: *Proc. of IEEE Int. Symp. Circuits Syst.*, Vol. 2020-Octob, 2020.

[33] S. Allamy and A. L. Koerich, “1D CNN Architectures for Music Genre Classification”, In: *Proc. of 2021 IEEE Symp. Ser. Comput. Intell. SSCI 2021 - Proc.*, 2021.

[34] I. Mitiche, A. Nesbitt, S. Conner, P. Boreham, and G. Morison, “1D-CNN based real-time fault detection system for power asset diagnostics”, *IET Gener. Transm. Distrib.*, Vol. 14, No. 24, pp. 5816–5822, 2020.

[35] L. N. Do, H. J. Yang, H. D. Nguyen, S. H. Kim, G. S. Lee, and I. S. Na, “Deep neural network-based fusion model for emotion recognition using visual data”, *J. Supercomput.*, Vol. 77, No. 10, pp. 10773–10790, 2021.

[36] S. A. Tuama, “Ear and Tongue Multi Biometric Identification System Using Convolutional Neural Network”, *Iraqi Commission for Computers*, 2020.

[37] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, and Z. Ambadar, “The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression”, *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, Vol. 4, No. July, pp. 94–101, 2010.

[38] C. Gautam and K. R. Seeja, “Facial emotion recognition using Handcrafted features and CNN”, *Procedia Comput. Sci.*, Vol. 218, No. 2022, pp. 1295–1303, 2023.

[39] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with Gabor wavelets”, In: *Proc. of 3rd IEEE Int. Conf. Autom. Face Gesture Recognition, FG 1998*, pp. 200–205, 1998.

[40] K. Shan, J. Guo, W. You, D. Lu, and R. Bie, “Automatic facial expression recognition based on a deep convolutional-neural-network structure”, In: *Proc. of 2017 15th IEEE/ACIS Int. Conf. Softw. Eng. Res. Manag. Appl. SERA 2017*, pp. 123–128, 2017.