ACTA IMEKO
ISSN: 2221-870X
June 2022, Volume 11, Number 2, 1 -7

Video-based emotion sensing and recognition using convolutional neural network based kinetic gas molecule optimization

Kasani Pranathi[1], Naga Padmaja Jagini[2], Satish Kumar Ramaraj[3], Deepa Jeyaraman[4]

[1] *Dept of Information Technology, VR Siddhartha Engineering College, Kanuru, Vijayawada-520007 Andhra Pradesh, India*
[2] *Dept of Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad-501218, Telangana , India*
[3] *Sengunthar College of Engineering, Tiruchengode, Tamilnadu, India*
[4] *Dept of Electronics and Communication engineering, K. Ramakrishnan College of technology, Trichirapalli-621112, Tamilndadu, India*

ABSTRACT
Human facial expressions are thought to be important in interpreting one's emotions. Emotional recognition plays a very important part in the more exact inspection of human feelings and interior thoughts. Over the last several years, emotion identification utilizing pictures, videos, or voice as input has been a popular issue in the field of study. Recently, most emotional recognition research focuses on the extraction of representative modality characteristics and the definition of dynamic interactions between multiple modalities. Deep learning methods have opened the way for the development of artificial intelligence products, and the suggested system employs a convolutional neural network (CNN) for identifying real-time human feelings. The aim of the research study is to create a real-time emotion detection application by utilizing improved CNN. This research offers information on identifying emotions in films using deep learning techniques. Kinetic gas molecule optimization is used to optimize the fine-tuning and weights of CNN. This article describes the technique of the recognition process as well as its experimental validation. Two datasets such as video-based and image-based datasets, which are employed in many scholarly publications, are also investigated. The results of several emotion recognition simulations are provided, along with their performance factors.

Section:RESEARCH PAPER

Keywords: Artificial intelligence; convolutional neural network; kinetic gas molecule optimization; images; video-based emotion recognition

Citation: Kasani Pranathi, Naga Padmaja Jagini, Satish Kumar Ramaraj, Deepa Jeyaraman, Video-based emotion sensing and recognition using convolutional neural network based kinetic gas molecule optimization, Acta IMEKO, vol. 11, no. 2, article 13, June 2022, identifier: IMEKO-ACTA-11 (2022)-02-13

Section Editor: Francesco Lamonaca, University of Calabria, Italy

Received October 2, 2021; **In final form** June 3, 2022; **Published** June 2022

Corresponding author: Kasani Pranathi, e-mail: pranathivrs@gmail.com

1. INTRODUCTION

Emotional recognition is considered a major theme in machine learning and Artificial Intelligence (AI) [1] in recent years. The huge upsurge in the creation of advanced interaction technologies between humans and computers has further encouraged progress in this area [2]. Facial actions convey emotions that transmit the character, the mood, and the intentions of a person, in turn. Emotions and moods swiftly lead to the identification of the human mind. The psychologist says that emotions are mostly short and that mood is milder than emotion [3]. Human emotions can be detected in different ways, such as verbal or voice responses, physical reactions or the languages of the body, autonomous responses, and so on [4]. In a person, the basic types of emotions are pleasing, normal, surprised, frightened, angry, disgusting or sad. While other expressions like dislike, amusement, or pride, dislike and honesty in humans are very difficult to find by expression of the face [5]-[6] is easy to detect emotions like happiness, normal, disgust, and fear. As we know, people identify emotion by combining difficult multimodal information and tend only to pay attention to significant information in various ways. For example, some people always talk while they keep a smile, while others talk loudly but not angrily [7]. Consequently, we deliberate those human beings do not detect emotions based on modal alignment. The objective of emotional awareness can be generally achieved by means of visual or sound techniques. The field of human-computer interaction has changed by Artificial Intelligence, providing many machine learning techniques to achieve our goal

[7]. The extraction of representative modal features using profound learning technology has grown easier. For example, the neural network [8] includes advanced CNN - Convolutional Neural Network - (AlexNet, VGG, ResNet, SENets, etc.) tuning process is extremely useful for capturing features of fine-grained facial expression; and Long Short-Term Memory Unit (LSTMs) are another profound learning technology to store info with short-term interaction of time-step memory functionality [9]. On the basis of these technology of learning, there is much work to be done to define dynamic multimodal interactions by matching the relevance of each LSTM memory unit.

Video-based emotional recognition is multidisciplinary, covering areas such as psychology, affective computing and interaction between humans and computers. The main element of the message is the expression of the face which makes up 55% of the overall impression. In order to create an appropriate model for the recognition of video emotions, proper feature frames of face expression must be provided within the scope. Deep learning offers a diversity in terms of accuracy, learning rate and forecasting, rather than using standard techniques. CNN has offered support and platform for the analysis of visual imaging, among the in-depth learning methodologies. Convolution is the basic application of a filter in an action to an input that results. Reusing a related filter to an input creates an enactment map known as a feature map that shows the areas and the quality of, for instance, an identified element in an image. The growth of neural systems of convolution is thus the ability to learn skills with an enormous number of filters equating explicitly to a training dataset according to the needs of, for example, an image characterisation. The result is deeply explicit highlights which are distinguishable in the input images anywhere. Deep learning accomplished a major achievement in the recognition of emotions and CNN is the renowned profound way of learning with exceptional image processing performance. This work aims to develop video-based emotional awareness through optimal CNN, as detailed in the next section.

The remaining part of the paper is arranged: The related study on video emotional recognition is presented in Section 2. Section 3 provides an explanation for the suggested optimized CNN. Section 4 discusses the validation of the methodology suggested with its current techniques. Finally, Section 5 presents the conclusion of this study with its future work.

2. LITERATURE REVIEW

A number of disciplines, such as spam filtering, audio recognition, facial identification, classifying documents and processing of natural languages are addressed by machine learning algorithms. Classification is one of the most frequently used domains of machine learning. Video-based face-motion research in the computer vision community recently attracted notice. Different kinds of input data, including facial expressions, voice, physiological indicators, and corporal motions, are utilised in emotional recognition. The work of Michel Healy[10], who describes a video feeding system in real time and uses a support vector machine for fast and reliable classification, offers several ways to the detection of emotion by facial expressions. The 68-point facial features used in [10] are symbols. The application was taught to detect six emotions by monitoring changes in the expressions of the face.

The work of Dennis Maier [11] currently uses neural networks via TensorFlow to train image features and then achieves classification through fully connected neural layers. The advantage of image features over facial landmarks is the larger information space, where the spatial formation of landmarks gives a viable method for analysing facial expressions. However, this is also accompanied by a higher computing power requirement. The structure provides for an outsourced classification service that runs on a server with a GPU. Images of faces are brought to the service in real time, which can perform a classification within a few milliseconds. In the future, this approach will be extended to include text and audio features and conversation context to boost accuracy. Another approach uses CNNs with TensorFlow. An example of using TensorFlow.js with the Sense-Care KM-EP is discussed in [10], which deploys a web browser and Node Server. 93% rely on non-verbal's (facial expressions as 55%, sound: 38%) and 7% rely on verbal language in terms of human emotional understanding. That is why various efforts have been carried out to recognize Facial Expression (FER) and Acoustic Emotion (AER) tasks. Most of these works use Deep Learning (DL) skills to extract computer features to get recognition of high emotions. Pramerdorfer et al., [12] used and confirmed contemporary DNN (VGG, ResNet, Inception) architectures to extract aspects of facial expression to enhance FER performance. On the other hand, the most typically employed features include pitch, log-Mel filters, and cepstral coefficients for MFCCs, as far as AER tasks are concerned. Huang et al. [13] used four kinds to extract more complete emotional characteristics using Log-Mel.

Multiple Spatial-Time Fusion Feature Framework (MSFF) was proposed by Lu et al. [14]. They have improved the pre-trained model for photos of facial expression to draw on facial expression characteristics and have applied the VGG-19 and BLSTM models to extract audio emotional aspects. However, the interactions between different modes were not taken into account. Zadeh et al. [15], on the other hand, considered consistency and attributes complementary to the diverse modal information, proposing a memory fusion network which model modal and multimodal interactions through time to capture more efficacious emotional characteristics in the CMU-MOSI dataset. Liang et al. [16] have presented the neural model Dynamic Fusion Graph to shape multimodal interactions, to capture one, two and three modal interactions, and, based on the importance, dynamically adjust multimodal dynamics of individual fusions. Although [16] is able to dynamically collect interactions in several modalities, different modalities must be aligned with the word utterance time interval through the average of their modalities. The word-based alignment technique can nonetheless miss the chance to capture more active relationships between modes.

3. PROPOSED METHODOLOGY

This section provides a description of the overall architecture for the development of a deep learning algorithm as a video-based emotional recognition model. In addition, architectural diagrams are briefly described together with various operations before and after processing. The system overview with CNN displays the suggested training and testing method in Figure 1. The video input must pass through a number of procedures before CNN takes action.

3.1. Pre-processing

This is the first procedure used for the video sample input. Emotions are typically classified as happiness, sadness, anger, pride, fear and surprise. Frames should therefore be removed from the video input. The number of frames depends on complexity and computational time for different researchers.

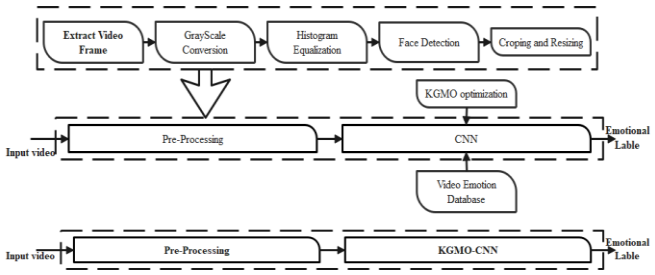

Figure 1. Proposed workflow (top) and proposed training (bottom) testing processes.

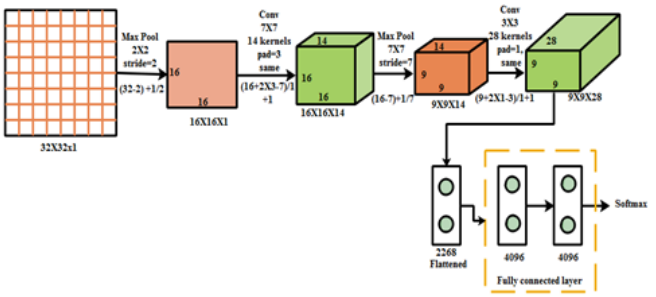

Figure 2. Typical architecture of some of the pre-trained deep CNN networks used in the study

The pictures are transformed to the greyscale. The frame is rather black and white or grey monochrome after grey scaling. The contrast with low intensity leads to grey and white with high intensity. The histogram equalization of the frames is monitored by this step. Histogram is a computer image management strategy to improve photographic contrast. This is achieved by extending, for instance by loosening the intensity of the image, the most successive intensity estimates. The intensity of a picture is shown by the histogram and it is the number of pixels for each intensity value deliberated in simple terms.

3.2. Face detection

Emotions are usually characterized by the face. It is therefore important to detect the face for processing and recognition. Many face detection algorithms are used by several investigators such as OpenCV, DLIB, Eigenfaces, the local histograms of binary patterns (LBPH) and Viola-Jones (VJ). Conventional procedures included face recognition work in which facial highpoints are distinguished from the face image by extracting highlights or milestones. The calculation may, for example, survey the shape and size of the eyes, nose size and their relative position with the eyes in order to delete face highlights. The cheekbones and mastic may also be dissected. These highlights extracted would be used to view different images with matching features. The industry has gone deeply into learning throughout the years. CNN was recently used to improve the accuracy of calculations for facial recognition. These controls accept an image as information and concentrate on a very complex arrangement of features. These features include facial width, facial stature, nose width, lips, eyes, width proportion, skin shading, and surface. Basically, a CNN divides a huge sum of highlights from a picture. This is then synchronised with the highlights in the database

3.3. Image cropping and resizing

During this phase, the face of the facial detection procedure is trimmed so that the facial image looks broader and clearer. Cropping is the ejection from the photographic or graphical image of unwanted exterior parts. The technique often consists of expelling a section of the outermost regions of a picture, expelling the image's incidental waste, improving its surroundings, changing its perspective, or highlighting and disentangling the subject. The size of the images varies after the frames have been cropped. Those photographs are therefore subject to resize, say 80 to 80 pixels for instance, in order to achieve homogeneity. A digital picture is only a quantity of information displaying a variety of red, green and blue pixels in a certain location. We notice these pixels more often than not as smaller than normal pixels on the PC screen wedged together. The frame size determines how long it takes to process. Resizing is therefore very important if processing time is to be reduced. In addition, techniques of better resizing should be used to maintain image attributes following resizing. Whether the features represent the expression well or not depends on the accuracy of the classification. The optimization of the selected features, therefore, improves classification precision automatically.

3.4. Classification

In this section, the classification includes learning rate optimization for CNN using Kinetic Gas Molecule Optimization (KGMO) is described briefly. Initially, the CNN is explained as follows: CNN is a neural transmission network with several layer feeders, comprising several sorts of layers, including convolution layer and ReLU, pooling layers and fully connected output layers. Figure 2 shows the architecture of CNN, which is intended to recognize visual characteristics such as borders and forms.

3.4.1. CNN

CNN employs the vector $\boldsymbol{X}$ of the trained samples as the input for the associated target group y to support the back propagation technique for training information. Learning is performed by comparing the desired target with each CNN output, and a learning error occurs in the difference between the two. Taking mathematical responsibility for the future CNN,

$$E(\omega) = \frac{1}{2}\sum_{\rho=1}^{p}\sum_{\mathrm{j}=1}^{N_{\mathrm{l}}}(o_{\mathrm{j,p}}^{\mathrm{l}} - y_{\mathrm{j,p}})^2 \ . \tag{1}$$

Our goal is the cost function lessening of $E(\omega)$, discovery a minimizer $\tilde{\omega} = \tilde{\omega}^1, \tilde{\omega}^2, \ldots, \tilde{\omega}^{\mathrm{v}} \in \mathbb{R}^{\mathrm{v}}$, where $\mathrm{v} = \sum_{\mathrm{k}=1}^{\mathrm{L}} \mathrm{WeightNum}\ (k)$ and indicate that the space of weight $\mathbb{R}^{\mathrm{v}}$ is equal to the number of weights $(\mathrm{WeigtNum}(.))$ of the CNN network at each k layer of total L layers.

$$\nabla E_{\mathrm{i}}(\omega_{\mathrm{i}}) = \left(\frac{\partial E_{\mathrm{i}}}{\partial \omega_{\mathrm{i}}^{1}}, \ldots, \frac{\partial E_{\mathrm{i}}}{\partial \omega_{\mathrm{i}}^{\mathrm{v}}}\right) \tag{2}$$

$$\omega_{\mathrm{i}+1} = \omega_{\mathrm{i}} - n\,\nabla E_{\mathrm{i}}(\omega_{\mathrm{i}})\,, \tag{3}$$

where n is the learning rate (step) value. CNN is adapted to the video-based emotion detection, but how fast CNN is adapted can be controlled by this learning rate. More training epochs are acquired for smaller learning rates that give only small changes to the weights during each update. On the other hand, fewer training epochs are required for larger learning rates. Specifically, the learning rate is a configurable hyper-parameter used in the training of neural networks that has a small positive value, often in the range between 0.0 and 1.0. To find the optimized learning rate value, this work uses the KGMO algorithm. Here, the n has

been designated with the assistance of the KGMO technique, which is explained as follows.

3.4.2. KGMO

The Boyle laws introduced gas, which are based on unproven descriptions to label gas glimmer's macroscopic plants. The motion of the gas molecule is based on certain characteristics such as pressure, temperature and gas volume. Five suggestions and a cinematic molecular system of spectacular blasts were used to indicate the fauna of the gas molecules:

- A gas involves the movement in straight-line movement of tiny molecules. The processing of the movement presents according to Newton's law.
- There is no capacity for the gas glimmer. It's like a subject.
- No repulsive or attracting force exists between molecules, the average molecular kinetic energy of 3 k $T/2$ is described as T and Boltzmann constant is described as k, Income which has 1.38×10^{-23} m^2 kg s^{-2} K^{-1}.

Take on, that the scheme contains M particles. The locality of j^{th} the agent is labelled by (4)

$$Y_j = \left(y_j^1, \ldots \ldots y_j^d, \ldots \ldots y_j^m\right) \text{ for } (j = 1, 2, \ldots, M), \tag{4}$$

where Y_j^d defines the position of the j^{th} agent, that in dth dimension. The velocity of the j^{th} agent, it stated as below in (5)

$$V_j = \left(v_j^1, \ldots \ldots, v_j^2, \ldots \ldots v_j^m\right) \text{ for } (j = 1, 2, \ldots, M), \tag{5}$$

where V_j^d represents the j^{th} agent velocity, that in the d^{th} dimension.

Units are modified by the circulation of Boltzmann. Random element yield speed is related to the kinetic energy of the unit. Equation (6) is the kinetic force of the atom as,

$$\begin{aligned} k_j^d(u) &= \frac{3}{2} Mb\, T_j^d(u), K_j \\ &= \left(k_j^1, \ldots \ldots, k_j^d, \ldots \ldots k_j^m\right) \text{ for } (j = 1, 2, \ldots, M), \end{aligned} \tag{6}$$

where the number of atoms is represented Mb as the Boltzmann relentless and the high temperature of an j^{th} agent in the d^{th} dimension at a time u is characterized as $T_j^d(u)$. The molecules rate is updated by (7)

$$\begin{aligned} v_j^d(u+1) &= T_d^j(u) w V_j^d(u) \\ &+ D_1 \text{rand}_j(u) \left(gbest^d - y_j^d(u)\right) \\ &+ D_2 \text{rand}_j(u) \left(pbest_j{}^d(u) - y_j^d(u)\right), \end{aligned} \tag{7}$$

where $T_j^d(u)$ for the meeting molecules reduce the exponentially over time and is planned in (8).

$$T_j^d(u) = 0.95 \times T_j^d(u-1). \tag{8}$$

The mass m of individual sub-division is a random number and it consumes a range of $0 < m \leq 1$. Once the mass is well-known, the whole algorithm remains unchanged because only one type of gas is taken into consideration at any time. A random number shall be used to claim places in diverse types to produce various executions of the procedure. The constituent part is expressed as equation, based on the gesticulation equation (9),

$$y_{u+1}^j = \frac{1}{2} a_j^d (u+1) u^2 + v_j^d (u+1) u + y_j^d(u), \tag{9}$$

where acceleration of the j^{th} agent is in the d^{th} dimension exemplified as a_j^d. Centered on the acceleration dismiss achieve in (10).

$$a_d^j = \frac{(\mathrm{d}v_j^d)}{\mathrm{d}u}. \tag{10}$$

Built on the gas specks law is signified in (11).

$$\mathrm{d}k_d^j = \frac{1}{2} m (\mathrm{d}v_j^d)^2 \Rightarrow \mathrm{d}v_j^d = \sqrt{\frac{2(\mathrm{d}k_j^d)}{m}}. \tag{11}$$

From (10) and (11), the acceleration is mentioned in (12)

$$a_d^j = \frac{\sqrt{\frac{2(\mathrm{d}k_j^d)}{m}}}{\mathrm{d}u}. \tag{12}$$

The acceleration reckoning is re-written depends on the duration interval Δu which is shown in (13)

$$a_d^j = \frac{\sqrt{\frac{2(\Delta k_j^d)}{m}}}{\Delta u}. \tag{13}$$

In a unit time interval, the acceleration would be (14)

$$a_d^j = \sqrt{\frac{2(\mathrm{d}k_j^d)}{m}}. \tag{14}$$

From (9) and (14), the section of the particle is computed by uttered by (15)

$$\begin{aligned} y_{u+1}^j &= \frac{1}{2} a_j^d (u+1) \Delta u^2 + v_j^d (u+1) \Delta u + y_j^d(u) \Rightarrow \\ y_{u+1}^j &= \frac{1}{2} \sqrt{\frac{2(\Delta k_j^d)}{m}} (u+1) \Delta u^2 + v_j^d (u+1) \Delta u \\ &+ y_j^d(u). \end{aligned} \tag{15}$$

In the past, the assumption is that the molecular mass is a random element in all rules exercising, but the execution is same in all particles. The location, which is expressed in (16). is sporadically reorganized to make the approach easier.

$$y_{u+1}^j = \sqrt{\frac{2(\Delta k_j^d)}{m}} (u+1) + v_j^d (u+1) + y_j^d(u) \tag{16}$$

The lowest appropriateness utility is firm by using resulting (17).

$$\begin{aligned} pbest_j &= f(y_j), \text{if } f(y_j) < f(pbest_j) \\ gbest_j &= f(y_j), \text{if } f(y_j) < f(gbest_j). \end{aligned} \tag{17}$$

In the position (x_j^d) of that, each element is studied by consuming space that amid the current situation and the current position among in the space and $gbest_j$. The next section will show the validation of proposed methodology with existing techniques.

4. RESULTS AND DISCUSSION

All of our results were created on a single NVIDIA GeForce RTX 2080 Ti GPU. All the code was applied using PyTorch2.

4.1. Datasets Description

4.1.1. Image-Based Emotion Recognition in the Wild

In this work, we have selected appropriate data sets to train the face extraction model. The data sets must address environments in the wild, in which numerous factors, such as obstruction, poses, lighting, etc., are uncontrolled. AffectNet [17] and RAF-DB [18] are the most extensive datasets that meet these criteria by far. The photographs in the data sets are acquired using emotional keywords on the internet. Experts note emotion labels to ensure trustworthiness. AffectNet has two kinds of data, manual and automatic, with over 1,000,000 photos marked with 10 categories of emotions and dimensional emotions (valence and arousal). We only utilized photos in the manual group of seven fundamental categories of emotions. For example, we used 283901 training photos and 3500 validation images. The RAF-DB dataset comprises of approximately 30,000 facial photographs in basic emotional categories that have taken lighting variations, arbitrary poses, and occlusion under in-the-wild situations. In this study, we selected 12,271 training images and 3068 evaluation images, all of them from the basic set of emotions.

4.1.2. Video-Based Emotion Recognition in the wild

We used the AFEW Dataset [19] to assess our work to determine face emotions in video clips. The video samples in the collection are obtained through uncontrolled occlusion, lighting and head positions from films and TV shows. Each video clip was selected based on its label including emotional keywords that reflect the emotion shown by the main topic. Using this information, we have assisted to deal with the challenge of temporality in the wild. We have used 773 trainings with the AFEW dataset and 383 validation video clips with labels for the seven basic types of emotion (anger, happiness, neutrality disgust, fear, sadness, and surprise).

4.2. Evaluation Parameters

As quantitative measurements in this investigation, we employed precision (Acc.) and the F 1 score. We also employed the average $Mean_{Acc}$ depending on the major diagonal of the standardized M_{norm} confusion matrix, for the performing results to be evaluated as in [18]. These measurements are derived accordingly.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{18}$$

$$F_1 = 2\frac{Precision.Recall}{Precision + Recall} \tag{19}$$

$$Mean_{Acc} = \frac{\sum_{i=1}^{n} g_{i,j}}{n} \tag{20}$$

$$Std_{Acc} = \sqrt{\frac{\sum_{i=1}^{n}(g_{i,j} - Mean_{Acc})^2}{n}}, \tag{21}$$

where $g_{i,j} \in \text{diag}(M_{\text{norm}})$ is the ith diagonal value of the normalized confusion matrix M_{norm}, n is the size of M_{norm}, and TP, TN, FP, and FN, respectively, are true positive, false positive, true negative, and false negative. Table 1 shows the confusion matrix for proposed methodology.

Table 1. Confusion matrix of Proposed Model with KGMO on Image-Based Emotion Recognition Evaluation.

Overall classification with Manual ID		Ground Truth					
		anger	happi-ness	neutrality disgust	fear	sad-ness	surprise
Predicted	anger	**93.30**	0.93	4.19	2.33	1.40	1.86
	happiness	2.79	**83.93**	5.89	2.65	2.33	2.33
	neutrality disgust	6.51	8.35	**87.37**	2.65	7.44	6.05
	fear	3.26	2.72	0.93	**88.19**	4.19	3.72
	sadness	2.79	3.14	4.36	1.86	**86.28**	7.91
	surprise	1.26	1.69	72.56	2.65	11.16	**84.70**

Table 2. Validation of Proposed Model with KGMO on Image-Based Emotion Recognition Evaluation.

CNN-Model	Acc (%)	F_1 (%)	$Mean_{Acc} \pm Std$
ConvNet	56.26	56.38	56.23 ± 11.18
DenseNet	61.51	61.50	61.51 ± 10.40
ResNet	61.57	61.46	61.57 ± 10.79
ConvNet-KGMO	81.23	81.79	77.08 ± 08.10
DenseNet-KGMO	83.64	83.81	76.96 ± 11.12
ResNet-KGMO	**87.22**	**87.38**	**82.45 ± 09.20**

4.3. Evaluation of proposed model with KGMO for two different datasets.

In this section, the different CNN models include ConvNet, DenseNet and ResNet are compared with and without KGMO techniques in terms of accuracy, F1-score and mean accuracy with standard metrics on two different datasets such as image-based and video-based datasets. Table 2 and Figure 3 shows the results of projected model with KGMO on image-based datasets.

In the accuracy experiments, the ConvNet, DenseNet and ResNet without KGMO achieved 56.26 %, 61.51 % and 61.57 %, where these techniques are implemented with KGMO technique and achieved 81.23 %, 83.64 % and 87.22 %. These results proved that ResNet with KGMO achieved better accuracy than other models. In the F1-score analysis, the ConvNet, DenseNet and ResNet without KGMO achieved 56.38 %, 61.50 % and 61.46 %, where these techniques are implemented with KGMO technique and achieved 81.79%,

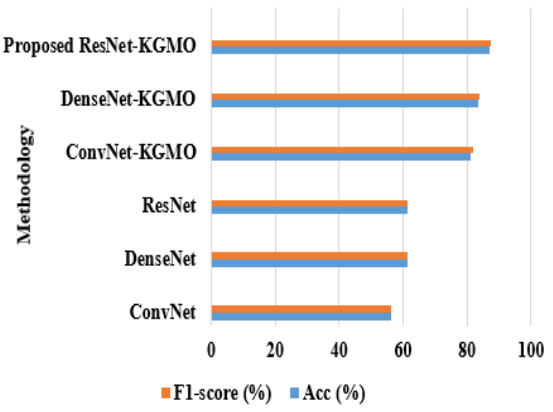

Figure 3. Graphical Representation of Proposed Model with KGMO in terms of accuracy and F1-score on image-based dataset.

Table 3. Validation of Proposed Model with KGMO on Video-Based Emotion Recognition Evaluation.

CNN-Model	Acc (%)	F_1 (%)	$Mean_{Acc} \pm Std$
ConvNet	51.70	46.17	46.51 ± 34.38
DenseNet	52.22	48.26	47.33 ± 31.73
ResNet	54.05	50.78	48.98 ± 32.28
ConvNet-KGMO	55.87	52.76	51.21 ± 29.87
DenseNet-KGMO	56.14	54.61	52.35 ± 25.53
ResNet-KGMO	**58.66**	**58.50**	$\mathbf{56.25 \pm 15.63}$

Table 4. Comparative analysis of proposed model with existing techniques on Video-based Dataset

Author	Technique	Accuracy (%)
Vielzeuf et al. [20] (2018)	Max Score Selection with Temporal Pooling	52.20
Fan et al. [21] (2018)	Deeply-Supervised CNN (DSN) Weighted Average Fusion	48.04
Duong et al. [22] (2019)	CNN Features with LSTM	49.30
Li et al. [23] (2019)	VGG-Face Features with Bi LSTM	53.91
Meng et al. [24] (2019)	Frame Attention Networks (FAN)	51.18
Lee et al. [25] (2019)	CAER-Net	51.68
Kumar et al. [26] (2019)	Noisy Student Training with Multi-level attention	55.17
Proposed (2021)	**CNN-KGMO**	**58.66**

83.81% and 87.38%. The mean accuracy of each technique without KGMO achieved nearly 56% to 61% with standard deviation of 11. But, when these techniques are implemented with KGMO, they achieved nearly 77 % to 82 % of mean accuracy with standard deviation of 9.20 on proposed model with KGMO. Table 3 and Figure 4 shows the results of proposed model, existing techniques by implementing with and without KGMO on video-based datasets.

In the accuracy experiments, the ConvNet, DenseNet and ResNet without KGMO achieved 51.70 %, 52.22 % and 54.05 %, where these techniques are implemented with KGMO technique and achieved 55.87 %, 56.14 % and 58.66 %. These results proved that ResNet with KGMO achieved better accuracy than other models, however the proposed technique achieved less performance than image-based dataset. In the F1-score analysis, the ConvNet, DenseNet and ResNet without KGMO achieved 46.17 %, 48.26 % and 50.78 %, where these techniques are implemented with KGMO technique and achieved 52.76 %, 54.61 % and 58.50 %. The mean accuracy of each technique without KGMO achieved nearly 46 % to 48 % with standard deviation of 32. But, when these techniques are implemented with KGMO, they achieved nearly 52 % to 56 % of mean accuracy with standard deviation of 15.63 on proposed model with KGMO. The reason for the less performance is that it is difficult to identify the proper emotion recognition while the video is continuously playing.

4.4. Comparative Evaluation of Proposed Technique with existing techniques

The following Table 4 shows the comparative analysis of proposed technique with various existing techniques in terms of accuracy. Figure 5 shows the graphical representation of proposed model on video-based dataset in terms of accuracy.

The existing techniques [20]-[21] such as DSN and CNN features with LSTM achieved nearly 49 % of accuracy. The

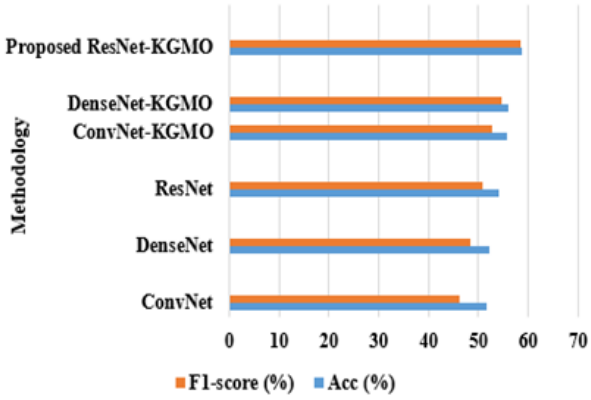

Figure 4. Graphical Representation of Proposed Model with KGMO in terms of accuracy and F1-score on video-based dataset.

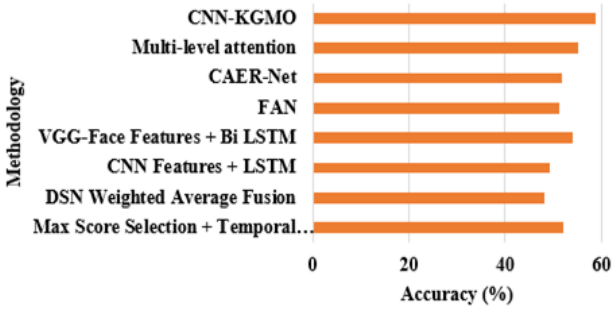

Figure 5. Graphical Representation of Proposed Model with KGMO in terms of accuracy on video-based dataset.

existing FAN [23] and CAER-Net [24] achieved only 51 % of accuracy, where the existing techniques achieved nearly 52 % to 53 % of accuracy. Noisy student training with multi-level attention [26]-[30] achieved 55.17 % of accuracy, where the proposed model achieved 58.66 % of accuracy. The reason is that the CNN is implemented with optimization technique called KGMO for fine-tuning and weight optimization.

5. CONCLUSION

Computer vision research on facial expression analysis has extensively studied in the past decades. In recognition of emotions, the success of this method has improved greatly over time. Our work has shown the general architectural model for developing a deep learning recognition system. The objective is to examine pre- and post-process methods. Using KGMO Algorithm, the smoothing and weight of CNN is optimized. This report also included the data sets available for academics in this subject on pictures and video. In different study, the advancement in this area is measured by different performance measures. The testing is performed on the various datasets, in which ResNet-KGMO has achieved an accuracy of 58.66 % in the video dataset, and an accuracy of 87.22 % for the image-based dataset. There is a very attractive scope of future developments in this area. Various deep learning multimodal and various architectures can be employed to increase the performance parameters. Besides the realization of the feelings alone, the intensity level can be raised further. This can contribute to forecasting the intensity of the feeling. In future works multi-medians may also be used; for example, the construction of a model with multi-data sets can be used with video and audio.

REFERENCES

[1] D. L. Carnì, E. Balestrieri, I. Tudosa, F. Lamonaca, Application of machine learning techniques and empirical mode decomposition for the classification of analog modulated signals, Acta IMEKO, vol. 9, 2020, no. 2, pp. 66–74. DOI: 10.21014/acta_imeko.v9i2.800

[2] P. C. Vasanth, K. R. Nataraj, Facial Expression Recognition using SVM Classifier, Indonesian Journal of Electrical Engineering and Informatics (IJEEI), vol. 3, no. 1, pp. 16-20, 2015. DOI: 10.11591/ijeei.v3i1.126

[3] Anurag De, Ashim Saha, A comparative study on different approaches of real time human emotion recognition based on facial expression detection, 2015 International Conference on Advances in Computer Engineering and Applications, ICACEA, Ghaziabad, India, 19-20 March 2015, pp. 483–487. DOI: 10.1109/ICACEA.2015.7164792

[4] M.A. Ozdemir, B. Elagoz, A. Alaybeyoglu, R. Sadighzadeh, A. Akan, Real time emotion recognition from facial expressions using CNN architecture, TIPTEKNO 2019, Izmir, Turkey TeknolKongresi 3-5 October 2019, pp. 1–4. DOI: 10.1109/TIPTEKNO.2019.8895215

[5] D. Sokolov, Patkin M, Real-time emotion recognition on mobile devices. In: and others, editor. Proc - 13th IEEE IntConfAutom Face Gesture Recognition, vol. 787. 2018. DOI: 10.1109/FG.2018.00124

[6] H. Kaya, F. Gürpınar, A. A. Salah, Video-based emotion recognition in the wild using deep transfer learning and score fusion, Image and Vision Computing. 2017, 65, 66–75. DOI: 10.1016/j.imavis.2017.01.012

[7] G. Betta, D. Capriglione, M. Corvino, A. Lavatelli, C. Liguori, P. Sommella, E. Zappa, Metrological characterization of 3D biometric face recognition systems in actual operating conditions, Acta IMEKO, vol. 6, 2017, no. 1, pp.33-42, 2017. DOI: 10.21014/acta_imeko.v6i1.392

[8] A. S. Volosnikov, A. L. Shestakov, Neural network approach to reduce dynamic measurement errors, Acta IMEKO, vol. 5, 2016, no. 3, pp. 24-31. DOI: 10.21014/acta_imeko.v5i3.294

[9] Y. Xie et al., Deception detection with spectral features based on deep belief network, ACTA Acustica, vol. 2, 2019, pp. 214-220.

[10] M. Healy, R. Donovan, P. Walsh, H. Zheng, A Machine Learning Emotion Detection Platform to Support Affective Well Being, in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 3-6 December 2018, pp. 2694–2700. DOI: 10.1109/BIBM.2018.8621562

[11] D. Maier, Analysis of technical drawings by using deep learning, M.Sc. Thesis, Department of Computer Science, Hochschule Mannheim, Germany, 2019

[12] C. Pramerdorfer, M. Kampel, Facial expression recognition using convolutional neural networks: state of the art. arXiv preprint arXiv:1612.02903 (2016). DOI: 10.48550/arXiv.1612.02903

[13] C. Huang, S. Narayanan, Characterizing types of convolutions in deep convolutional recurrent neural networks for robust speech emotion recognition, arXiv preprint arXiv:1706.02901 (2017). DOI: 10.48550/arXiv.1706.02901

[14] C. Lu, W. Zheng, C. Li, Chuangao Tang, S. Liu, S. Yan, Y. Zong, Multiple spatio-temporal feature learning for video-based emotion recognition in the wild, Proceedings of the International Conference on Multimodal Interaction. ACM, Boulder, CO, USA, 16-20 October 2018, 646–652. DOI: 10.1145/3242969.3264992

[15] A. Zadeh, P. Pu Liang, N. Mazumder, S. Poria, E. Cambria, L.-P. Morency, Memory fusion network for multi-view sequential learning, Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2-7 February 2018, 9 pp. DOI: 10.48550/arXiv.1802.00927

[16] P. Liang, R. Salakhutdinov, L. P. Morency, Computational Modeling of Human Multimodal Language: The MOSEI Dataset and Interpretable Dynamic Fusion, 2018.

[17] A. Mollahosseini, B. Hasani, M. H. Mahoor, AffectNet: A Database for Facial Expression, Valence, and Arousal Computing, in the Wild. IEEE Trans. Affect. Comput. 2019, 10, 18–31. DOI: 10.48550/arXiv.1708.03985

[18] S. Li, W. Deng, J. Du, Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017, pp. 2584–2593. DOI: 10.1109/CVPR.2017.277

[19] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Collecting large, richly annotated facial-expression databases from movies. IEEE Multimed. 2012, 19, 34–41. DOI: 10.1109/MMUL.2012.26

[20] V. Vielzeuf, C. Kervadec, S. Pateux, A. Lechervy, F. Jurie, An Occam's razor View on Learning Audiovisual Emotion Recognition with Small Training Sets, 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018, pp. 589–593. DOI: 10.48550/arXiv.1808.02668

[21] Y. Fan, J. C. K. Lam, V. O. K. Li, Video-based Emotion Recognition Using Deeply-Supervised Neural Networks, 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018, pp. 584–588. DOI: 10.1145/3242969.3264978

[22] D. H. Nguyen, S. Kim, G. S. Lee, H. J. Yang, I. S. Na, S. H. Kim, Facial Expression Recognition Using a Temporal Ensemble of Multi-level Convolutional Neural Networks, IEEE Trans. Affect. Comput. 2019, 33, 1. DOI: 10.1109/TAFFC.2019.2946540

[23] S. Li, W. Zheng, Y.Zong, C. Lu, C. Tang, Bi-modality Fusion for Emotion Recognition in the Wild, International Conference on Multimodal Interaction, Jiangsu, China, 14–18 October 2019, pp. 589–594. DOI: 10.1145/3340555.3355719

[24] D. Meng, D. Peng, Y. Wang, Y. Qiao, Frame Attention Networks for Facial Expression Recognition in Videos, IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019, pp. 3866–3870. DOI: 10.48550/arXiv.1907.00193

[25] J. Lee, S. Kim, S. Kim, J. Park, K. Sohn, Context-aware emotion recognition networks, IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019, pp. 10142–10151. DOI: 10.48550/arXiv.1908.05913

[26] V. Kumar, S. Rao, L. Yu, Noisy Student Training Using Body Language Dataset Improves Facial Expression Recognition. In Computer Vision—ECCV 2020 Workshops, A. Bartoli, A. Fusiello, Eds., Springer International Publishing: Cham, Switzerland, 2020, pp. 756–773. DOI: 10.48550/arXiv.2008.02655

[27] F. Vurchio, G. Fiori, A. Scorza, S. A. Sciuto, Comparative evaluation of three image analysis methods for angular displacement measurement in a MEMS microgripper prototype: a preliminary study, Acta IMEKO, vol.10, 2021, no.2, pp.119-125. DOI: 10.21014/acta_imeko.v10i2.1047

[28] H. Ingerslev, S. Andresen, J. Holm Winther, Digital signal processsing functions for ultra-low frequency calibrations, Acta IMEKO, vol.9, 2020, no.5, pp. 374-378. DOI: 10.21014/acta_imeko.v9i5.1004

[29] M. Florkowski, Imaging and simulations of positive surface and airborne streamers adjacent to dielectric material, Measurement, vol. 186, 2021, pp.1-14. DOI: 10.1016/j.measurement.2021.110170

[30] G. Ke, H. Wang, S. Zhou, H. Zhang, Encryption of medical image with most significant bit and high capacity in piecewise linear chaos graphics, Measurements, vol. 135, 2021, pp. 385-391. DOI: 10.1016/j.measurement.2018.11.074