Hybridization of Deep Sequential Network for Emotion Recognition Using Unconstraint Video Analysis

P. Naga Bhushanam, Selva Kumar S. *

School of Computer Science and Engineering, VIT-AP University, Andhra Pradesh, India
Emails: nagabhushnam.22phd7012@vitap.ac.in; selvakumar.s@vitap.ac.in

Abstract

The reliable way to discern human emotions in various circumstances has been proven to be through facial expressions. Facial expression recognition (FER) has emerged as a research topic to identify various essential emotions in the present exponential rise in research for emotion detection. Happiness is one of these basic emotions everyone may experience, and facial expressions are better at detecting it than other emotion-measuring methods. Most techniques have been designed to recognize various emotions to achieve the highest level of general precision. Maximizing the recognition accuracy for a particular emotion is challenging for researchers. Some techniques exist to identify a single happy mood recorded in unrestricted video. Still, they are all limited by the processing of extreme head posture fluctuations that they need to consider, and their accuracy still needs to be improved. This research proposes a novel hybrid facial emotion recognition using unconstraint video to improve accuracy. Here, a Deep Belief Network (DBN) with long short-term memory (LSTM) is employed to extract dynamic data from the video frames. The experiments conducted uses decision-level and feature-level fusion techniques are applied unconstrained video dataset. The outcomes show that the proposed hybrid approach may be more precise than some existing facial expression models.

Keywords: Unconstrained video; emotion recognition; prediction; network model; feature representation

1. Introduction

The study of emotion identification has grown to be an exciting field with many potential uses. It may be used, for instance, in the advertising sector to comprehend clients' emotions [1]. Recognizing facial expressions and physiological states in criminal justice and judicial systems can help detect lies. In medical applications, it can be helpful in the diagnosis of various illnesses such as Parkinson's and anxiety. We may also use emotion detection systems on the web and linked world to define the viewers' emotions and moods when recommending books, movies, or products in online recommender systems [2]. For recognizing human emotional states, different emotion identification systems are available. These include physiological emotion recognition, verbal emotion recognition and facial expression recognition (FER) [3]. Multimodal systems that can recognize human emotions may also be created by combining these systems. More so than other non-verbal cues, facial expressions are a crucial predictor of someone's emotions. When asked to express a feeling, the standard FER systems were designed to identify just a high level of accuracy of over 97% for the lab-controlled databases of face expressions [4]. The precision was brought down to a very low accuracy by using these techniques on more challenging real-world datasets acquired from film or television. With Acted Facial Expressions in the Wild (AFEW), the best was 40%, as an illustration. Several variables affect how well emotion recognition systems work, including diverse backdrops, extreme head posture fluctuation, changes in lighting, and numerous occlusions and sounds [5].

The most prevalent facial emotion people use in daily communication is a smile or a chuckle. A joyful emotion recognition module may define numerous applications, much like FER systems. For instance, investigating whether a grin is real or staged can help us predict further features and behaviours [6]. Single emotion detection's solitary accuracy is insufficient for practical applications even though specific remarkable algorithms have been built for identifying with comparatively high average accuracy, delight, surprise, contempt, rage, sadness, and

Doi : https://doi.org/10.54216/JCIM.130209
Received: January 05, 2024 Revised: Mrach 09, 2024 Accepted: May 04, 2024

fear. All six basic emotions are identified. In this study, we practice employing focused optimum techniques for a single emotion to identify just the joyful expression from the unconstrained films to improve accuracy. After more research in this area, we have discovered a couple of reliable, independent algorithms for joyful emotion identification from videos [7]. As a result, the methods now in use, which were created to study the perfect faces taken in lab photography and cannot be applied in real-world scenarios, are challenged by the fact that films of people in the field often have considerable variances in their head postures. Due to the face's imperfect availability, various view angles for photographing human faces provide substantial difficulty in removing face features, such as texture and landmark characteristics [8]. Therefore, it is crucial to create systems that can recognize emotions in any situation from the unrestricted recordings ordinarily recorded in the wild [9] – [10].

This work presents a novel hybrid method known as the DBN with Long Short Term Network ($DBN - LSTM$) recognition technique that uses visual data by extracting two separate deep feature types. First, textural characteristics cannot detect facial emotions when looking at unfinished human faces captured from arbitrary viewing angles. Additionally, face landmarks are required since they are crucial in systems that recognize expressions. As a result, this study considers the complementing data from landmarks and facial texture. Second, we apply deep learning methods by using hybrid network consisting of DBN and LSTM; computer vision is employed to extract textured and spatial-temporal data from subsequent video frames. Using a location on the face as a reference, measure the distance between the mouth and eye landmarks and a time series of precise facial changes may be captured. The time series of facial changes based on landmarks tracks muscle movements between successive frames and can display changes throughout an emotion. These discriminative characteristics are extracted, and their fusion at the system's effectiveness is evaluated at the feature and decision levels. Using feature-level fusion, the last aspect of a vector is created, and an utterly interconnected layer divides the films into joyful and sad categories. The evolutionary-based weighted sum and weighted mean are used to recognize emotions in decision-level fusion to merge the outcomes.

All six fundamental emotions are recognized by FER systems most similar to ours. Our suggested system can significantly increase accuracy by, for instance, selecting action units exclusively pertinent to a joyful mood and intends to illustrate the single (for example, happy) emotion detection utilizing specialized optimization tactics. The following are some ways in which it differs from these studies: To identify pleasant emotions, DBN and LSTM are exclusively significant for that emotion. The deep neural network architecture that is being suggested is more straightforward but more accurate than the existing one. Results from experiments on the AFEW unconstrained video datasets show that the anticipated strategy has higher level of accuracy. Following is an overview of the work's contribution:

1) This work presents a distinct, lighter, quicker, independent method for identifying happy emotions in videos that differs from the previous approaches that primarily rely on recognizing all six basic emotions through facial expressions.

2) This work provides a distinctive $DBN - LSTM$ technique that creates a deep modelling framework for recognition based on spatial and temporal data. A recurrent neural network with a hierarchy is called LSTM, which is used for handling the visual input matched with landmark trends to enhance system performance and produce more accurate results.

3) This work creates a recognition algorithm that can successfully identify expressions in footage shot human faces may not always be fully seen in the outdoors due to unpredictable angles) taking into account several frame apexes.

4) Using the findings from the experiment, two unrestricted datasets, and a sizeable multi-party chat dataset, we test the trained system to show that our suggested $DBN - LSTM$ technique works accurately and efficiently.

The work is drafted as follows: Section 2 details the broader analysis of diverse approaches. The methodology is outlined in section 3. The experimentation is conducted in section 4, with outcomes in section 5.

2. Related works

The various backdrops, differences in lighting, and other disturbances caused the FER systems' accuracy to be low when applied to datasets from the actual world. Traditional approaches cannot disregard the output feature vector's changing backdrop or illumination because the features they extract are contextually dependent [11]. As a result, various essential pre-processing techniques are used to unite the films. In contrast, many deep neural networks have a significant capacity for learning novel patterns and extracting fine-grained without pre-processing. As a result, they have been used in several research fields, including photo classification and facial recognition [12]. LSTM is a specific recurrent neural network (RNN), convolutional neural networks (CNN),

Doi : https://doi.org/10.54216/JCIM.130209
Received: January 05, 2024 Revised: Mrach 09, 2024 Accepted: May 04, 2024

and, among other deep architectures, have been shown to perform better at processing consecutive frames. There are several deep FER systems available. To extract features from movies, a hybrid technique that combines geometric relations and features like CNN and DSIFT has been created [13]. The system's effectiveness is then evaluated using spontaneous datasets for practical applications. Stacks of CNN blocks are employed to categorize the feelings. Layers using convolution, batch normalization, and dropout are used in eight blocks once the input photos are pre-processed to grayscale. Despite the positive results, it could not be used with unrestricted datasets [14] – [15].

The features were extracted using three distinct CNN frameworks. The temporal characteristics were also extracted using RNN and rectified linear hidden units (ReLu). Combining the features results in more accuracy. However, all three presented networks had an over-fitting issue [16]. An attention network is suggested to identify local and global traits. The technique uses bilinear attention pooling and certain convolutional filters to identify the emotions in the pictures. To monitor two fundamental problems, [17] supplies an STN (Spatial Transformer Network Block)- equipped deep neural network architecture. The size of the input picture impacts CNNs and has different input images with various sizes. The model uses several VGG16 networks to scale all inputs to a specified size before identifying the emotions in the photos. It might be challenging to choose a scale value that applies to all images [18]. Within the GoogleNet network, the Inception layers have been developed. These kinds of layers combine many convolution filters to extract face characteristics at different sizes. There are several ways that Inception architectures are employed for facial expression recognition. Another CNN that combines inception and residual blocks is available [19]. Its wide and deep suggested architecture eliminates unnecessary filters. It speeds up training, but the issue is that a regular CNN cannot assemble the temporal characteristics of succeeding frames. As a result, 3D-CNN is adopted including video emotion recognition and 3D object identification, to gather temporal and spatial information [20].

RNNs can map temporal dynamic activity utilizing predetermined hidden neurons and internal states (memory). Even though they show a decent capacity for trait extraction, they have shown poor memory and retention of lengthy sequences [21]. As a result, by including additional remembering and forgetting units, the LSTMs are used to follow their disappearing slopes. For identifying the emotions in the videos, tuned CNN and LSTM networks have been merged in some FER systems [22]. The retrieved VGG16 features are combined with an LSTM layer to create a hybrid framework [23]. Their technique may record both temporal and spatial data in video sequences. The network based on the CNN-LSTM is suggested and to get better results, movies with significant head posture fluctuation and occlusions have been deleted. The only datasets on which this approach can attain high accuracy are those from laboratories [24] – [25].

3. Methodology

As discussed earlier, a lot of deep-learning architecture has been used to identify facial emotions. As the functionality of LSTM and Deep Belief Network demonstrated superior performance in recognizing face expressions individually, several consecutive video frames are used in this work to extract spatial-temporal properties. The anticipated recommendation approach is depicted in Fig 1. This work is motivated to investigate the prospect of using such a system to train the computer to demonstrate genuine enjoyment, allowing for the development of specific strategies. The dynamics of time aspects are essential for conveying emotions in video processing. So, towards the conclusion of the network design, this work supplements the extracted characteristics using the dynamics features extracted using LSTM unit. Additionally, the significance of action units in face expression recognition systems must be considered. Therefore, this work may use a DBN to extract deep facial landmarks' characteristics. Ultimately, this work hybridizes two models at the feature and decision levels. These characteristics are combined using a straightforward feature fusion technique. Thus, a fully integrated system can distinguish between joyful and sad movies. The final class labels are determined following the independent classification of the features.

Doi : https://doi.org/10.54216/JCIM.130209
Received: January 05, 2024 Revised: Mrach 09, 2024 Accepted: May 04, 2024

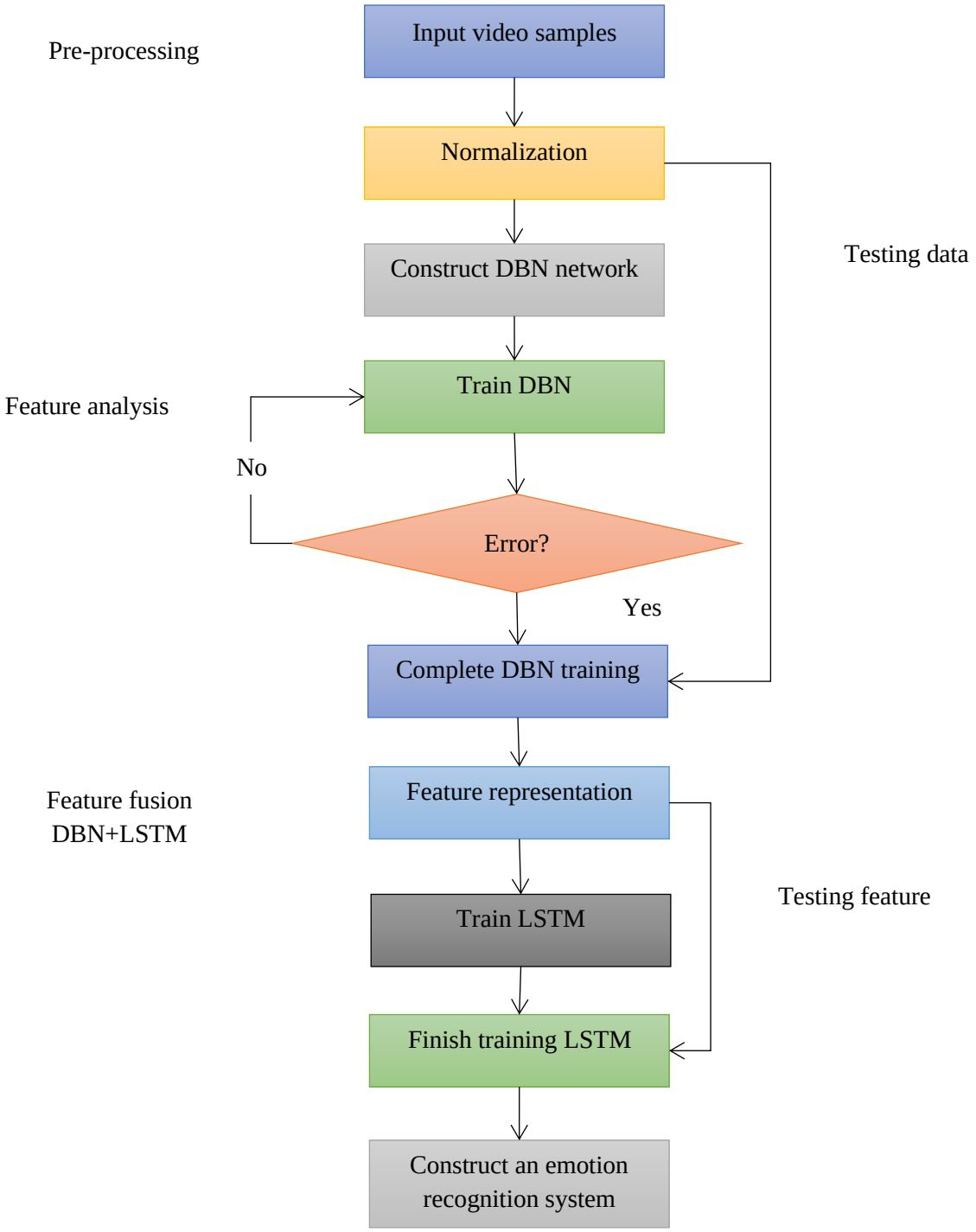

Figure 1: Flowchart construction

3.1. Deep Belief Networks (DBN)

DBNs are effectively adopted for variety of recognition tasks, including the identification of handwritten numbers, the recognition of objects, and the modelling of human motion. Several restricted Boltzmann machines (RBMs) are stacked to create DBNs (See Fig 2), which are probabilistic generative models. RBMs are two-layered shallow networks consisting of "visible" units representing the input data and a layer of "hidden" units learning to specify qualities. In RBM design, no connections exist among two units on the similar layer. Yet, every visible unit on each hidden unit on the hidden layer is connected to the visible layer. The hidden and transparent binary-valued units of the conventional form of RBM allow each unit to only exist in one of two

Doi : https://doi.org/10.54216/JCIM.130209
Received: January 05, 2024 Revised: Mrach 09, 2024 Accepted: May 04, 2024

states, "0" or "1". A unit's bias, connection weights, and the states of other units form a sigmoid function that determines the likelihood of setting it to "1".

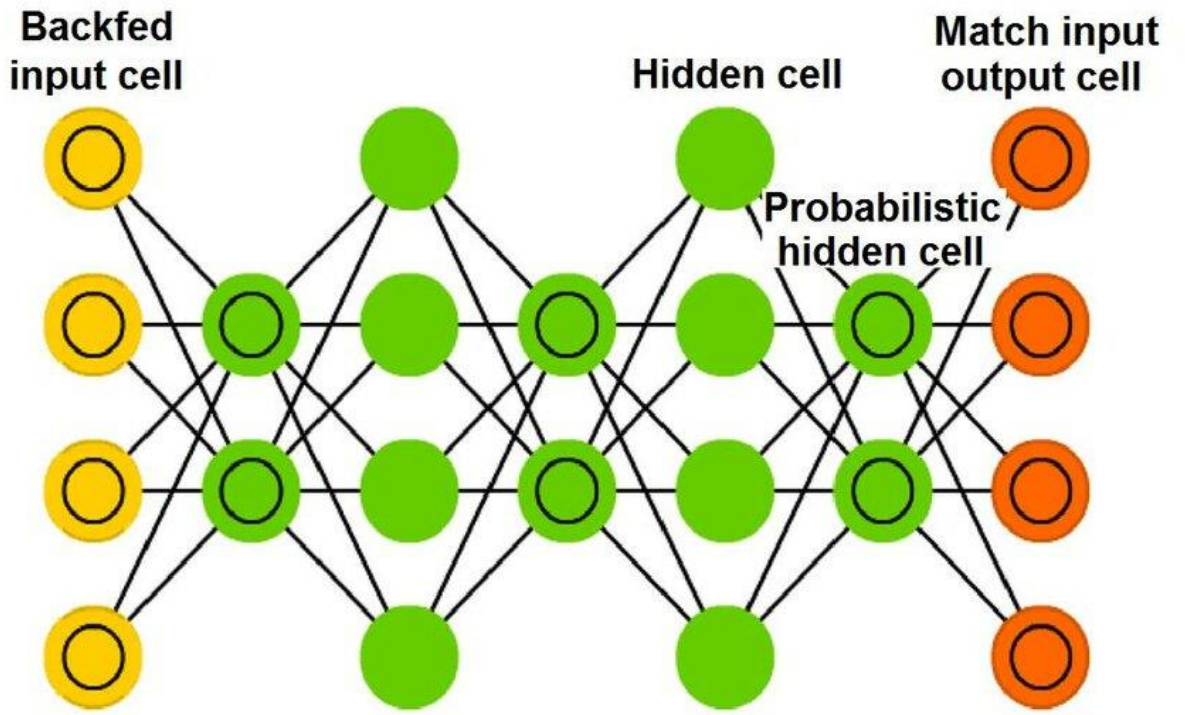

Figure 2: DBN architecture

The preliminary stage is pre-processing, where the input is acquired from the data, and the successive layers are formed from the RBM using training to attain the local optimal parameters. Then, output is fed to third layer, and then third and second layer includes RBM. The parameters are trained with unsupervised learning, and RBM is piled layer by layer to attain the local optimal solution. The weighted factor approximation is done, which results in the diminishment of training time. The under-fitting problems are resolved during the training process. The successive stage is the fine-tuning stage, where the back-propagation is utilized to fine-tune the entire network parameters to attain global optimum using the up-down algorithm, an enhanced version of training DBN with labelled data. The label collection is applied over the top layer in the learning process to determine the network boundaries. The algorithm does not suffer from averaging issues, resulting in low recognition weights. Diverse RBM can be stacked into DBN, and the probability distribution function is expressed as:

$$p\,(v,\theta) = \sum_{h} \frac{e^{-E\,(v,h,\theta)}}{\sum_{v,h} e^{-\,(v,h,\theta)}} \tag{1}$$

$$p\,(v,\theta) = \frac{1}{z(\theta)} \sum_{h} \exp(v^T wh + b^T v + a^T h) \tag{2}$$

$$p\,(v,\theta) = \frac{1}{z(\theta)} e^{(b^T v)} \prod_{j=1}^{F} \left(1 + \exp\left(a_j + \sum_{i=1}^{D} w_{ij} v_i\right)\right) \tag{3}$$

Here, $p\,(v,\theta)$ is the probability distribution function, $E\,(v,h,\theta)$ represents the energy function, $z(\theta)$ represents the regularization factor, v is the input vector, h represents output vector, and w represents weighted vector, a *and* b are biases of visible and hidden layers.

3.2. Long Short-Term Memory

The enhanced Recurrent Neural Network (RNN) model introduces the concept of time notation, which is comparatively better than the conventional feed-forward neural network. It can relate data and time. When the time sequences rise, layer-wise deepness is encountered in the back-propagation hidden layer. It poses a

Doi : https://doi.org/10.54216/JCIM.130209
Received: January 05, 2024 Revised: Mrach 09, 2024 Accepted: May 04, 2024

vanishing gradient or gradient explosion problem during the training process. Thus, RNN cannot recollect the association between long-time and current information in time sequence for long-time sequential data. The memory cell concept is introduced in the advanced LSTM architecture to resolve these issues, as in Fig 3. The memory cells intend to substitute the intermediate hidden layers encountered in conventional RNN, composed of three gate structures: input, output, and forget gate. Thus, LSTM is more appropriate than the traditional RNN for predicting time series-based video sequences. The LSTM input comprises prior time output h_{t-1} and present time input x_t, which is provided as the input to memory cells to show the information related to forget gate (f_t). It deals information from prior memory cell which is to be terminated and expressed as:

$$f_t = \sigma\left(w_{fx}x_t + w_{fh}\,h_{t-1} + b_f\right) \tag{4}$$

Here, σ specifies the sigmoid function, w_{fh} and w_{fx} represent weight from input to hidden layer of forgot gate and weight from prior hidden layer to hidden layer of forget gate at $'t'$, and b_f specifies the bias of forget gate respectively. The memory cell has to notify revised information, which is merged with $i_t s\, f_t, g_t$ and c_{t-1}. The expression is utilized to demonstrate the process where the memory cells depict the revised information:

$$i_t = \sigma(w_{ix}x_t + w_{ih}h_{t-1} + b_i) \tag{5}$$
$$g_t = \sigma(w_{gx}x_t + w_{gh}h_{t-1} + b_g) \tag{6}$$
$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \tag{7}$$

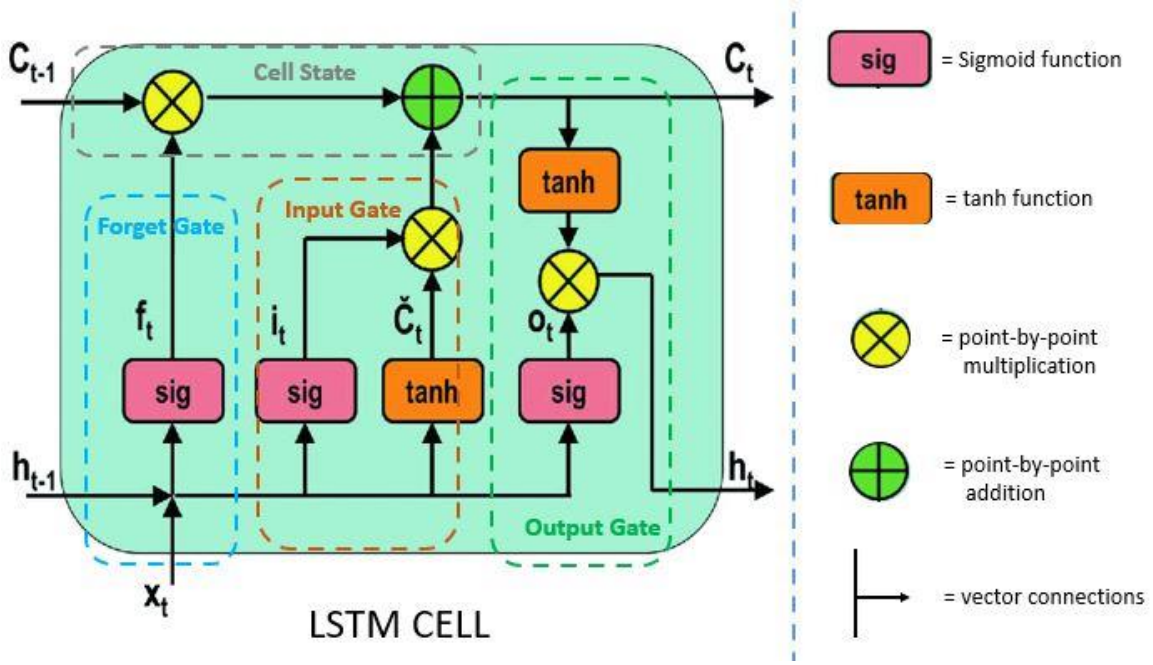

Figure 3: LSTM architecture

Here, w_{ix} and w_{ih} specify the hidden and input layer of the input gate and weight of the prior hidden layer to the hidden layer of input data at $'t'$ time, b_i and b_g specify the bias of candidate and input gate, c_t represents cell state at the current time, $\tanh(.)$ which defines the hyperbolic tangent function and $\otimes$ represents element-wise multiplication. In the provided memory cell, the network determines the present state c_t output and output gate o_t, is depicted as:

$$o_t = \sigma(w_{ox}x_t + w_{oh}h_{t-1} + b_o) \tag{8}$$
$$h_t = o_t \otimes \tanh(c_t) \tag{9}$$

Here, w_{ox} and w_{oh} represent the weight from the hidden layer input to the output gate and the weight from the prior hidden layer to the output gate's hidden layer at $'t'$ time, and b_o represents the bias of the output gate, respectively.

Doi : https://doi.org/10.54216/JCIM.130209
Received: January 05, 2024 Revised: Mrach 09, 2024 Accepted: May 04, 2024

3.3. Hybridization

The prediction performance decreases with time gradually. The extracted feature needs to retain the required information of input raw data. DBN poses a more robust non-linear mapping performance to haul out the patterns required. Video collected from the $'t'$ time step, the gathered data T is pre-processed to [0 1] normalization based on $t^* = (t - t_{min})/(t - t_{max})$ where t represents raw data and t_{max} and t_{min} means the maximal and minimal raw data. After training DBN, the normalized training set T^* is provided to train LSTM. The trained DBN is acquired by fine-tuning the normalized video dataset. At last, the features (h_n) are provided as:

$$\begin{cases} h_1 = \sigma(W_1 t + b_1) \\ h_2 = \sigma(W_2 t_1 + b_2) \\ \vdots \\ h_n = \sigma(W_n t_{n-1} + b_n) \end{cases} \quad (10)$$

Here, W_i and b_i are DBN parameters ($i = 1, 2, \dots, n$), x specifies raw video signal, σ represents the DBN activation function, n represents hidden layer and h_i represents hidden layer output. It is essential to merge the features via a suitable technique to use the extracted feature. The LSTM comprises input, output and hidden layers and maps the high dimensional data to a 2D structure. Thus, the LSTM is provided to merge the features and classification to acquire the benefits of degraded information. After determining LSTM, the trained feature h_n is adopted to train LSTM. Then, trained LSTM acquires the competitive formula for feature training. At last, the feature is provided to LSTM for the construction of $h_i = ||f - m_{bmu}||$ where h_i represents time features at $'t'$ time, f represents layer input features, and m_{bmu} represents the weighted vector of finest matching unit of provided input. As explained before, LSTM shows lower prediction performance with smaller faulty datasets. Therefore, the learned knowledge from the input source is initiated to enhance LSTM performance over the targeted domain. Usually, the source dataset is diverse; however, it is related to the targeted domain. The target and source dataset is expressed as:

$$Y_t = \{Y_t^{train}, Y_t^{test}\} \quad (11)$$
$$Y_s = \{f_1^s, f_2^s, \dots, f_n^s\} \quad (12)$$

Here, Y_s specifies source data, Y_t represents target data, Y_t^{train} and Y_t^{test} represent training and testing data. The source data is provided to pre-train LSTM. Then, knowledge acquired from the source assist the entire target task via the learning function. The LSTM's learned parameters are supplied to the target model. At last, the target training data is used to fine-tune LSTM parameters to fit the targeted data. The learned parameters from the source are effectually used for target optimization. It utilizes the features from the source domain to enhance learning tasks. It can improve LSTM's prediction accuracy for both small and large training data.

4. Numerical results and discussion

Here, this work examines the suggested technique using unconstraint video datasets collected in the field and shows its efficacy by contrasting it with several facial expression detection techniques. The datasets are briefly introduced, followed by an explanation of the network parameters and the outcomes of combining at the decision and feature levels. The results of applying certain qualities are also examined to highlight the significance of each feature vector.

4.1. Dataset description

To assess the universality of the suggested strategy on various difficulty levels, we have chosen three benchmark facial expression datasets from a wide range: AFEW. Naturalistic facial reaction recordings were recorded in everyday situations. It was required of the international participants to view video commercials. They permitted the webcam to capture their face while they watched videos. The server has now received 1044 webcam videos as a consequence. 545 of them were personally tagged with the smile facial action units. Since the locations where participants shot their films weren't specified, each video's illumination was unique. Most films featured frontal views, although several featured other head poses. Almost all grin films include three sorts of frames (onset, apex, and offset). Some movies are lengthy (around 5-6 minutes), and the first five minutes need more meaningful information since people record their videos. The majority of the faces are seen from the front, and

Doi : https://doi.org/10.54216/JCIM.130209
Received: January 05, 2024 Revised: Mrach 09, 2024 Accepted: May 04, 2024

there are films with frames' onset, apex, and offset, as well as different films' expressions of happiness, which vary somewhat from one another (small intra-class distance).

We may get unbalanced data if we group joyful films into one category and all other emotions into another. Data is selected randomly from all the emotions (except cheerful) to create three distinct training and validation sets for our model. The experimentation has picked 24 frames per second even though the movie lengths range from 0.75 to 4 seconds. A frame order for movies with fewer than 24 frames is maintained. To verify that the movie had at least 24 frames, we duplicated the initial several and last several frames. Since nearly every frame of the videos contains the pinnacle expressions, the max voting method is then utilized to categorize the videos into happy and unhappy groups. Due to the significant intra-class disparity amongst joyful videos, it is classified as the second most challenging dataset. Numerous performers of different racial and ethnic backgrounds convey the feelings. It's possible that onset, apex, and offset frames will only sometimes appear in the videos. Individually, sets 1743, 163, and 402 films are separately given a pleasant feeling classification. Each statement was given a feeling and emotion classification. Happiness, disgust, sorrow, fury, fear, and surprise are the six fundamental feelings used to describe emotions, whereas there are three types of sentiments—neutral, positive, and negative. Selecting films that reflected a range of emotions allowed us to establish two unpleasant training groups at random, just as how the AFEW data were prepared.

4.2. Pre-processing

Recognizing faces in video frames, the network resizes them to meet their input size before our system is trained to distinguish pleasant emotions. The MATLAB 2020a toolkit has been used to recognize the faces using the proposed $DBN - LSTM$, or multi-task convolutional neural network. A piecewise affine was also developed to precisely align faces and evaluate the accuracy of the landmarks. To account for 24 frames per second of 24×199×199×3, we scaled the aligned faces to 199×199. It has allowed us to align video frames with a size comparable to a network input for the $DBN - LSTM$ device. The OpenFace facial identification module's convolutional expert's constrained local model (CE-CLM) was trained at various head angles; however, despite this, we found that the system occasionally failed to recognize any landmarks or mark the face with the new landmarks. When the position of the head varied greatly, it happened. A few failed instances of incorrectly recognized landmarks are shown. It's crucial to locate the proper 3D facial markers in the suggested method since we want to create a system that can realize pleasant emotions accurately for unconstrained footage. The 2D/3D facial landmarks are located using the proposed approach, which took the role of a recommended for the hierarchical, parallel, and multi-scale blocks (four distinct architectural blocks in the network). The precise landmarks found using the technique are shown. We selected 24-3D landmarks along with the lips and eyes; we gathered 16 3D facial contour points from the 68 facial landmarks. The size of the image-like matrix, $24\ x\ 120$, corresponds to the recommended landmarks supplied by the CNN network. Due to the lack of independence, we use a 10-fold CV during validation, training and test sets to verify the generalizability of the methodology. Ten data groups are created, nine of which are training sets, and one is a testing set. To avoid problems brought on by unbalanced data, we take into account a nearly equal expression in both groups' distributions. The dataset's average results from ten runs are then given. For the AFEW dataset, this work created balanced validation and training sets and as a consequence, we also explained the average findings.

4.3. Network parameters

$DBN - LSTM$ networks must be trained in the suggested method to extract characteristics from faces and architectural aspects accurately. The proposed method may be trained using a total of 409,068,581 trainable parameters. Network initialization is crucial to training because it prevents over-fitting and premature convergence. Initially set at 0.005, the learning rate dropped to 0.0003 for the 20 epochs, 0.0002 during the next 20, and 0.00001 during the training procedure's last 20 epochs. Activation of Rectifier Linear Unit (ReLu) was utilized to fire all layers other than pooling and final layer, and to integrate the results of the individual layers, a linear function was used. The sigmoid function decided the outcome. This work adopts Adam optimizer to optimize the training process, and the loss was calculated using binary cross-entropy. The process was continued until the model reached the smallest lost mistake. On a machine running the GeForce GTX 1080 Ti GPU from NVIDIA, the training phase was carried out with a batch size 16. The suggested method's training and validation are displayed in Table 1. We demonstrate how long it takes for feature-level fusion to combine features using weighted mean, and DBN employs the decision-level method. The suggested technique for feature-level fusion was completed about the AFEW dataset in 3, 2, and 30 and a half hours based on how big each dataset is. The decision-level fusion models for the AFEW datasets took the longest because of the complexity. The decision-level models using weighted means ran for the corresponding times for the AFEW datasets are 3, 2, and 30 and a half hours.

Doi : https://doi.org/10.54216/JCIM.130209
Received: January 05, 2024 Revised: Mrach 09, 2024 Accepted: May 04, 2024

4.4. Discussion

As was previously said, we investigated the recommended course of action at the fusions of the feature and decision levels. The AFEW datasets were used to assess the trained models. The values of validation and training losses for three datasets are described after 200 iterations. By achieving the lowest loss, it is clear that the three datasets' training processes have converged. Table 2 displays average accuracy of assessing the trained models using fully connected layers (feature-fusion) and weighted sum (decision fusion) of AFEW dataset. We included extracted feature and single feature vector formats while presenting the results. In this instance, the last fully connected layer, followed by the DBN classifier, received distinct feeds from each feature vector. Although action units and grin classifications have been used to identify the videos in AFEW, we only considered the smile classification throughout the assessment stage. After the extracted features are fused at the feature level, Table 3 shows that the most remarkable accuracy for the AFEW datasets is 94.89%, respectively. Combining characteristics gives the system supplementary knowledge to fend off potential disturbances.

For the training cycle that results in the greatest, Fig 3 shows the reliability of AFEW variations throughout 195 epochs. Due to the concentration of faces and frontal views in most movies, with some minor variations, AFEW's accuracy is more accurate than the two other datasets. The scenes weren't picked since the AFEW dataset's main objective was based on the text's description of emotions just for their faces and low-resolution aligned faces; it is evident that the AFEW dataset's findings are superior to those of the dataset. It is an excellent example of a dataset collected in the field where faces were far from the camera and not concentrated in the image. Consequently, low-resolution faces were created by identifying the faces and aligning them, as shown in Fig 4. It illustrates the suggested system's robustness and well-generalized nature, and the results were satisfactory and significant-good. The AFEW dataset (92.76%) showed improved accuracy when the features were classified individually. However, when we used the fully linked layers for the classifier and the fusion, the accuracy decreased, and the concatenated features produced superior accuracy. It demonstrates that feature-level fusion beats decision-level based on AFEW datasets with 94.89%, respectively, compared to other approaches.

Table 3 demonstrates how we used the weighted mean to produce the decision-level model's optimum weights. When employing $DBN - LSTM$ with AFEW, the recognition accuracy is 94.52%, higher than other approaches using weighted mean decision levels. We discovered a similar result for the AFEW dataset; additionally, [the weighted mean decision-level model's accuracy is somewhat greater (93.74%) than other model's (93.29%). $DBN - LSTM$ has compared its accuracy to those in Table 4 and prediction outcomes even if it has improved accuracy (by around 2%). Utilizing $DBN - LSTM$ to optimize the weights helps obtain more accuracy while sacrificing a tolerable efficiency. It is usually preferable to use a slower technique with more accuracy than a quicker one with a lower rate when using the pre-trained models, just like in any practical application.

Table 1: Training and validation

	Angry	**Disgust**	**Fear**	**Happy**	**Sad**	**Surprise**	**Neutral**	**Total**
Training	132	73	80	149	116	75	145	774
Validation	65	39	45	64	60	47	64	384

Table 2: Average accuracy comparison with baseline LSTM methods

Methods	**AFEW**
Standard LSTM	90.5%
Hybrid LSTM with SVM	92.5%
BiLSTM + SVM	93.1%
Weighted features with GA	93.7%
LSTM with weighted mean analysis	94.6%
CNN with LSTM	93.2%
DBN − LSTM	**95.6%**

Table 3: Comparison with existing approaches

Methods	**AFEW**
Fuzzy system	87%
LBP	88%
HOG	89%
ELM	91%
SIFT	90%

Doi : https://doi.org/10.54216/JCIM.130209
Received: January 05, 2024 Revised: Mrach 09, 2024 Accepted: May 04, 2024

RNN	92%
DBN − LSTM	**95.6%**

Table 4: Performance analysis based on AFEW

Emotions	**Prediction accuracy**
Angry	86%
Disgust	28%
Fear	55%
Happy	78%
Neutral	54%
Sad	55%
Surprise	20%

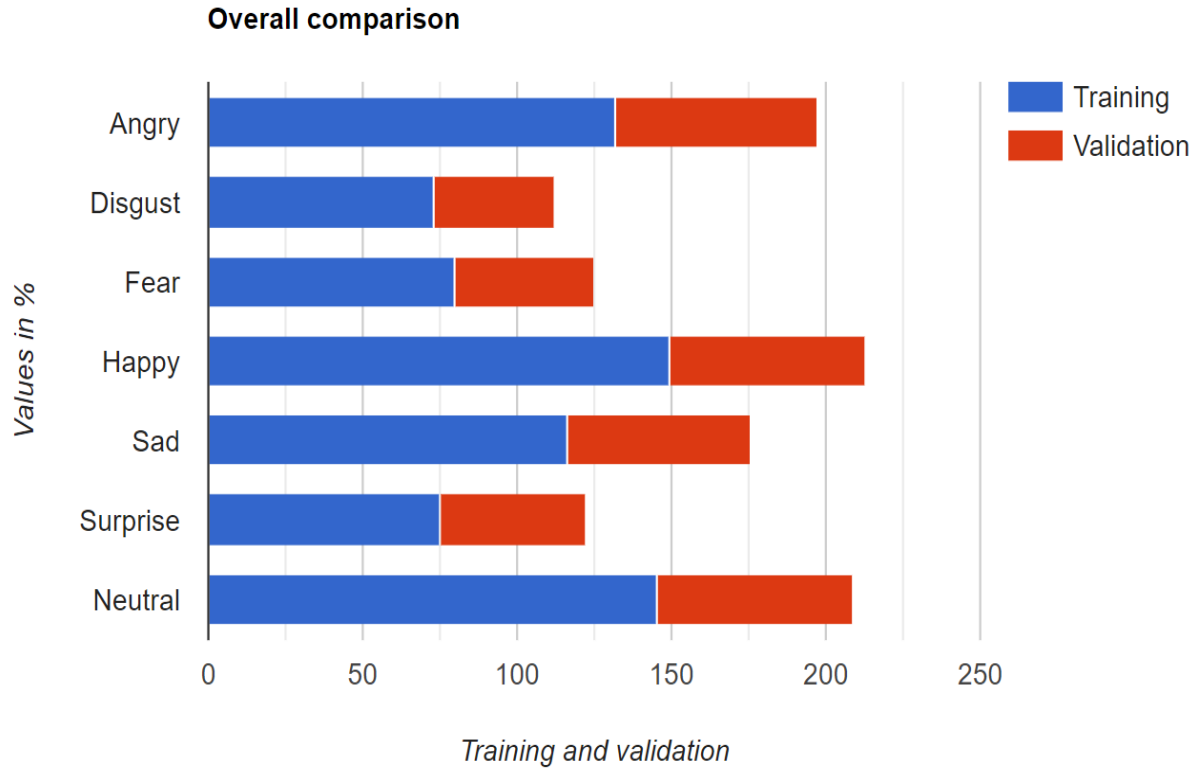

Figure 4: Training and validation

Doi : https://doi.org/10.54216/JCIM.130209
Received: January 05, 2024 Revised: Mrach 09, 2024 Accepted: May 04, 2024

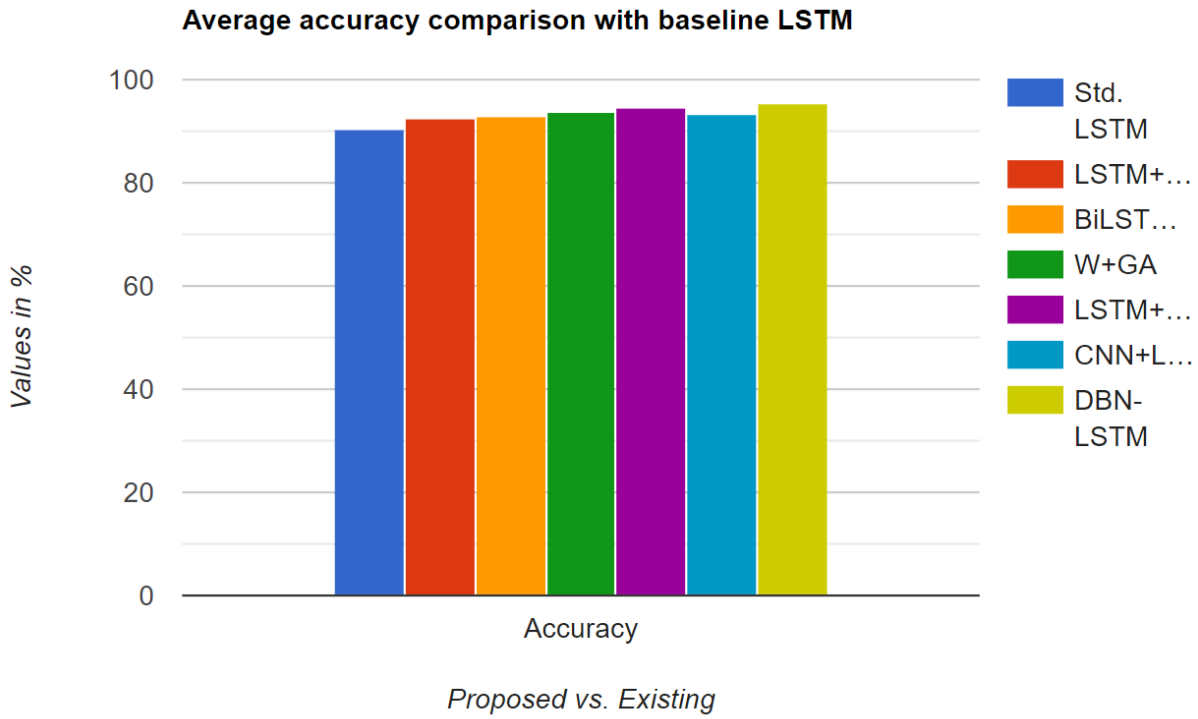

Figure 5: Average accuracy comparison with baseline LSTM

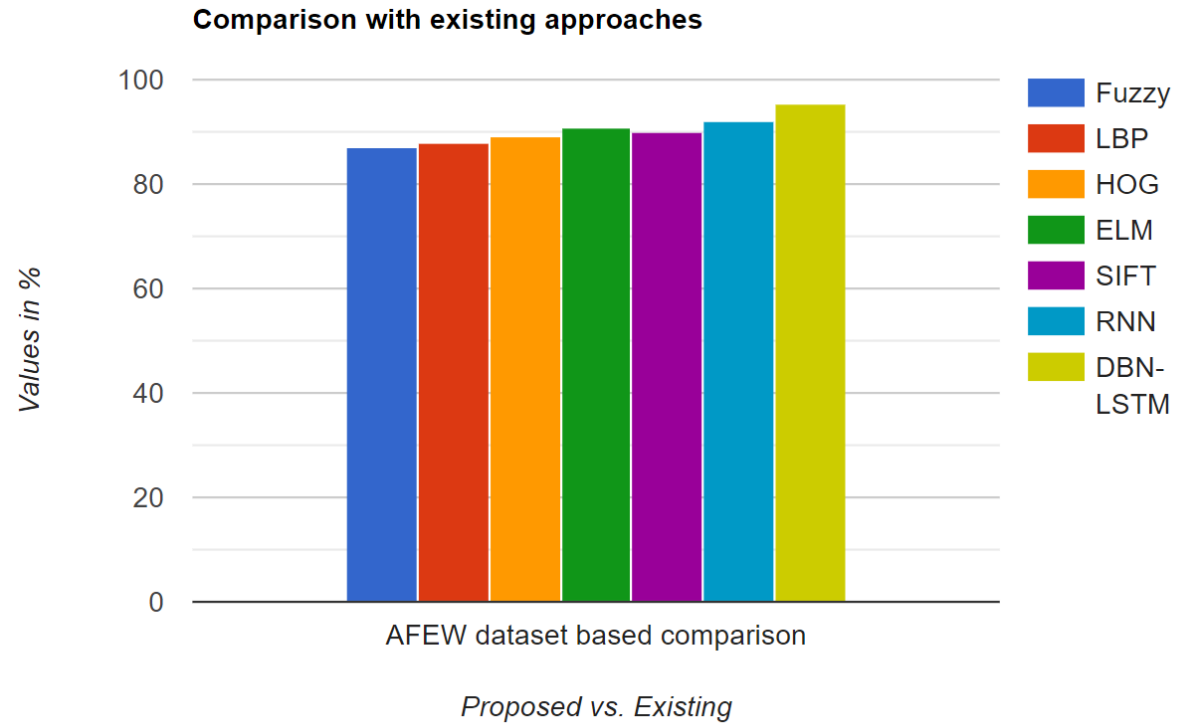

Figure 6: Comparison with existing approaches

Doi : https://doi.org/10.54216/JCIM.130209
Received: January 05, 2024 Revised: Mrach 09, 2024 Accepted: May 04, 2024

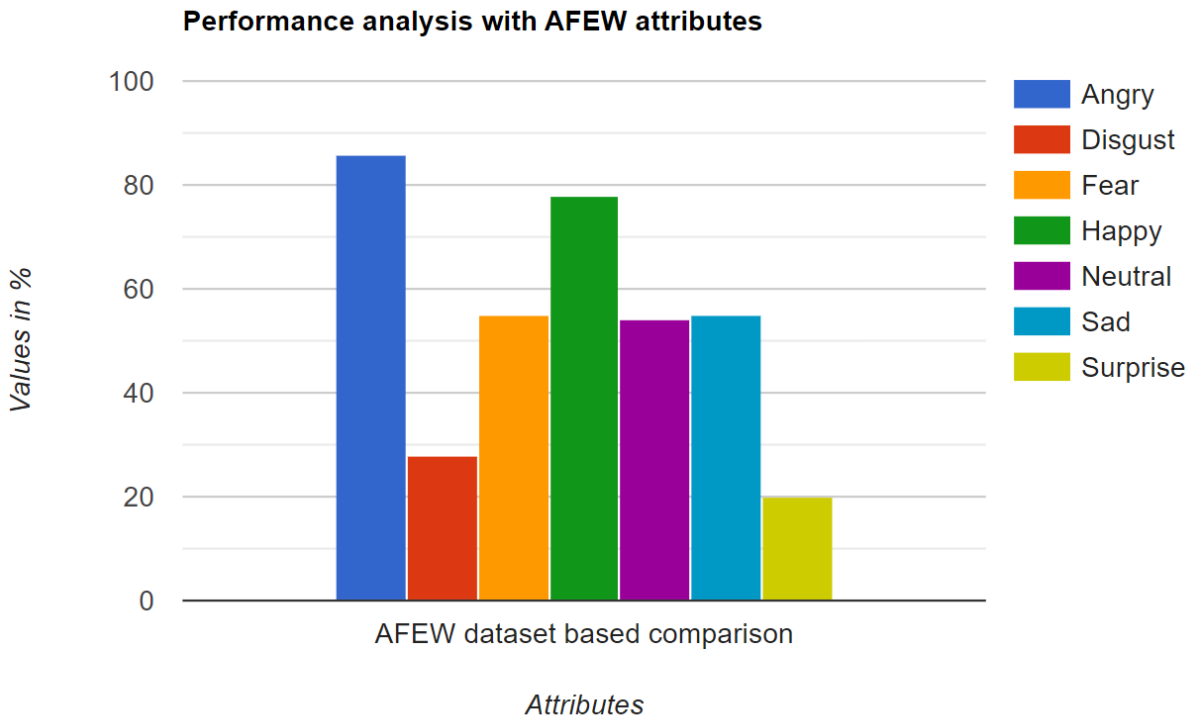

Figure 7: Performance analysis with AFEW attributes

To show the usefulness of the suggested strategy, the results are contrasted with those from other techniques for gauging happiness from Table 4. When applied to AFEW, the best-advised approach (fusion at the feature and decision levels) has remarkably obtained employing feature-level fusion, 95.97%. It is better than the starting point (reaching 87.9%) and all other comparable techniques. Concatenating faces, landmarks, characteristics, and spatial-temporal in movies made among the competition by dramatically and robustly enhancing performance. To obtain a close accuracy of 92.61%, the approach incorporated landmarks and textural elements with a CNN structure similar to the work, which was good enough for third place. With an accuracy of 91.76%, the HOG-TOP features (conventional methods) and the CNN combination (based on deep learning) performed well and finished fourth. After all deep learning techniques, scale-invariant features (SIFT) and local binary pattern (LBP) are employed together to produce a pretty high accuracy of 90.33%. LBP has the lowest accuracy of any approach after the baseline method at 88.2%. As a result, it is inapplicable to actual applications since handling many frames takes a lot of work. They came from movies and television programs where the actors and actresses exhibited sincere feelings and included all the real-world circumstances. Table 3 shows that our suggested technique, with accuracy values of 94.89% for AFEW, surpasses all other approaches, including traditional and deep learning.

Data are classed using feature representation; SIFT and LBP are traditional and supplementary suggested feature extraction techniques. Even though their system placed second with 91.84% accuracy on the AFEW test set, it fared noticeably worse on the AFEW dataset with an accuracy level of around 84.15%. This is because the system was not adapted to unknown data. The work showed high accuracy on selected datasets for their analysis, as in Fig 5 to Fig 7. The manual frame selection issue is the name for this issue. Using the parameters and appropriately unifying the data, we achieved a superior result using the AFEW test results, with 79% versus 76% for $DBN - LSTM$. The LSTM layer performs better on movies since, as was already established, the RNN cannot recall lengthy sequences. Because the films were recorded in unrestricted environments, 2D-CNN performed the lowest, with accuracy scores of 64.81% and 65.28%. This is because it cannot monitor all the problems, such as the wide range of head postures. The suggested system beats at least seven cutting-edge methods, notably under challenging conditions and with real-world datasets. In addition, the information in Table 3 indicates that the kind of features and the classifier used significantly impact the rate at which pleasant emotion is correctly identified. The same classifier can attain near-accuracy levels using several feature extraction techniques. For instance, switching the features from LBP and HOG might result in greater accuracy for the ELM classifier. Additionally, introducing an extra CNN feature might result in greater accuracy than just employing a single CNN network.

Doi : https://doi.org/10.54216/JCIM.130209
Received: January 05, 2024 Revised: Mrach 09, 2024 Accepted: May 04, 2024

In conclusion, the findings show that our suggested technique has significantly outperformed modern FER systems in terms of accuracy, with scores of 94.89% using the AFEW dataset, respectively. In general, it has been shown that accuracy is increased by combining both textural and landmark properties. The decision- and feature-level results were provided with feature-level fusion yielding the most remarkable considerable accuracies. We tested our suggested strategy using a variety of real-world datasets that had challenging situations, such as dramatic changes in lighting and substantial variations in head attitude. Additionally, there was racial variety among the performers shown in the videos of their unscripted facial expressions. The results of these datasets show that the suggested strategy is well-generalized [26] – [27]. Solving the head posture issue could attain great accuracy and was typically lighting-invariant. Since these are un-posed emotions, identifying the unprompted happy/smile expressions as opposed to the posed ones shown in the lab-controlled datasets is also a significant success. The suggested method achieves comparable outcomes on comparable real-world datasets under challenging circumstances when various angles are used to capture faces.

5. Conclusion

To detect positive emotions in unconstraint video, this study presents the unique $DBN - LSTM$ technique which hybridizes $DBN - LSTM$. This work adopts the improved version of the architecture to extract spatial-temporal information because of LSTM frameworks for facial expression recognition. To account for the temporal dynamic qualities in the following frames, an LSTM layer is added to the recovered features. DBN is employed to haul out deep features from time series of facial distances to distinguish facial expressions using the geometric qualities given by face landmarks. Our approaches distinguished between the happy and dissatisfied groups by combining these feature vectors at both decision and feature levels. The experiment on the unrestricted video datasets detects happy emotions more accurately than several competing methods, with an accuracy of 95.6% for the AFEW dataset. Due to the generally low accuracy produced by emotion recognition systems, developing a successful single emotion recognition strategy for expressing happiness can stimulate to continue working on the emotion recognition for describing other emotions. To uncover additional significant face components for emotion detection, it is vital to remember that the proposed technique's performance may still be enhanced by including certain attention blocks in the framework we previously outlined. Future research may consider evaluating and refining the suggested model on other massive, complex, unconstrained datasets.

References

[1] Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, “The faces of engagement: Automatic recognition of student engagement from facial expressions,” IEEE Transactions on Affective Computing, vol. 5, no. 1, pp. 86–98, 2014.

[2] Soleymani and M. Pantic, “Emotionally aware TV," in Proceedings of TVUX-2013: Workshop on Exploring and Enhancing the User Experience for TV at ACM CHI, 2013.

[3] Cockburn, M. Bartlett, J. Tanaka, J. Movellan, M. Pierce, and R. Schultz, “Smilemaze: A tutoring system in real-time facial expression perception and production in children with autism spectrum disorder,” in ECAG 2008 Workshop Facial and Bodily Expressions for Control and Adaptation of Games. Citeseer, 2008.

[4] Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems, p. 2012.

[5] Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," 2016.

[6] Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, “Hierarchical committee of deep convolutional neural networks for robust facial expression recognition,” Journal on Multimodal User Interfaces, pp. 1–17, 2016.

[7] Du, Y. Tao, and A. M. Martinez, “Compound facial expressions of emotion,” Proceedings of the National Academy of Sciences, vol. 111, no. 15, pp. E1454–E1462, 2014.

[8] Krishna, V. K. Deepak, K. Manikantan, and S. Ramachandran, “Face recognition using transform domain feature extraction and PSO based feature selection,” Appl. Soft Comput., vol. 22, pp. 141–161, Sep. 2014.

[9] Siddiqi, R. Ali, A. M. Khan, Y.-T. Park, and S. Lee, “Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields,” IEEE Trans. Image Process., vol. 24, no. 4, pp. 1386–1398, Apr. 2015.

[10] Senechal et al., “Facial action recognition combining heterogeneous features via multikernel learning,” IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 42, no. 4, pp. 993–1005, Aug. 2012.

[11] Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, “Learning multi-scale active facial patches for expression analysis,” IEEE Trans. Cybern., vol. 45, no. 8, pp. 1499–1510, Aug. 2015

Doi : https://doi.org/10.54216/JCIM.130209
Received: January 05, 2024 Revised: Mrach 09, 2024 Accepted: May 04, 2024

[12] Sanchez-Mendoza, D. Masip, and A. Lapedriza, "Emotion recognition from mid-level features," Pattern Recognition Letters, vol. 67, pp. 66–74, 2015.
[13]Sathya Preiya V, Kumar VDA. Deep Learning-Based Classification and Feature Extraction for Predicting Pathogenesis of Foot Ulcers in Patients with Diabetes. Diagnostics. 2023; 13(12):1983. https://doi.org/10.3390/diagnostics13121983.
[14]Balakrishnan C, Ambeth Kumar VD. IoT-Enabled Classification of Echocardiogram Images for Cardiovascular Disease Risk Prediction with Pre-Trained Recurrent Convolutional Neural Networks. Diagnostics (Basel). 2023 Feb 18;13(4):775. doi: 10.3390/diagnostics13040775. PMID: 36832263; PMCID: PMC9955174.
[15] Yu et al., "Is interactional dissynchrony a clue to deception? Insights from automated analysis of nonverbal visual cues," IEEE Trans. Cybern., vol. 45, no. 3, pp. 492–506, Mar. 2015.
[16] Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas, "Time-delay neural network for continuous emotional dimension prediction from facial expression sequences," IEEE Transactions on cybernetics, vol. 46, no. 4, pp. 916–929, 2016.
[17] Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013.
[18] Hemamalini, Selvamani, and Visvam Devadoss Ambeth Kumar. 2022. "Outlier Based Skimpy Regularization Fuzzy Clustering Algorithm for Diabetic Retinopathy Image Segmentation" Symmetry 14, no. 12: 2512. https://doi.org/10.3390/sym14122512.
[19] Kumar, V.D.A., Sharmila, S., Kumar, A. et al. A novel solution for finding postpartum haemorrhage using fuzzy neural techniques. Neural Comput & Applic 35, 23683–23696 (2023). https://doi.org/10.1007/s00521-020-05683-z
[20] V. D. A. Kumar, M. Raghuraman, A. Kumar, M. Rashid, S. Hakak and M. P. K. Reddy, "Green-Tech CAV: Next Generation Computing for Traffic Sign and Obstacle Detection in Connected and Autonomous Vehicles," in IEEE Transactions on Green Communications and Networking, vol. 6, no. 3, pp. 1307-1315, Sept. 2022, doi: 10.1109/TGCN.2022.3162698.
[21] Preeti Singh, Khyati Chaudhary, Gopal Chaudhary, Manju Khari, Bharat Rawal, A Machine Learning Approach to Detecting Deepfake Videos: An Investigation of Feature Extraction Techniques, Journal of Journal of Cybersecurity and Information Management, Vol. 9 , No. 2 , (2022) : 42-50 (Doi : https://doi.org/10.54216/JCIM.090204)
[22] Zhou, F. Fei, G. Zhang, J. D. Mai, Y. Liu, J. Y. Liou, and W. J. Li, "2d human gesture tracking and recognition by the fusion of MEMS inertial and vision sensors," IEEE Sensors J., vol. 14, no. 4, pp. 1160–1170, 2014.
[23] Poularakis and I. Katsavounidis, "Low-complexity hand gesture recognition system for continuous streams of digits and letters," IEEE T CYBERNETICS, vol. 46, no. 9, pp. 2094–2108, 2016
[24] Fan, C. Ma, Z. Gu, Q. Lv, J. Chen, D. Ye, J. Huangfu, Y. Sun, C. Li, and L. Ran, "Wireless hand gesture recognition based on continuous wave Doppler radar sensors," IEEE Micro Theory, vol. 64, no. 11, pp. 4012–4020, 2016.
[25] Zhang, Z. Tian, and M. Zhou, "Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor," IEEE Sensors J, vol. 18, no. 8, pp. 3278–3289, 2018.
[26] Galka, M. Master, M. Zaborski, and K. Barczewska, "Inertial motion sensing glove for sign language gesture acquisition and recognition," IEEE Sensors J., vol. 16, no. 16, pp. 6310–6316, 2016.
[27] Gupta, H. S. Chudgar, S. Mukherjee, T. Dutta, and K. Sharma, "A continuous hand gestures recognition technique for human-machine interaction using accelerometer and gyroscope sensors," IEEE Sensors J., vol. 16, no. 16, pp. 6425–6432, 2016.
[28] Wu, K. Chen, and C. Fu, "Natural gesture modelling and recognition approach based on joint movements and arm orientations," IEEE Sensors J., vol. 16, no. 21, pp. 7753–7761, 2016.
[29] Zhou, F. Fei, G. Zhang, J. D. Mai, Y. Liu, J. Y. Liou, and W. J. Li, "2d human gesture tracking and recognition by the fusion of MEMS inertial and vision sensors," IEEE Sensors J., vol. 14, no. 4, pp. 1160–1170, 2014.
[30] Aymen Hussein, S. Ahmed, Shorook K. Abed, Noor Thamer, Enhancing IoT-Based Intelligent Video Surveillance through Multi-Sensor Fusion and Deep Reinforcement Learning, Journal of Fusion: Practice and Applications, Vol. 11 , No. 2 , (2023) : 21-34 (Doi : https://doi.org/10.54216/FPA.110202)

Doi : https://doi.org/10.54216/JCIM.130209
Received: January 05, 2024 Revised: Mrach 09, 2024 Accepted: May 04, 2024

[31] P. Sherubha, P Amudhavalli, SP Sasirekha, "Clone attack detection using random forest and multi-objective cuckoo search classification", International Conference on Communication and Signal Processing (ICCSP), pp. 0450-0454, 2019.
[32] S. Dinesh, K. Maheswari, B. Arthi, P. Sherubha, A. Vijay, S. Sridhar, T. Rajendran, and Yosef Asrat Waji, "Investigations on Brain Tumor Classification Using Hybrid Machine Learning Algorithms", Hindawi Journal of Healthcare Engineering, Volume 2022.

Doi : https://doi.org/10.54216/JCIM.130209
Received: January 05, 2024 Revised: Mrach 09, 2024 Accepted: May 04, 2024