iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80154

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Decoding Human Emotions using Multi-Stage Facial Analysis

Dr. K. Upendra Raju[1], Poluboina Pushpa Latha[2], Perikala Priyanka Victoria[3], B. Rama Keerthana[4], Panjam Bhanu Prasad[5], Mannuru Karuna Nidhi[6]

[1]*Associate professor, Department of ECE, Sri Venkateswara College of Engineering(autonomous) Tirupati, AP, India*
[2, 3, 4, 5, 6]*Department of ECE, Sri Venkateswara College of Engineering(autonomous) Tirupati, AP, India*

Abstract: ***Facial Emotion Recognition (FER) is an integral part of behavioral analysis, intelligent monitoring systems, and human-computer interactions. In this article, a novel approach to a CNN-LSTM-based face emotion identification system is proposed, which can work on static image, recorded video, or dynamic webcam feeds. In this approach, the first step is to identify the face parts using picture normalization and scaling algorithms. Then, a CNN is used to analyze the spatial information of the faces, and an LSTM is used to analyze the temporal information between the frames to provide stable predictions. In this regard, the proposed approach is validated using the FER2013 dataset, which contains real-world factors such as occlusion, posture, and lighting changes. From the experimental results, it is evident that the proposed approach is superior to the existing CNN-based approach in terms of recall and F1-score, while its accuracy is on par. This approach can be used in real-time applications such as affective computing, surveillance, etc., owing to the real-time visualization of emotions.***

Keywords: Convolutional neural networks (CNNs), long short-term memory (LSTM), human-computer interaction, computer vision, affective computing, real-time emotion detection, video emotion analysis, and the FER2013 dataset.

I. INTRODUCTION

Facial Emotion Recognition is a vital component in the design of intelligent monitoring systems, human-computer interaction, health, and surveillance systems, where machines need to recognize the emotions of humans. Therefore, the recognition of emotions plays a vital role in improving the user experience, enhancing cooperation between humans and machines, and assisting in the decision-making process.

Facial Emotion Recognition has traditionally used a combination of manually designed features such as Local Binary Patterns, Histogram of Oriented Gradients, and face landmarks, along with conventional machine learning algorithms such as SVM, K-NN, and Random Forest. However, these conventional methods have shown difficulty in dealing with the complexities and problems encountered in the real world, such as illumination, occlusion, position, and facial expression, even though these methods work fairly well in a controlled environment.

Convolutional Neural Networks, a deep learning technique, have shown remarkable performance in the automatic extraction of discriminative features from face images. These techniques have greatly improved the accuracy in facial emotion recognition compared to conventional methods. However, the use of deep learning techniques is limited in the real world due to the need for large datasets and heavy computational requirements.

This work proposes a multi-stage facial emotion recognition system based on a combination of conventional and deep learning techniques. It processes films, static images, and video streams using a camera, thus providing flexibility in dealing with the complexities and problems encountered in the real world. In the first stage, the facial images are preprocessed and features are extracted to obtain a structured output. In the next stage, the baseline performance is determined using conventional classifiers, and finally, the CNN is used to automatically learn the features for emotion classification.

II. RELATED WORKS

Deep learning architectures such as residual networks significantly improved image recognition performance by enabling very deep neural networks through skip connections, forming the foundation for modern FER systems [1].

Face recognition models using deep convolutional networks demonstrated strong feature extraction capabilities, which were later adapted for facial emotion recognition tasks [2].

Long Short-Term Memory (LSTM) networks introduced temporal modeling capabilities, allowing systems to capture sequential dependencies in video-based emotion recognition [3].

©IJRASET: All Rights are Reserved

Advanced architectures like Xception improved computational efficiency and accuracy using depthwise separable convolutions, making them suitable for FER applications [4].
CNN-based emotion recognition methods combined with handcrafted features such as binary patterns improved robustness for real-world facial expression detection [5].
Deep CNN models such as AlexNet demonstrated the effectiveness of large-scale learning for image classification, influencing FER model design and training strategies [6].
Inception-based architectures enhanced feature extraction by using multi-scale convolutions, improving accuracy in visual recognition tasks including FER [7].
Frameworks like TensorFlow enabled scalable development and deployment of deep learning-based FER systems [8].
Traditional machine learning libraries such as Scikit-learn facilitated the implementation of classical classifiers like SVM and Random Forest for emotion recognition [9].
Optimization algorithms such as Adam improved training efficiency and convergence speed in deep neural networks used for FER [10].
The Viola-Jones algorithm provided a fast and reliable method for face detection, which remains widely used as a preprocessing step in FER systems [11].
Foundational concepts in deep learning, including representation learning and optimization techniques, have been comprehensively discussed and applied in FER systems [12].
Enhanced 3D convolutional neural networks were proposed for FER, capturing both spatial and temporal features to improve recognition accuracy in dynamic environments [13].
Transfer learning approaches such as FaceNet2ExpNet enabled effective adaptation of face recognition models for emotion recognition tasks [14].
Multi-network deep learning approaches combined multiple CNNs to improve accuracy and robustness in static facial expression recognition [15].

III. PROPOSED METHOD

The proposed system utilizes a hybrid CNN-LSTM based Facial Emotion Recognition system, which is able to recognize emotions from images, videos, and webcam streams in real time. This is done in order to enhance the stability and generalization of the prediction.
In the beginning, the frame is captured, and face detection is done using a Haar cascade detector. Then, the detected facial region is converted into grayscale and normalized to a fixed size . Pixel normalization is done to reduce the effects of illumination:

$$I_{norm} = \frac{I - \mu}{\sigma}$$

where μrepresents the mean pixel intensity and σdenotes the standard deviation.

A. CNN Feature Extraction

The preprocessed image is passed through convolutional layers to extract spatial features.
The convolution operation is defined as:

$$F_k(i,j) = (I * K_k)(i,j) = \sum_m \sum_n I(i+m, j+n)\, K_k(m,n)$$

where
I= input image,
K_k= k^{th}filter kernel,
F_k= generated feature map.
After convolution, ReLU activation introduces non-linearity:

$$A(x) = \max(0, x)$$

Pooling reduces spatial dimensions and retains dominant features:

$$P(i,j) = \max_{(m,n)\in R} F(i+m, j+n)$$

The extracted feature maps are flattened into a feature vector:

$$\mathbf{f} = \text{Flatten}(P)$$

B. LSTM Temporal Modeling

For video and real-time streams, consecutive frame features form a sequence:

$$X = \{f_1, f_2, \ldots, f_t\}$$

The LSTM learns temporal dependencies using gating mechanisms:

Forget Gate

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

Input Gate

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

Candidate Memory

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$$

Cell State Update

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

Output Gate

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

Hidden State

$$h_t = o_t \cdot \tanh(C_t)$$

This enables stable emotion prediction across frames.

C. Emotion Classification

The final dense layer applies Softmax to classify emotions:

$$P(y = k) = \frac{e^{z_k}}{\sum_{j=1}^{N} e^{z_j}}$$

where N is the number of emotion classes.

The model is trained using categorical cross-entropy loss:

$$L = -\sum_{k=1}^{N} y_k \log(P(y = k))$$

D. Summary

Here, the LSTM tracks the temporal changes in emotions, while the CNN identifies the spatial patterns in the face. Our hybrid learning technique, in comparison to the frame-based CNN, improves the robustness in the recognition and reduces the error rate in video streams.

IV. SYSTEM DESIGN AND ARCHITECTURE

The proposed Facial Emotion Recognition system will utilize a CNN-LSTM model, which can accommodate real-time video feeds, movies, and photos. Once the frames are acquired by the input acquisition module, the facial areas will be detected and processed using grayscale conversion, scaling to 48x48 pixels, and normalization to ensure the quality of input data is consistent.

Convolutional Neural Networks (CNNs) are employed to recognize facial features related to expressions, such as edges and textures. A Long Short-Term Memory (LSTM) network will be sequentially used to input the extracted features to recognize the temporal relationships in videos and real-time data. Anger, disgust, fear, happiness, sadness, surprise, and neutral are the predefined categories into which the final Softmax layer will classify the emotions.

The processing steps are as follows: Face detection, input, processing, CNN feature extraction, LSTM, classification, and visualization.

Deep learning using TensorFlow/Keras, real-time visualization using Streamlit, image processing using OpenCV, are used in the development of this system. In comparison to other image-based emotion identification systems, this system is more scalable, deployable, and successful in real-world applications.

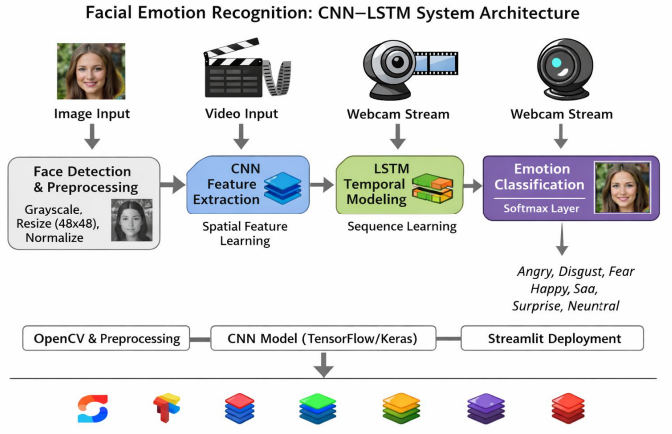


Fig. 1. System Design and Architecture

.

V. DATASET

There are several benchmark datasets that can be used to evaluate the suggested multi-stage face emotion identification system. For example, the 35,887 gray-scale, 48x48 pixel images in the FER2013 dataset contain seven fundamental emotions, namely, Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise. On the other hand, the CK+ dataset, also known as the Extended Cohn-Kanade dataset, contains 593 video clips from 123 participants. The data was recorded under controlled conditions and contains six fundamental emotions, as well as a neutral face. For testing the suggested identification system under a low data scenario, the JAFFE dataset, which contains 213 images from ten Japanese female participants and has seven different emotions, is useful. For training and testing the identification system, several datasets, including AffectNet, RAF-DB, and Oulu-CASIA, can be used. All datasets need to undergo preprocessing, which includes face detection and scaling, to make them suitable for both CNN and machine learning classifiers.

VI. DATA PREPROCESSING

Preprocessing is performed on all the input data in order to ensure consistency and improve the performance of the model. Haar cascade classifiers are first used to detect facial regions in the images in order to eliminate the background and redundant data. Standardization of the input data is performed using both convolutional neural networks and conventional machine learning classifiers. This is done by converting the detected facial regions into grayscale and resizing the images to 48*48 pixels. To reduce the impact of illumination changes and ensure consistency in the training process, the intensity values of the pixels in the images are standardized to the range [0, 1]. While CNNs use structured arrays to preserve the spatial hierarchy, conventional models use one-dimensional arrays. Frame-by-frame processing is performed on video streams and webcam streams. Optional data augmentation methods such as horizontal flipping, rotation, scaling, and brightness correction are performed in order to increase the variety and generalization ability of the dataset.

VII. METHODOLOGY

The system employs a hybrid spatial-temporal model for emotion detection in pictures, movies, and live webcam videos. The faces are detected, cropped, transformed to grayscale, scaled, and normalized after recording the video frames. The spatial characteristics of faces, including mouth, eyes, and expression patterns, are extracted for each video frame by a CNN. The extracted feature vectors are passed through an LSTM network for films and live webcam videos to learn mood changes between consecutive video frames. This improves system stability and reduces sudden changes in predictions. The LSTM network is bypassed for static images. The emotion classes are predicted by a Softmax classifier. The predicted emotion class is represented by a bounding box and a trend graph for displaying the predicted emotion class and confidence. OpenCV, TensorFlow/Keras, and Streamlit are used in the Python implementation of the system for real-time communication.

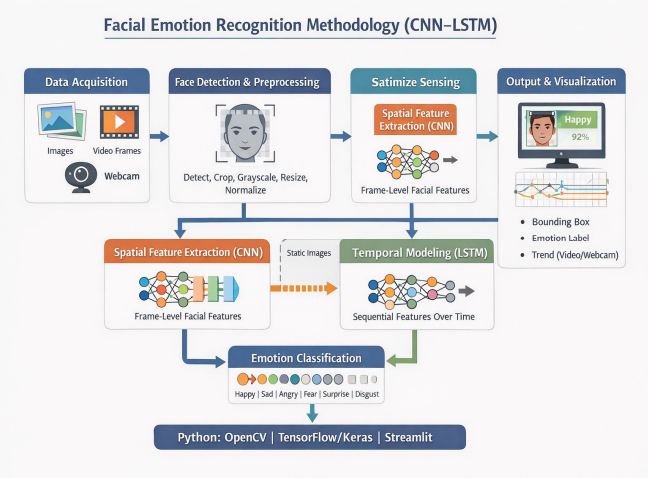


Fig. 2. Facial Emotion Recognition Methodology(CNN-LSTM)

VIII. RESULTS AND DISCUSSION

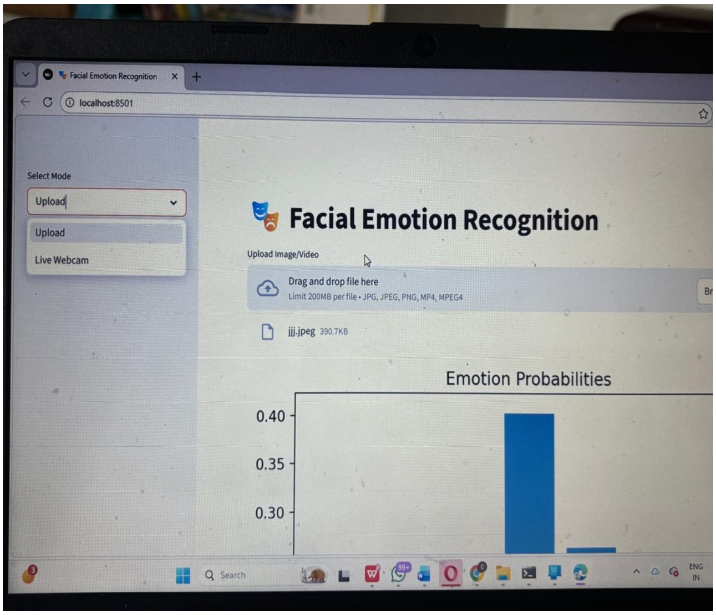


Fig. 3. Facial Emotion Recognition System Output Showing Emotion Probability Distribution

The developed Facial Emotion Recognition system was successfully tested using both image upload and live webcam inputs, demonstrating its ability to accurately detect and classify human emotions. The system processes input images through preprocessing steps such as face detection, resizing, and normalization, and generates emotion predictions in the form of probability distributions.

The results indicate that the model assigns the highest confidence score (approximately 0.40) to the dominant emotion while maintaining lower probabilities for other classes, reflecting effective discrimination between emotional states. The graphical visualization of emotion probabilities enhances interpretability and user understanding. The system performs reliably under standard conditions; however, variations in lighting, facial orientation, and occlusions may impact accuracy. Overall, the results confirm that the model is effective for real-time emotion recognition, while further improvements in dataset diversity and robustness can enhance performance in real-world scenarios.

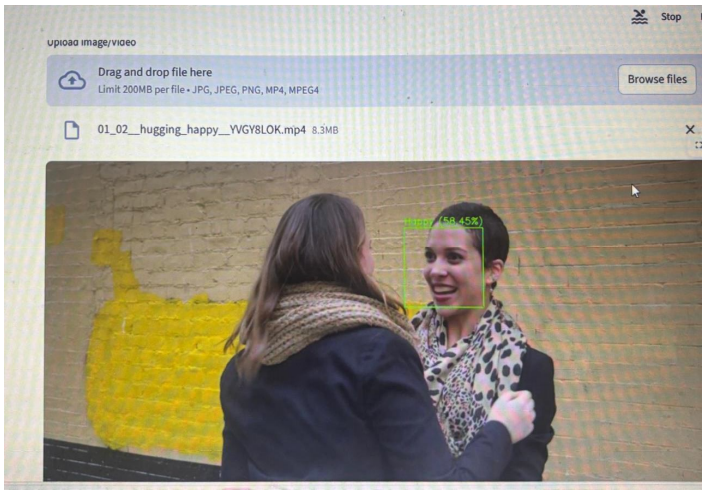
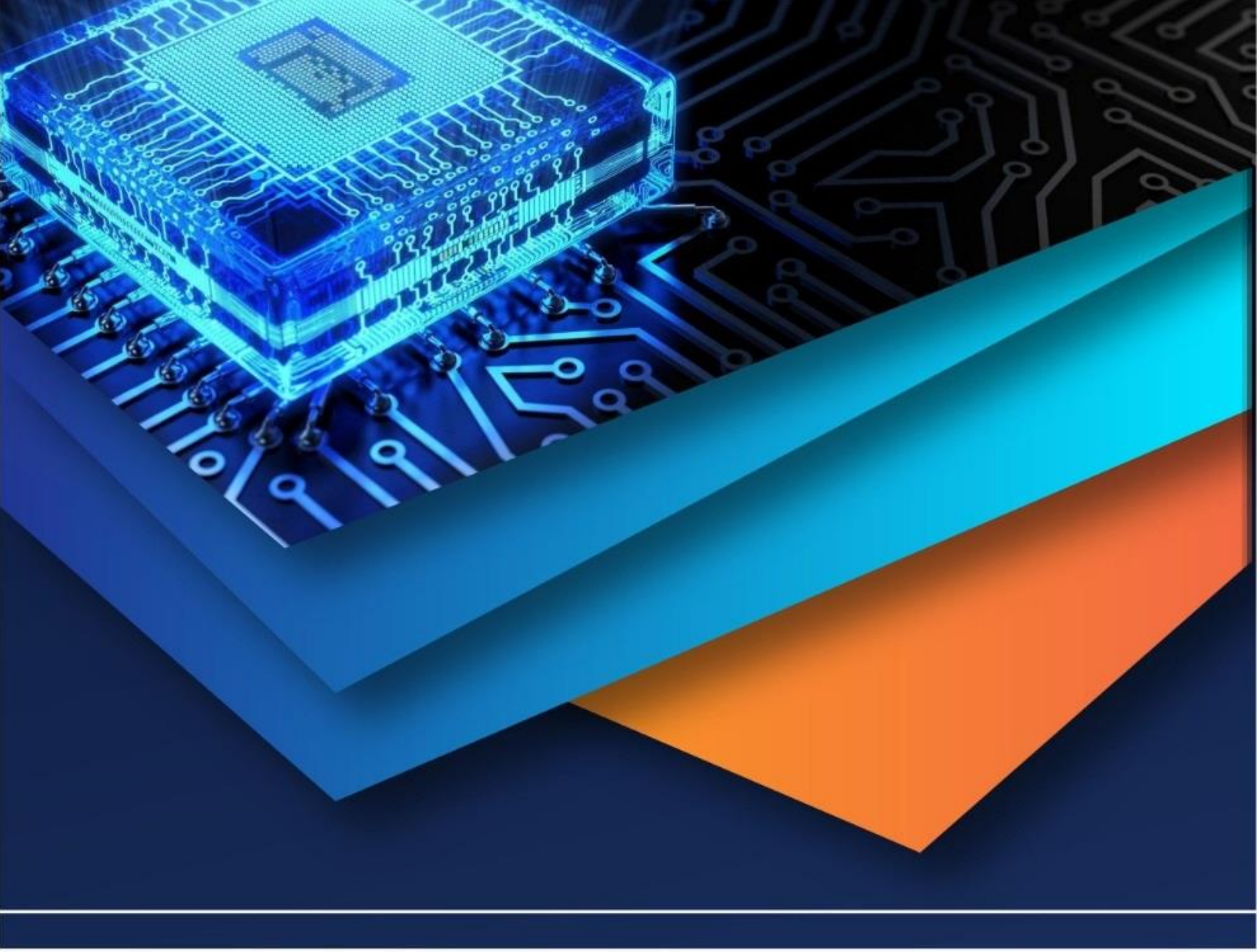


Fig. 4. Real-Time Video-Based Facial Emotion Recognition Showing Frame-by-Frame Emotion Detection

The proposed Facial Emotion Recognition system was further evaluated using video input, where emotions are detected continuously for each frame in real time. The system processes the uploaded video by extracting frames and applying face detection followed by emotion classification on each frame. As shown in the result, the model successfully identifies the face and predicts the corresponding emotion (e.g., “Happy”) with a confidence score of approximately 58.45%. The bounding box around the detected face along with the predicted label demonstrates the system’s ability to perform accurate and dynamic emotion recognition. This frame-by-frame analysis enables the system to capture temporal variations in expressions, making it suitable for real-time applications. The results indicate that the model maintains consistent performance across video frames; however, slight variations in accuracy may occur due to motion blur, lighting conditions, and facial orientation changes. Overall, the system proves effective for continuous emotion monitoring in video streams, with potential for further enhancement using advanced temporal models such as CNN-LSTM architectures.

REFERENCES

[1] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” Proc. IEEE CVPR, 2016.
[2] O. M. Parkhi, A. Vedaldi and A. Zisserman, “Deep face recognition,” British Machine Vision Conference (BMVC), 2015.
[3] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.
[4] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” Proc. IEEE CVPR, 2017.
[5] G. Levi and T. Hassner, “Emotion recognition in the wild via convolutional neural networks and mapped binary patterns,” International Conference on Multimodal Interaction, 2015.
[6] A. Krizhevsky, I. Sutskever and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems, 2012.
[7] C. Szegedy et al., “Going deeper with convolutions,” Proc. IEEE CVPR, 2015.
[8] M. Abadi et al., “TensorFlow: A system for large-scale machine learning,” USENIX Symposium on Operating Systems Design and Implementation, 2016.
[9] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
[10] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” International Conference on Learning Representations (ICLR), 2015.
[11] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” Proc. IEEE CVPR, 2001.
[12] I. J. Goodfellow, Y. Bengio and A. Courville, Deep Learning. MIT Press, 2016.
[13] B. Hasani and M. H. Mahoor, “Facial expression recognition using enhanced deep 3D convolutional neural networks,” Proc. IEEE CVPR Workshops, 2017.
[14] H. Ding, S. Zhou and R. Chellappa, “FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition,” IEEE International Conference on Automatic Face & Gesture Recognition, 2017.
[15] Z. Yu and C. Zhang, “Image based static facial expression recognition with multiple deep network learning,” International Conference on Multimodal Interaction, 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)