



e-ISSN:2582-7219



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 8, August 2024



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.521



6381 907 438



6381 907 438



ijmrset@gmail.com



www.ijmrset.com



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Way Approach for Human Emotion Detection

Dr. Shashidhar. V¹, Adi Reddy Y², Smita Rani³, Sonashree C⁴

Assistant Professor, Department of Computer Science & Engineering, Rajarajeswari College of Engineering,
Bangalore, Karnataka, India

Department of Computer Science & Engineering, Rajarajeswari College of Engineering, Bangalore, Karnataka, India

ABSTRACT: Emotion detection has become a crucial area Research with a spectrum of applications, spanning various fields. human-computer interaction to mental health assessment. Traditional methods have primarily focused on analyzing individual modes such as text or audio, limiting the depth of emotional understanding. This paper introduces a novel multi-modal framework for emotion detection that integrates text, audio, and video analyses to capture the intricate nuances of human emotional expression. By leveraging state-of-the-art Architectures for deep learning, such as recurrent neural networks (RNNs)convolutional neural networks (CNNs), and attention mechanisms, our approach effectively combines information from diverse modalities. We present empirical results on standard datasets, demonstrating the superior participate of our method compared to single-modal approaches. Additionally, we explore the interpretability of the acquired representations, Offering understanding into the fundamental mechanisms below the surface. governing emotional expression across different modalities. These searching is not only advance the field of multi-modal emotion detection but also lay the foundation for more sophisticated and robust emotion-aware systems applicable across the various domains, including affective computing, human-computer interaction, and psychological balance assessment.

I. INTRODUCTION

Emotion detection has risen as a crucial area of exploration within affective computing, with the objective of understanding and interpreting human emotional[1][3] states across various mediums like text, audio, and video. The surge in digital communication platforms and multimedia content has underscored importance of comprehending emotions[5] conveyed through these channels, impacting applications ranging from human-computer interaction to mental health monitoring.

Advancement in the The realms of machine learning and deep learning. technique have facilitated substantial progress in discerning emotions from textual, auditory, and visual signals. Foundational studies [1] [2] laid the groundwork by introducing fundamental methodologies for sentiment analysis in text and speech recognition, setting the stage for subsequent developments in emotion detection. These seminal works provided a framework for evolution of multimodal approaches aimed at amalgamating information from diverse sources to enhance accuracy and robustness Given the intricate nature of human emotions, advanced computational models are essential in capturing nuanced expressions across different modalities. Text-based emotion detection traditionally relies on the natural language processing (NLP) methods,[3] [4], which introduced lexicon-based and machine learning-based approaches for the sentiment analysis. In contrast, in audio-based emotion detection have employed techniques such as acoustic Extracting features and performing classification. algorithms[5][6] to decipher emotional states from speech signals.

Simultaneously, the field of video-based emotion detection has witnessed significant advancements. Seminal contributions [7] introduced methodologies For the analysis of facial expressions and affect Recognition within video content. streams. These foundational studies laid the groundwork for more sophisticated approaches leveraging deep learning architectures For example, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [9] to extract spatiotemporal features and infer emotions from dynamic visual content.

Despite the progress achieved, there are ongoing challenges in multimodal emotion detection, such as dealing with diverse data sources, Incorporating data from various modalities, and accounting for cultural differences in emotional expression. Meeting these challenges necessitates cooperation. among experts in computer science, psychology,



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

linguistics, and neuroscience to create comprehensive models capable of understanding emotions Spanning a broad spectrum of contexts.

This paper presents a comprehensive review of state-of-the-art techniques in text, audio, and video-based emotion detection. It delves into key methodologies, challenges, and future directions in each modality while exploring emerging trends such as deep learning-based multimodal fusion. From synthesizing insights by existing literature and identifying research gaps, the goal is to provide roadmap for advancing in field of multimodal emotion detection[6] and its applications.

II. RELATED WORK

Emotion detection has emerged as a vital area of research across text, audio, and video modalities due to its wide-ranging applications in fields like human-computer interaction, affective computing, and mental health monitoring. This review aims to consolidate key contributions and trends in these modalities, offering insights in the evolution of techniques, challenges faced and future directions in multimodal emotion analysis. Text-Based Emotion Detection:

Initially, text-based emotion detection methods predominantly utilized lexicon-based sentiment analysis and machine learning classifiers. For instance, [1] introduced techniques that employed pre-established dictionaries of emotion-rich terms to deduce sentiment from textual information. Similarly, Support Vector Machines (SVM) and Recurrent Neural Networks (RNN) were concurrently employed to autonomously discern patterns from annotated text [4].

Recently, there has been a notable transformation in text-based emotion detection, driven by the emergence of the deep learning architectures. Architectures like Convolutional Neural Networks (CNNs) [9] and Transformer models [10] have demonstrated remarkable efficacy in capturing nuanced semantic connections within text, surpassing the effectiveness of conventional methods in terms of accuracy and scalability. Additionally, researchers have delved into domain adaptation techniques [5] to enhance model generalization across various textual domains, mitigating a significant challenge in this area of study.

Audio-Based Emotion Detection:

Audio-based emotion detection involves extracting the acoustic feature from the voice signals to discern underlying emotional states. Early studies, such as [6] [7], introduced techniques like Mel-Frequency Cepstral Coefficients (MFCCs) and prosodic Characteristics for classifying emotions, emphasizing The significance of acoustic signals such as pitch and intensity in distinguishing between emotions conveyed through speech.

In recent years, audio-based emotion detection have seen the rise pertaining to deep learning. methodologies. Long Short-Term Memory (LSTM) networks [8] alongside Convolutional Neural Networks (CNNs). have been employed to automatically extract selective features from raw audio signals, eliminating the need for handcrafted features. Additionally, researchers have explored the fusion of acoustic and linguistic features [10], recognizing the complementary nature of different modalities in enhancing emotion recognition accuracy.

Video-Based Emotion Detection:

The procedure of video-based emotion detection involves analyzing facial expressions and body language to deduce emotional states. Initial investigations, exemplified by works such as [11] [14], introduced methodologies such as Active Appearance Models (AAMs) and Using the Facial Action Coding System (FACS) for extracting facial features and recognize expressions. These pioneering efforts set the stage for further progress in computer vision-based emotion analysis.

In recent studies on video-based emotion detection, there has been a notable transition towards employing deep learning methodologies adept at extracting spatiotemporal features from the video sequences. Convolutional Neural Networks (CNNs) [9] and Recurrent Neural Networks (RNNs) [11] have showcased exceptional proficiency in recognizing facial expressions, outperforming conventional techniques. Moreover, researchers have investigated multimodal fusion strategies [7] to combine data from facial expressions, voice, and contextual cues, resulting in enhanced accuracy in emotion recognition.



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. METHODOLOGY

Proposed Architecture

Our web application comprises login and signup pages where users must enter correct credentials to access their accounts. When users log in, they are greeted with a user interface offering three options for emotion detection: video-based, audio-based, and text-based systems. Additionally, we have integrated a chatbot to serve as a voice assistant, enhancing the person’s experience with various project features. The application offers personalized recommendations based on detected emotions, including music, videos, articles, movies, healthcare resources, and more. These recommendations are available In various languages, such as English, Hindi, Kannada, Telugu, and Tamil.

Facial Expression Detection:

I. Proposed Methodology

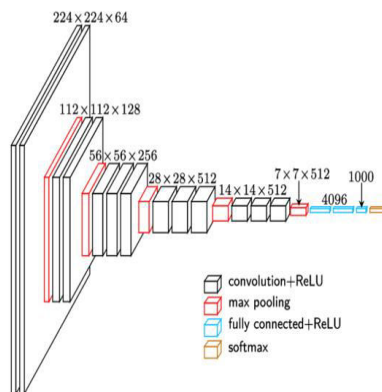
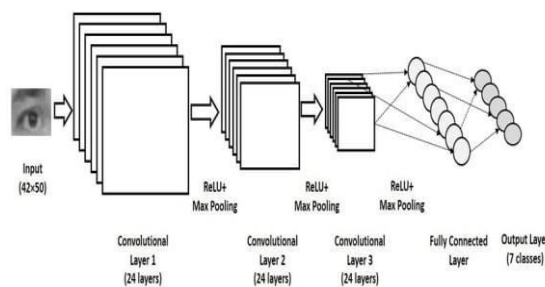


Fig :VGG16 CNN Model Diagram

There are various steps and process in VGG16 CNN model which finally leads to the required specific format of the images.

Convolutional Layer

The Convolutional Layer serves as the initial stage in CNN architecture. It involves taking a 3x3 part of the input matrix, which is acquired from the High-pass filter. This section is then multiplied with the corresponding filter matrix, and the resulting sums are situated within their respective positions within the output matrix, as described in the accompanying diagram. Subsequently, this output undergoes further reduction in the pooling layer. The Convolutional Layer process is depicted in the figure below.





International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Classification of CNN in Image Classification:

1. Input Layer:

The input layer of the CNN takes in the raw image data as input. The images depicted typically represented as matrices of pixel values. The dimensions of input layer correspond to the size of input images i.e; height, width and color channels.

2. Convolutional Layers:

Convolutional layers is responsible for feature extraction. They consist of filters that are convolved from the input images to capture relevant patterns and features. These layers learn to detect edges, textures, shapes and other important visual elements.

3. Pooling Layers:

Pooling layers reduce the structural dimensions of feature maps produced by convolutional layers. They perform down sampling operations i.e., max pooling to retain the most salient information while discarding unnecessary details. This helps in achieving translation invariance and reducing computational complexity.

4. Fully Connected Layers:

The result of last pooling layer is flattened and connected to at least one fully connected layers. These layers function as traditional neural network layers and classify the extracted features. Fully connected layers learn complex relationships between the features and output class probabilities or predictions.

5. Output Layer:

The output layer represents the last layer of the CNN. It consists of neurons equal to number of distinct classes in a classification task. The output layer provides each class's classification probabilities or predictions, indicating the likelihood of input image belonging to a specific class.

VGG16 Algorithm

Step 1:Input: Receive input image data of fixed size.

Step 2: Convolutional Layer: Apply a 3x3 convolutional filter with stride 1 and padding, followed by ReLU activation. Repeat this step for 2 times.

Step 3:Max Pooling Layer: Performs 2x2 max pooling with the stride 2 for downsampling.

Step 4:Flatten: Flatten the output feature maps from the last convolutional block into a vector.

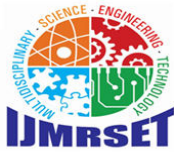
Step 5: Fully Connected Layer: Connect all neurons from the flattened vector to fully connected layer utilizing ReLU activation. Repeat this step again.

Step 6: Output Layer: Connect the final fully connected layer to the the output layer employing softmax activation. for classification.

Mathematical Model

The ImageNet dataset comprises images with a standardized size of 224x224 pixels and includes RGB channels. Consequently, our input tensor is structured as (224, 224, 3). This model analyzes the input image and generates a vector consisting of 1000 values as its output.

The classification vector depicts the probabilities assigned to each class by the model. For instance, if the model predicts that an image belongs to class 0 with a probability of 1, class 1 with a probability of 0.05, class 2 with a probability of 0.05, class 3 with a probability of 0.03, class 780 with a probability of 0.72, class 999 with a probability of 0.05, and assigns a probability of 0 to all other classes, the resulting classification vector reflects these probabilities.:



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

$$\hat{y} = \begin{bmatrix} \hat{y}_0 = 0.1 \\ 0.05 \\ 0.05 \\ 0.03 \\ \cdot \\ \cdot \\ \cdot \\ y_{780} = 0.72 \\ \cdot \\ \cdot \\ y_{999} = 0.05 \end{bmatrix}$$

To ensure that these probabilities sum up to 1, the softmax function is applied. This function is defined as follows:

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}}$$

Following utilizing the softmax function, we select the five most probable candidates and include them in the vector.

$$C = \begin{bmatrix} 780 \\ 0 \\ 1 \\ 2 \\ 999 \end{bmatrix}$$

The true data vector is given as follows.:

$$G = \begin{bmatrix} G_0 \\ G_1 \\ G_2 \end{bmatrix} = \begin{bmatrix} 780 \\ 2 \\ 999 \end{bmatrix}$$

Then the Error function is the following:

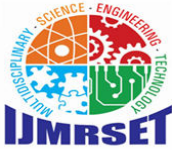
$$E = \frac{1}{n} \sum_k \min_i d(c_i, G_k)$$

It calculates the minimum gap among the each ground truth class and the predicted candidates, where the distance function d is defined as:

d=0 if $c_i = G_k$

d=1 otherwise

Since, all these categories present in the true data are in Predicted top 5 matrix, therefore loss becomes 0.



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Text Based Emotion Detection:

II. Proposed Methodology

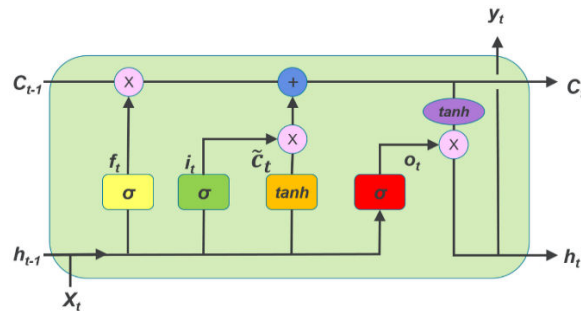


Fig :LSTM Model Diagram

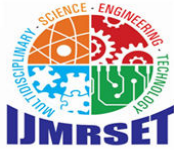
Long Short-Term Memory (LSTM) is a neural network with recurrence (RNN) architecture specifically tailored for sequence prediction tasks like forecasting time series, processing natural language. (NLP), and speech recognition. Unlike traditional RNNs, LSTMs are engineered to mitigate the vanishing gradient issue, enabling them to effectively capture long-term dependencies in sequential data.

Sequential Input: LSTM networks process arrangement of data. Each element in the sequence (e.g., a word in a sentence or a data point in a time series) is inputted into the network one at a time.

Memory Cells: Memory cells are the fundamental components belonging to an LSTM network., serving the crucial function of preserving data over time. Designed with a specialized structure, these cells possess the capability to retain data for extended durations, thereby facilitating the network in grasping long-term dependencies.

Gates: LSTM networks utilize unique types of gates for regulating information flow. within the network:

- Forget Gate: This gate decides which information to omit from the cell state, based on the previous hidden state and the current input as input and generates A numerical value ranging from 0 to 1 assigned to each element within the cell state. A value of 0 signifies complete forgetting, while a value of 1 indicates complete retention.
- Input Gate: This gate identifies the new information for retention within the cell state. It comprises two components: a sigmoid layer that explains which values to update and a tanh layer that generates a a vector containing new candidate values to incorporate into the state.
- Cell State Update: This stage involves modifying the cell state by integrating the information from the forget gate (indicating what to discard) and the input gate (indicating what to add).
- The Output Gate: This component decides which information to output by assessing the updated cell state. It functions as a filter for the cell state, producing the LSTM cell's output for the present timestep..
- Long-Term and Short-Term Memory: Within the LSTM architecture, there exists the capability to selectively retain or discard information across time. This functionality equips the network to comprehend prolonged relationships within the information while also adapting to immediate changes.
- LSTM Algorithm
- Step 1: Input: At each time interval. t , the LSTM receives an input vector x_t .
- Step 2. Hidden State: The LSTM maintains a hidden state vector h_t that represents the network's memory of the sequence up to time step t .
- Step 3. Cell State: Furthermore, alongside the hidden state, LSTMs have cell state vector C_t that acts as an internal memory. This cell state can store the information over a long periods of time.
- Step 4. Gates: LSTMs use three gates to prevent the flow of information:
 - Forget Gate:
 - $f_t = \sigma(W_f [h_{t-1}, x_t] + b_f)$
 - Input Gate:
 - $i_t = \sigma(W_i [h_{t-1}, x_t] + b_i)$



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- $C_t = \tanh(W_c [h_{t-1}, x_t] + b_c)$
- Update Cell State:
- $C_t = f_t \cdot C_{t-1} + i_t \cdot C_t$
- Output Gate:
- $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$

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

Output Gate: Decides which information from the cell state to utilize as the hidden state (h_t) .

Audio Based Emotion Detection

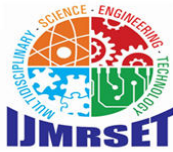
- **Librosa:** Librosa stands as a Python package tailored for the analysis and manipulation of audio and music. It furnishes a comprehensive suite of tools for tasks like audio loading, feature extraction, manipulation, and playback. Here's a breakdown of the functionalities offered by Librosa
- **Audio Loading:** With Librosa, you can effortlessly import audio files in various formats (e.g., WAV, MP3) into Python, represented as NumPy arrays, facilitating seamless integration of audio data within your Python environment.
- **Feature Extraction:** A core capability of Librosa lies in its capacity to extract an extensive array of audio features crucial for audio analysis and machine learning endeavors. These features encompass:
- **Mel-frequency cepstral coefficients (MFCCs):** These serve as prevalent features for depicting the spectral envelope of audio signals.
- **Spectrograms:** Librosa facilitates the computation of spectrograms, visual depictions showcasing The temporal distribution of frequency content in an audio signal. **Chroma features:** These illustrate the energy distribution of pitch classes (i.e., musical notes) within an audio signal.
- **Tempogram and Beat Tracking:** Librosa acquire users with tools for estimating tempo and identifying beat locations within audio recordings.
- **Signal Processing:** Librosa offers functions for essential signal processing tasks like resampling, filtering, and windowing, frequently required for preprocessing audio data prior to analysis.

Datasets

Datasets used for the training and testing are the most crucial process in any deep learning methods. It provides the required accuracy with proper methodology and number of datasets. True labels and predicted labels are taken in below diagram. Ratio of two emotions together found in a single face is also provided in below table 1.

Table 1. Confusion Metric of emotions

True ↓	Angry	82.7	1.24	0.23	2.58	9.37	1.54			2.31
	Disgust	1.14	93.0	0.1	2.24	2.01	0.77			0.67
	Fear	0.57	0.77	94.6	1.04	1.81	0.94			0.27
	Happy	2.0	0.97	0.57	80.5	10.68	2.24			3.04
	Neutral	6.43	1.74	0.8	11.05	71.74	4.32			3.92
	Sad	1.54	0.84	0.54	2.81	8.03	85.21			1.04
	Surprise	3.05	1.04	0.44	5.23	10.09	1.68			78.47
		Surprise	Angry	Disgust	Fear	Happy	Neutral	Sad		
	Predicted Labels →									



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. CONCLUSION

Our review underscores the significance of multimodal approaches, which integrate in enhancing the precision and dependability of emotion recognition systems, visual, auditory, and physiological signals play a crucial role. Utilizing the complementary aspects of these modalities, more refined and dependable emotion detection models could be developed. Empirical investigations conducted across various datasets have validated the feasibility and efficiency of employing machine learning and deep learning techniques to automatically recognize emotions from video data. These methods vary from fundamental facial expression recognition to the analysis of complex affective states, demonstrating promising outcomes in diverse applications such as human-computer interaction, affective computing, and mental health assessment. However, despite the advancements, several challenges and limitations persist. These include the necessity for larger and more diverse datasets, particularly concerning cultural and demographic variations in emotional expression. Moreover, addressing issues regarding model interpretability, adaptation to real-world scenarios, and ethical considerations surrounding privacy and consent remains crucial.

REFERENCES

- [1] CHENGHAO ZHANG AND LEI XUE "Autoencoder With Emotion Embedding for Speech Emotion Recognition" in IEEE con, March 30, 2021
- [2] MINJIA LI 1, LUN XIE 1, ZEPING LV2 , JUAN LI, AND ZHILIANG WANG "Multistep Deep System for Multimodal Emotion Detection With Invalid Datasets in the Internet of Things" in IEEE con, October 23, 2020.
- [3] L. H. Nakamura, P. H. Gomes, G. L. Libralon, R. I. Meneguete, G. P. R. Filho, G. T. Giancristofaro, G. Pessin, B. Krishnamachari, and J. Ueyama, "Utilizing IoT technologies to improve health-oriented smart homes by implementing patient identification and emotion recognition., Comput. Commun." , Sep. 2016.
- [4] Azimi, T. Pahikkala, A. M. Rahmani, H. Niela-Vilén, A. Axelin, and P. Liljeberg, "Missing data resilient decision-making for healthcare IoT through personalization: A case study on maternal health, Future Gener. Comput. Syst".Jul. 2019.
- [5] P. Gong, H. T. Ma, and Y. Wang, "Emotion recognition derived from the multiple physiological signals", Jun. 2016.
- [6] M. Singh, K. Yadav, A. Kumar, H. J. Madhu, and T. Mukherjee, "Method and device for non-invasive monitoring of physiological parameters". Apr. 20, 2017.
- [7] C.-M. Kuo, S.-H. Lai, and M. Sarkis, "A streamlined deep learning model for resilient facial expression analysis. recognition." Jun. 2018.
- [8] O. Krestinskaya and A. P. James, "Facial emotion identification using min max similarity classifier", Sep. 2017.
- [9] Sujata, M. Trivedi, and S. K. Mitra, "A modular method for recognizing facial expressions using Euler principal components analysis" Dec. 2018.
- [10] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions". Apr. 2013.
- [11] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016). TensorFlow: A system for large-scale machine learning. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI) 16(pp. 265-283).
- [12] McFee, B., Raffe, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., ... & Casper, J. (2015). librosa: Audio and music signal analysis in python. In the 14th Proceedings python in science conference (Vol. 8, pp. 18-25).
- [13] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."
- [14] Simonyan, K., & Zisserman, A. (2014). Verydeep Convolutional networks designed for recognizing images on a large scale.. arXiv preprint arXiv:1409.1556.
- [15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Machine Learning Research Journal., 12(Oct), 2825-2830.
- [16] Grinberg, M. (2018). Flask web development: Developing web applications with Python. " O'Reilly Media, Inc."
- [17] Taiba Majid Wani, Hasmah Mansor, Teddy Surya Gunawan, Mira Kartiwi, Syed Asif Ahmad Qadri Nanang Ismail. "Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks" Published in: IEEE 2017.
- [18] C. Jain, K. Sawant, M. Rehman and R. Kumar, "Emotion Detection and Characterization using Facial Features," 2018 3rd Conference and Workshops on an International Scale. Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, India, 2018, pp. 1-6.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com