International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Multimodal emotion recognition from audio and video

Nithyasri S *, Hemavarthini B and Bharathi N. Gopalsamy

Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani, Chennai, India.

International Journal of Science and Research Archive, 2024, 12(01), 142–149

Publication history: Received on 14 March 2024; revised on 26 April 2024; accepted on 29 April 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.12.1.0723

Abstract

As humans we want to interact with a machine as we would with a person, in a way that it understands us, advises us, and looks after us with no human supervision. Current systems lack empathy and user understanding in spite of very effective logical reasoning. By predicting the emotions of the users, we are able to identify their needs and cater to them as best as possible. Emotion recognition in video and audio has many potential applications, including conversational agents, recommendation systems as well as systems for smart homes, mental illness care, virtual reality games, remote physical training, education and car-hailing services. The aim of the project is to develop an automatic emotion detection system based on voice and facial expression. We propose a model that highlights contextual, multimodal information for emotion detection and recognition. If systems can understand emotions and respond accordingly to behavioral patterns, we can anticipate artificial agents becoming our cognitive-consulting partners in our daily lives. Additionally, the project can be expanded to make interactions more natural and better suited to handle complex situations.

Keywords: Multimodal Emotion Recognition; Computer Vision; Deep Learning; Audio and Video; OpenCV; Librosa

1. Introduction

The objective of this project is to develop a method of identifying emotions from both sound and video inputs. This will involve creating an ensemble model that combines data gathered from both a microphone (audio) and a camera (video) in a clear and easily understandable manner. Recognizing emotions is an ongoing process that is sensitive to a person's current state, meaning that the emotions associated with a particular action may differ from person to person. To ensure meaningful communication, it is important. Accurately interpreting emotions in real-time is crucial. In everyday life, emotional recognition is vital for social interactions, and emotions are a significant factor in determining human behavior. It is something that comes naturally to most humans. Emotions have a big effect on human decision making. Advanced systems like conversational agents and recommendation systems, that involve human users are better off integrating emotions in the loop. This not only enhances system performance but also helps in the integration of various components.

2. Literature Review

The literature review below provides an overview of the research on face detection, evaluation metrics for imbalanced classification, deep learning for multimodal integration, data fusion, and emotion recognition using different modalities.

Face detection is an essential task in computer vision and has numerous applications.

Brownlee [2] provides a comprehensive tour of various evaluation metrics that can be used to evaluate the performance of classifiers on imbalanced data sets Imbalanced datasets are datasets where the number of samples in each class is not equal. This can be problematic for classifier models because they tend to be biased towards the majority class and

* Corresponding author: Nithyasri S

perform poorly on the minority class. Therefore, it is important to evaluate the performance of these models on imbalanced datasets using appropriate metrics.. Evaluation metrics for imbalanced classification is an important area of research that deals with the problem of imbalanced datasets in classification tasks.

Dimitri [3] provides a short survey of the applications, future perspectives, and challenges of deep learning for multimodal integration. The paper does not provide a detailed technical analysis of the deep learning models used for multimodal integration. This may limit its usefulness for readers who are interested in the technical aspects of these models. Deep learning for multimodal integration is an emerging area of research that deals with the integration of information from multiple modalities such as text, image, and audio.

Haileleol [4] provides an introduction to data fusion, covering the basic concepts, types of data fusion, and applications. There are two main types of data fusion: "low-level" fusion and "high-level" fusion. Low-level fusion involves combining raw data from multiple sources, such as sensor data or imagery, while high-level fusion involves combining processed data or information that has already been analyzed. Data fusion is another area of research that deals with the integration of data from multiple sources to produce more accurate and reliable results.

Somashekhar and Chiranth [5] propose a real-time human emotion recognition system using facial expressions. Emotion recognition using different modalities has received significant attention in recent years. In this paper, the authors focus on facial expressions as a modality for emotion recognition. Emotion recognition systems may exhibit bias, as they rely on data that may be limited in diversity and representativeness. This can result in inaccurate and unfair predictions, especially towards underrepresented groups.

Zexu et al. [6] propose a multi-modal attention mechanism for speech emotion recognition. The authors evaluate their proposed multi-modal attention mechanism on two benchmark datasets: the EmoReact dataset and the IEMOCAP dataset. The proposed multi-modal attention mechanism consists of two main components: a visual attention module and an audio attention module. One potential limitation is that the proposed model requires both audio and visual inputs to perform emotion recognition, which may not always be available or feasible in real-world settings.

Zhang et al. [7] propose a real-time video emotion recognition system based on reinforcement learning and domain knowledge. The proposed system relies on pre-defined facial feature extraction methods, which may not be optimal for all types of data. This could limit the generalizability of the system to different datasets and applications.

Zhao et al. [8] review the state-of-the-art in multimodal emotion recognition using deep learning. The authors discuss the various modalities that can be used for emotion recognition, including facial expressions, speech, text, and physiological signals such as electroencephalography (EEG) and electrocardiography (ECG). They also discuss the challenges associated with each modality, such as the need for high-quality sensors for physiological signals and the difficulty of accurately interpreting text-based data.

Chen et al. [9] propose a multimodal approach to emotion recognition using audio and video information. The hybrid approach used in the paper combines the benefits of both early and late fusion techniques. In this approach, the features extracted from both the audio and video modalities are first processed separately using separate models. The authors use two deep neural networks (DNNs) to extract features from the audio and video data respectively. The DNNs are pre-trained using a large amount of data and fine-tuned on the emotion recognition task.

Fan et al. [10] propose an attention-based multimodal feature fusion for emotion recognition. The proposed approach consists of two main components: a multimodal feature extraction network and an attention-based feature fusion network. The multimodal feature extraction network processes the input data from each modality separately and extracts a set of features that capture the relevant information. These features are then fed into the attention-based feature fusion network, which selectively combines the features from different modalities using attention weights. The resulting fused features are then used to predict the emotional state of the person.

Yang et al. [11] propose a novel deep learning-based framework for multimodal emotion recognition. The framework consists of three main components: a feature extractor, a modality-specific module, and a fusion module. The framework was evaluated on several benchmark datasets, including the AffectNet, IEMOCAP, and MSP-IMPROV datasets. The experimental results showed that the proposed framework outperformed several state-of-the-art methods for multimodal emotion recognition, demonstrating the effectiveness of the proposed approach.

Khorrami et al. [12] propose a deep multimodal fusion of acoustic and text features for speech emotion recognition. The authors use a deep neural network architecture that consists of two branches: one for processing acoustic features and one for processing text features. The acoustic branch processes the speech signal using a convolutional neural network (CNN) and a long short-term memory (LSTM) network. The text branch processes the transcription of the speech using a bidirectional LSTM network. The two branches are then fused using a concatenation layer and a fully connected layer.

In conclusion, the literature review above provides an overview of the research on face detection, evaluation metrics for imbalanced classification, deep learning for multimodal integration, data fusion, and emotion recognition using different modalities. It highlights the different approaches, challenges, and future directions of these areas of research.

3. Dataset Description

Dataset used for audio modality are:

- Ryerson Audio Visual Database of Emotional Speech and Song (RAVDESS)-768 audio files representing 4 emotions
- Surrey Audio Visual Expressed Emotion (SAVEE)-240 audio files representing 4 emotions
- Toronto emotional speech set(TESS)-1600 audio files representing 4 emotions

Dataset used for video modality is:

- Extended Cohn-Kanade dataset(ck+48)-501 images representing 4 emotions

The four emotions are ‘angry’,’ happy’, ‘sad’ and ‘fear’.

4. Challenges

This project faces three challenges, which are as follows:

4.1. Multimodality

The modalities being merged in this project may contain contradicting or duplicative information within the same frame. The task at hand is to pinpoint the context by understanding how to detect the interaction within a single modality over time while also capturing the interaction across modalities. Additionally, selecting relevant features poses an aspect of the problem that we intend to address.

4.2. Data complexity

The input data for this project comes from various modalities, making it naturally heterogeneous with distinct structures. Furthermore, the temporal aspect adds high-dimensionality to the data. The data is also prone to noise, and distinguishing noisy sources from outliers and preprocessing them present another problem that we intend to solve. One solution is to replace the noisy modality with a proxy feature vector obtained from the other modalities.

4.3. Emotion representation

Two emotion models are being considered: discrete and continuous. In the discrete approach, emotions are categorized into distinct states, such as anger or sadness, whereas, in the continuous approach, emotions are placed in a continuous space where the primary dimensions are more ambiguous. Discrete emotions are easier to comprehend, while the continuous space can represent a broad range of emotions. For the sake of simplicity, we have chosen the discrete model for the remainder of the work, but there is potential for further research in the continuous approach. Our objective is to create an algorithm for discrete emotion recognition that can extract relevant features, handle conflicting and redundant ones, perform dimensionality reduction and multimodal fusion of the extracted features, and predict an emotional representation that is realistic and easy to describe mathematically.

5. System Design

The system is designed in such a manner so as to incorporate the architecture, modules, interfaces, and data to satisfy specified requirements. On a high level, it is split into two modalities, audio unimodal and video unimodal.

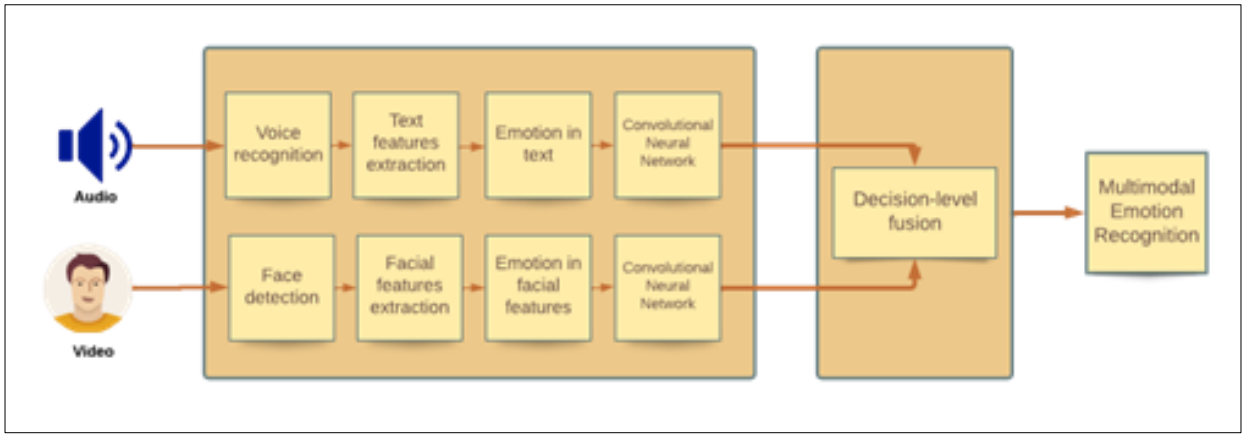

Figure 1 High Level application architecture for multimodal emotion recognition

The video and audio unimodal training is carried out individually which is then fused using late fusion method also known as decision level fusion.The predicted output post decision level fusion is the output of the multimodal considering both audio and video inputs.The purpose of decision level fusion is that it is considered to work better in case of varying input modalities.In addition to that fusion in general is used to come to a conclusion in predicting the emotion in case of contradictory(differ slightly) emotion exhibited by an individual verbally and non-verbally.

An overview of the application has also been laid out for a specific use-case involving a use and an admin.

Figure 2 Use case diagram for multimodal emotion recognition

A user has access to audio unimodal, video unimodal and multimodal options. Upon choosing the audio unimodal option, the user is redirected to a page where their microphone is enabled and the audio is recorded in real-time and analyzed automatically within seconds. But this goes through different stages internally. Once the audio is recorded, features are extracted from the input and stacked horizontally in the form of an array using the Librosa library and the model that is pre-trained by the admin is consulted and a prediction is returned. Similarly, for the video unimodal, the video is recorded in real-time by capturing frame by frame from the camera and face is detected using Haar Cascade methods provided by the python OpenCV library. This is done so by cropping a region of interest and using Haar features to detect the face.Haar features consist of a series of square shape functions that have been rescaled and are comparable to the convolution kernels utilized in Convolution Neural Networks. These features are employed on all pertinent facial regions to identify a human face. Once detected, relevant features are extracted, and the pre-trained model designated by the administrator is consulted to make a prediction, which is subsequently returned. When a user goes for the multimodal option, both of the above mentioned operations are performed simultaneously and the two outputs received undergo a data fusion where the weighted average is taken. Based on the output of the above process, an emotion is returned to the user.

6. Feature Detection and Extraction

To detect features in the video modality, Haar Cascade is implemented as an algorithm capable of detecting objects in images at any scale and location, running in real-time without complexity. The model is saved in XML files, and OpenCV methods can read it. Moreover, the algorithm includes boilerplate models for detecting faces, eyes, upper and lower body parts, and license plates. For the audio modality, Librosa, a Python package for analyzing audio and music files, is employed to extract features. This involves applying various techniques such as ZCR, Chroma_STFT, and MFCCs to audio input snippets, resulting in a stacked feature array. The zero-crossing rate of an audio time series is computed by ZCR. Chroma_STFT leverages short-term Fourier transformation to compute Chroma features, which represents audio's tonal aspects for pitch and signal structure classification. In Librosa, MFCCs provides an API for feature extraction and data processing in Python, simplifying the process of obtaining MFCCs by allowing arguments to set the number of frames, hop length, and number of MFCCs. Based on these arguments, a 2D array is returned. Finally, the feature array undergoes Mel Spectrogram, which logarithmically displays frequencies above a certain threshold and is stacked horizontally.

7. Modeling

After the feature extraction process, Convolutional Neural Networks (CNNs) are utilized to construct the model, which includes three layers: convolutional layers, max pooling layers, and fully connected layers. The initial layer is the convolutional layer, responsible for identifying the different features from the input images. Following the convolutional layer, a pooling layer is applied to decrease the computational burden by reducing the connections between layers and conducting independent operations on each feature map. In the max pooling method, the feature map's most prominent element is chosen.The convolutional layer's extracted features are then generalized, and the network can recognize them independently, significantly decreasing computations in the network. The fully connected layer, which is typically located towards the end of the CNN architecture, just before the output layer, is comprised of neurons, weights, biases, and two primary operations: flatten and dropout. In the flatten operation, the input from previous layers is compressed into a single vector and fed into the fully connected layer. Mathematical operations are generally performed at this stage, which initiates the classification process. Two fully connected layers are preferred over a single connected layer, although they are computationally expensive. To reduce overfitting with respect to the training dataset, a dropout layer is added, dropping some neurons during the training process, resulting in a smaller model size. Activation functions are also used, adding non-linearity to the network and approximating continuous and complex relationships between variables. As this is a multi-class classification into four emotions, a softmax activation function is utilized. The CNN's layers reduce human supervision, and the dropout layer enhances the model's performance.

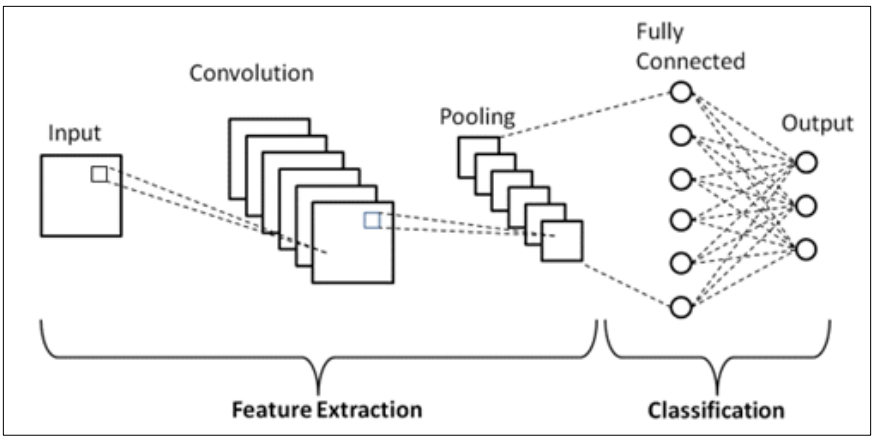

Figure 3 Architecture of a CNN

8. Evaluation Metric, Loss Function And Optimiser

The evaluation metric to assess the model is accuracy and the loss function chosen is categorical cross-entropy with Adam optimizer. On choosing accuracy, we ensure that the features chosen for the purpose of modeling are the right ones and complement the output that is predicted. When dealing with two or more output labels in multi-class

classification models, categorical cross-entropy is commonly utilized. The output label is assigned a one-hot category encoding value, where categorical labels are represented as a series of 0s and 1s. Adam optimizer is also being made use of and is an extension of stochastic gradient descent (SGD) which finds applications in deep learning such as in computer vision and natural language processing. Adam converges rapidly to a “sharp minima” whereas stochastic gradient descent is computationally heavy and converges to a “flat minima” but generally performs well on the test data.

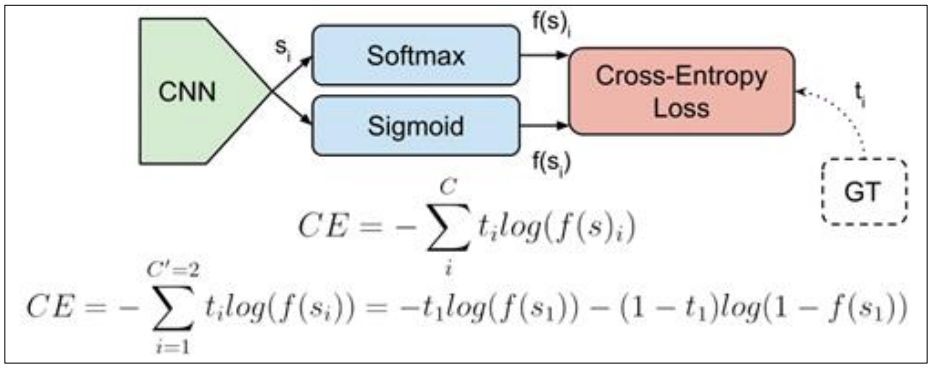

Figure 4 Formula for calculating categorical cross-entropy

9. Results And Conclusion

Video unimodal attained a training accuracy of 99.33% with an accuracy of 91% on the testing data.The audio unimodal attained an accuracy of 98.31%, with an accuracy of 86.9491% on the test data.

Decision level fusion (weighted average) is used for obtaining the final result.

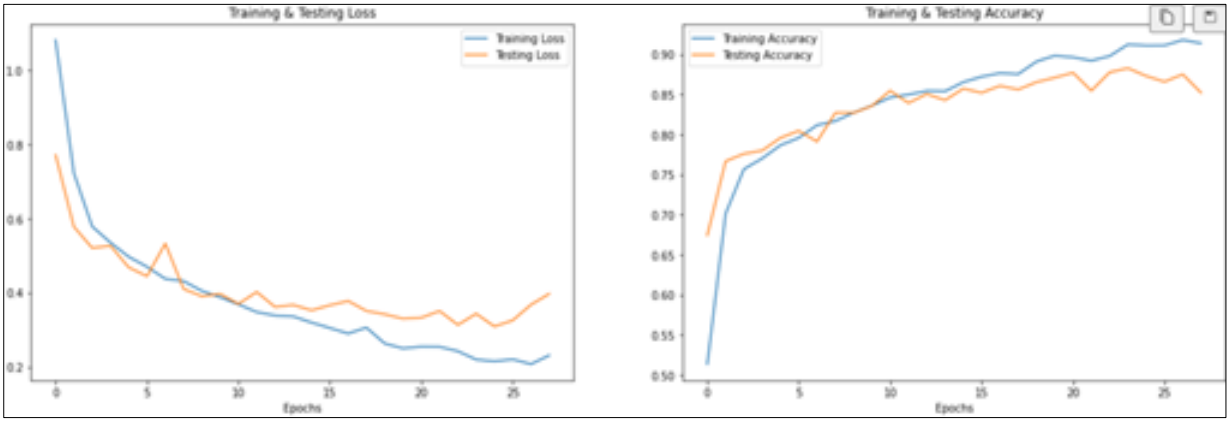

Figure 5 Training and Testing Loss and Training and Testing Accuracy

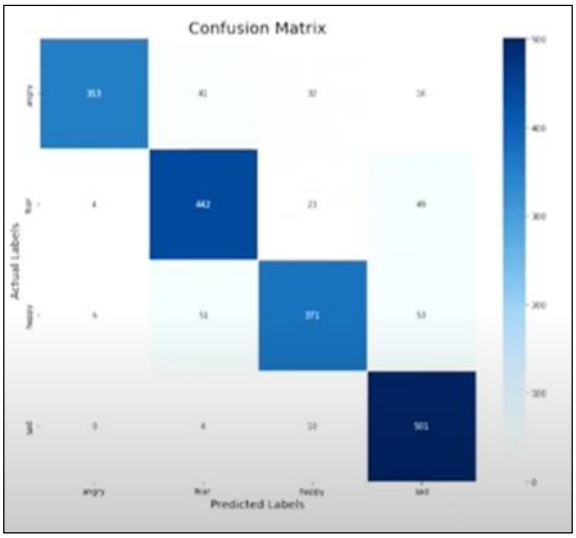

Figure 6 Confusion matrix of the audio unimodal

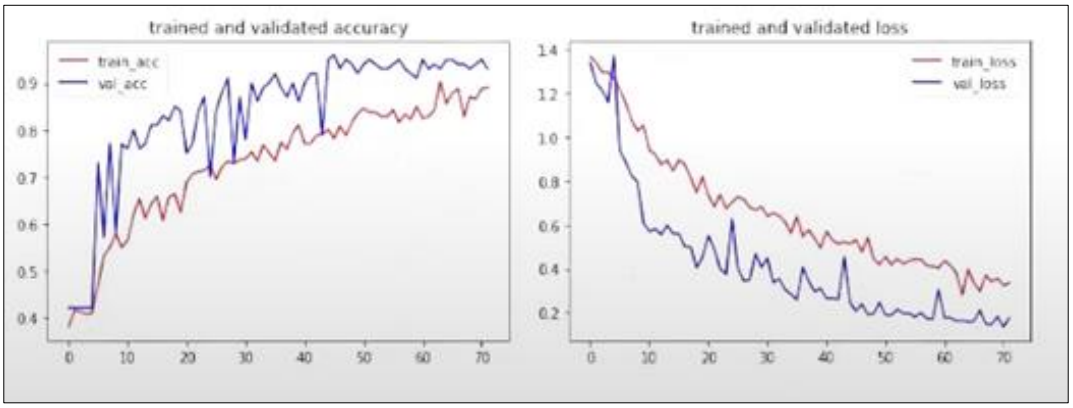

Figure 7 Training and Testing Loss and Training and Testing Accuracy

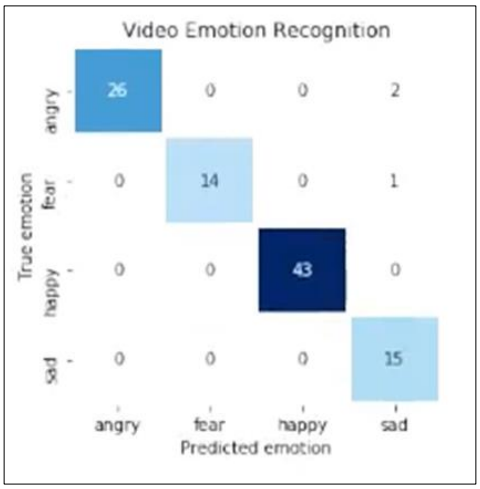

Figure 8 Confusion matrix of the video unimodal

The decision level fusion of the audio and video modalities is being performed with the purpose of predicting emotion being depicted verbally and non-verbally. Emotion recognition is a dynamic process that is constantly evolving as the emotional state of the person is being targeted. Additionally, the emotions corresponding to each individual are different. So there is a huge potential for expansion in different directions, like adding a further few modalities to detect hand gestures, posture, movements and so on. The scope of the project can also be enhanced in the field of IOT by including biological signals in order to incorporate multi-dimensional inputs.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Behera, Girija Shankar. Face Detection with Haar Cascade. Exploring a bit older algorithm which... | by Girija Shankar Behera. Towards Data Science, 24 December 2020.

[2] Brownlee, Jason. Tour of Evaluation Metrics for Imbalanced Classification - MachineLearningMastery.com. Machine Learning Mastery, 8 January 2020.

[3] Dimitri, G. M. A Short Survey on Deep Learning for Multimodal Integration: Applications, Future Perspectives and Challenges..

[4] Haileleol, Haylat T. INTRODUCTION TO DATA FUSION. multi-modality | by Haylat T | Haileleol Tibebu. Medium, 29 January 2020.

[5] Somashekhar, B. M., and N. L. Chiranth. REAL TIME HUMAN EMOTION RECOGNITION. IJRTI, 2018.

[6] Zexu, Pan, et al. [2009.04107] Multi-modal Attention for Speech Emotion Recognition. arXiv, 9 September 2020.

[7] K. Zhang, Y. Li, J. Wang, E. Cambria and X. Li, Real-Time Video Emotion Recognition Based on Reinforcement Learning and Domain Knowledge, in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 3, pp. 1034-1047, March 2022, doi: 10.1109/TCSVT.2021.3072412.

[8] Zhao, Y., Liu, Y., & Feng, G. (2021). Multimodal emotion recognition using deep learning: A review. IEEE Transactions on Affective Computing, 12(1), 1-1. doi: 10.1109/TAFFC.2021.3102672

[9] Chen, H., Huang, Y., & Liu, J. (2020). A multimodal approach to emotion recognition using audio and video information. Multimedia Tools and Applications, 79(27), 19387-19406. doi: 10.1007/s11042-020-09419-3

[10] Fan, W., Wang, S., & Li, J. (2020). Multimodal emotion recognition using attention-based multimodal feature fusion. IEEE Access, 8, 163361-163372. doi: 10.1109/ACCESS.2020.3026525

[11] Yang, H., Liu, Y., & Jiang, Y. (2019). A novel deep learning-based framework for multimodal emotion recognition. IEEE Transactions on Neural Networks and Learning Systems, 30(5), 1365-1376. doi: 10.1109/TNNLS.2018.2844739

[12] Khorrami, P., Le Paine, T., Brady, D., & Leung, T. (2019). Speech emotion recognition using deep multimodal fusion of acoustic and text features. In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7370-7374). doi: 10.1109/ICASSP.2019.8683336