PREPROCESSING CHALLENGES FOR REAL WORLD AFFECT RECOGNITION

Karishma Raut and Sujata Kulkarni

Electronics and Telecommunication Engineering, Sardar Patel Institute of Technology, Mumbai, India

ABSTRACT

Real world human affect recognition requires immediate attention which is a significant aspect of human-computer interaction. Audio-visual modalities can make a significant contribution by providing rich contextual information. Preprocessing is an important step in which the relevant information is extracted. It has a crucial impact on prominent feature extraction and further processing. The main aim is to highlight the challenges in preprocessing real world data. The research focuses on experimental testing and comparative analysis for preprocessing using OpenCV, Single Shot MultiBox Detector (SSD), DLib, Multi-Task Cascaded Convolutional Neural Networks (MTCNN), and RetinaFace detectors. The comparative analysis shows that MTCNN and RetinaFace give better performance in real world data. The performance of facial affect recognition using a pre-trained CNN model is analysed with a lab-controlled dataset CK+ and a representative wild dataset AFEW. This comparative analysis demonstrates the impact of preprocessing issues on feature engineering framework in real world affect recognition.

KEYWORDS

Affect recognition, preprocessing, audio-visual features, human-computer interaction

1. INTRODUCTION

Affect recognition is relevant in many fields such as emerging human-computer interfaces, assistive robotics, healthcare, and security. Artificial intelligence and machine learning techniques can be used to analyze human behavior to help society in many ways. Human-computer interfaces (HCI) can be developed to help autistic children interpret other people's facial expressions [1]. In a connected healthcare framework, automatic feedback about a patient's health by recognizing the patient's emotions provides a remote facility for diagnosis or primary screening and telemedicine [2]. For AI techniques to be applied and accepted in clinical settings, the interpretability of the features they learn and the justification of the conclusions they make are essential [3].

The automatic extraction of affect cues from an uncontrolled real world environment is an extremely challenging task and multidisciplinary task that involves various aspects of signal processing and computer vision along with social psychology. A more accurate representation of human affect is provided through multiple modalities, hence exploring methods to process multimodal data and smart fusion strategies are of utmost importance. Despite the fact that there are multiple input modalities for affect recognition, the most significant contributions can be made by audio-visual data. These two input modes are more expressive than other input modalities and can be recorded non- invasively. For example, visual data plays an important role to recognize happiness whereas audio data is crucial for understanding anger [4]. Simultaneous audio-visual cues can provide better discrimination of emotions, such as fear, disgust, and

surprise. The recognition rate and robustness depend on the way features are extracted from these two signals and the fusion between them. Our contributions to the field are:

- The challenges in preprocessing real world data are highlighted through experimental testing and comparative analysis using different detectors on uncontrolled and real world visual data.
- Its impact on feature extraction and prominent feature selection is discussed by analyzing results from the pre-trained CNN model on CK+ and AFEW datasets.
- The relative importance of audio modality is discussed.

2. REAL WORLD AUDIO-VISUAL AFFECT RECOGNITION

Generally, the audio-visual emotion recognition approach consists of four steps: (1) audio-video dataset, (2) preprocessing, (3) feature extraction, and (4) fusion and classification. The general pipeline for audio-visual emotion recognition is presented in Figure 1.

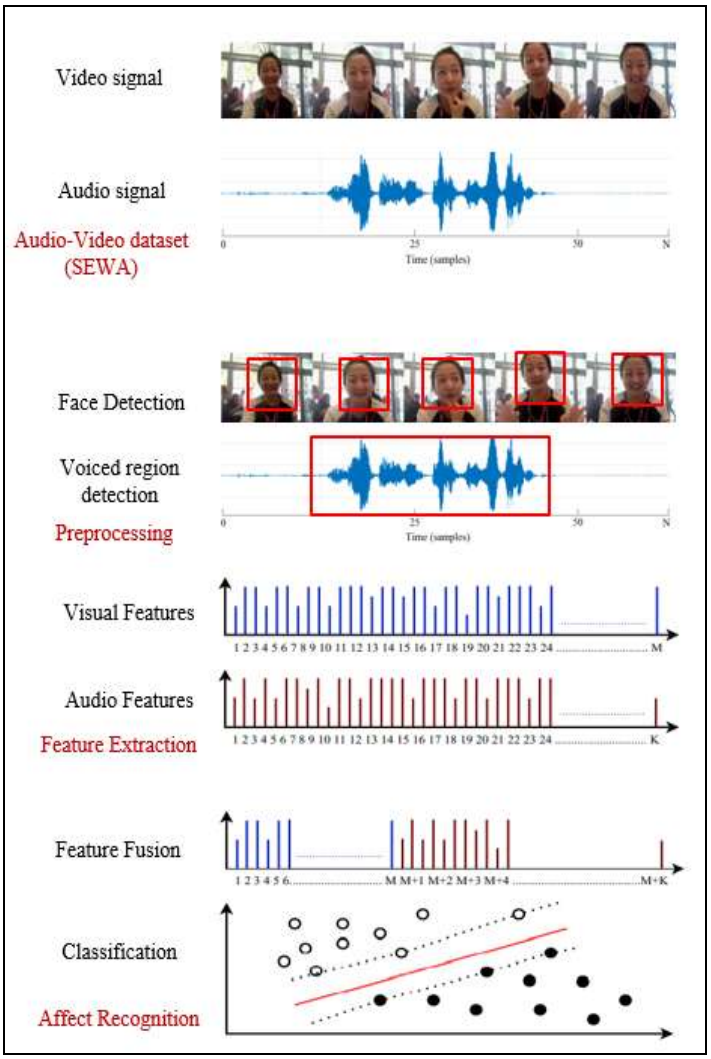

Figure 1. The general Multimodal Affect recognition framework.

The datasets play an important role to develop precise models. The application area demands more realistic, spontaneous displays of affect which are limited in number. The majority of datasets that are currently available and appropriate for audio-visual affect research have been gathered in laboratories or under controlled circumstances, with maintained noise level and reverberation, frequently with minimal verbal content, illumination, and calibrated cameras [5]. Such situations are completely different from real-world applications, and tools developed using such information typically have poor performance when applied to behavioral recordings taken in the wild. [6], [7]. From the last decade, the research has seen more emphasis on real, spontaneous displays of affect but is limited in number due to the lack of data. These real world databases were small in size and had fewer participants. Also, they are said to be cultural-specific, e.g., the RECOLA database consists of French-speaking participants only [8]. It is necessary to make use of databases that would make it possible to conduct an extensive study on how culture affects people's ability to recognize and express their emotions.

Figure 2. (a) Lab based Datasets (CK+). (b) Real world datasets (AFEW) [5,9].

To record the whole spectrum of expressions that people make in their daily lives, one option would be to incorporate data from the wild (for instance, from various Internet sources). The Acted Facial Expressions in the Wild (AFEW) is the first database with realistic temporal video scenarios that covers toddler, child, and teenager subjects. The database includes videos showing natural scenarios, close-to-real world illumination, with multiple subjects, occlusions [9]. The most recently updated database Affect-in-the-wild (Aff-Wild2) contains total 558 videos with around 2.8 million frames. These are YouTube videos having large variations in subject pose, age, and illumination conditions [10]. The Automatic Sentiment Analysis in the Wild (SEWA) database was created utilizing conventional web cameras and microphones in entirely uncontrolled, real world settings. It contains audio and video recordings of uncontrolled conversations between people of various ages, genders, and cultural backgrounds [11].

Preprocessing is a significant step in which the relevant information is extracted from the audio and video modalities. It includes face detection and tracking method, voice/non-voice region detection method, and peak stage facial expressions and speech segment selection method. Human face detection and tracking in video frames are extensively studied in computer vision [12]. Voice/non-voice region identification in audio has also been well investigated in signal processing. The peak stage is the time point where people indicate the maximum variation in facial muscles and vocal tract to express their feelings. The peak stage frames also referred as keyframes of audio-video signals provide the most valuable facial expressions and speech information for emotion recognition. Existing keyframe selection strategies are broadly categorized into three different types: (1) dissimilarity analysis, (2) divergence analysis, and (3) cluster analysis. The dissimilarity analysis method assumes that the contents of the peak frame have the maximum distance from other frames. The divergence analysis method computes the Euclidean distance between the visual frames and the transformed neutral face subspace. The cluster analysis methods have grouped the visual frames into n number of distinct clusters. Each cluster contains a similar type of visual frame with minimum distance among them.

Feature extraction is an important step to successfully detecting human emotions in audio-video signals. The feature extraction methods transform the audio-video signals into the feature vector that inputs the classification model [13]. Many methods use spatial features to express emotions from video frames, which restricts their capacity to encode rich dynamic information. Some researchers have extracted spatiotemporal data to increase estimation accuracy. Though expression variations happen over a large range, these approaches only use one range of temporal information from video. Table 1 shows the most useful audio and visual features [14], [15].

Similarly, Mel-frequency cepstral coefficients (MFCCs), a popular spectral feature for speech identification, only convey the signal's short-term spectral qualities, leaving out crucial temporal behaviour information [16]. The existence of Spectro-temporal (ST) receptive fields can extend up to hundreds of milliseconds and are responsive to time-frequency domain modulations. The characteristics of the speech signal are captured in both spectral and temporal terms by the modulation spectrum, which is based on frequency analysis of the temporal envelopes (amplitude modulations) of various acoustic frequency bins. Regarding the feature space's dimension and the signal fusion, these factors must be taken into account [17], [18].

Table 1. Audio-visual features

Sr.No.	Visual Features	Description
1.	Appearance-based	It relates to facial texture changes
2.	Geometry-based	It is about facial configurations in which face shape is characterized through a set of facial fiducial points.
3.	Dynamic texture-based	The spatial appearance and dynamic motions in a video sequence are both simultaneously modelled. Ex: LBP-TOP, HOG-TOP
4.	Deep features	Extracted by deep models
Sr.No.	Audio Features	Description
1.	Prosodic features	The vocal folds have an impact on them and provide crucial indications about the speaker's emotions. Ex: Zero Crossing Rate (ZCR), Short Time Energy (STE)
2.	Spectral features	They communicate the frequency content of the voice stream and are regulated by the vocal tract. Ex: Mel frequency cepstral coefficient (MFCC) and Linear Prediction Cepstral Coefficients (LPCC)
3.	Spectrogram	It is the visual representation of signal strength over time at different frequencies.
4.	Cochleagram	It is a variation of Spectrograms and closely mimics the behavior of the human ear compared to the Spectrogram.
5.	Modulation spectral feature (MSF)	It is an auditory spectro-temporal representation of the long-term dynamics of the speech signal.

Multimodal fusion can improve model robustness and accuracy since people express and perceive their affective states in social interactions via a variety of different modalities [19]. It is important to study and analyse the way modalities need to be fused [14].

Fusion can be accomplished early in the model development process, near to the raw sensor data, or afterwards by integrating separate models. Early or feature-level fusion allows the model to capture correlations between the modalities because features are retrieved individually, concatenated, and then learned as a joint feature representation. The outcomes of separate recognition models are combined through late or decision-level fusion [20]. Due to the large dimensionality of the raw data and varying temporal dynamics between the modalities, multimodal feature learning is difficult. Regarding audio-visual content, one of the difficulties is that typically there isn't a perfect alignment between the two data channels in terms of emotion expression. For instance, facial emotions can convey happiness at first, although the matching time-slices in the audio channels are not yet relevant. However, the following audio time slices may offer important information. [21], [22], [23].

Table 2. Fusion strategies

Fusion Strategy	Feature level or early stage	Decision level or later stage	Intermediate
Approach	The features are extracted independently close to the raw sensor data and then concatenated. The model can learn correlations between the modalities.	Aggregates the results of independent recognition models optimally	Data fusion at different stages of model training.
Example	Principal component analysis (PCA) Canonical correlation analysis Independent component analysis	score weighing Bayes rules, max-fusion and average-fusion	Fusion layer or a shared representation layer.

It implies that emotional expressiveness may rise or fall gradually, peaking at particular times. As a result, the amount of time needed for accurate emotion recognition may vary. Therefore, these modifications could be exploited efficiently while designing the framework [24].

3. EXPERIMENTAL ANALYSIS

Preprocessing is a significant step in which the relevant information is extracted from the audio and video modalities. It includes face detection and tracking method, voice/non-voice region detection method, and peak stage facial expressions and speech segment selection method. It is of most importance in real world situations, where alignment, head poses variations, and noise will be observed.

The experimental analysis is carried out using the DeepFace python framework for face detection on the AFEW dataset to understand the impact of preprocessing on real world affect recognition. The default VGG-FACE model configuration is used to detect, align and normalize face. Different detectors like OpenCV, SSD, DLib, MTCNN, and RetinaFace are used for face detection and alignment. The results are shown in Figure. 3. The complex head position and illusion have a severe impact on the performance of OpenCV, SSD, and DLib detectors. These detectors are not able to detect the face in dynamic conditions. In phases of detection and alignment under dynamic conditions, the RetinaFace and MTCNN appear to outperform other algorithms. As per experimental analysis, 5% of images are not detected by any detector and in such condition, audio modality can play a significant role in real world affect recognition. These

preprocessing issues are one of the major reasons for the poor recognition rate in real world conditions. This can be analyzed by comparing the recognition rate of a lab-controlled dataset and a real world dataset for a particular framework.

Figure 3. Face detection using the Deepface framework. (a)Frontal face detected by all detectors. (b) The Face is detected by all but not properly aligned by OpenCV and SSD. (c) Variation in a head pose is only detected by MTCNN, Retinaface. (d) MTCNN, Retinaface detectors are only successful on Head pose variation and illumination effects.

The performance of facial affect recognition using pre-trained CNN model is analysed with a lab-controlled dataset CK+ and a representative wild dataset AFEW. The Face Emotion Recognizer (FER) is an open-source Python library that uses a convolution neural network trained on FER 2013 dataset [25].

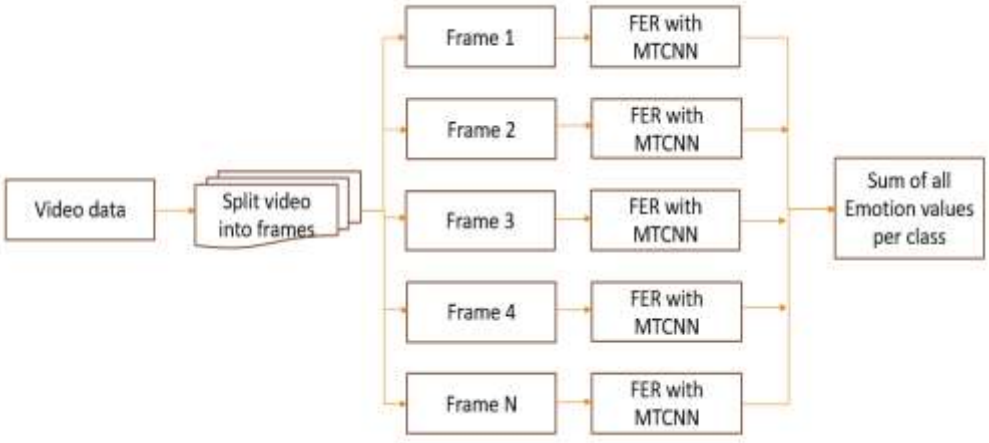

Figure 4. Facial emotion detection framework.

The model detects faces using OpenCV's Haar Cascade Classifier by default. According to the preprocessing results, real-world applications with more visual changes of human faces suffer greatly from performance degradation. As an alternative, a Multi-Cascade Convolutional Network (MTCNN) that is more sophisticated must be utilized [26]. It uses a three-stage, cascading deep convolutional network architecture to predict face and landmark locations from coarse to fine. The frames are extracted from the AFEW and CK+ dataset and each frame is analysed for emotion detection. The decision is made by taking individual emotion values per frame that were recognized by the model and finding which emotion was dominant across the entire video. The experimental analysis shows following results.

Table 3. CK+ Confusion matrix

Class	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	72	0	0	0	32	31	0
Disgust	139	16	0	8	7	7	0
Fear	7	0	8	8	9	36	7
Happy	0	0	0	207	0	0	0
Neutral	0	0	0	11	33	10	0
Sad	2	0	10	0	3	69	0
Surprise	10	0	15	0	14	1	209

All the samples of the CK+ dataset are detected and processed to recognize particular emotion. The frontal posed faces give a good performance. The recognition rate of happy, sad, and surprise is excellent. The average performance is observed for angry and neutral emotions. The most misclassified emotions are fear and disgust.

Table 4. Evaluation parameters CK+

Class	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Precision	0.533	0.090	0.106	1	0.611	0.821	0.839
Recall	0.313	1	0.242	0.884	0.336	0.448	0.967
F1-score	0.394	0.165	0.148	0.938	0.434	0.579	0.898
Accuracy	62.58%						

The average recognition rate in terms of true positives is 62.58% for CK+. The error on out of sample test data is referred as the generalization error.

Table 5. AFEW Confusion matrix

Class	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	52	0	3	11	34	16	8
Disgust	19	0	1	12	18	20	3
Fear	15	0	7	2	21	20	11
Happy	19	0	5	79	22	16	6
Neutral	20	0	2	4	56	43	12
Sad	12	0	4	11	21	55	6
Surprise	9	0	6	6	20	15	13

The 5% data of the AFEW dataset was not recognized by the model due to diverse conditions. The emotions of fear and surprise are poorly recognized and disgust is completely misclassified. The average recognition rate in terms of true positives for AFEW is 35.65%.

Table 6. Evaluation parameters AFEW

Class	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Precision	0.356	0	0.250	0.632	0.291	0.297	0.220
Recall	0.419	0	0.092	0.537	0.408	0.504	0.188
F1-score	0.385	0	0.134	0.580	0.340	0.374	0.203
Accuracy	35.65%						

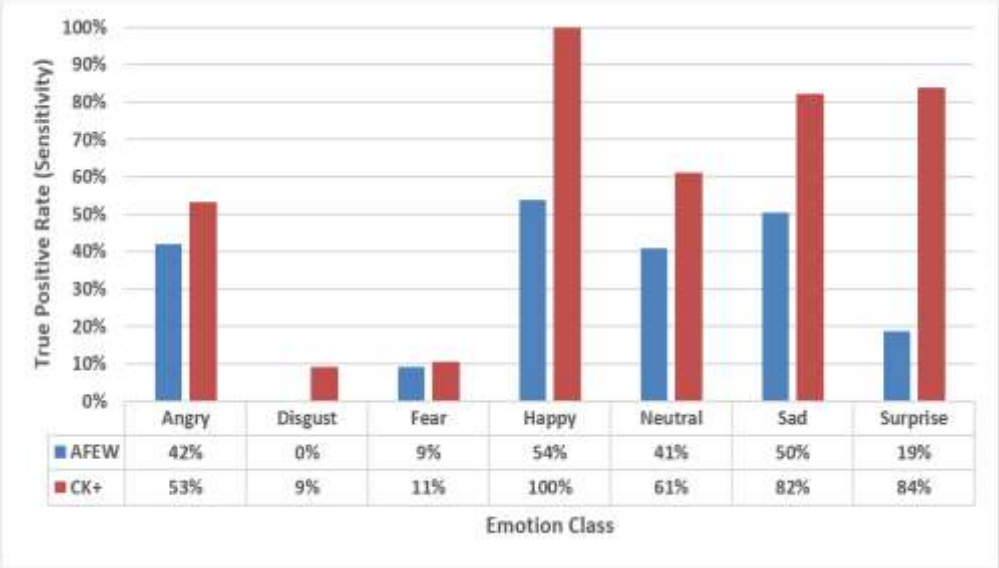

Figure 5. Sensitivity analysis of AFEW and CK+.

The cross-corpus examination shows how difficult it may be to identify an emotional state in a natural setting as opposed to a controlled setting. It highlights the reason for utilizing multiple modalities to provide supplementary information.

4. CONCLUSION

The real world situations are more diverse and it is more challenging to detect and align faces as compared to controlled situations. The original facial expression may appear differently due to occlusion and non-frontal head posture. It has a severe impact on affect recognition. It is one of the reasons for the misclassification of anger, fear, surprise, and disgust. The experimental results showed that the model provides an average recognition rate of 36% for the real world dynamic AFEW dataset, and 62.58% for CK+, a limited environment dataset. The comprehensive study and experimental analysis demand considering these preprocessing challenges while developing precise models for real world affect recognition. The techniques developed must assess the relative importance of specific segments in a coarse to fine manner. The cascaded networks are seen to be promising and must be capable of face alignment, pose normalization, and illumination normalization. In the future, the performance can be improved by employing deep models for feature engineering and by utilizing audio modality to provide supplementary information.

REFERENCES

[1] P. V. Rouast, M. T. P. Adam, and R. Chiong.: Deep Learning for Human Affect Recognition: Insights and New Developments. IEEE Transactions on Affective Computing. vol. 12, no. 2, pp. 524–543, (2021).

[2] M. S. Hossain and G. Muhammad.: An Audio-Visual Emotion Recognition System Using Deep Learning Fusion for a Cognitive Wireless Framework. IEEE Wireless Communication. pp. 62–68, (2019).

[3] T. Hassan et al.: Automatic Detection of Pain from Facial Expressions: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 43, no. 6. pp. 1815–1831, (2021).

[4] R. Srinivasan and A. M. Martinez.: Cross-Cultural and Cultural-Specific Production and Perception of Facial Expressions of Emotion in the Wild. IEEE Transactions on Affective Computing. vol. 14, no. 8, pp. 1–15, (2018).

[5] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010, June). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In 2010 ieee computer society conference on computer vision and pattern recognition-workshops (pp. 94-101). IEEE.

[6] S. Li and W. Deng.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Transactions on Image Processing. vol. 28, no. 1, pp. 356–370, (2019).

[7] R. Guetari, A. Chetouani, H. Tabia, and N. Khlifa.: Real time emotion recognition in video stream, using B-CNN and F-CNN. International Conference on Advanced Technologies for Signal and Image Processing. IEEE, pp. 5–10, (2020).

[8] Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013, April). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG) (pp. 1-8). IEEE.

[9] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon.: Collecting large, richly annotated facial-expression databases from movies. IEEE Multimedia. vol. 19, no. 3, pp. 34–41, (2012).

[10] D. Kollias and S. Zafeiriou.: Aff-Wild2: Extending the Aff-Wild Database for Affect Recognition. (2018). [Online]. Available: http://arxiv.org/abs/1811.07770.

[11] J. Kossaifi et al.: SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 43, no. 3, pp. 1022–1040, (2021).

[12] Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. IEEE transactions on affective computing.

[13] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari.: Audio-Visual Emotion Recognition in Video Clips. IEEE Trans. Affect. Comput., vol. 10, no. 1, pp. 60–75, (2019).

[14] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska.: Audiovisual emotion recognition in wild. Machine Vision and Applications. vol. 30, no. 5, pp. 975–985, (2019).

[15] W. Carneiro de Melo, E. Granger, and A. Hadid.: A Deep Multiscale Spatiotemporal Network for Assessing Depression from Facial Dynamics. IEEE Transactions on Affective Computing. vol. 3045, no. AUGUST 2019, pp. 1–1, (2020).

[16] Kudakwashe Zvarevashe and Oludayo Olugbara.: Ensemble Learning of Hybrid Acoustic Features for Speech Emotion Recognition. Algorithms. vol. 13, no. 3, (2020).

[17] Z. Peng, J. Dang, M. Unoki, and M. Akagi.:Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech. Neural Networks. vol. 140, pp. 261–273, (2021).

[18] A. R. Avila, Z. Akhtar, J. F. Santos, D. Oshaughnessy, and T. H. Falk.: Feature Pooling of Modulation Spectrum Features for Improved Speech Emotion Recognition in the Wild. IEEE Transactions on Affective Computing. vol. 12, no. 1, pp. 177–188, (2021).

[19] A. Salah, H. Kaya, and F. Gurpnar.: Video-based emotion recognition in the wild. Multimodal Behavior Analysis in the Wild. Elsevier , (2018).

[20] Y. T. Lan, W. Liu, and B. L. Lu.: Multimodal Emotion Recognition Using Deep Generalized Canonical Correlation Analysis with an Attention Mechanism. Proceedings of the International Joint Conference on Neural Networks, (2020).

[21] L. Stanciu and A. Albu.: Analysis on emotion detection and recognition methods using facial microexpressions. a review. 7th E-Health and Bioengineering Conference. pp. 21–24, (2019).
[22] A. Birhala, C. N. Ristea, A. Radoi, and L. C. Dutu.: Temporal aggregation of audio-visual modalities for emotion recognition. 43rd International Conference on Telecommunications and Signal Processing. pp. 305–308, (2020).
[23] E. Ghaleb, M. Popa, and S. Asteriadis.: Multimodal and Temporal Perception of Audio-visual Cues for Emotion Recognition. 8th nternational Conference on Affective Computing and Intelligent Interaction. pp. 552–558, (2019).
[24] E. Ghaleb and M. Popa.: Metric Learning-Based Multimodal Audio-Visual Emotion Recognition. IEEE Multimedia. pp. 37–48, (2020).
[25] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. (2013, November). Challenges in representation learning: A report on three machine learning contests. In International conference on neural information processing (pp. 117-124). Springer, Berlin, Heidelberg.
[26] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters, 23(10), 1499-1503.

AUTHORS

Karishma Raut is a research scholar at S.P.I.T. Mumbai, India. She has received Master of Engineering from University of Mumbai. She is currently working as Assistant Professor at VIVA Institute of Technology, India. Her research interests include deep learning, affective computing, HCI and related applications.

Dr. Sujata Kulkarni is academician with 22 years of experience. She is currently working as Associate Professor at S.P.I.T. Mumbai, India. Her research interests include Pattern recognition, Communication and Networking, Wireless communication networks, Embedded system.