Expression Recognition Survey Through Multi-Modal Data Analytics

[1]Kummari Ramyasree and [2]Ch. Sumanth Kumar

[1]Research Scholar, [2]Professor
[1.2] Department of E&ECE, GITAM Deemed to be University, Visakhapatnam, AP, India
[1] Assistant Professor, Guru Nanak Institutions Technical Campus, Hyderabad, TS,India

Abstract

In computer vision, capturing human expression from a video is essential for various time-sensitive applications, including driver safety and education. Because there are multiple expression models (e.g., speech, face, gesture, etc.), Human Expression Recognition can be done in various methods. In this work, we look at multiple strategies for recognizing facial expressions. We primarily concentrated on two models in this review: speech signal and facial image. The entrance survey is conducted in two steps, according to the generic approach of human expression recognition.The methods we used to extract features were classified into two groups: audio features and facial features. The review found different classifiers helped to decide the best framing of the Markov chain for a model. We also looked at specific multimodal expression recognition algorithms that used the multimodal notion at various stages, such as feature fusion and decision fusion. We also looked at several databases that have been used by previous studies in addition to these approaches. A full-fledged comparison is also provided to show the review ultimately.

Keywords: *Expression recognition, speech, face, MFCC, Prosodic, LBP, Machine Learning, Databases.*

1. Introduction

Emotion analysis has grabbed researchers' interest in recent years due to substantial improvements in Human-Computer Interaction (HCI) technology [1, 2]. Its broad applicability in diverse scenarios has sparked researchers' curiosity. The primary goal of the Emotion Recognition job is to connect humans and technology by recognizing emotional states. The issue starts with the abilities of artificial intelligence to recognize emotions through various stages, such as joy, boredom, and anger. Important to human-computer interaction (HCI) is the power of the machine to identify the user's mood and respond accordingly. Furthermore, it can manage the HCI system's actions regarding human emotion. The interactions between the personal computer, humans, smartphones, IPADs, and robots have risen due to regular daily acts such as voice, language, and Facial Expression. Emotion recognition technology is the key to developing the perfect computer personality with improved interactions between humans and their computers.

Emotion can be revealed by mixed signals such as facial expression, language, voice, and gesture. The varied study has given different viewpoints on multimodal emotion recognition based on previous findings. Mehrabian [4] discovered through an analysis that facial expressions account for 55 percent of overall expression, voice for 38 percent, and semantics (word, gesture, etc.) for the remaining 7%. As a result of this investigation, facial expression can be considered a primary model for accurately detecting emotions. Speech can be used as an auxiliary or supportive data model in an HCI system to identify human emotions. Speech signals have been widely exploited in emotion analysis and recognition systems [6-9]. An emotion identification system that relies entirely on a single model input, on the other hand, has several drawbacks. As a result, a fascinating study topic must be investigated, in which multiple models of input data for recognizing emotions must be examined to improve recognition performance. These numerous models combine speech signals, facial images, and gesture movements, among other things.

This review paper looks at a single (speech signals, facial expressions, and gesture motions, for example) and multiple models for emotion recognition that have previously been established (Combat ion of more than two models).The input signal model divides the emotion recognition system into two types: unimodal and multimodal. Following that, depending on the type of input data assessed, the prior strategies are classed as voice signal-based methods, facial image-based methods, and fusion-based approaches. The first two methods use voice as input data, the third way uses facial pictures as input data, and the fourth method uses a mix of face and speech as input data. This research looks into several methods for recognising emotions. The survey is separated into two sections: methods for extracting features and methods for classifying them. We focus on the methods used to extract features from input signals in the feature extraction

Manuscript received June 5, 2022
Manuscript revised June 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.6.74

methods section. Following that, we go over the several conventional algorithms used for classification, notably machine learning techniques, in the classification phase. Furthermore, we examine various standard datasets that have been used and generated by previous researchers for the aim of experimental validation. Finally, in the conclusion section, the possibilities for furthering emotion recognition research and the limitations of the present study are discussed.

The following is the rest of the paper: Section II delves into the specifics of the emotion recognition system model. The recognition system model is currently being finished in three stages: pre-processing, extraction, and classification. The full details of the Literature Survey are presented in Section III. Finally, in the conclusion section, I discuss possibilities for emotion recognition research and the limitations of the present analysis.

2. System model

Pre-processing, feature extraction, and classification are the three key components of an emotion identification system. The input in a unimodal system is just one model, whereas the information in a multimodal approach is several models. A comprehensive system model of uni- and multimodal emotion recognition is shown in Figure 1.

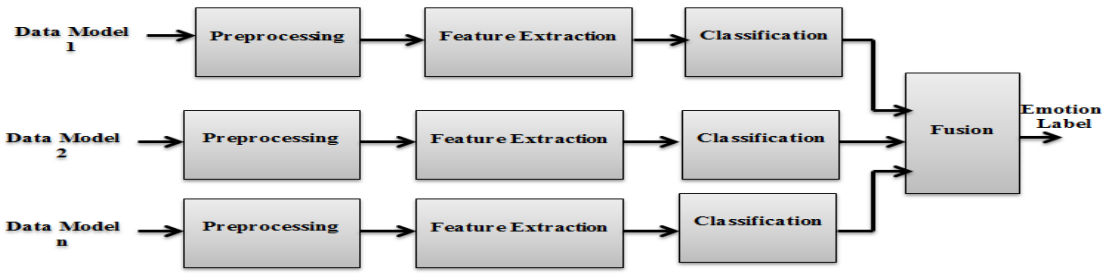

Figure.1 general schematic of the multimodal emotion recognition system

In the initial Pre-Processing phase, the input signal (speech or face) is converted into a suitable format. The Input signal is converted to a compatible format, such as numeric to symbolic or vice versa, during the Pre-Processing step. Only data that has been pre-processed is suitable for further processing. The technique to Pre-Processing will vary depending on the signal. In the case of facial expression image data, the Pre-Processing stage focuses on removing extraneous sounds and visual artefacts. The Pre-Processing process of a facial picture aids in removing external sounds, resulting in a more quality visual image. This step entails noise removal, contrast adjustments, and artifact removal, among other things. Next, for speech signals, the pre-processing stage removes sounds, extracts non-voice components, and performs windowing, among other things. The goal of a pre-processing step is to increase the signal quality of a voice signal, making it possible for the minor and major components of the speech to be recognized more clearly.

An emotion identification system's feature extraction stage extracts features from pre-processed data. The method for extracting characteristics varies depending on the signal. Several ways to efficiently extract feature sets from multimodal data have been developed in the past with diverse purposes. Some of these approaches were concerned with the entire set of traits, while others were only concerned with a subset of them. Furthermore, depending on the data size (one-dimensional, two-dimensional, or three-dimensional data), the feature set changes. The 1-D DWT, for example, is enough to filter out the signals' High- and Low-pass frequencies. In a facial picture or gesture movement, a 2-D DWT is necessary to filter out intermediate frequencies while considering low and high frequencies. As a result, developing a feature extraction strategy that allows portability between many signals without affecting the recognition system's overall accuracy performance is difficult when using multimodal signals for emotion recognition. Finally, the classifier stage uses the features gleaned from the second stage to identify emotions. Many ways are recommended at this stage to attain optimal performance. Machine learning and data mining techniques were at the forefront of many of the attempts. Other machine learning algorithms include Support Vector Machines (SVM), Artificial Neural Networks (ANN), Nearest Neighbour (NN), Decision Trees, and others. Furthermore, the database employed for experimental validation has a significant impact on the system's performance in the instance of emotion recognition. For facial photographs, speech signals, and speech and gesture movements, there are several data sources. For instance, the German Speech

database VAM [10] is a sample database for speech-based emotions that has 947 utterances in less than 12 hours. RML [11], a separate database, comprises 500 movies in six different languages plus six different emotional types (such as disgust, anger, happiness, fear, surprise, and sadness). Some authors may choose to use pre-made datasets, while others may choose to create their own. Table 1 provides a comprehensive overview of the databases utilised for emotion recognition.

Table.1. Standard emotion-related databases

Name of Database	Year	No. of Subjects used	Acquired process
CUAVE [22]	2002	20 speaker pairs and 36 individual speakers	Tilted head face, back-and-forth and side-to-side
interface '05 [20]	2006	2 (34 male, 8 females from 14 different nations)	Six basic emotions (Surprise, sad, happiness, fear, disgust, and anger)
TUM AVIC [19]	2007	21	Five different modes (laughter, hesitation, garbage, consent, breathing)
RML [11]	2008	8	Six basic emotions (Surprise, sad, happiness, fear, disgust, and anger)
SAVE [18]	2009	4 Male	Seven primary emotions (Neutral, Surprise, sad, happy, fear, disgust, and anger)
SEMAINE [17]	2010	150	5 (emotional intensity, expectation, power, activation and valence)
MHMC [16]	2012	7	4 emotions (neutral, angry, sad, happy)
A FEW [15]	2012	330 (multiple and single subjects per each emotion, age limit is 1-70 years)	Seven primary emotions (Neutral, Surprise, sad, happy, fear, disgust, and anger)
MAHNOB [14]	2013	22 (10 female and 10 male)	Speech, speech laughter, pose and pose laughter, & some other vocalizations
AVDLC [13]	2013	292 (ager from 18-63)	Different kinds of depression (minimal, mild, moderate, and severe)
RECOLA [12]	2013	46 (27 female and 19 male)	Five (Rapport, performance, engagement, dominance, agreement); valence and arousal
OuluVS2 [21]	2015	50	Five different views spanned frontal and profile views
CK+ [108]	2010	97	Both posed and non-posed (spontaneous) expressions.

3. Literature Survey

Based on the described system model in Section II, the Emotion recognition system comprises three phases. Many researchers developed various methods depending on multiple tactics to reach the primary intention of emotion recognition. Some focused on feature extraction, some mainly concentrated on pre-processing the input signal, and some on classifier designing through various algorithms like data mining and machine learning methods. A quick survey of previously developed emotion recognition algorithms is provided here, based on their focus. The entire survey is divided into two categories: data extraction and classification. Furthermore, feature extraction approaches are classified into three categories based on the input signal model: speech, face expression, & gestures.

3.1. Feature Extraction

In an emotion recognition system, feature extraction is critical. The system can analyze the input signal characteristics in-depth, utilizing the features, giving it more information about the signal. In recent years, numerous ways have been presented to achieve efficient outcomes in the emotion recognition system. Based on these findings, we discovered that most previous techniques focused primarily on signals as input, such as speech and facia images. Furthermore, we found that speech and speech signals are two distinct models that can communicate emotion; we divided the feature extraction methods into two categories based on these two models.

3.1.1. Audio Features

Choosing the fundamental factor for emotion identification from speech is a significant difficulty. In recent years, auditory features and prosodic qualities of emotional audio signal have been explored [24-28]. Prosodic characteristics are used to capture the major advantages of influencing content in spoken messages, and global prosodic features have really been reliably and widely used for speech-based emotion recognition [29], [30], and [32].

K. S. Rao et al. [23] suggested an emotion recognition system. To recognise emotions, they used pitch, energy, and duration values to represent the prosodic elements of the input voice signal. The gross statistics of minimal, maximal, average, variance, and slope of contours of features are explained by global prosodic features. The temporal dynamics of prosodic features are then shown using local prosodic features.

Idris et al. [31] used a number of voice features in their emotion recognition system, including quality, hybrid, and prosodic aspects. They use a total of five categories of emotional components to describe speech signals. Hybrid features account for two, voice quality features for two, and prosodic features for one. For experimental validation, they employed the Berlin dynamic database.

Jacob et al. [33] investigated the outcomes in emotion recognition based on prosodic features by extracting 1050 segmental and 1400 supra segmental features from English spoken wave files. The study looked at seven emotions: neutral, happy, sad, surprise, disgust, and fear, as well as six fundamental emotions: angry, cheerful, unhappy, shock, hatred, and anxiety. For this, they enlisted the cooperation of ten Indian female English speakers to compile the database. This study looked at

pitch, length, and intensity, all of which were statistically assessed.

Because both prosodic and spectral data contain emotional information, the combining of prosodic and spectral data is believed to improve the performance of the emotion identification system. Yu, Zhou, et al. [34] proposed associating the prosodic and spectral features in their emotion recognition system as a result of this. To arrive at a final choice, the combination of prosodic and spectral information is presented as a data fusion problem.

Next, Monorama Swain et al. [35] created an emotion recognition system for 4 Indian languages: Odia, Sambalpuri, Cuttack, & Berhampur, which was text and speaker independent. A dialect is an identifiable language spoken by a group of people. Prosodic features from speech, also including pitch, duration, energy, as well as fundamental frequency, are extracted and utilised to classify emotions. To test the system's performance for prosodic features, the orthogonal forward selection (OFS) technique is used for substantial feature selection. Morrison et al. [26] went on to explain the link among emotions and prosodic traits.

Spectral and Cepstral components such formants and "Mel-frequency Cepstral coefficients (MFCC's)" as well as spectral and Cepstral components such formants and "Mel-frequency Cepstral coefficients (MFCC's)" were usually applied and researched for emotion focus [35-40]. The MFCCs were used as acoustic features that could acquire local & temporal characteristics in [41, 43, 44]. Metallinou et al. [43] used a 39 D feature vector by merging three separate sets of features, such as 12MFCCs, Energy, and the related first and second derivatives, to create a 39 D feature vector. Many features have been introduced to the MFCCs in order to improve emotion identification performance. "In [45], features based on auditory characteristics of emotional speech are extracted using the Sub Band based Cepstral Parameter (SBC) and MFCC techniques.

Zhang et al. [46] investigated how the amount and statistical values of speech characteristics affect Emotions in Speech Recognition Accuracy. Based on the Gaussian mixture model (GMM), they distinguish two compelling features: MFCC's and Auto Correlation Function Coefficients (ACFC) are directly retrieved from speech signals. They use a GMM super vector formed from the values of MFCCs, delta MFCCs, and ACFCs to run tests with the Berlin emotions datasets. For emotion recognition, they looked at six primary emotions: disgust, rage, happiness, fear, sadness, and neutral.

Speech features can be classified into two types, according to model attributes [47, 49]: local (frame level) features and global (utterance level) features. The local characteristics are retrieved from speech features depending on the unit of speech "frame." Global features, on the other hand, are created utilising statistics from all speech features extracted from the entire "Utterance" [49]. Local features include energy LLDs (e.g., Loudness, Energy), voice LLDs (e.g., F0, Shimmer, and jitter), and spectral LLDs (e.g., MFCC's and Mel-Filter bank-(MFB)). Global characteristics include the collection of functions generated from the LLD, such as maximum, minimum, mean, standard deviation, duration, and linear predictive coefficients (LPC) [48].

3.1.2. Facial Features

According to previous research, the most commonly used facial features can be divided into two groups: geometric features and appearance features [51], [52]. Geometric characteristics depict the placement and shape of face features (Ex: Eyes, Mouth, Eyebrows, ears, etc.). Furrows, bulges, and wrinkles are examples of appearance traits that depict facial texture. The Active Appearance Model (AAM) introduced by Saatci and Town [53] is the most widely used method for extracting appearance features. It was used to determine gender as well as emotional states. Wilhelm et al. [54] studied and explored the differences between the use of "Independent component analysis (ICA)" and AAM through multiple classifiers for the identification of facial expressions, gender, identity, and age. Then, based on the divergent shape, Zhou et al. [55] suggested a Bayesian inference solution. A Bayesian Tangent shape modal is used to create the approximation. Lee and Elgammal [56] propose a new nonlinear generative model based on manifold embedding and maps of empirical Kernels for the description of facial expressions. The complicated nonlinear deformation of the look and forms in facial expressions is accommodated by this method. It creates a precise emotional harmony. Lie bel et al. [57] offer a multilayer interactive technique for AAM [61] that is suitable for 2D picture alignment and 3D shape alignment. Antonin et al. [58] use AAM to develop a set of high-level features known as "Expression Descriptive Units (EDU)" that may be used to compare the performance of different classifiers in identifying the six main emotions. Mathew [59] created a system for classifying emotional states based on still photographs of the face. This method comprises applying AAM [66], [67] to face photos from a publicly available database in order to capture contour and texture variation, which is critical for expression identification. The parameters of the AAM are used as features in a classification algorithm that successfully recognises faces linked with the six universal emotions. Based on AAM, Wen Chao et al. [60] described a method for recognising facial expressions utilising textural information from specific locations and the geometry of facial feature points. They started by looking at the physical significance of AMM's key parameters, and then discovered that the texture and shape parameters can express more emotion-related information. After that, a machine learning

classification technique was used to classify expressions using these two attributes. The authors then utilised AAM to extract 68 labelled "facial feature points (FFPS)" from five facial visual features, including the brows, eyes, nose, mouth, and face contour, in [62, 63]. Due to its strong local appearance descriptors, "Local Binary Patterns (LBPs)" [72-75] have been widely used for face expression recognition in recent years. LBPs were utilised as baseline features for the recent challenges in the AVEC 2011-2014 Challenges [68-71]. Schuler et al. [68], [69], for example, used LBP appearance descriptors to describe emotional aspects in a facial image. After detecting face and eye regions, the acquired face region was normalised based on eye locations. The facial image was then partitioned into tiny sections using binary comparisons, and histograms of LBP features were retrieved using 8 neighbour pixels in each area. Finally, the LBP properties of each tin region were blended into a single feature histogram vector to depict a facial image. Anusha Vuppature et al. [76] proposed an efficient pre-processing algorithm to perform emotion recognition from the facial images. They combined it with LBP to categorize emotions through "Kullback Leibler (K.L.) divergence." Initially, they employed Viola Jones (V.J.) algorithm to detect facial regions like eye pairs. This region helps in the extraction of an effective face region and helps in the nullification of the illumination effect. Then they employed LBP to determine facial emotions' texture properties, represented through a histogram vector. Using training images for seven basic expressions, a template histogram is created, which is then compared to test the best histogram match through the dissimilarity measure and K.L. divergence. A multi-dimensional LBP-based automatic expression recognition method was proposed by Krystian Radlak et al. [77]. The originality of the contribution stems from the use of the Random from Algorithm [79] for gene selection in microarray research and the high-dimensional LBP characteristics as input for Support Vector Machines (SVM) classifiers for facial expression categorization. They tested their method using the Static facial features in the wild (SFFW) database [80], which contains face images from movies that are significantly closer to real-time.

3.1.3. Fusion Approaches

We used feature fusion in this part to focus on strategies that primarily aim to improve the performance of emotion recognition. The features of the facial photos and the speech signal are fused in this method to create a new and composite feature set. Because these qualities offer so many benefits, various data fusion algorithms have been developed [85-88].Through feature level fusion, face and speech features are integrated to create a composite feature vector, which is then represented by a single classifier for emotion recognition [81-83]. Metallinou et al. [81] focused on the audiovisual detection of the emotional content of Spontaneous Emotional Conversations at the utterance level. They looked at Context Sensitive Systems for Emotion Identification in a multimodal, hierarchical framework, using "Hierarchical Hidden Markov Model Classifiers (HMMs)," "Bidirectional Long short-term memory (BLSTM)" [84], and hybrid HMM/BLSTM classifiers for modelling emotion evaluation within an utterance and between utterances throughout a dialogue. In addition, [82] presents a mixed bimodal feature verdict fusion technique to increase the performance of estimating emotions in impulsive speech interactions. To begin, a feature vector is created by merging audio information from the full speech phrase with video features from the numerous keyframes that comprise that sentence. After that, a decision level fusion of Predictions is used to generate the final estimation from all of the associated structures. [83] describes a linear log model dubbed "Bimodal Log-Linear Regression (BLLR)" that allows for interaction between the two modalities' characteristics. This strategy aimed at the performance improvisation of a laughter vs. speech discrimination problem because laughter and speech are natural and audiovisual phenomena. Fusion at the feature level employing a simple blend of audiovisual characteristics has been effectively applied in numerous applications. The high-dimensional feature set, on the other hand, is prone to data sparsity also fails to account for feature interactions. As a result, the benefits of combining auditory and visual cues will be limited at the feature level. To overcome the limitations of feature-level fusion approach, a large portion of data fusion research has concentrated on decision-level fusion [43, 81, 75, 89], in which various signals can be modelled by the appropriate classifier first. The recognition outcomes from each classifier are then combined in the end by combining voice and image data. "Error Weighted Combination (EWC)" [43, 81] is one of the greatest examples of a decision level fusion method. By merging actual evidence with past judgments, EWC [81] employed the Bayesian framework to integrate various inputs. GMM was utilised to prepare the facial data and HMM was used to train the voice modality. Finally, to efficiently find the emotion, weighted sums of individual decisions are calculated. However, because facial and vocal aspects have been described as complementary to one another in emotional expression [90], the assumption of conditional independence across multimodalities is incorrect at the decision level. The relationship between visual and aural modalities is improper in this case. The relationship between visual and aural modalities must be taken into account here.A model-level fusion technique was created to overcome this challenge [91, 41, 63, 16, 50, 92, 93] by focusing correlation data among different modalities and analysing the temporal link between audio and visual

information streams. There are numerous instances in this manner, including "Coupled HMM (CHMM)" [93, 94], "Triple HMM (T-HMM)" [50], "Semi Coupled HMM (SC-HMM)," and "multi-stream fused HMM (MFHMM)" [92].

Yan J et al. [111] introduced the MCFER, a multi-cue fusion emotion recognition framework based on three multimodal cues: audio signal, facial landmark, and facial texture. To capture the dynamic changes in the facial surface, they used "Convolutional Neural Network (CNN)" and "Bi-directional Recurrent Neural Network (BRNN)." Face landmarks are specifically represented by facial muscles. The emotion-related patterns are retrieved using CNN and SVM. To extract low-level acoustic properties from an audio source, CNN is used. At both the decision and feature levels, fusion is used. K.P. Seng and colleagues [112] proposed an audiovisual emotion recognition system based on a combination of rule and machine learning methods. "Bi-directional Principal Component Analysis (BDPCA)" is used to build the optical route, while "Least Square Linear Discriminant Analysis (LSLDA)" is used to reduce dimensionality. The "Optimized Kernel-Laplacian Radial Basis Function (OKL-RBF) neural classifier" is used to process the collected features. The audio path is then built using prosodic and spectral data. A feature-level fusion module is used to combine the derived features. The sound - visual elements are finally combined. Juan D.S. Ortega and colleagues [113] developed an audiovisual fusion-based emotion identification system for predicting valence and arousal levels. To extract the characteristics from video frames, this method used a pre-trained CNN and transfer learning. For audio content, a minimal collection of metrics such as prosodic, spectral, vocal tract, and stimulation properties are retrieved. The results are combined and sent into a Single Support Vector Regressor (SVR). The RECOLA database and the predicting of natural and impulsive emotions are used in the simulation. Cai et al. [114] proposed two new methods to fuse the features extracted from audio and video data models. Open Smile toolkit is used for audio and video features extraction; they employed LBP-TOP, bi-directional LSTM, and an ensemble of CNNs. Model-level fusion and feature-level fusion are employed for fusion. AmotioW2018 and AFEW datasets were used for simulation.

3.2 Classification

After extracting the relevant features from the input data models, they are processed for training through different machine learning algorithms. Here the categorization can be done using machine learning and data mining methodologies. Euclidean distance, correlative measures, KLD [6], length through power distribution Law, and entropy are required to accomplish classification in data mining methodologies. In the case of Machine Learning methodologies, ANN [31], HMM [63] and [93], SVM [79], [101] and [102], etc., are the most used techniques to classify the input data into its corresponding expression.M. Sinith et al. [100] used speech information, primarily pitch and energy statistics and spectral features, to construct an SVM-oriented expression detection system. With binary tree classification algorithms, two kernels, linear and Gaussian radial basis functions, are evaluated, one version one and one vs the others. [101] also developed a scheme based on texture analysis of all emotions and emotional elements that was manually categorised and explained. The levels of emotion intensity of all evocative words and content are assessed here. Using a threshold-based fusion of textual and auditory information, the same emotional state is estimated. SVM is employed in the recognition of emotional states. For facial emotion identification, [102] mainly concentrates on various learning methods, and several techniques are implemented like Deep Boltzmann Machine (DBM) and SVM at the classification phase. Recently, deep learning [95-99] has gained considerable interest in research because it has much tractability in data analytics through its analysis of the profound characteristics of input data. Based on the Convolution Neural Networks (CNNs), [95] proposed a new emotion recognition framework. This method starts with a new emotion recognition framework.This method starts with a face detection module that employs an ensemble of three cutting-edge face detectors, followed by a classification module that employs an ensemble of multiple deep CNNs. Each CNN model is pre-trained on a larger dataset provided by the "Facial Expression Recognition (FER) challenge 2013" and is randomly initialised. Pre-trained models were tweaked using the SFEW 2.0 training set. Log-likelihood loss minimization and hinge loss minimization are two strategies developed for combining different CNN models to obtain the ensemble weights of network responses. A novel approach for identifying emotions based on statistical face scans is offered, based on CNN's success. On test the performance analysis, the suggested method is used to the standard face web database CASIA. M. Goyani and N. Patel [104] developed a new version of LBP called "Local Mean Binary Pattern (LMBP)" for the extraction of texture features of the facial image. Unlike the encoding process of LBP, which seeks the magnitude of center pixels, the LMBP seeks the importance of the mean of the block cantered with a center pixel for encoding. For classification purposes, they used "Histogram Normalized Absolute Difference (HNAD)," which classifies the facial test image into one emotion class by comparing its LMBP features with the LMBP features of trained facial images. However, the main disadvantage of LBP and its subsequent methods is huge information loss and limited accuracy at recognition.

Min Guo et al. [105] suggested an expanded version of LBP called "Extended Local Binary Pattern (ELBP)," which was integrated with "The k-L transform (KLT)" to address LBP's drawbacks. The ELBP was first used to extract facial features, which were then subjected to covariance computation in order to reduce the size of the resultant element. Finally, they used the SVM method for categorization. A. Muqeet and R. S. Holambe [106] created an interconnected emotion identification system by combining LBP characteristics using Interpolation-Based Directional Wavelet Transform (DIWT)" components. To determine the frequency bands, the DIWT is used first, and then each band is treated to an LBP process that allows for adaptive direction selection. The LBP of top-level DIWT bands is then used to derive local descriptive characteristics. M. Revina and W. R. S. Emmanuel [107] developed a new descriptor termed "Dominant Gradient Local Ternary Pattern (DGLTP)" as part of their FER system. EDLTP is a "Local Ternary Pattern (LTP)" [30] extension that extracts the local dominating texture elements of a facial image. To eliminate noise in the pre-processing stage, the "Enhanced Modified Decision Based Unsymmetric Trimmed Median Filter (EMDBUTMF)" is used. After pre-processing, DGLTP extracts histogram characteristics, which are then input to an SVM classifier for expression categorization. The LTP, on the other hand, requires extra bits to pixel representation. Furthermore, directed neglect harmed recognition performance significantly. Hossain et al. [115] proposed an emotion recognition system based on deep learning and large amounts of video and audio emotional data. The audio signal is first processed to identify MFCC, after which it is passed to CNN for classification. After that, the video stream is divided into frames and sent to two ELMs in a row. The system's output is sent into SVM as an input. Along with the techniques described above, several other approaches have been presented previously to produce successful recognition outcomes in the case of various emotions. Based on earlier proposed methods, a simple comparative analysis is explained in Table.2.

Table.2. Literature survey comparison on the emotion recognition With Multimodal data

Author Name	Year	Methods	Features	Emotions	Database	Drawback
Song et al. [50]	2008	T-HMM and HMM	1. Facial Animation Parameters. 2. 16 formats and 48 prosodic features.	Neutral, Sad, Fear, Anger, Joy, Surprise	Own created	HMM three times introduces an enormous complexity
Zeng et al. [92]	2008	HMM, MFHMM	1. 12 facial movement units 2. Pitch energy	Seven basic emotions in 4 cognitive states	Own created	Markov process is not robust for facial features
Metallinouet al. [43]	2008	Bayesian Classifier, GMM, and Weighting scheme	1. Separation of the facial region into smaller regions 2. 39-dimensional MFCC	Sad, Neutral, Happy, and Angry	IEMOCAP	MFCC has poor performance in the presence of background noise.
Paleariet al. [91]	2009	SVM, N.N., and N.N. based on Evidence Theory (NNET)	1. Absolute Facial movements 2. 1st five formants and 48 prosodic features	Sad, Happy, Fear, Disgust and Angry	eNTERFACE'05	Less efficiency at decreased annotation
Nicolaou et al. [94]	2010	SVM, HMM, and Panic particle Filter based tracking	1. Corners of Chin (1 point), Mouth (4 points), nose (3 points), Eyes (8 points), Eyebrows (4 points), 2. MFCC	Arousal, Valence	Sensitive Artificial Listener (SAL)	Several filters influence MFCC efficiency
Metallinou et al. [103]	2010	Bayesian Fusion, GMM, and HMM	1. Facial Marker Coordinates 2. MFCC	Sad, Neutral, Happy, Angry	IEMOCAP	Coordinates won't explore texture features on the face
Jiang et al. [41]	2011	DBN and HMM	1. 18 facial features from 7 facial action units 2. 42-MFCC	Surprise, Sad, Happy, Fear, Disgust, and Angry	eNTERFACE'05	Less number of units are segmented in face image
Lu et al. [93]	2012	HMM and BCHMM	1. 10 geometric distance features and 2. 12-MFCC, Energy, pitch, and F0	Activation and Valence	Own-Created	Geometric features are sensitive to illuminations
Rosas et al. [74]		SVM with linear kernel	1. Gaze at the camera, smile duration 2. loudness, intensity, pitch, pause duration	Negative and Positive	Spanish Multimodal Opinion	Focused only on one emotion
Wu et al. [63]	2013	HMM and 2HSCHMM	1. 30 FAPs 2. 1st five formants, Energy and pitch	Neutral, Anger, Sad, Happy	SEMAINE and MHMC	SCHMM introduces uncertainty in the fusion process
Idris et al. [31]	2014	MLP-NN	1. Prosodic Features 2. Voice quality features 3. Hybrid features	Disgust and Angry	Berlin emotional database	Slow convergence of MLP-NN
Dhall et al. [80]	2015	GEM	1. Local phase quantization 2. Histogram of gradient	Happy intensity levels	Own created	HoGs are susceptible to external noises
M. Goyani and N Patel [104]	2017	LMBP and HNAD	1. Local mean of a bock in facial image 2. Texture features	Neutral, Sad, Surprise, Disgust, Angry, Fear, Contempt	C.K., and JAFFE	Information loss in each neighborhood
Guo et al. [105]	2017	ELBP, KLT, and SVM	Covariance features	Neutral, Sad, Surprise, Disgust, Angry, Fear, Contempt	C.K., and JAFFE	KLT has a more considerable computational complexity
M. A. Muqeet et al. [106]	2017	LBP, DIWT	Wavelet ad texture features	Neutral, Sad, Surprise, Disgust, Angry, Fear, Contempt	C.K., and JAFFE	Not effective in Scale and rotational features
I. M. Raveen et al. [107]	2018	DGLTP, SVM	Texture features and Histogram features	Neutral, Sad, Surprise, Disgust, Angry, and Fear,	JAFFE	Directional information of an emotion is not calculated
Bhupendra Singh et al. [109]	2019	LBP and SVM	Facial texture features	Sad, surprised, disgusted, feared, angry, happy	JAFFE	Information loss is high due to LBP
Yan, J et al. [111]	2019	CNN and BRNN	the audio signal, facial landmark, and facial texture	Basic emotions	Standard	Landmark-based LBP constitutes a dimensionality issue in fusion
Seng, K.P et al. [112]	2019	CNN and SVM	BDPCA and LSLDA	Basic emotions	Standard	PCA is not robust for nonlinear combinations

Ortega, Juan D.S et al. [113]	2019	pre-trained CNN and transfer learning	as prosodic, spectral, vocal tract, and excitation features	valence and arousal	RECOLA	Less effective for new and different types of signals like whispered speech
Cai et al. [114]	2019	bi-directional LSTM and an ensemble of CNNs	LBP-TOP	Neutral, surprised, sad, happy, fearful, disgusted, and angry	AmotioW2018 and AFEW	Loss of directionality information
Hossain, M et al. [115]	2019	CNN, ELM, and SVM	MFCC and Segmented frames	Basic emotions	Standard	Three ML algorithms make the system very slow.
Kamal A. El Dahshan [110]	2020	DBN and QPSO	Pixel intensities	Sad, surprised, disgusted, fearful, angry, happy	JAFFE	Only pixel intensities are not enough to provide discrimination between emotions

IV. CONCLUSION AND DISUSSION

In this paper, we reviewed different methods mainly aimed at emotion recognition. Since the emotions of a human being are related to several applications, it has gained a considerable research interest. So many researchers have put effort and developed many methods with different methodologies. In general, emotions can be expressed through speech and facial expressions; most research happened by considering those two models as input data. An emotion identification system's typical method consists of two phases: feature extraction and categorization. Feature extraction attempts to extract features from input data that can reveal emotion-related characteristics. These characteristics aid the system in distinguishing between emotions. However, different models have distinct features. Prosodic and Cepstral features are the most basic emotional indicators that researchers consider for speech signals. The top features analysed by researchers for facial photographs, on the other hand, are texture, look, geometric aspects, and so on. We reviewed almost all the features for both speech and facial images and found that the texture features are more effective in the case of facial images. In contrast, prosodic features had shown maximum performance for speech signals. Next, the classifier design also has a significant role in creating the emotion recognition system. In general, Machine learning algorithms are employed for classification and are different for different models. From the review, we found that the popular and effective classifier for speech signals is HMM while for facial images, it is SVM. Even though extensive research has been carried out on emotion recognition, a gap exists. One possible direction is to focus on multiple models at a time for emotion recognition. However, we found some issues while designing multi-modal-based emotion recognition. They are.

1. In the pre-processing phase, the data model is subjected to an initial screening process, i.e., noise removal, quality enhancement, size compatibility settings, etc. However, most of the authors won't focus much on this direction because they employed mostly standard datasets for validation which does not require additional pre-processing. However, in real-time, the input data is noisy and has many artifacts. Especially in facial images, the high-frequency components can be treated as noise components.

2. In the feature extraction phase, the feature extractors are different for different data models and are incompatible. A feature extraction method developed for speech signals won't support facial images and vice versa. Even though some procedures can provide compatibility for both data models, different settings must be done at their inner properties. 3. Next, the critical aspect of feature extraction is the design of a descriptor that can provide perfect discrimination between different expressions. The data descriptor's structure must be so that it can describe even the hidden properties of data. Then only the expressions recognition system recognizes faces accurately. Even though most researchers concentrated on this direction, every method has pros and cons. 4. In the last phase, classifier selection is essential. At this phase, the classifier needs to be chosen or develop a new classifier that is compatible with different data models. Earlier approaches focused only on a few classifiers like support vector machines. However, the SVM can't support all times and has limited performance in multiple data models.

References

[1] Sebastian Loth and Jan P. DeRuiter, "Editorial: Understanding Social Signals: How Do We Recognize the Intentions of Others?" Front. Psychol. 7:281, 2016.

[2] Lukasz Piwek, Frank Pollick and Karin Petrini, "Audiovisual integration of emotional signals from others' social interactions," Front. Psychol. 6:611, 2015.

[3] Enrique M. Albornoz, Diego H. Milone, Hugo L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," Comput. Speech Lang. 25 (2011) 556–570.

[4] Cowie, R. *et al.*: "Emotion recognition in human-computer interaction," IEEE Signal Process. Mag., 18 (2001), 33–80.

[5] Fragopanagos, N.; Taylor, J.G.: "Emotion recognition in human-computer interaction," Neural Netw., 18 (2005), 389–405.

[6] Ayadi, M.E.; Kamel, M.S.; Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit., 44 (2011), 572–587.

[7] Schuller, B.; Batliner, A.; Steidl, S.; Seppi, D.: Recognising realistic emotions and affect in speech: state of the art and lessons learned from the first challenge. Speech Commun. 53 (2011), 1062–1087.

[8] Sumathi, C.P.; Santhanam, T.; Mahadevi, M.: Automatic facial expression analysis a survey. Int. J. Computer. Sci. and Eng. Survey. (IJCSES), 3 (2013), 47–59.

[9] Pantic, M.; Bartlett, M.: Machine Analysis of Facial Expressions. Face Recognition. *I-Tech Education and Publishing*, Vienna, Austria, 2007, 377–416.

[10] Grimm, M.; Kroschel, K.; Narayanan, S.: The Vera am Mittag German audiovisual emotional speech database, in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2008, 865–868.

[11] Wang, Y.; Guan, L.: Recognizing human emotional state from audiovisual signals. IEEE Transactions on Multimedia, 10 (2008), 936–946.

[12] Ringeval, F.; Sonderegger, A.; Sauer, J.; Lalanne, D.: Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. 2nd Int. Workshop on Emotion Representation, Analysis, and Synthesis in Continuous Time and Space (EmoSPACE), in *Proc. IEEE Face & Gestures*, 2013, 1–8.
[13] Valstar, M. *et al.*: AVEC 2013 – The Continuous Audio/Visual Emotion and Depression Recognition Challenge, *ACM Multimedia*, 2013.
[14] Petridis, S.; Martinez, B.; Pantic, M.: The MAHNOB laughter database. Image Vis. Computer., 31 (2013), 186–202.
[15] Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T.: Collecting large, richly annotated facial-expression databases from movies. IEEE Multimedia, 19 (2012), 34–41.
[16] Lin, J.C.; Wu, C.H.; Wei, W.L.: Error weighted semi-coupled hidden Markov model for audiovisual emotion recognition. IEEE Trans. Multimedia, 14 (2012), 142–156.
[17] Mc Keown, G.; Valstar, M.; Pantic, M.; Cowie, R.: The SEMAINE corpus of emotionally coloured character interactions, in *Proc. IEEE Int. Conf. Multimedia & Expo*, 2010, 1–6.
[18] Haq, S.; Jackson, P.J.B.: Speaker-dependent audiovisual emotion recognition, in *Proc. Int. Conf. Auditory-Visual Speech Processing*, 2009, 53–58.
[19] Schuler, B.Muller, R.; Hornler, B.; Hothker, A.; Konosu, H.; Rigoll, G.: Audiovisual recognition of spontaneous interest within conversations, in *Proc. 9th Int. Conf. Multimodal Interfaces (ICMI), Special Session on Multimodal Analysis of Human Spontaneous Behaviour*, ACM SIGCHI, 2007, 30–37.
[20] Martin, O.; Kotsia, I.Macq, B.; Pitas, I. The eNTERFACE'05 audiovisual emotion database, in *Int. Conf. Data Engineering Workshops*, 2006.
[21] IrynaAnina, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis", *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition* (F.G.), 2015.
[22] E.K Patterson, "CUAVE: A new audiovisual database for multimodal human-computer interface research", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002
[23]K. S. Rao, Shashidar K koolagudi, Ramu Reddy Vempada, "Emotion recognition from speech using global and local prosodic features," International Journal of Speech Technology, Volume 16, Issue 2, pp 143–160, June 2013.
[24] Wu, C.H.; Yeh, J.F.; Chuang, Z.J.: Emotion Perception and Recognition from Speech, Affective Information Processing, *Springer*, New York, 2009, 93–110.
[25] Wu, C.H.; Liang, W.B.: Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. IEEE Trans. Affective Computer. 2 (2011), 1–12.
[26] Morrison, D.; Wang, R.; De Silva, L.C.: Ensemble methods for spoken emotion recognition in call-centers. Speech Commun. 49 (2007), 98–112.
[27] Murray, I.R.; Arnott, J.L.: Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J. Acoustic. Soc. Am., 93 (1993), 1097–1108.
[28] Scherer, K.R.: Vocal communication of emotion: a review of research paradigms. Speech Commun. 40 (2003), 227–256.
[29] Luengo, I.; Navas, E.; Hernaez, I.; Sanchez, J.: Automatic emotion recognition using prosodic parameters. *INTERSPEECH*, 2005, 493–496.
[30] Kooladugi, S.G.; Kumar, N.; Rao, K.S.: Speech emotion recognition using segmental level prosodic analysis, in *Int. Conf. Devices and Communications*, 2011, 1–5.
[31] Inshirah Idris, MdSah Hi Salam, "Emotion detection with hybrid voice quality and prosodic features using Neural Network," Fourth World Congress on Information and Communication Technologies (WICT), 2014.
[32] Rode Snehal Sudhakar, Manjar Chandra Prabha Anil, "Analysis of Speech Features for Emotion Detection: A Review," Computing Communication Control and Automation (ICCUBEA), 2015 International Conference on
[33] Agnes Jacob, P Mythili, "Prosodic feature-based speech emotion recognition at segmental and suprasegmental levels," IEEE International Conference on Signal Processing, Informatics, Communication, and Energy Systems (SPICES), 2015
[34] Yu Zhou, Yanqing Sun, Jianping Zhang, Yonghong Yan, "Speech Emotion Recognition Using Both Spectral and Prosodic Features," International Conference on Information Engineering and Computer Science, 2009, ICIECS 2009.
[35] Monorama Swain, AurobindaRoutray, P Kabisatpathy, Jogendra N Kundu, "Study of prosodic feature extraction for multidialectal Odia speech emotion recognition," IEEE Region 10 Conference (TENCON), 2016.
[36] Ayadi, M.E.; Kamel, M.S.; Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognition. 44 (2011), 572–587.
[37] Schuller, B.; Batliner, A.; Steidl, S.; Seppi, D.: Recognizing realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun. 53 (2011), 1062–1087.
[38] Schuller, B.; Steidl, S.; Batliner, A.: The INTERSPEECH 2009 emotion challenge, in *Proc. Interspeech*, 2009, 312–315.
[39] Schuller, B. *et al.*: The INTERSPEECH2010 paralinguistic challenge, in *Proc. INTERSPEECH*, 2010, 2794–2797.
[40] Schuller, B. *et al.*: The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism, in *Proc. Interspeech*, 2013, 148–152.
[41] Jiang, D.; Cui, Y.; Zhang, X.; Fan, P.; Gonzalez, I.; Sahli, H.: Audio visual emotion recognition based on triple-stream dynamic Bayesian network models, in *Proc. Affective Computing and Intelligent Interaction*, 2011, 609–618.
[42] Busso, C. *et al.*: IEMOCAP: interactive emotional dyadic motion capture database. J. Lang. Resources Eval., 42 (2008), 335–359.
[43] Metallinou, A.; Lee, S.; Narayanan, S.: Audiovisual emotion recognition using Gaussian mixture models for face and voice in *Proc. Int. Symp. Multimedia*, 2008, 250–257.
[44] Rudovic, O.; Petridis, S.; Pantic, M.: Bimodal log-linear regression for fusion of audio and visual features, in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, 789–792.
[45] K.V.Krishna Kishore, P Krishna Satish, "Emotion recognition in speech using MFCC and wavelet features," IEEE 3rd International Advance Computing Conference (IACC), 2013.
[46] Qingli Zhang, Ning An, Wang, Fuji Ren, Lian Li, "Speech emotion recognition using a combination of features," Fourth International Conference on Intelligent Control and Information Processing (ICICIP), 2013.
[47] Huang, Y.; Zhang, G.; Li, X.; Da, F.: Improved emotion recognition with novel global utterance-level features. Int. J. Appl. Math. Inf. Sci., 5 (2011), 147–153.
[48] Schuller, B.; Steidl, S.; Batliner, A.; Schiel, F.; Krajewski, J.: The INTERSPEECH 2011speaker state challenge, in *Proc. Interspeech*, 2011, 3201–3204.
[49] Ayadi, M.E.; Kamel, M.S.; Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognition. 44 (2011), 572–587.
[50] Song, M.; You, M.; Li, N.; Chen, C.: A robust multimodal approach for emotion recognition, Neuro-computing, 71 (2008), 1913–1920.
[51] Sumathi, C.P.; Santhanam, T.; Mahadevi, M.: Automatic facial expression analysis a survey. Int. J. Computer. Sci. and Eng. Survey. (IJCSES), 3 (2013), 47–59.
[52] Pantic, M.; Bartlett, M.: Machine Analysis of Facial Expressions. Face Recognition. *I-Tech Education and Publishing*, Vienna, Austria, 2007, 377–416.
[53] Saatci Y., Town C., Cascaded Classification of Gender and Facial expression using Active Appearance Models, The 7th Conference on Automatic Face and Gesture Recognition FGR'06, 2006.
[54] Wilhelm, T., Bohne, H.-J., Gross, H.-M., Classification of Face Images for Gender, Age, Facial Expression, and Identity, ICANN 2005, LNCS 3696, 569-574, 2005.
[55] Zhou Z.-H. and Geng X., Projection Functions for Eye Detection, State Key Laboratory for Novel Software Technology, NU, China, 2002.
[56] Lee C.S. and Elgammal A, Nonlinear Shape and Appearance Models for Facial Expression Analysis and Synthesis, Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06) - Volume 01, 497 – 502.
[57] Liebelt J., Xiao J., Yang J., Robust AAM Fitting by Fusion of Images and Disparity Data, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, June 17-22, Vol. 2, 2483-2490, 2006.
[58] Antonini, G., Sorci, M., Bierlaire, M., Thiran, J.-P., Discrete Choice Models for Static Facial Expression Recognition, ACIVS'06, 710-721, 2006.
[59] Matthew S Ratliff, E. Patterson, "Emotion recognition using facial expressions with active appearance models," Proceeding HCI '08 Proceedings of the Third IASTED International Conference on Human-Computer Interaction Pages 138-143, Innsbruck, Austria — March 17 - 19, 2008.
[60] Wenchao Zheng, Cuicui Liu, "Facial expression recognition based on texture and shape," Wireless and Optical Communication Conference (WOCC), 2016 25th
[61] [93] Cootes, T.F.; Edwards, G.J.; Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell., 23 (2001), 681–685.
[62] Lin, J.C.; Wu, C.H.; Wei, W.L.: Error weighted semi-coupled hidden Markov model for audiovisual emotion recognition. IEEE Trans. Multimedia, 14 (2012), 142–156.
[63] Wu, C.H.; Lin, J.C.; Wei, W.L.: Two-level hierarchical alignment for semi-coupled HMM-based audiovisual emotion recognition with the temporal course. *IEEE Trans. Multimedia*, 15 (2013), 1880–1895.
[64] Hans van Kuilenburg, Marco Wiering, and Marten den Uy, "A Model-Based Method for Automatic Facial Expression Recognition," European Conference on Machine Learning, 2005.

[65] Yong-Hwan Lee, "Virtual representation of facial avatar through weighted emotional recognition," IJIPT, Vol. 10, issue 1, 2017.
[66] Kwang-EunKo, Kwee-Bo Sim, "Emotion recognition in facial image sequences using a combination of AAM with FACs and DBN," International Conference on Intelligent Robotics and Applications, 2010.
[67] Whitehill J, Omlin C (2006) Haar features for FACS AU recognition. In: Proceedings of the IEEE international conference on face and gesture recognition.
[68] Schuller, B.; Valstar, M.; Eyben, F.; McKeown, G.; Cowie, R.; Pantic, M.: AVEC 2011 the first international audio/visual emotion challenge, in *Proc. First Int. Audio/Visual Emotion Challenge and Workshop (ACII)*, 2011, 415–424.
[69] Schuller, B.; Valstar, M.; Eyben, F.; Cowie, R.; Pantic, M.: AVEC 2012 – the continuous audio/visual emotion challenge, in *Proc. of Int. Audio/Visual Emotion Challenge and Workshop (AVEC)*, ACM ICMI, 2012.
[70] Valstar, M. *et al.*: AVEC 2013 – The Continuous Audio/Visual Emotion and Depression Recognition Challenge, *ACM Multimedia*, 2013.
[71] Valstar, M.*et al.*: AVEC 2014 – 3D dimensional affect and depression recognition challenge, in Proc. AVEC 2014, held in conjunction with the 22nd ACM Int. Conf. Multimedia (MM 2014), 2014.
[72] Shan, C.; Gong, S.; Mcowan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Computer., 27 (2009), 803–816.
[73] Ahonen, T.; Hadid, A.; Pietikainen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell., 28 (2006), 2037–2041.
[74] Rosas, V.P.; Mihalcea, R.; Morency, L.-P.: Multimodal sentiment analysis of Spanish online videos. IEEE Intell. Syst., 28 (2013), 38–45.
[75] Ramirez, G.A.; Baltrušaitis, T.; Morency, L.P.: Modeling latent discriminative dynamic of multi-dimensional affective signals, in *Proc. Affective Computing and Intelligent Interaction*, 2011, 396–406.
[76] AnushaVuppaturi, SukadevMeher, "Facial Expression Recognition using Local Binary Patterns and KullbackLeibler divergence," *International Conference on Communications and Signal Processing (ICCSP),* 2015.
[77] KrystianRadlak, Bogdan Smolka, "High dimensional local binary patterns for facial expression recognition in the wild," *18th Mediterranean Electrotechnical Conference* (MELECON), 2016.
[78] CaifengShan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *Image and Vision Computing,* Volume 27, Issue 6, 4 May 2009, Pages 803-816.
[79] Li HD, Xu QS, Liang YZ, "Random frog: an efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification," *Analytica Chimica Act*a, Volume 740, 31 August 2012, Pages 20-26
[80] A Dhall, R Goecke, T Gedeon, "Automatic Group Happiness Intensity Analysis," *IEEE transactions on Affective Computing,* 2015.
[81] Metallinou, A.; Wollmer, M.; Katsamanis, A.; Eyben, F.; Schuller, B.; Narayanan, S.: Context-sensitive learning for enhanced audiovisual emotion classification*. IEEE Trans. Affective Computer*., 3 (2012), 184–198.
[82] Sayedelahl, A.; Araujo, P.; Kamel, M.S.: Audiovisual feature decision level fusion for spontaneous emotion estimation in speech conversations*. Conf. Multimedia and Expo Workshops*, 2013, 1–6.
[83] Rudovic, O.; Petridis, S.; Pantic, M.: Bimodal log-linear regression for fusion of audio and visual features, in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, 789–792.
[84] TriasThireou, Martin Reczko, "Bidirectional Long Short-Term Memory Networks for Predicting the Subcellular Localization of Eukaryotic Proteins," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Volume: 4, Issue: 3, July-Sept. 2007.
[85] Maxim Sidorov, EvgenjiSopov, lliaLvanov, Wolfganag Minker, "Feature and decision level audiovisual data fusion in emotion recognition problem," *12th International Conference on Informatics in Control, Automation and Robotics (ICINCO),* 2015.
[86] Gaffary, Y., Eyharabide, V., Martin, J. C., & Ammi, M. (2014). The impact of combining kinesthetic and facial expression displays on user emotion recognition. *International Journal of Human-Computer Interaction*, *30*(11), 904–920
[87] Biswas, P., & Langdon, P. (2015). Multimodal intelligent eye-gaze tracking system. *International Journal of Human-Computer Interaction*, *31*(4), 277–294,
[88] Bosch, N., Chen, H., D'Mello, S., Baker, R., & Shute, V. (2015). Accuracy vs. availability heuristic in multimodal affects detection in the wild. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)* (pp. 267–274).
[89] Wollmer, M.; Kaiser, M.; Eyben, F.; Schuller, B.; Rigoll, G.: LSTM modeling of continuous emotions in an audiovisual affect recognition framework, in Image and Vision Computing (IMAVIS). Spec. Issue Affect Anal. Constant Input, 31 (2013), 153–163.
[90] Picard, R.W.: Affective Computing. *MIT Press*, 1997.
[91] Paleari, M.; Benmokhtar, R.; Huet, B.: Evidence theory-based multimodal emotion recognition, in *Proc. 15th Int. Multimedia Modelling Conf. Advances in Multimedia Modelling*, 2009, 435–446.
[92] Zeng, Z.; Tu, J.; Pianfetti, B.M.; Huang, T.S.: Audiovisual affective expression recognition through multi-stream fused HMM. IEEE Trans. Multimedia, 10 (2008), 570–577.
[93] Lu, K.; Jia, Y.: Audiovisual emotion recognition with boosted coupled HMM, in *Int'lConf. Pattern Recognition (ICPR)*, 2012, 1148–1151.
[94] In Int, Nicolaou, M.; Gunes, H.; Pantic, M.: Audiovisual classification and fusion of spontaneous affective data in likelihood space*. Conf. Pattern Recognition (ICPR)*, 2010, 3695–3699.
[95] Z. Yu and C. Zhang, "Image-based static facial expression recognition with multiple deep network learning," in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI '15, (New York, NY, USA), pp. 435–442, ACM, 2015.
[96] B. Kim, J. Roh, S. Dong, and S. Lee, "Hierarchical committee of deep convolution neural networks for robust facial expression recognition," Journal on Multimodal User Interfaces, pp. 1–17, 2016.
[97] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in Proc. ACM International Conference on Multimodal Interaction (ICMI), November 2015.
[98] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," CoRR, vol. abs/1411.7923, 2014.
[99] S. Ouellet, "Real-time emotion recognition for gaming using deep convolutional network features," CoRR, vol. abs/1408.3750, 2014.
[100] M.S. SInith, E. Aswathi, T.M. Deepa, "Emotion recognition from audio signals using Support Vector Machine," IEEE Recent Advances in Intelligent Computational Systems (RAICS), 2015.
[101] SHilpi Gupta, Anu Mehra, Vinay, "Speech emotion recognition using SVM with thresholding fusion," 2nd International Conference on Signal Processing and Integrated Networks (SPIN), 2015.
[102] Ma Xiaoxi, Lin Weisi, Huang Dongyan, Dong Minghui, Haizhou Li, "Facial emotion recognition," IEEE 2nd International Conference on Signal and Image Processing (ICSIP), 2017.
[103] Metallinou, A.; Lee, S.; Narayanan, S.: Decision level combination of multiple modalities for recognition and analysis of emotional expression, in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 2010, 2462–2465.
[104] M. M. Goyani, and N. Patel, "Recognition of facial expressions using local mean binary pattern", *Electronic Letters on Computer Vision and Image Analysis,* Vol. 16, No. 1, pp. 54–67, 2017.
[105] M. Guo, X. Hou, Y. Ma, "Facial expression recognition using ELBP based on covariance matrix transform in KLT", *Multimedia Tools and Applications,* Vol. 76, pp. 2995–3010, 2017.
[106] Mohd. Abdul Muqeet, Raghunath S. Holambe, "Local binary patterns based on directional wavelet transform for expression and pose invariant face recognition," *Applied Computations And Informatics*, Vol. 15, No.2, pp. 163-171, 2017.
[107] I. M. Revina, and W.R. Sam Emmanuel, "Face expression recognition using LDN and Dominant Gradient Local Ternary Pattern descriptors," *Journal of King Saud University –Computer and Information Sciences*, Vol. xx, pp.1-7, 2018.
[108] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," In *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Workshops*, San Francisco, CA, USA, pp. 94–101, 2010.
[109] B. Singh, R. K. Sharma, R. K.Saxena, Ms. Ragini Malviya, "Facial Expressions Recognition Based Using LBP and SVM Classifier," *International Journal of Innovative Research in Science, Engineering, and Technology*, Vol. 8, Issue 8, pp.8508-8517, 2019.
[110] Kamal A. El Dahshan, Eman K. Elsayed, Ashraf Aboshoha, Ebeid A. Ebeid, "Recognition of Facial Emotions Relying on Deep Belief Networks and Quantum Particle Swarm Optimization," *International Journal of Intelligent Engineering and Systems*, Vol.13, No.4, pp.90-101, 2020.
[111] Yan, J.; Zheng, W.; Cui, Z.; Tang, C.; Zhang, T.; Zong, Y., Multi-cue fusion for emotion recognition in the wild. *Neurocomputing,* 309, 27–35, 2018.
[112] Seng, K.P.; Ang, L.M.; Ooi, C.S., "A Combined Rule-Based & Machine Learning Audio-Visual Emotion Recognition Approach," *IEEE Trans. Affect. Comput.*, 9, 3–13, 2018.
[113] Ortega, Juan D.S., Cardinal, Patrick, Koerich, Alessandro L., "Emotion recognition using a fusion of audio and video features," *In CVPR.* arXiv:1906. 10623v1, 2019.
[114] Cai, Jie, Meng, Zibo, Khan, Ahmed Shehab, Li, Zhiyuan, O'Reilly, James, Han, Shizhong, Liu, Ping, Chen, Min, Tong, Yan, "Feature-level and model-level audiovisual fusion for emotion recognition in the wild," *In CVPR.* arXiv:1906.02728v1, 2019.

[115] Hossain, M. Shamim, Muhammad, Ghulam, "Emotion recognition using deep learning approach from audio–visual, emotional big data," *Inf. Fusion*, 49, 2019.

KummariRamyasree is a Research Scholar at Department of Electrical, Electronics and Communications Engineering, GITAM Institute of Technology, GITAM (Deemed to be University), Visakhapatnam and working as an Assistant Professor in the Department of Electronics and Communication Engineering at Guru Nanak Institute of Technical Campus, Telangana, India since 2013. She is awarded B.Tech in ECE from JNTUH and M.Tech in JNTUH. She has 10 years of academic teaching experience and has presented and published several papers at International Conferences and International Journals. Her areas of Research Interests are in Image Processing, Signal Processing,Artificial Intelligence and Machine Learning, Deep Learning.

Dr. Ch. Sumanth Kumar is a professor at Department of Electrical, Electronics and Communication Engineering, GITAM School of Technology, GITAM (Deemed to be University), Visakhapatnam, India. He awarded B.Tech. from Nagarjuna University in 1995, M.E from M.S University of Baroda in 1997 and Ph.D. from Andhra University in 2013. He has 25 years of teaching experience, and published several research papers in international and national journals. His areas of Research Interests are in VLSI Signal Processing, Image processing and Communications.