

Acoustic and visual geometry descriptor for multi-modal emotion recognition from videos

Kummari Ramyasree, Chennupati Sumanth Kumar

Department of Electrical, Electronics and Communication Engineering, GITAM Deemed to be University, Visakhapatnam, India

Article Info

Article history:

Received Jul 8, 2023

Revised Dec 7, 2023

Accepted Dec 25, 2023

Keywords:

Acoustic feature

Audio and video

Decision level fusion

Geometric features

Key frames

Multimodal emotion recognition

ABSTRACT

Recognizing human emotions simultaneously from multiple data modalities (e.g., face, and speech) has drawn significant research interest, and numerous research contributions have been investigated in the affective computing community. However, most methods concentrate less on facial alignment and keyframe selection for audio-visual input. Hence, this paper proposed a new audio-visual descriptor, mainly concentrating on describing the emotion through only a few frames. For this purpose, we proposed a new self-similarity distance matrix (SSDM), which computes the spatial, and temporal distances through landmark points on the facial image. The audio signal is described through an asset of composite features, including statistical features, spectral features, formant frequencies, and energies. A support vector machine (SVM) algorithm is employed to classify both models, and the final results are fused to predict the emotion. Surrey audio-visual expressed emotion (SAVEE) and Ryerson multimedia research lab (RML) datasets are utilized for experimental validation, and the proposed method has shown significant improvement from the state of art methods.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Chennupati Sumanth Kumar

Department of Electrical, Electronics and Communication Engineering, GITAM Deemed to be University

Visakhapatnam-530045, AP, India

Email: schennup@gitam.edu

1. INTRODUCTION

The emotions reveal a person's thoughts. Emotions can aid self-regulate jobs, including car safety, marketing, therapies, autonomous health feedback, assessments, games, and monitoring. Human-computer interaction (HCI) [1] and affective computing help automate tasks from humans to machines. HCI emphasizes human-machine interaction. Knowing human emotions makes the machine more effective. A system that makes decisions based on human mood would be great. Because congested roads make drivers angry, the car may take a quieter, longer route.

Different methods can recognize emotions. Speech, faces, eye gaze, gestures, and physiological signals like electroencephalograms convey emotion. Each data model has pros and cons. For instance, illumination, viewpoint, scale, and orientation concerns affect face image emotion recognition. Ambient noise and speaker voices also affect speech emotion recognition. Multi-modal analysis improves emotion recognition. Humans emphasize speech with body, head, arm, and facial expressions [2]. The fundamental reason is that 93% of human communication is non-verbal, including body positions, facial emotions, and eye gazes. Multiple-modality emotion recognition systems have been developed. From facial images, P. Ekman introduced action units to identify the six basic emotions (disgust, fear, anger, happiness, sadness, and surprise). Most methods used face as a main model and speech, gestures, as cooperative models.

Automatic facial emotion detection requires face identification, tracking, feature extraction, and recognition. First, identify and track the face in a multi-frame video sequence. Piecewise bezier volume deformation tracker, resilient faces identification algorithm, AdaBoost learning algorithm, improved Kanade-Lucas-Tomasi tracker method, and ratio template tracker. The following step extracts features from the detected faces that aid in recognizing expressions. The feature extraction methods are broadly classified as geometry and appearance-based approaches [3]. The first category uses face shapes and landmark points to convey expressions. Appearance-based approaches portray face expressions using classic computer vision algorithms.

Problem statement: most methods have focused on the static face. But, currently, expression recognition from videos has gained significant interest. Unlike a static face image, the video shows facial muscles and voice. The main challenge is compactly representing images and speech to characterize video emotions accurately. To solve this issue, acoustic and visual geometry descriptors for multi-modal emotion recognition from videos. This paper proposes a face-speech-based multi-modal emotion recognition framework is proposed. The contribution of the proposed work is specified below.

This framework first finds landmark points on non-frontal faces in a video to align them. Self-similarity distance matrix (SSDM) computes spatial and temporal Euclidean distances to identify key frames. A visual geometric descriptor represents an emotion over keyframes. We represent emotions in audio data using 68 features from different categories. Support vector machine (SVM) classified both models. Combining classifier outputs yields the final prediction. The structure of this article is specified below. The second portion describes the video data emotion recognition framework. The fourth portion shows experimental analysis details, and 5th section presents the conclusion of the paper.

Researchers have improved multi-modal expression recognition based on audio and visual input modalities. Multi-modal fusion is divided into decision level, data model/feature level, kernel level, score level, and hybrid levels [4]. Recent and other multi-modal expression recognition algorithms are included in this survey. Video and text investigation of speech emotion recognition. Audio samples are transformed into spectrograms and then fed to modified AlexNet for feature extraction. They used bidirectional encoder representations from transformers (BERT) embedding for text data and long term short memory (LSTM) for categorization. Surrey audio-visual expressed emotion (SAVEE) and ryerson audio-visual database of emotional speech and song (RAVDESS) datasets are utilized for simulation experiments. A model-level fusion-based multi-modal emotion recognition framework combining video and audio data [5]. After modeling audio data with mel frequency cepstrum coefficients (MFCCs) and video with spatiotemporal features, a deep learning feature and extractor are created for both data models. Experimental validation uses SAVEE and RAVDESS. Emotion recognition, and Javaid combined voice and infrared images [6]. In the first stage, two convolutional neural networks (CNNs) are trained with infrared (IR) and visible images. Transfer learning was utilized to fuse feature vectors supplied to SVM for classification. To train the artificial neural network (ANN), they employed another CNN model to understand speech emotions from audio spectrograms. The final class label is predicted by combining SVM and third CNN outputs. Experimental validation used RAVDESS. A multi-modal neural architecture that uses audio and video input for bidirectional long short term memory (BiLSTM) networks and two CNNs [7]. The MFCCs, energy, and spatio-temporal audio and video aspects are combined and sent to Bi-LSTM for outputs. For the final categories, they used a softmax classifier. Ryerson RML, SAVEE, and RAVDESS validate multi-modal emotion recognition [8]. They used the powerful attention mechanism to represent the audio-video sequence. They suggested an emotion-labeling audio-visual time windows architecture. Attention computes weights, and fusion predicts the result.

Three data models to discern emotion: audio, text, and video. MFCCs, standard deviation, interquartile ranges, quartile ranges, arithmetic mean, quadratic root mean, amplitude mean, pitch, and voice intensity were retrieved from audio using CNN with different layers [9]. After testing on the interactive emotional dyadic motion capture (IEMOCAP) and CMU multi-modal opinion sentiment and emotion intensity (CMU-MOSI) datasets, decision and feature level fusions are used. A multi-modal expression recognition system: multi-level factorized bilinear pooling (MLFBP) [10]. They initially used a 1-dimensional fully convolutional network (FCN) for the audio stream. Audio-visual information is fused in next global FBP. To calculate fusion weights for different modalities, they created an adaptive mechanism for FBP. Experimental validation using IEMOCAP.

An informed segmentation and labeling approach (ISLA) for multi-modal emotion recognition using voice signals and facial regions [11]. The pitch helps predict upper and lower-face emotions. Experimental validation uses IEMOCAP and SAVEE datasets. A histogram of oriented gradient son three orthogonal planes (HOG-TOP) [12] to represent emotion from video data. They used a composite feature set of 38 low-level and 21 functional descriptors for audio data. Feature level fusion was provided to the SVM algorithm for emotion classification [13]. The purpose of the Moodle platform and Zoom video conference mechanism is to influence Zoom and Moodle on learning skills [14]. This mechanism presents the scope of the cross-

section's correlational, valued, non-observational approach. Additionally, reliability is executed with Cronbach's Alpha, receiving the value of 0.875. The emotion detection scheme discovered via body gestures uses PoseNet to generate emotional data for every student. The recognition results are treated and exhibited on an information system as a website [15]. Using images from the internet will evaluate the models to discover the best model for recognizing human emotions and detecting faces [16]. The high-level motion feature frames are forwarded to the pre-trained CNN to distinguish the 17 emotions in the Geneva multimodal emotion portrayals dataset [17].

2. PROPOSED FRAMEWORK

2.1. Overview

Here in this section, we explore the complete details of the proposed multimodal emotion recognition framework. The proposed framework considers two data models, video and audio, and recognizes the emotion. This system benefits from the fusion of different classifier outputs, each focusing on an individual data model. The overall schematic of the proposed framework is depicted in Figure 1.

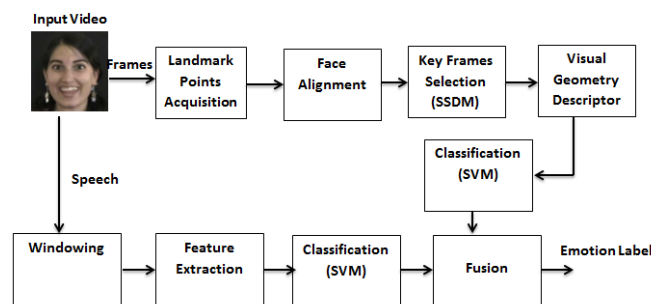


Figure 1. The overall architecture of the proposed framework

In this framework, we adopted decision-level fusion, which was performed after getting the classification results from individual classifiers. The confidence scores of individual classifiers are used to accompany the decision-level fusion. Next, we propose efficient features for each model through which the recognition system can discriminate between emotions. For video input, the proposed system initially extracts the frames and then extracts landmark positions for the face in each frame. Based on the obtained landmark positions, the faces are aligned in each frame and then subjected to keyframe selection. Then, the obtained keyframes are described through a visual geometry descriptor based on the landmark positions of the face in each frame. Here, we propose a new SSDM metric to identify the keyframes. This matrix is inspired by the self-similarity matrix (SSM), which was generally employed to find the similarity between pixels at the same positions in different frames. As another model, we considered speech; each speech sample is represented with a set of 88-dimensional feature vectors. We used common classifiers, and the obtained confidence scores obtained were used to get the final expression label.

2.2. Visual features

Several frames need to be processed to recognize an emotion from a video. Moreover, in most cases, similar facial expressions appear within the same video. Hence, we intended to represent an emotion with only a few sets of keyframes. If a video is processed entirely by the recognition system, the system misclassifies the emotions due to the similarity between the semantics of the faces in different videos. Generally, in any video, the starting frames have neutral faces, the emotion lies in only a few frames, and they generally lie in the center of the video. Hence, key frame selection is necessary to reduce misclassification and processing complexity.

2.2.1. Landmarks

Geometry features can discover 68 landmark points on the face image, as shown in Figure 2, to identify crucial frames in a film. FERA 2015 code [18] detects and tracks video facial features. Acquiring landmark locations in a video is difficult, which might lead to head pose issues. We considered just recorded video frames. Each frame of a video has 68 landmark points. The first 17 landmark points are the face contour, 18 to 22 are the left eyebrow, 23 to 27 are the right eyebrow, 37 to 42 are the left eye, and 43 to 48

are the right eye. Landmark points 28 to 30 symbolize the nose, and 49 to 68 indicate the mouth. 68 landmark points divide the face into six areas [19].

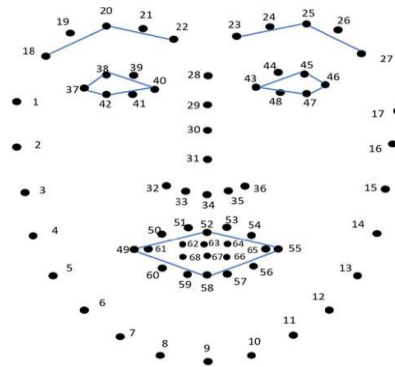


Figure 2. Landmark points on the face image

2.2.2. Face alignment

Face alignment has a significant role in the recognition of expressions from facial images. For a non-frontal view, encoding the strongest features of expression is impossible. Hence, the non-frontal view facial images in the video need to be aligned. To align the non-frontal views, we consider the landmark points assessed above. We use only three landmark points for the facial alignment: the left eye inner corner (LEIC), right eye inner corner (REIC), and nasal spine point (NSP) above the mouth. The landmark points used are 40 LEIC, 43 REIC, and 34 NSP. Since these have less impact on recognition and are stable, we considered them landmark points for alignment. Each landmark point is represented by two coordinates (x, y). Consider (LE_x, LE_y), (RE_x, RE_y) REy, and (NS_x; NS_y) are the three representations of landmark points such as LEIC, REIC, and NSP based on these three Landmark points. Initially, we construct a rotation matrix R of size 3×2 as follows (1) to (4);

$$R = \begin{bmatrix} R_{1x} & R_{1y} \\ R_{2x} & R_{2y} \\ R_{3x} & R_{3y} \end{bmatrix} \tag{1}$$

$$R_{1x} = \frac{RE_x - LE_x}{\sqrt{(RE_x - LE_x)^2 + (RE_y - LE_y)^2}} \& R_{1y} = \frac{RE_y - LE_y}{\sqrt{(RE_x - LE_x)^2 + (RE_y - LE_y)^2}} \tag{2}$$

$$R_{2x} = \frac{NS_x - LE_x}{\sqrt{(NS_x - LE_x)^2 + (NS_y - LE_y)^2}} \& R_{2y} = \frac{NS_y - LE_y}{\sqrt{(NS_x - LE_x)^2 + (NS_y - LE_y)^2}} \tag{3}$$

$$R_{3x} = \frac{RE_x - NS_x}{\sqrt{(RE_x - NS_x)^2 + (RE_y - NS_y)^2}} \& R_{3y} = \frac{RE_y - NS_y}{\sqrt{(RE_x - NS_x)^2 + (RE_y - NS_y)^2}} \tag{4}$$

Based on the values in the rotation matrix, the new coordinates of three landmark points are calculated using (5) to (7). (LE'_x, LE'_y), (RE'_x, RE'_y) and (NS'_x, NS'_y) represent the new LEIC, REIC, and NSP coordinates in the aligned face.

$$(LE'_x, LE'_y) = (LE_x, LE_y) \times R^T \tag{5}$$

$$(RE'_x, RE'_y) = (RE_x, RE_y) \times R^T \tag{6}$$

$$(NS'_x, NS'_y) = (NS_x, NS_y) \times R^T \tag{7}$$

2.2.3. Key frames selection

Keyframes contain the most critical face expression data. Most frames in long videos are superfluous and should be deleted to portray an expressive video. The video starts with neutral frames and an on-set frame where the expression begins. The peak frame is where the expression peaks and the offset frame

is where it stops. Keyframes are onset, offset, and peak frames. We calculate SSDM using Euclidean distances between landmark locations in the frame to identify crucial frames. First, the landmark points in each frame are compared to others. They are aggregated to calculate frame distance. Hence, we first create a distance matrix;

$$d_{ij} = \begin{bmatrix} d_{11} & d_{11} & \dots & d_{1N} \\ d_{21} & d_{22} & \dots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \dots & d_{NN} \end{bmatrix} \quad (11)$$

where d_{ij} represents the aggregated distance between two frames at i and j instances, the aggregated distance is obtained by accumulating distances between individual landmark points at different frames. Consider (x_t, y_t) , and (x_p, y_p) to be the coordinates of a single landmark point in two different frames located at time instances t and p , respectively; the Euclidean distance between them is calculated as:

$$d_{pt} = \sqrt{(x_p - x_t)^2 + (y_p - y_t)^2} \quad (12)$$

Then, we found the frames with the highest deviation using the distance matrix. After applying the maximum rule to the first row of the distance matrix, we get one value whose index indicates that the corresponding frame has a maximum distance from the first frame. This operation is performed across the distance matrix, yielding N frames. The input video has N frames. The (13) determines it.

$$F_i^j = \max_j(d_{ij}) \quad (13)$$

F_i^j consists of the j^{th} frame with a maximum deviation from the i^{th} frame. The keyframes are now identified from F_i^j by applying the sorting rule. The sorting rule sorts the distances in descending order. Finally, we select the keyframes based on the mean and maximum deviations. Consider μ_F as the mean of deviation of frames in F_i^j and M_F as the maximum deviation; a threshold T is calculated as follows. The threshold are finally selected as keyframes:

$$F_i^j = \max_j(d_{ij}) \quad (14)$$

$$T = M_F - \mu_F \quad (15)$$

2.2.4. Visual descriptor

After completing the keyframe selection, they are described through a geometric descriptor. For this purpose, we compute the Euclidean distances between consecutive landmark points in each frame. Consider two consecutive landmark points l_i , l_i and l_{i+1} . The Euclidean distance between them is calculated as (16). After that, they are normalized by performing division operations through the length as (17).

$$d(l_i, l_{i+1}) = \sqrt{(l_{i+1,x} - l_{i,x})^2 + (l_{i+1,y} - l_{i,y})^2} \quad (16)$$

$$\hat{d}(l_i, l_{i+1}) = \frac{d(l_i, l_{i+1})}{\sum_j d(l_j, l_{j+1})} \quad (17)$$

According to the landmark points shown in Figure 2, j varies from 18 to 26 if the normalization is intended on the eyebrow region. Next, j varies from 37 to 41 or 43 to 46 for the eye region. Next, j varies from 49 to 59, 30 to 35, and 6 to 11 if the normalization is intended on mouth, nose, and chin regions, respectively. In addition, we also compute the angles between two lines connected by two landmark points with a common landmark point. Table 1 explains a landmark point at different regions and triplets. Consider such kind of triplet as $l_i - l_j - l_k$, then the angle is computed as (18).

$$l_j = \arccos\left(\frac{d(l_i, l_j)^2 + d(l_i, l_k)^2 - d(l_j, l_k)^2}{2d(l_i, l_j)d(l_i, l_k)}\right) \quad (18)$$

Many investigations found that the mouth and ocular areas were crucial to every expression. Normalization gives the system scale invariance. Eyebrows, eyes, and lips are shown in the Table 1. Normalized distances and calculated angles form the final descriptor.

Table 1. Landmark points at different regions and triplets

Region	Landmark points	Region	Triplets
Chin	6, 7, 8, 9, 10, 11, 12	Mouth	52-55-58, 53-49-58
Nose	28, 29, 30, 31, 32, 33, 34, 35, 36	Eyes	45-43-47, 43-45-46, 45-46-47, 38-40-42, 37-38-40, 38-37-42
Mouth	49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68	Eyebrows	23-25-27, 18-20-22
Eyebrows	18, 19, 20, 21, 22, 23, 24, 25, 26, 27		
Eyes	37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48		

2.3. Audio features

Audio also conveys video emotion. Pitch frequency, high energy, and speech tempo can describe anger. Most writers used frequency band energies (FBEs), MFCCs, spectral energy distribution, duration, intensity, and pitch to recognize audio-based emotions. These traits represent prosodic patterns to distinguish speakers. These patterns affect intonation, pitch, range, phrasing, speaking tempo, and accentuation. Short voice samples are used to extract spectral characteristics. We analyze 11 features: pitch, intensity, zero crossing rate (ZCR), autocorrelation, standard deviation MFCC, ΔMFCC, FBEs, cepstrum coefficients (CCs), formant frequencies (FFs), and harmonic to noise ratio (HNR). We explain features here.

- Intensity: it computes the peak of a syllable, representing the speech signal's loudness. In general, the peak of a syllable lies at the center, and it is a vowel. For a given input speech sample x , the intensity feature is computed as (19), H_w^n is the hamming window, and it is computed as (20).

$$I_i(x) = \frac{\sum_{n=1}^N (x_{i+n}) \cdot H_w^n}{\sum_{n=1}^N H_w^n} \tag{19}$$

$$H_w^n = 0.54 - 0.46 \cos\left(\frac{2\pi n}{L}\right), 1 \leq n \leq N - 1 \tag{20}$$

- Pitch: it can be determined in frequency or time domains. For the speech signal x the pitch is computed as (21). Where ln denotes the length of x .

$$\zeta(s) = DFT\{\log|DFT(x \cdot H_w^n \cdot ln)|\} \tag{21}$$

- Autocorrelation ($r(\tau)$): $r(\tau)$ is measured concerning delay τ , and it maximizes the inner product of $x(n)$ by its shifted variation $x(n + \tau)$.
- FBEs: FBEs and their derivatives are calculated with the help of the 1st order finite impulse response (FIR) filter. For a given input speech signal $x(n)$, the output $y(n)$ is calculated as (22). Where $a_i = h(i)$ and M denote the filter function order, here, $m = 0, 1, \dots, N - l$ and L denote the pulse length. Then, the FBEs are calculated as;

$$y(n) = \sum_{i=0}^M a_i x(n - 1) + \sum_{j=1}^N b_j y(n - j) \tag{22}$$

$$y(m) = \sum_{\theta=0}^{L-1} h(\theta) x[(m - \theta) \text{mod}(N)] \tag{23}$$

- HNR: itsive feelings can be expressed through HNR. The H for a rough voice of younger speakers is approximately 20 DB, which means only one percent is noise, and the remaining 99% is a periodic signal.
- CCs: these coefficients separate the original signal from the filter signal. Sometimes, the signal needs to be truncated to fetch the spectra details. For instance, the vocal tract can be analyzed through low coefficients calculated by finding the DFT of the log magnitude of the DFT of the signal.
- FFs: these frequencies are used to describe the resonating frequency of the speaker's vocal tract. Consider F_s as the sampling frequency; the FFs are calculated as (24). Where $F(x)$ is the transformed speech signal, $Re_{F(x)}$ is the real part, and $Im_{F(x)}$ is the imaginary part. Here, we considered the 3rd and 4th FFs and computed their mean, minimum, maximum, and standard deviation.

– Δ MFCCs are calculated as the DCT version of the FBEs log.

$$FFs = \frac{F_s}{2\pi} \arctan \frac{Re_{F(x)}}{Im_{F(x)}} \quad (24)$$

3. SIMULATION RESULT

Here, in the current section, we discuss the details of simulation experiments of the proposed framework. For simulation purposes, we used two datasets, namely SAVEE and RML. Initially, we explore the details of datasets and then the results obtained. Finally, we explore the effectiveness of the proposed framework by comparing its performance with several existing methods.

3.1. Datasets

SAVEE: based on British English utterances, the SAVEE database was built. Four male actors aged 27 to 31 in the visual media lab helped create this audio-visual database. This database includes surprise, sadness, neutral, pleasure, fear, disgust, and fury. There are 480 native British English utterance files at 441.1 kHz, and 16 bits exist in this database. Sixty samples per emotion and 120 for neutral. The participant was positioned in front of the monitor to record text cues. Three images and one video accompany each emotion request. Each emotion has three written suggestions to avoid tiredness. This dataset has ten subjects. The current simulation used all samples but just 60 neutral samples to maintain emotion consistency. So, $60 \times 7 = 420$ samples were tested.

RML: this dataset is constructed at the ryerson multimedia lab, which includes 720 audio-visual samples. There are six basic emotions: surprise, sadness, happiness, fear, disgust, and angry. For recording, they used a digital camera and captured it in a bright environment with a plain background. The total subjects used for recording this dataset are eight, and they spoke different languages like Chinese, Italian, Persian, Punjabi, Urdu, Madarin, and English, along with different accents. The samples were recorded at 22,050 MHz through a 16-bit single channel. The sampling rate of videos is kept at 30 frames per second (FPS), and the duration of each video is between three and six seconds.

3.2. Results

Here, we conduct a three-phase simulation by varying the data models. In the first phase, we used only visual geometric descriptors from video files to train the system. Next, the second phase considers only audio features extracted from the audio files. Finally, the third phase simulation considers both data models, and the obtained results are fused to produce the final results. We formulate the confusion matrix at every simulation based on the results obtained. They are explored below. After the simulation of the proposed mechanism on SAVEE and RML datasets, the results are shown as a confusion matrix. Tables 2 to 4 show the confusion matrix of the simulation study through only audio, visual, and fused features, respectively.

Table 2. Confusion matrix of audio descriptor with SVM on SAVEE

	Neutral	Surprise	Sad	Happiness	Fear	Angry	Disgust	Total
Neutral	56	0	2	0	2	0	0	60
Surprise	0	58	0	1	0	0	1	60
Sad	2	0	53	0	3	0	2	60
Happiness	1	0	0	56	0	2	1	60
Fear	2	0	4	0	54	0	0	60
Angry	0	2	0	2	0	55	1	60
Disgust	1	1	1	0	2	0	55	60
Total	62	61	60	59	61	57	60	420

Table 3. Confusion matrix of visual geometric descriptor with SVM on SAVEE

	Neutral	Surprise	Sad	Happiness	Fear	Angry	Disgust	Total
Neutral	60	0	0	0	0	0	0	60
Surprise	0	58	0	1	0	0	1	60
Sad	1	0	55	0	2	0	2	60
Happiness	0	1	0	58	0	1	0	60
Fear	1	0	3	0	56	0	0	60
Angry	0	0	1	0	1	58	0	60
Disgust	1	0	1	0	2	0	56	60
Total	63	59	60	59	61	59	59	420

Table 4. Confusion matrix of audio-visual geometric descriptor with SVM on SAVEE

	Neutral	Surprise	Sad	Happiness	Fear	Angry	Disgust	Total
Neutral	60	0	0	0	0	0	0	60
Surprise	0	60	0	0	0	0	0	60
Sad	1	0	57	0	2	0	0	60
Happiness	0	0	0	60	0	0	0	60
Fear	1	0	2	0	57	0	0	60
Angry	0	0	0	1	0	58	1	60
Disgust	0	0	1	0	1	0	58	60
Total	62	60	60	61	60	58	59	420

Similarly, the results of the RML dataset with audio, visual, and audio-visual features are shown in Tables 5 to 7, respectively. Both case studies show that the maximum recognized samples are observed in the simulation of audio-visual features. Next, the visual features showed better contribution than the audio features.

Table 5. Confusion matrix of audio descriptor with SVM on RML

	Neutral	Surprise	Sad	Happiness	Fear	Angry	Disgust	Total
Neutral	116	1	3	1	1	0	120	116
Surprise	0	106	0	7	3	4	120	0
Sad	3	0	112	0	3	0	120	3
Happiness	1	6	1	108	1	3	120	1
Fear	4	0	4	0	110	2	120	4
Angry	0	4	1	4	1	110	120	0
Disgust	124	117	121	120	119	119	720	124
Total	116	1	3	1	1	0	120	116

Table 6. Confusion matrix of visual geometry descriptor with SVM on RML

	Neutral	Surprise	Sad	Happiness	Fear	Angry	Disgust	Total
Neutral	116	0	2	0	2	0	120	116
Surprise	0	110	1	6	1	2	120	0
Sad	2	0	116	0	0	2	120	2
Happiness	1	5	0	112	1	1	120	1
Fear	2	0	2	0	116	0	120	2
Angry	1	3	0	3	1	112	120	1
Disgust	122	118	121	121	121	117	720	122
Total	116	0	2	0	2	0	120	116

Table 7. Confusion matrix of audio-visual descriptor with SVM on RML

	Neutral	Surprise	Sad	Happiness	Fear	Angry	Disgust	Total
Neutral	120	0	0	0	0	0	120	120
Surprise	0	114	0	3	1	2	120	0
Sad	0	0	120	0	0	0	120	0
Happiness	1	3	0	114	0	2	120	1
Fear	2	0	2	0	116	0	120	2
Angry	0	2	0	2	0	116	120	0
Disgust	123	119	122	119	117	120	720	123
Total	120	0	0	0	0	0	120	120

The audio-visual features have good performance for dull emotions like sad and fear compared with other emotions. For example, the number of true positives (TPs) of sad in Table 2 is 53, while the same expression has 57 TPs in Table 4. Similarly, the TPs of fear emotion in Table 5 are observed as 108, while in Table 7, their count is increased to 114. Due to the consideration of multiple features to describe an emotion, the recognition system gains sufficient knowledge about the characteristics of emotion and recognizes them properly. Figure 3 shows the emotion recognition rates and Figure 4 shows F1-score for different emotions of SAVEE.

Audio-visual descriptors perform best, while audio descriptors perform worse. The visual descriptor recognized certain emotions 100%, but not others. Especially when melancholy and fear have similar muscle movements, it confuses them. Hence, audio-visual descriptors outperform separate descriptors because they improve the recognition rate. Auditory, visual, and audio-visual descriptions are recognized at 92.1429%, 95.4762%, and 97.6190%, respectively. Figure 4 shows the average F1-score of audios, visual, and audio-visual descriptors as 92.1517%, 95.4777%, and 97.6169%.

Figures 5 and 6 show the recognition rate and F1-score of RML dataset emotions at different descriptions. Surprise and happiness have 100% recognition rates. The proposed visual descriptor perfectly distinguishes them from other emotions due to their unique muscle movements.

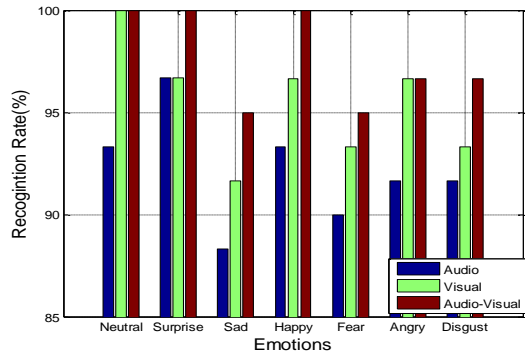


Figure 3. The recognition rate for different emotions of SAVEE

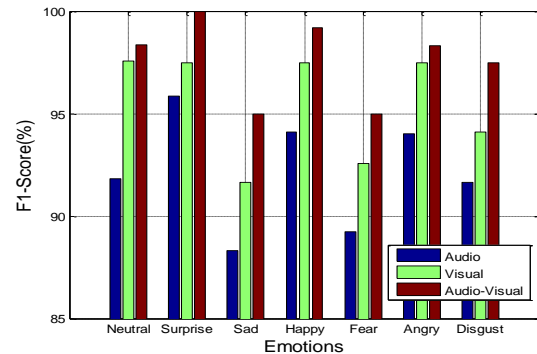


Figure 4. F1-score for different emotions of SAVEE

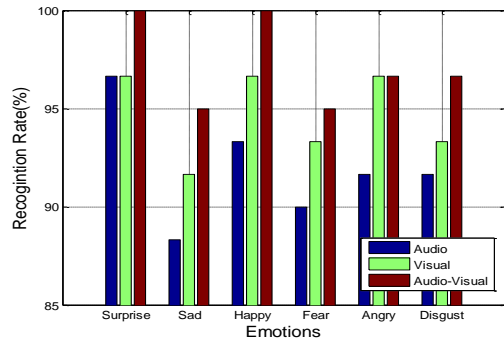


Figure 5. The recognition rate for different emotions of RML

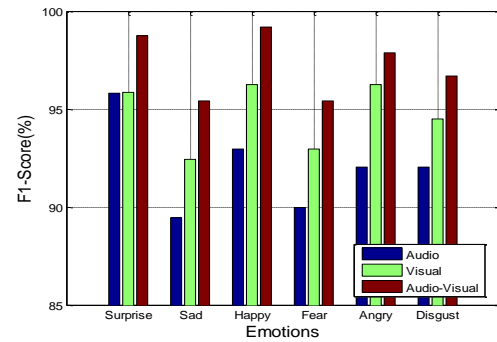


Figure 6. F1-score for different emotions of RML

Sad and fear emotions had the lowest F1-scores for audio-visual descriptors, 95.33% and 93.33%, respectively. Happy has the highest F1-score, 99.53%. The proposed fusion description worked for all emotions. Auditory, visual, and audio-visual recognition rates are 91.0205%, 94.7123%, and 97.2315%, respectively. The average F1-score of audio, visual, and audio-visual descriptors is 90.2144%, 993.4578%, and 96.4471%. Table 8 compares proposed and existing SAVEE and RML dataset techniques. LSTM and multimodal classification to distinguish emotions from audio-video footage. Several data models yielded 83.7% accuracy for audio-textual data. Video emotion using facial landmark points and speech signal emotion using spectral and prosodic features. They predicted emotion using feature-level fusion and classifiers.

Table 8. Comparative analysis

Reference	Method	Dataset	Accuracy
Begadi [20]	Modified Alexnet, LSTM BERT embedding	SAVEE	83.7000
Abdulmohsin <i>et al.</i> [21]	Geometrical, prosodic and spectral features, PCA, and LDA	SAVEE	84.000
Rahdari <i>et al.</i> [22]	Facial landmarks, prosodic and spectral features, and fuzzy rough neural network	SAVEE	91.6000
Chen <i>et al.</i> [23]	K-means clustering and KCCA	SAVEE	93.0600
Yang <i>et al.</i>	Fusion of Kernel matrix	RML	82.2200
Fadil <i>et al.</i> [24]	Deep networks with multi-layer perceptron	RML	79.7200
Seng <i>et al.</i> [25]	BDPCA, LSLDA, OKL, and RBF-SVM	RML	90.8300
Proposed	Geometrical and composite features and SVM	SAVEE	97.1600
		RML	97.2200

Figure 7 explains the accuracy of SAVEE and RML at different descriptors. FRNN was classified with 91.60% accuracy. K-means clustering [26] and Kernel canonical correlation analysis (KCCA) to recognize emotions from audio-visual characteristics. SAVEE dataset accuracy was 93.06%. Using 420 samples, the proposed strategy outperformed these algorithms with 97.16% accuracy. For the RML dataset, the proposed strategy achieved 97.2% accuracy, far higher than existing methods.

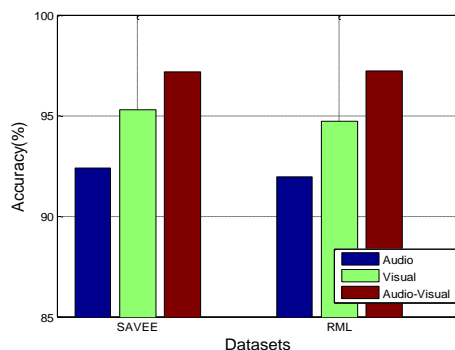


Figure 7. Accuracy of SAVEE and RML at different descriptors

4. CONCLUSION

This paper introduced an efficient multimodal emotion recognition system that considers audio and video data models as inputs and recognizes the emotion in audio-video clips. Audio feature descriptor includes MFCCs, statistical features, formant frequencies, and energy features. The visual descriptor includes landmarks acquisition, non-frontal view faces alignment, key frame selection, and visual geometry based on self-similarity distance. For classification, SVM is employed. Finally, at fusion, this work applied decision-level fusion and combined the confidence scores of individual classifiers to get the final prediction. Simulation experiments on two datasets, SAVEE and RML, explore the effectiveness. The approximate accuracies over the mentioned datasets are 97.16% and 97.20%, respectively. The average improvement from the existing methods is approximately 4% for SAVEE and 7% for RML.




REFERENCES

- [1] R. Beale and C. Peter, "The role of affect and emotion in HCI," in *Affect and Emotion in Human-Computer Interaction*, vol. 4868 LNCS, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 1–11.
- [2] K. Karthika, S. Dhanalakshmi, S. M. Murthy, N. Mishra, S. Sasikala and S. Murugan, "Raspberry Pi-enabled wearable sensors for personal health tracking and analysis," *International Conference on Self Sustainable Artificial Intelligence Systems*, pp. 1249–1253, 2023, doi: 10.1109/ICSSAS57918.2023.10331909.
- [3] G. Hemalatha and C. Sumathi, "A study of techniques for facial detection and expression classification," *International Journal of Computer Science and Engineering Survey*, vol. 5, no. 2, pp. 27–37, Apr. 2014, doi: 10.5121/ijcses.2014.5203.
- [4] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 3 PART1, pp. 597–607, Jun. 2012, doi: 10.1109/TMM.2012.2189550.
- [5] A. I. Middy, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio-visual modalities," *Knowledge-Based Systems*, vol. 244, p. 108580, May 2022, doi: 10.1016/j.knosys.2022.108580.
- [6] M. F. H. Siddiqui and A. Y. Javaid, "A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images," *Multimodal Technologies and Interaction*, vol. 4, no. 3, p. 46, Aug. 2020, doi: 10.3390/mti4030046.
- [7] M. Sharafi, M. Yazdchi, R. Rasti, and F. Nasimi, "A novel spatio-temporal convolutional neural framework for multimodal emotion recognition," *Biomedical Signal Processing and Control*, vol. 78, p. 103970, Sep. 2022, doi: 10.1016/j.bspc.2022.103970.
- [8] E. Ghaleb, J. Niehues, and S. Asteriadis, "Multimodal attention-mechanism for temporal emotion recognition," in *Proceedings - International Conference on Image Processing, ICIP*, Oct. 2020, vol. 2020-October, pp. 251–255, doi: 10.1109/ICIP40778.2020.9191019.
- [9] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification," *International Journal of Multimedia Information Retrieval*, vol. 9, no. 2, pp. 103–112, Dec. 2020, doi: 10.1007/s13735-019-00185-8.
- [10] H. Zhou, J. Du, Y. Zhang, Q. Wang, Q.-F. Liu, and C.-H. Lee, "Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2617–2629, 2021, doi: 10.1109/TASLP.2021.3096037.
- [11] Y. Kim and E. M. Provost, "ISLA: temporal segmentation and labeling for audio-visual emotion recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 196–208, Apr. 2019, doi: 10.1109/TAFFC.2017.2702653.
- [12] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition in video with multiple feature fusion," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 38–50, Jan. 2018, doi: 10.1109/TAFFC.2016.2593719.




- [13] D. Kamińska, T. Sapiński, and G. Anbarjafari, "Efficiency of chosen speech descriptors in relation to emotion recognition," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2017, no. 1, Feb. 2017, doi: 10.1186/s13636-017-0100-x.
- [14] L. Andrade-Arena, W. R. Perez, and C. Y. Arias, "Moodle platform and Zoom videoconference: learning skills in the virtual modality," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 31, no. 1, pp. 337–349, Jul. 2023, doi: 10.11591/ijeecs.v31.i1.pp337-349.
- [15] H. Artanto and F. Arifin, "Emotions and gesture recognition using affective computing assessment with deep learning," *IAES International Journal of Artificial Intelligence (IJAI)*, vol. 12, no. 3, pp. 1419–1427, Sep. 2023, doi: 10.11591/ijai.v12.i3.pp1419-1427.
- [16] A. Y. Nawaf and W. M. Jasim, "A pre-trained model vs dedicated convolution neural networks for emotion recognition," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 1, pp. 1123–1133, Feb. 2023, doi: 10.11591/ijece.v13i1.pp1123-1133.
- [17] V. Sekar and A. Jawaharlalnehru, "Semantic-based visual emotion recognition in videos-a transfer learning approach," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 4, pp. 3674–3683, Aug. 2022, doi: 10.11591/ijece.v12i4.pp3674-3683.
- [18] T. Baltrusaitis, "FERA-2015," *GitHub*, 2015. <https://github.com/TadasBaltrusaitis/FERA-2015>.
- [19] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 60–75, Jan. 2019, doi: 10.1109/TAFFC.2017.2713783.
- [20] K. R. Bagadi, "A comprehensive analysis of multimodal speech emotion recognition," *Journal of Physics: Conference Series*, vol. 1917, no. 1, p. 12009, Jun. 2021, doi: 10.1088/1742-6596/1917/1/012009.
- [21] H. A. Abdulmohsin, H. B. Abdulwahab, and A. M. J. Abdulhossen, "A new proposed statistical feature extraction method in speech emotion recognition," *Computers and Electrical Engineering*, vol. 93, p. 107172, Jul. 2021, doi: 10.1016/j.compeleceng.2021.107172.
- [22] F. Rahdari, E. Rashedi, and M. Eftekhari, "A multimodal emotion recognition system using facial landmark analysis," *Iranian Journal of Science and Technology - Transactions of Electrical Engineering*, vol. 43, no. S1, pp. 171–189, Oct. 2019, doi: 10.1007/s40998-018-0142-9.
- [23] L. Chen, K. Wang, M. Wu, W. Pedrycz, and K. Hirota, "K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 10250–10254, 2020, doi: 10.1016/j.ifacol.2020.12.2756.
- [24] C. Fadil, R. Alvarez, C. Martinez, J. Goddard, and H. Rufiner, "Multimodal emotion recognition using deep networks," in *IFMBE Proceedings*, vol. 49, Springer International Publishing, 2015, pp. 813–816.
- [25] K. P. Seng, L. M. Ang, and C. S. Ooi, "A combined rule-based & machine learning audio-visual emotion recognition approach," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 3–13, Jan. 2018, doi: 10.1109/TAFFC.2016.2588488.
- [26] R. Raman, V. Sujatha, C. B. Thacker, K. Bikram, M. B. Sahaai, and S. Murugan, "Intelligent parking management systems using IoT and machine learning techniques for real-time space availability estimation," *International Conference on Sustainable Communication Networks and Application*, pp. 286–291, 2023, doi: 10.1109/ICSCNA58489.2023.10370636.

BIOGRAPHIES OF AUTHORS



Kummari Ramyasree    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 DECS from 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 image processing, signal processing, artificial intelligence and machine learning, deep learning. She can be contact at email: kummariramyasreesree@yahoo.com.



Dr. Sumanth Kumar Chennupati    is a professor at Department of Electrical, Electronics and Communications Engineering, GITAM Institute of Technology, GITAM (Deemed to be University), Visakhapatnam, India. He awarded B.Tech. from Nagarjuna University in 1995, M.E. from MS University of Baroda in 1997 and Ph.D. from Andhra University in 2013. He has 25 years of teaching experience, and several research scholars are working towards their doctoral degree under his esteemed guidance. Published several international journals and conferences. His areas of intrests include image processing, signal processing and VLSI design. He can be contact at email: schennup@gitam.edu.