Multimodal modeling of human emotions using sound, image and text fusion

Seyed Sadegh Hosseini  
Islamic Azad University, Lahijan Branch

Mohammad Reza Yamaghani (✉ o_yamaghani@liau.ac.ir)  
Islamic Azad University, Lahijan Branch

Soodabeh Poorzaker Arabani  
Islamic Azad University, Lahijan Branch

Research Article

Keywords: Multimodal emotion, fusion, LSTM, CNN, RNN, machine learning, audio processing, text processing, video processing

Posted Date: February 20th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2579610/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Multimodal modeling of human emotions using sound, image and text fusion

Seyed Sadegh Hosseini, Mohammad Reza Yamaghani, Soodabeh Poorzaker Arabani

sadeghhosseini@malayeru.ac.ir, o_yamaghani@liau.ac.ir, arabani@liau.ac.ir

Abstract

Multimodal emotion recognition and analysis is considered a developing research field. Improving the multimodal fusion mechanism plays a key role in the more detailed recognition of the recognized emotion. The present study aimed to optimize the performance of the emotion recognition system and presented a model for multimodal emotion recognition from audio, text, and video data. First, the data were fused as a combination of video and audio, then as a combination of audio and text as binary, and finally the results were fused together. The final output included audio, text, and video data taking common features into account. Then, the convolutional neural network, as well as long-term and short-term memory (CNN-LSTM), were used to extract audio. Next, the Inception-Res Net-v2 network was applied for extracting the facial expression in the video. The output fused data were utilized by LSTM as the input of the softmax classifier to recognize the emotion of audio and video features fusion. In addition, the CNN-LSTM was combined in the form of a binary channel for learning audio emotion features. Meanwhile, a Bi-LSTM network was used to extract the text features and softmax was used for classifying the fused features. Finally, the generated results were fused together for the final classification, and the logistic regression model was used for fusion and classification. The results indicated that the recognition accuracy of the proposed method in the IEMOCAP data set was 82.9.

Keywords: Multimodal emotion, fusion, LSTM, CNN, RNN, machine learning, audio processing, text processing, video processing
1. Introduction

Human emotions are considered a complicated psychological and physiological phenomenon. Emotion recognition as one of the bases of human-computer emotional interaction affects the development of artificial intelligence technology. The limited emotional information available in a single mode is one of the challenges of emotion recognition. Since humans naturally communicate and express their emotions differently, the result of one-dimensional information is not very accurate compared with real emotions. One-dimensional information evaluates the emotions merely from one dimension, which may not reflect real emotions because of noise and non-involvement of other dimensions of emotion recognition. Thus, modern studies focus on multimodal fusion in emotion recognition and analysis.

The different methods of collecting the expressed emotional information can be divided into two main groups. The first group is called touch methods which collect information through sensors by connecting the sensors directly to the person's body including EEG, ECG, EMG signals, etc. The second group is called non-touch method in which the expressed emotional information is collected and processed by using physical non-touch sensors such as microphones, cameras, etc. In daily communication, humans receive emotional information such as facial changes, audio tone, body movements, spoken words, etc. by non-touch method. In a conversation, a person's emotional state can be shown by [1] words, audio, tone of audio, facial expressions, and body movements. In everyday conversations, the fusion of textual, audio and visual information [2] can provide more information on each dimension separately. Based on the fusion of non-touch methods, emotional computing can recognize subtle emotions and complicated emotions more accurately. The appropriate selection of unimodal emotion features and multimodal fusion strategies are considered as two key components of multimodal emotion analysis systems which often operate better than unimodal emotion recognition systems. In such strategies, different unimodal emotion recognition methods are combined with data fusion at feature level or decision level, resulting in a considerable improvement in their performance. Regarding the
significant improvement in the obtained models, the current studies perform better in using a single emotion recognition method which cannot fully use complex nonlinear relationships. Another problem is the need for labeled data for each method, which is difficult and may not always be available or might be accompanied with noise. Deep learning has recently been considered as a powerful technique for many applications where several layers are used for processing at various levels. These layers are formed of simple but non-linear modules for training highly complex functions. As a result, deep learning techniques have a good performance in solving various types of problems. The problems which were not solvable using traditional techniques and learning techniques were solved using deep learning techniques and achieved high accuracy.

2. Related Studies

Many researchers are currently engaged in research about multimodal compound emotion recognition. However, most of them use traditional feature extraction methods such as Hidden Markov Model (HMM) and Gaussian Mixed Models (GMM) [3] [4]. An adaptive GMM model is used to recognize emotions and extract the statistical features of spatial distribution, which has high accuracy [5]. An HMM model is proposed for facial expression classification and recognition in videos. In addition, random features are improved by using Gaussian kernel method. Nevertheless, the emotion recognition process is more based on ideal environments while influencing factors are not considered well. [6] Presented an emotion recognition method based on auditory features. Cepstrum coefficients of Gammatone frequency and cepstrum coefficients of GammChirp frequency are used as definite methods and the HMM model is used as a classifier to achieve more accurate emotion recognition. However, these models can only model limited contextual information. The features of slow changes in human emotions and severe dependence on contextual information cannot be fully applied [7]. Deep neural networks can effectively reveal the internal and hidden structure in data and extract useful features for emotion recognition compared to traditional feature extraction methods [8].

By considering the intrinsic relationship between different data in an emotional state and the emotional features extracted in different ways such as text and audio which are highly
interconnected, they can be fused and more effective features can be used for emotion recognition [9]. The main objective of deep learning is obtaining the representation of effective features from the data automatically through the fusion of different features, which are ultimately entered into the classifier model and determine the result [10]. Convolutional neural networks (CNN) as the main model of deep learning along with other network models such as recurrent neural network (RNN) and long short-term memory network (LSTM) are extensively used in emotion recognition, providing many corresponding optimization models and good research results [11]. Reference [12] focused on the extraction of features with high differences in the recognition process and provided a strategy for feature coordination recognition. As a result, ResNet101 and training iteration increased the entropy and realized high-accuracy video performance recognition [13]. To better achieve the emotional features in the audio signal, the researcher proposed a parallel convolutional RNN with spectral features for audio emotion recognition, among which several features were fused and the SoftMax classifier was used for emotion classification, resulting in better emotion recognition. However, the impact of unimodal mode on emotion recognition is still low [14] proposed an RNN-based emotion recognition method for the emotional states of musical instruments, having a better performance than the machine learning classification algorithm. However, the environment for the application of this method is somehow specific and cannot be generalized. [15] raised an RNN-based emotion recognition framework with parameter deviation and recorded some expressions for recognizing emotions by combining the coordinates of the angles and the angular velocity of human joints. Then, the researcher presented the compound structure of various states in unconscious behavior to improve the classification performance of the proposed method. Deep neural network models can be used in multimodal sentiment recognition to learn the feature of fused multimodal data [16]. The deep neural network imitates the information processing of the neural tissue in the human brain. In this regard, the original data are in the structure of the deep network, and then they are used in feature engineering of multifaceted sentiment analysis [17]. Traditional feature extraction methods such as HMM (Hidden Markov Model) and GMM (Gaussian Mixture Model) [19], [18] can only model a limited amount of contextual information. However, such methods ignore the fact that human emotions change slowly and are very much
dependent on contextual information. Unlike traditional feature extraction methods, deep neural networks can effectively reveal the hidden internal structure in data and extract high-level features which are beneficial for emotion classification. For instance, [20] used LSTM to obtain multi-classification for text emotional features and realized that LSTM is better in emotion analysis of long sentences than conventional RNN (recurrent neural network). Other studies applied CNN [21] and LSTM [22] in audio emotion recognition to obtain high-level features from raw audio clips. LSTM is considered as an advanced RNN with strong computing capacity and storage capabilities which can effectively solve common RNN problems. Audio is a kind of non-linear time series transformation signal while textual information is closely related to time context and both of them are common in time. Thus, LSTM network is appropriate for extracting and learning audio and text features and helps to learn feature relationships. Nevertheless, there is no intermediate nonlinear hidden layer in LSTM to increase the variation in hidden mode factors [23]. On the contrary, CNN network can decrease the variance of the input frequency and record the local information without considering the general features. In conclusion, both CNN and LSTM modeling capabilities are limited [24]. CNN can be fused to LSTM to automatically learn the best representation of the audio signal directly. Multimodal emotion recognition studies have made good progress based on the effective fusion of unimodal information such as audio, face, and text. Features from different modes such as text and audio can be combined due to the internal correlation between different modes to obtain more effective emotional features for emotion recognition. The deep neural network model can be fused and applied for feature learning in multivariate data. The original high-dimensional heterogeneous data can be changed into a non-linear mode in the same feature space for several times and then used to extract and select multimodal features. Pouria et al [25] considered all emotional features in the category of MKL (multi-kernel learning) classifier. Furthermore, they used the open source software openSMILE to extract low-level audio features and the results were considered in the classifier. [26] used a deep network with 50 layers to extract features from face information. In model [27], the final CNN layer calculated the weighted sum of all information extracted from the input. After extracting features from CNNs, [28] applied three-layer deep neural network to fuse multimodal features. In addition, [29] used RNN to extract features and proposed a new multi-objective
method to focus on specific parts containing the strong emotional information of audio data. Furthermore, [30] applied Bi-LSTM to classify audio and visual emotional information in the IEMOCAP database. The obtained experimental results showed that the Bi-LSTM framework works better than the traditional HMM framework. [31] used a RAVEN network for multi-level emotion recognition and used LSTM to extract unimodal features. The above-mentioned studies implemented the deep neural network model and the results had a better performance than the traditional methods. Moreover, all of these studies used the handmade features in the text or audio content. However, some focused on unimodal emotion recognition and the improvement of the limited presented models. Moreover, in the multimodal model, all of them use CNN or RNN as trainable feature extractors and do not simultaneously focus on the temporal information and ignore the temporal features of audio and text information.

3. The proposed method

Based on the conducted studies and their findings that were discussed above, the idea of using audio as a common feature in expressing the emotions of audio and video, and audio and text in identifying multimodal emotion recognition was presented, and a method for combining features with the shared element of audio was proposed. In this method, the common features of audio and video, as well as the features of audio and text, are extracted and classified, and their output will be classified at the end.

3.1. The Architecture of the Proposed Method

Fig.1 shows the architecture of the proposed method.
3.1.1. Extraction of Features for Audio and Video Method

The overall architecture of the multimodal compound emotion recognition method for audio and video expression based on deep learning is displayed in Fig. 2. In general, it is divided into feature extraction, feature selection, and emotion classification modules.
The received input data cannot be used directly for emotion classification. Thus, the data mode should be changed. The separation module transforms the video data into audio and video data, and then transforms the audio data into text.

The feature extraction module is the basis of the multimodal emotion recognition method which aims to receive the input for information processing and transforms it into the features which can be used by the model [32]. Various schemes are considered in the feature extraction module for extracting features. The audio uses the long-term and short-term neural memory network (CNN-LSTM) while the video uses the Inception-Res Net-v2 network to extract the related data.

Multimodal emotion data attempts to capture a user's current feelings from different angles. On the one hand, various methods assign different weights to describe emotions at a certain moment, causing data redundancy and noise. On the other hand, unimodal features may cause redundant data and noise information in the feature extraction process. If the redundancy and noise in data are not corrected, not only it increases the difficulty of training the model, but also it leads to the waste of resources. More unimodal features are extracted when the scale of dataset is relatively large. When several methods are fused, it might lead to "curse of dimensionality". Therefore, a feature selection module is added to decrease the features of each state. In a multifaceted emotional state, there is a correlation between features and LSTM structure can be used for temporal dependence recognition from time series. In this regard, there is an informational relationship between different methods. These relationships can be used as shared and hidden emotional cues between unimodal methods to help make decisions on the final emotion. LSTM can capture the relationship between different methods and make emotion recognition more accurate when there is little information or the recognition results of two or more emotions are inconsistent. The extracted unimodal features are sent to the LSTM unit to capture the interdependence. After the feature selection process, the unimodal features are combined to obtain a compound feature. Then, the LSTM structure is added to extract the interdependence between the methods.

### 3.1.1.1. Extraction of audio signal features

Audio is a non-linear time sequence signal that is closely related to time. Meanwhile, LSTM network is appropriate for extracting and learning audio features. In context-sensitive
information modeling, it is beneficial to learn the temporal information of audio emotion features. However, there is no intermediate nonlinear hidden layer in LSTM network, leading to an increase in the hidden mode coefficient. Overall, the capabilities of both CNN and LSTM are limited. CNN and LSTM networks are applied for audio emotion recognition. For this reason, a combined network of CNN and LSTM is used here to learn audio emotional features. The structural features of CNN network are combined in CNN-LSTM network and an LSTM layer is added. To transform the input data from the convolution layer and then for re-learning, we enter the input to the LSTM layer. The correlation part includes two convolution layers and a maximum aggregation layer. The parameters of the convolution layer are the same as the CNN network parameters and the activation function is a linear unit (ReLU) [33].

Convolutional Neural Network (CNN) is considered as one of the most common deep learning neural networks. CNN infrastructure involves the convolutional layer, pooling layer, and dense layer. In addition, CNN can extract some advanced features automatically. Sharing the layer weight reduces the network complexity and overfitting while the pooling operations reduce the number of neurons. The net input of the convolution layer is measured as follows:

$$Z^p = W^p \otimes X + b^p = \sum_{d=1}^{D} W^{p,d} \otimes X^d + b^p,$$

where $W$ represents the convolution kernel, $X$ implies the feature mapping, and $b$ shows the bias term. The mapping of output characteristic $Y^p$ is achieved after passing the net input through the nonlinear activation function:

$$Y^p = f(Z^p),$$

where $f(\cdot)$ represents the nonlinear activation function. CNN network can reduce the input frequency variance and retain the local information regardless of general features and contexts. The output of a set of CNNs directly enters into LSTMs in CNN-LSTM shared network. In this way, the advanced information including both local information and long-term contextual dependencies can be obtained. Nevertheless, CNNs focus on local information and discard a lot of data. A method using CNN and Bi-LSTM channels was presented to prevent the loss of valuable data. A number of four one-dimensional convolutional layers were used in CNN channel with
different number of filters for one-dimensional temporal input (audio features). In addition, the maximum pooling layer and the general average pooling layer were applied for conducting the maximum pooling operations and the average pooling operation for the data. A set of Bi-LSTM cells with arguments 256 was placed in Bi-LSTM channel to extract the information of long-term text dependencies and the attention mechanism was added to find more effective features. Finally, the data of the two channels were interconnected and the output was placed in a dense layer. After the nonlinear change of the dense layer, the correlation between these features was extracted and eventually mapped to the output space [34]. The structure of this method is displayed in Fig. 3.

3.1.1.2. Extraction of face features

The volume of video data is usually very high. Therefore, we need to find the contour of the face in each frame of the video and decrease the frame image size by cropping the face contour image, which is required for data processing and learning the next features. Here, the emotional information in the video is found by extracting short-term features by Fourier transform (STFT) and Inception-Res Net-v2 network to extract deep features. The Short-Time Fourier transform is a general tool for audio signal processing which defines a highly useful class of time-frequency distributions by specifying the amplitude of any signal which varies with time and frequency. The short-term Fourier transform is calculated by dividing a long-time signal into shorter parts of the same length and calculate the Fourier transform in each part [35]. Inception-Res Net-v2 refers to
a CNN developed by Google in 2016 by introducing the ResNet network based on the Inception model having an accurate diagnostic effect on very similar objects. This network is considered as a prototype of Inception V3 which uses the idea of connectivity in the Microsoft ResNet model so that the neural network can be trained more deeply. The Inception module can be significantly simplified by using multiple the convolutions of 3x3 instead of convolutions of 5x5 and 7x7. As a result, the computational complexity and dimensions of the parameters considerably reduce and the training speed of the network increases. The spatial complexity of this algorithm is higher compared to other networks.

The Inception-Resnet-v2 network involves six modules including stem, Inception-ResNet-A, Reduction-A, Inception-ResNet-B, Reduction-B, and Inception-ResNet-C. The output vectors of the convolution layer, Avg Pool layer, Dropout layer, and the fully connected layer, which are finally connected by the network are regarded as the depth features of the samples obtained after each layer of learning. The features of the fully connected layer with better recognition are selected as the extracted deep video features by classifying and comparing the output features through Inception-ResNet-v2 in the convolution layer, Avg Pool layer, Dropout layer, and the fully connected layer [37].

### 3.1.2. Selection of audio and video features for features fusion (A/V)

The LSTM-RNN structure is utilized to understand the dependence between the internal features of each mode. There is interdependence not only between the internal features of a single mode, but also between the features of different modes. When the features of one mode are few, the existing information in the other mode can help emotional decision-making [36]. In this regard, the LSTM structure is applied for obtaining the dependence between different modes [37], the structure of which is displayed in Fig. 4.
Fig. 4. RNN for preserving the internal dependencies of audio and video

By assuming that the unimodal is h-dimensional, the features in the multimodal can be represented by the feature vector while Rh and t represent sentence t in video i. All of the vectors in a video can be collected to obtain the vector matrix $X_i = x_{i,1}, x_{i,2}, \cdots, x_{i,L} \in \mathbb{R}^{L_i \times h}$. Li shows the total number of sentences in video and matrix $X_i$ can be used as the input of LSTM. This method uses a filtering method for feature selection to improve the generalizability of emotion recognition method and decreases the amount of calculation. The feature subsets are selected based on the general features of the data and statistical methods are used for assigning scores to the features. Then, such features are ranked in descending order based on their scores and the higher ranked features are kept in the feature subset, filtering out unrelated features. In addition, the Chi-Square test is used for showing the correlation between the feature and the corresponding class, in which a higher score means that the feature depends mostly on the corresponding class and the features with a lower score have less information and should be eliminated. Thus, the features with low scores have less information and should be eliminated. In this regard, redundant information, noise information, and noise are filtered [37].

$$\text{Eq.3}$$

$$\text{Chi}(f, c_i) = \frac{N \times (AD - BC)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}$$

$$\text{Chi}_{\text{max}}(f) = \max(\text{Chi}(f, c_i))$$

The chi-square test directly selects a subset of features from the main features and applies the chi-square value to sort the features and automatically filters the features having a score less
than a predetermined threshold $a$. The chi-square test can be defined as follows by assuming that the feature is independent of the final classification value:

$$\text{where } A \text{ represents the number of documents with feature } f \text{ belonging to category } c_i, B \text{ shows the number of documents with feature } f \text{ but not belonging to category } c_i. C \text{ shows the number of occurrences of } c_i \text{ without } f, D \text{ implies the frequency when neither } c_i \text{ nor } f \text{ appears, } N \text{ represents the total number of samples in the document set } N = A + B + C + D. \text{ The result of the chi test becomes zero if } f \text{ and } c_i \text{ are independent.}

### 3.1.3. Classification of features for audio and video method (A/V)

Audio features and facial expressions as the final feature of the multimodal signal are fused. In this method, softmax is used as a classifier for classifying emotions. Each channel can be considered a set of separate multimodal signals for classifying emotions. Fusion is conducted at the decision-making level for classification of the results generated by each softmax classifier.

### 3.2. Feature extraction scenario for audio and text features fusion (A/T)

As shown in Fig. 5, first the audio signal and text information are processed for extracting the low-level emotional features. Then, the audio features are entered into the model to display local and general information. Textual features are entered into Bi-LSTM neural networks to obtain high-level features. Then, the feature fusion method is used for the fusion of voice and text emotional features [34]. At the end, a deep neural network is used for learning and classifying the fusion features. In the multimodal method, the DNN network is applied for training and optimizing the fusion features.
3.2.1. **Audio and text feature extraction**

3.2.1.1. **Audio signal feature extraction**

First, the raw audio signal is preprocessed from the audio data in the database. Then, 34-dimensional low-level audio features are extracted from temporal domain, spectral domain and cepstral domain features including zero crossing rate, energy, energy entropy, spectral center, spectral expansion, spectral entropy, spectral flux, spectral launch, and MFCC mel frequency cepstral coefficients [38]. The maximum input of 100 frames are considered and the vector is obtained for each expression (100, 34). To represent the sensory features of hand-held acoustics, several features including zero-crossing rate, MFCC, energy, and spectral flux are considered.

3.2.1.2. **Text feature extraction**

Word embedding is a popular method in text feature extraction by using dense vectors to show the words and documents [39]. Word2vec is considered as one of the most common word embedding methods. It is a method for word matrix generation which combines the general and local statistical information of words to generate linguistic models and word vectors. Word2vec embedding method is used for text data to obtain the emotional feature vectors of the text. The 300-dimensional vectors from the Word2vec embedding with a maximum sequence length of 500 are used to obtain the vectors for each expression (300, 500). In Word2vec, non-linear activation functions (tanh, sigmoid, ReLu, etc.) are not used out of the softmax calculations in the last layer and outputs are sent as weighted combinations of inputs. In fact, we have:

$$Output \ from \ hidden \ node \ 'j' = u_j = \sum_{i=1}^{V} w_{i,j} x_i$$
where $u_j$ shows the input of the j-th hidden node, $w_{ij}$ represents the connection weight from input node $i$ to the j-th hidden node and $x_i$ means the value of input node $i$ (In case of word vectors, only one element of $x_i$ equals 1 but the other parameters are zero). The output layer has output nodes $V$, one node for each unique word ($V$ corresponds to word size), and the final output from each node is softmax.

3.2.2. Fusion of audio and text features

After entering the audio features into the model and the textual feature in the Bi-LSTM network, the high-level textual acoustic features $H = \{h_1, h_2, \ldots, h_t\}$ and $S = \{s_1, s_2, \ldots, s_t\}$ composed of general and local information can be obtained. Here, the fusion approach was used at the features level. The main advantage is that the emotional features extracted from different states are directly related to the final decision and can considerably retain the feature information needed in the final decision by merging the results. The final descriptor of the multimodal emotional features is the vector

$$V = [v_1, v_2, \ldots, v_t]$$

which is created using the regular concatenation of text $H$ and acoustic $S$ in the equation below:

$$V = [H, S].$$

Then, the feature vector of fusion $V$ enters into a deep neural network (DNN) containing three dense layers, in which the parameters are set as 1024, 512, 4, and a softmax layer to show the relationship between the features [34]. The output of the softmax classifier indicates the relative probability between different emotion classes:

$$P(x_a) = \text{softmax}(x_a) = \frac{\exp(x_a)}{\sum_a \exp(x_a)},$$

where $a$ shows the classes of emotions and $P(x_a)$ represents the probability of class $a$ in the classification of emotions.

3.1.3. Scenario of fusion in the outputs of two models (audio and video) and (audio and text) (A/V/T)
In the final stage, after generating the combined outputs of A/V and T/V, it is time to merge and produce the final result. In this section, the Multinomial Logistic Regression which is sometimes known as softmax regression is used in the final layer of neural networks classification. In multinomial logistic regression and audit analysis or linear audit analysis, the results of calculations made by k different linear functions are regarded as input to the function. According to the sample x and the vector of weights w, the predicted probability for the j-th class is as follows.

\[ P(y = j \mid x) = \frac{e^{x^T w_j}}{\sum_{k=1}^{K} e^{x^T w_k}} \]

The above-mentioned equation can be considered as a combination of K linear functions and smooth maximum function.

\[ x \mapsto x^T w_1, \ldots, x \mapsto x^T w_K \]

Note that \( x^T w \) represents the inner product of two vectors x and w.

4. Test and analysis

After preparing the scenarios, the methods are used to check the accuracy of feature extraction and the result of their fusion in a more accurate emotion recognition. IEMOCAP database is applied for conducting the test and following the above-mentioned steps to analyze the results.

4.1. IEMOCAP emotion database

The IEMOCAP emotion database [40] is used to evaluate the proposed method. The database includes almost 12 hours of audiovisual data such as video, audio, facial motion recording, and text transcription. In addition, 10 professional actors and actresses (five males and five females) perform two different scenarios of a text game and improvised dialogue in two-person conversations. The length of each conversation is about five minutes, which is divided into sentence levels. In this study, four emotional categories of angry, happy, sad, and neutral are used, similar to most of the previous studies which used such a dataset. It should be noted that "happy" and "excited" are fused into the "happy" class in the original annotation. In the tests, only the sentences with majority agreement are used which means that at least two out of three
evaluators of the emotional label have the same opinion. The distribution of classes is as follows: 20.0% angry, 19.6% sad, 29.6% happy and 30.8% neutral. Due to the low accuracy but high exaggeration in the implemented data, there is still a considerable gap between the implemented data and the normal emotions of people in daily communication. The improvised data are more valid in comparison to the implemented data. Considering the accuracy of data, the improvised data are used for voice emotion recognition.

<table>
<thead>
<tr>
<th>Name</th>
<th>Year</th>
<th>Components</th>
<th>Subjects</th>
<th>Type</th>
<th>Emotion Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEMOCAP [101]</td>
<td>2008</td>
<td>■ ■ ■ /</td>
<td>10</td>
<td>Acted</td>
<td>Happiness, Anger, Sad, Frustration and Neutral Activation-Valence-Dominance</td>
</tr>
</tbody>
</table>

4.2. Chi-Square test performance

In this method, first Chi-square test method is used for excluding redundant information and noise information. After feature selection processing, the unimodal features are found. Finally, these two methods are combined for multimodal emotion recognition. A series of tests were conducted to compare the model performance with the chi-square test and the feature without the chi-square test for testing the necessity of using the chi-square test to remove the redundant information and noise information in the multimodal emotion recognition method. The obtained result is presented in Fig. 6, among which regulating the parameters shows that the recognition effect is in the best state when the chi-square value is higher than 50%. To exclude the redundant information and noise data generated in the unimodal feature extraction process, the LSTM structure is used here to obtain the internal information dependence of each unimodal method. Further, the chi-square test feature selection method is added for eliminating the redundant features and noise.
As shown in Fig. 6, the accuracy of multimodal emotion recognition is significantly higher than that of unimodal emotion recognition. The multimodal recognition accuracy is about 77% while the unimodal accuracy is less than 73%. Moreover, the recognition accuracy of the model with chi-square test is significantly more than the model without chi-square test. The accuracy of emotion recognition increased by 3% by considering the multimodal model and adding chi-square test. As a result, the use of chi-square test to eliminate redundancy and noise from the information of several features is significant [37].

4.3. Comparing the results of the extracted features and their fusion by the above methods

4.3.1. Comparing the results of text and audio fusion

By combining CNN and Bi-LSTM networks, a better model is obtained for voice and text recognition with better accuracy, proving the validity of the fusion method. The results of the unimodal methods and the fused model are shown in Table 2.

Table 2. Confusion matrices of audio and text fusion

(A) Confusion matrix of audio unimodal model
### (B) Confusion matrix of text unimodal model

<table>
<thead>
<tr>
<th></th>
<th>angry</th>
<th>happy</th>
<th>neutral</th>
<th>sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>angry</td>
<td>69.87</td>
<td>8.9</td>
<td>14.02</td>
<td>6.9</td>
</tr>
<tr>
<td>happy</td>
<td>11.73</td>
<td>63.91</td>
<td>14.49</td>
<td>8.55</td>
</tr>
<tr>
<td>neutral</td>
<td>7.79</td>
<td>9.72</td>
<td>70.11</td>
<td>11.43</td>
</tr>
<tr>
<td>sad</td>
<td>9.2</td>
<td>5.02</td>
<td>15.2</td>
<td>69.85</td>
</tr>
</tbody>
</table>

### (C) Confusion matrix of text and audio fusion

<table>
<thead>
<tr>
<th></th>
<th>angry</th>
<th>happy</th>
<th>neutral</th>
<th>sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>angry</td>
<td>65.41</td>
<td>9.21</td>
<td>12.8</td>
<td>12.2</td>
</tr>
<tr>
<td>happy</td>
<td>14.83</td>
<td>60.22</td>
<td>14.32</td>
<td>9.76</td>
</tr>
<tr>
<td>neutral</td>
<td>10.09</td>
<td>14.81</td>
<td>62.14</td>
<td>11.64</td>
</tr>
<tr>
<td>sad</td>
<td>14.2</td>
<td>10.1</td>
<td>15.6</td>
<td>59.73</td>
</tr>
</tbody>
</table>

As shown in Tables 2(a) and 2(b), there is confusion in the emotional results of the unimodal model. Table 2(c) shows that the accuracy of most types of emotions is improved and the confusion of emotions is reduced by the fusion of audio and text emotional features. This indicates the validity of the fusion of the models which can effectively reduce emotional confusion and improve the recognition rate of emotional states.

#### 4.3.2. Results of audio and video fusion

Long- and short-term memory convolutional neural network and Inception-Res Net-v2 are used for extracting the feature data of facial expressions in video and audio, respectively. The features of the obtained models are simultaneously entered into the LSTM unit and the unit
methods are selected to obtain the effective features of fusion by the Chi-square test feature selection method. The output of LSTM feature is entered into softmax classifier to recognize the emotional states. Based on the IEMOCAP dataset, the emotion recognition performance is best when the network weight is around 0.57\cite{37} and the chi-square test is added. In this case, the feature selectivity of the fully connected layer of the Inception-Res Net-v2 network indicates the best performance as shown in Table 3.

**Table 3.** Confusion matrices of audio and video fusion

(A) Confusion matrix of audio unimodal method

<table>
<thead>
<tr>
<th>AUDIO</th>
<th>angry</th>
<th>happy</th>
<th>neutral</th>
<th>sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>angry</td>
<td>69.87</td>
<td>8.9</td>
<td>14.02</td>
<td>6.9</td>
</tr>
<tr>
<td>happy</td>
<td>11.73</td>
<td>63.91</td>
<td>14.49</td>
<td>8.55</td>
</tr>
<tr>
<td>neutral</td>
<td>7.79</td>
<td>9.72</td>
<td>70.11</td>
<td>11.43</td>
</tr>
<tr>
<td>sad</td>
<td>9.2</td>
<td>5.02</td>
<td>15.2</td>
<td>69.85</td>
</tr>
</tbody>
</table>

(B) Confusion matrix of image unimodal method

<table>
<thead>
<tr>
<th>Video</th>
<th>angry</th>
<th>happy</th>
<th>neutral</th>
<th>sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>angry</td>
<td>69.84</td>
<td>9.21</td>
<td>9.6</td>
<td>10.9</td>
</tr>
<tr>
<td>happy</td>
<td>12.81</td>
<td>68.15</td>
<td>10.72</td>
<td>8.02</td>
</tr>
<tr>
<td>neutral</td>
<td>9.57</td>
<td>10.21</td>
<td>71.86</td>
<td>7.86</td>
</tr>
<tr>
<td>sad</td>
<td>7.63</td>
<td>8.87</td>
<td>14.04</td>
<td>68.91</td>
</tr>
</tbody>
</table>

(C) Confusion matrix of audio and video fusion

<table>
<thead>
<tr>
<th>A&amp;V</th>
<th>angry</th>
<th>happy</th>
<th>neutral</th>
<th>sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>angry</td>
<td>71.89</td>
<td>2.02</td>
<td>11.55</td>
<td>14.17</td>
</tr>
<tr>
<td>happy</td>
<td>4.73</td>
<td>77.92</td>
<td>2.74</td>
<td>14.29</td>
</tr>
<tr>
<td>neutral</td>
<td>4.48</td>
<td>2.22</td>
<td>80.93</td>
<td>12.05</td>
</tr>
<tr>
<td>sad</td>
<td>2.91</td>
<td>9.87</td>
<td>12.16</td>
<td>74.62</td>
</tr>
</tbody>
</table>

4.3.3. Results of video, audio, and text fusion
After the fusion of (audio and text) and (audio and video) models, the regression is used for the final evaluation and comparison of the results of the mentioned models with the output of the video, audio and text model by using the confusion matrix, as shown in Table 4:

### Table 4. Confusion matrices of text, audio, and video fusion

#### (A) Confusion matrix of audio and video fusion

<table>
<thead>
<tr>
<th></th>
<th>angry</th>
<th>happy</th>
<th>neutral</th>
<th>sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>angry</td>
<td>71.89</td>
<td>2.02</td>
<td>11.55</td>
<td>14.17</td>
</tr>
<tr>
<td>happy</td>
<td>4.73</td>
<td>77.92</td>
<td>2.74</td>
<td>14.29</td>
</tr>
<tr>
<td>neutral</td>
<td>4.48</td>
<td>2.22</td>
<td>80.93</td>
<td>12.05</td>
</tr>
<tr>
<td>sad</td>
<td>2.91</td>
<td>9.87</td>
<td>12.16</td>
<td>74.62</td>
</tr>
</tbody>
</table>

#### (B) Confusion matrix of audio and text fusion

<table>
<thead>
<tr>
<th></th>
<th>Angry</th>
<th>happy</th>
<th>neutral</th>
<th>sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>angry</td>
<td>74.92</td>
<td>3.58</td>
<td>3.83</td>
<td>16.81</td>
</tr>
<tr>
<td>happy</td>
<td>1.92</td>
<td>73.25</td>
<td>2.08</td>
<td>22.03</td>
</tr>
<tr>
<td>neutral</td>
<td>2.81</td>
<td>2.51</td>
<td>80.85</td>
<td>13.39</td>
</tr>
<tr>
<td>sad</td>
<td>2.41</td>
<td>9.62</td>
<td>6.68</td>
<td>80.82</td>
</tr>
</tbody>
</table>

#### (C) Confusion matrix of text, audio, and video fusion

<table>
<thead>
<tr>
<th></th>
<th>angry</th>
<th>happy</th>
<th>Neutral</th>
<th>sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>angry</td>
<td>79.62</td>
<td>1.26</td>
<td>2.1</td>
<td>16.51</td>
</tr>
<tr>
<td>happy</td>
<td>1.62</td>
<td>82.8</td>
<td>0.3</td>
<td>14.9</td>
</tr>
<tr>
<td>neutral</td>
<td>4.07</td>
<td>1.48</td>
<td>80.94</td>
<td>13.22</td>
</tr>
<tr>
<td>sad</td>
<td>1.91</td>
<td>9.81</td>
<td>7.03</td>
<td>80.88</td>
</tr>
</tbody>
</table>

### 4.4. Result and comparison

#### 4.4.1. Result
• Some emotions are better recognized through audio such as sadness and fear while some others are best recognized through video such as anger and happiness. In voice emotion classification, anger and sadness are misrecognized as happy and neutral expressions, while anger may be mistaken for sadness, and happiness may be mistaken with neutral expression in facial expression classification.

• The text classifier very well recognizes anger, happiness, and neutral items. However, it is very difficult to distinguish between anger and sadness using text. One of the probable causes for this difficulty is that both of these classes are negative and numerous similar words are used for their expression.

• In the case of the audio model, better accuracy was observed for sad and neutral classes compared to text, but the difference was not as much as that for happy and angry classes. The classifier misclassified so many happy cases. Nevertheless, the classifier performed very well in distinguishing sadness and anger. In addition, some happy cases were misclassified as neutral.

• Although angry and sad faces can be classified effectively, the classifier indicated some confusion between angry and sad faces. The neutral classes were separated more accurately than the other classes, although some confusion was observed between happy and sad faces.

• In the case of voice and text fusion, as well as voice and image, the recognition accuracy is more than the unimodal model in all features, and confusion is reduced in the wrong classification.

• Bimodal and trimodal models performed better than unimodal models in different tests. In general, audio fusion performed well on all datasets.

4.4.2. Comparison to other methods

Mital et al. [40] suggested multiplicative multimodal emotion recognition (M3ER): First, feature vectors are extracted from three raw methods. Then, such features are transferred to the model evaluation step to retain the effective features and eliminate the ineffective features which are used for reconstructing the proxy feature vectors. Finally, the selected features are fused to
predict the six emotions based on the fusion of the final feature level with the attention module. M3ER achieved recognition accuracy of 82.7% in IEMOCAP. The results indicated that the recognition accuracy of the presented multimodal model is better than the unimodal model and better than other multimodal models in the dataset. This comparison is shown in Table 5.

Table 5. An overview of multimodal emotion recognition during 2020-2022 in the IEMOCAP database

<table>
<thead>
<tr>
<th>Publication</th>
<th>Year</th>
<th>Feature Representation</th>
<th>Classifier</th>
<th>Fusion Strategy</th>
<th>Database</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text-audio Emotion Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[38]</td>
<td>2020</td>
<td>A-DCNN, T-DNN Self-attention</td>
<td>FC</td>
<td>Feature-level</td>
<td>IEMOCAP</td>
<td>4 classes: 80.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4 classes: 79.22</td>
</tr>
<tr>
<td>[39]</td>
<td>2020</td>
<td>Acoustic features Word embeddings</td>
<td>Pooling Scalar weight fusion</td>
<td>Feature-level Decision-level</td>
<td>IEMOCAP</td>
<td>65.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>63.9</td>
</tr>
<tr>
<td>Our work</td>
<td>2022</td>
<td>Common Features</td>
<td>softmax</td>
<td>Feature-level</td>
<td>IEMOCAP</td>
<td>69.8</td>
</tr>
<tr>
<td>Visual-audio Emotion Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our work</td>
<td>2022</td>
<td>fusion</td>
<td>softmax</td>
<td>Feature-level</td>
<td>IEMOCAP</td>
<td>71.4</td>
</tr>
<tr>
<td>Visual-audio-text Emotion Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[40]</td>
<td>2020</td>
<td>Proxy and Attention Multiplicative fusion</td>
<td>FC</td>
<td>Feature-level</td>
<td>IEMOCAP</td>
<td>4 classes: 82.7</td>
</tr>
<tr>
<td>Our work</td>
<td>2022</td>
<td>Common feature and models combination</td>
<td>Regression softmax</td>
<td>Feature-level Decision -level</td>
<td>IEMOCAP</td>
<td>4 classes: 82.9</td>
</tr>
<tr>
<td>----------</td>
<td>------</td>
<td>--------------------------------------</td>
<td>-------------------</td>
<td>-------------------------------</td>
<td>---------</td>
<td>-----------------</td>
</tr>
</tbody>
</table>

5. Conclusion

In this study, a multimodal emotion recognition model based on video, audio, and text in the IEMOCAP database was proposed. First, the dual channel CNN and LSTM were used to learn acoustic emotion features. Furthermore, the Bi-LSTM was applied to extract the text features. In addition, a deep neural network was used to learn fusion features. The model used deep learning networks and artificial and advanced features and considered time and text information in the data. Moreover, L2 adjustment was used to optimize the model. A (A/V) bimodal fusion was simultaneously used based on deep learning using long-term and short-term memory convolutional neural network and 2Inception-Res Net-v to extract the feature data of facial expressions in audio and video. At the same time, the obtained features were entered into the LSTM unit, and the individual methods were inter-connected to obtain the fusion features using the chi-square test feature selection method. Then, the output feature data of the LSTM was entered into the softmax classifier for emotion recognition. The results on the IEMOCAP dataset indicated that the emotion recognition performance was the best when the network weight was about 0.57 and the chi-square test was added. In this case, the feature selection ability of the fully connected layer of the Inception-Res Net-v2 network was the best. In this method, attempts were made to eliminate the redundancy and noise of internal features. However, there were some complementary correlations between the information of modules. Here, all features of a unimodal model were mapped into a unified feature space before feature selection. Finally, the methods of (A/T) and (A/V) were fused, and the result of emotion recognition was extracted using the proposed method (A/T/V) at the accuracy of 82.9.
Funding

No grant was used to carry out this research, and it was done with personal investment.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Author Contribution

Seyed Sadegh Hosseini and Mohammad Reza Yamaghani carried out the experiment. Soodabeh Poorzaker Arabani wrote the manuscript with support from Seyed Sadegh Hosseini and helped supervise the project. Seyed Sadegh Hosseini and Mohammad Reza Yamaghani fabricated the models. Seyed Sadegh Hosseini conceived the original idea and supervised the project.

References:

neural networks for feature extraction and harmonizing for action recognition. \cite{Hossain2020}


