A Survey of Multi-modal Emotion Recognition Based on Deep Learning

Muhan Jia[1], Zijian Sun[2, *]

[1]School of Electrical Engineering, Southwest Jiaotong University, Chengdu, China

[2]School of Electrical Engineering, Hebei University of Technology, Tianjin, China

* Corresponding Author Email: pingshuoyear@ldy.edu.rs

Abstract. Multi-modal emotion recognition technology explores emotion recognition by integrating facial expression, voice intonation, text analysis and other multi-source data, so as to improve the naturalness and accuracy of human-computer interaction. Aiming at the emerging field of multi-modal emotion recognition, this paper introduces three single-modal emotion recognition methods of text, face and voice, especially the problem of multi-modal emotion fusion, and introduces the methods with high success rate of multi-modal emotion fusion recognition in recent years. Through comparative analysis, the conclusion is drawn that the current fusion methods are more complicated and the fusion success rate has been improved to some extent. However, the number of data sets on multi-modal emotion analysis is small, and the research on gesture and other modes of emotion recognition is also scarce. In the later stage, it is necessary to enrich the data set and add new modes to improve the accuracy and robustness of the multi-modal emotion recognition and analysis system.

Keywords: Multi-modal, emotion recognition, deep learning, fusion methods.

1. Introduction

In recent years, with the rapid progress of Internet technology, people's demand for intelligent life is also increasing. Human-computer interaction is the first step to achieving artificial intelligence, and emotion is the foundation of this process. As one of the key technologies of artificial intelligence, emotion recognition has been widely used in telemedicine, smart homes, crime prediction, and accident prevention [1]. In today's life, people produce a large amount of multi-modal data with rich emotions every day, such as voice, facial expressions and body movements. These data have important research value and significance for the analysis and recognition of emotion by multi-modal data. Generally speaking, researchers usually use text, voice, and face to predict emotion. However, when using single-mode information for emotion recognition, the information source is single, so there will be insufficient data set. Moreover, single-mode data may provide wrong information, thus affecting the final prediction result [2]. Therefore, the research on multi-modal emotion recognition is more important. This paper discusses three methods: dynamic convolution and residual gated multi-modal emotion recognition fusion, three-space representation multi-modal emotion fusion recognition network and multi-granularity multi-modal interaction-based emotion recognition fusion. After adding noise, the accuracy can still reach 94.6%. Finally, the improvement and enhancement of emotion recognition in data set shortage and multi-modal fusion are prospected [3] [4].

2. Single-modal Modal Emotion Recognition Based on Deep Learning

2.1. Text Emotion Recognition Based on Deep Learning

Conversations are easily influenced by emotions, which are often reflected in text, but recognizing emotions is a challenging task for machines. Customized sentiment detection methods based on Reema Goyal etc. to make the recognition process more realistic [5]. The flow of the model is determined by the input. If the input is in the form of a speech signal, then it attempts to classify

human emotions extracted from the automatic speech recognizer (ASR) as well as text processing. But if the input is text, then it can be classified directly from this model.

In the proposed model, PocketSphinx is used as a real-time speech translator, capable of quickly receiving speech input and recognizing acoustic events, such as coughs and exclamations. PocketSphinx selects the best pairing by matching possible word combinations and features noise suppression. Once the speech is recognized, it is further processed using the Word2Vec framework. The framework understands the meaning of words through their context and converts complete sentences into word vectors. Then, the word vectors are grouped by K-means, and the weighted emotion coefficient is calculated according to the assigned value of the cluster. Each sentence consists of two vectors: a weighted emotion score and a TF-IDF score. To analyze the identified emotions, calculate the dot product of the two vectors and obtain a TF-IDF score. K-means clustering was chosen because it can efficiently handle large-scale data, while TF-IDF assigns importance values to each word to highlight key content.

Table 1. Performance measurement table

Performance Measure	Word2Vec Model
Accuracy	0.802503
Precision	0.99498
Recall	0.799930
F1-Score	0.888608

This process builds a custom emotion detection architecture for emotion classification. The model receives user text input and computes expected emotions through a trained model. Table 1 shows an example of the output for training using the ISEAR database. Through the Word2Vec algorithm, the prediction accuracy reaches 80%, which is 10% higher than the unsupervised feature adaptive scheme of the standard model. This result shows that Word2Vec can produce the best results when combined with the right parameters. The final model had an overall accuracy of 0.99 and was particularly good at identifying negative emotions. The model had an F1 score of 0.88 and a recall rate of 80%, indicating that 80% of positive observations were accurately identified. Overall, the model performed well in negative emotion detection.

With the above results, as the data set becomes familiar with more user data, the classification of emotions in the text can be extended to the emotions of all users. In the current emotion recognition system, most of the models dealing with language are not well trained, but with the increase in the number of users, the corresponding set of text emotion data will become larger, and the trained models will be more accurate. In the future, it is hoped that it can be integrated into more powerful interactive machine learning, which can be used as a text analysis assistant and integrated into People's Daily lives.

2.2. Facial Emotion Recognition Based on Deep Learning

Emotion recognition is considered a major capability of machines in human-computer communication and can be based on voice and facial expressions as well as text. For human-computer interaction, facial expressions are very important because they can carry all kinds of information. In this regard, S. Dwijayanti, etc. proposed to use convolutional neural networks to realize real-time facial emotion recognition in humanoid robots [6].

The first stage involves the detection of faces using cameras to obtain images of faces for training datasets. In addition to face and emotion recognition, distances are measured to determine the location of objects. When recognizing faces and emotions, the frame involves four variables (x, y, w, and h), where x and y represent the coordinates of the top left and bottom right corner of the bounding box, respectively, while w and h represent the width and height of the bounding box. After processing, the final end x and end y values are obtained, thus determining the values of kdX and kdY. x and y coordinates are calculated as follows:

$$\mathrm{kdX} = \frac{\mathrm{StartX+EndX}}{2} \tag{1}$$

$$\mathrm{kdX} = \frac{\mathrm{StartY+EndY}}{2} \tag{2}$$

Where kdX is the x coordinate, kdY is the y coordinate, StartX is the starting point of the X axis in the bounding box, StartY is the starting point of the Y axis in the bounding box, EndX is the end point of the X axis in the bounding box, EndY is the end point of the Y axis in the bounding box.

$$\text{focal length} = \frac{\mathrm{w*d}}{\mathrm{W}} \tag{3}$$

$$\text{distance} = \frac{\mathrm{W*f}}{\mathrm{w}} \tag{4}$$

Where f is the focal length, w is the width in pixels, d is the distance in centimeters, and W is the width in centimeters.

Evaluations are conducted to determine the performance of the developed face and expression recognition system. Performance includes accuracy, as measured by the real-time recognition rate of faces and emotions by the proposed system. The formula for calculating the test accuracy is shown in formula (5):

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{5}$$

TP is true positive, TN is true negative, FP is false positive, and TN is true negative. This formula is used to obtain the accuracy of face recognition or emotion recognition. The calculation of the accuracy value shows the validity level of each class of the classification.

According to the results of the study, the VGG16 model (Model B) outperformed Model A and Model C in face and emotion recognition, with a success rate of 100% and 73%, respectively. At the same time, the minimum average loss of VGG16 was 0.011 (face recognition) and 0.1875 (emotion recognition). Although the training process of model B is longer because it has more layers than other models, its recognition effect is excellent. In contrast, model C, while slightly less accurate (87% for face recognition and 67% for emotion recognition), is faster to train because it has fewer parameters. Finally, models B and C were applied to the face and emotion recognition system of humanoid robots, showing the effectiveness of the two models in practical applications, while the study integrated the face and emotion recognition process into a unified framework.

This research shows that for humanoid robots, recognition can be achieved in real time and the distance between objects and humanoid robots can be measured well. Distance and lighting are also important factors in identification. There is also a model for video processing faces proposed by A.V.Savchenko etc., which, in AffectNet data sets, is about as accurate as the most advanced EfficientNet-B2 method, but its modal robustness is higher than the advanced methods[7]. Therefore, in future studies, in addition to expanding the existing data set, it is also important to consider training under different lighting conditions. For large data sets, the robustness and accuracy of the model recognition under different conditions can be improved. In addition, the proposed system still needs to be repaired and upgraded in future work. This includes training the pair of models to adapt to emerging data, such as the recognition of facial expressions in people with depression. It is also necessary to regularly check and upgrade the hardware of the robot system to ensure the stability and accuracy of the robot's operation.

2.3. Speech Emotion Recognition Based on Deep Learning

Linguistics focuses on exploring the underlying information in speech that represents the state of the speaker or acoustic intermediate, and speech emotion recognition (SER) can aid intelligent human-computer interaction through audio modes. However, the existing deep SER method divides the complete fragment into multiple parts, which inevitably leads to the loss of emotional details in the process. To solve this problem, P.Jiang etc. developed a convolutional RNN with multiple

attention mechanisms in SER, including two modules, CNN and LSTM. The model architecture is shown in Fig 1[8].

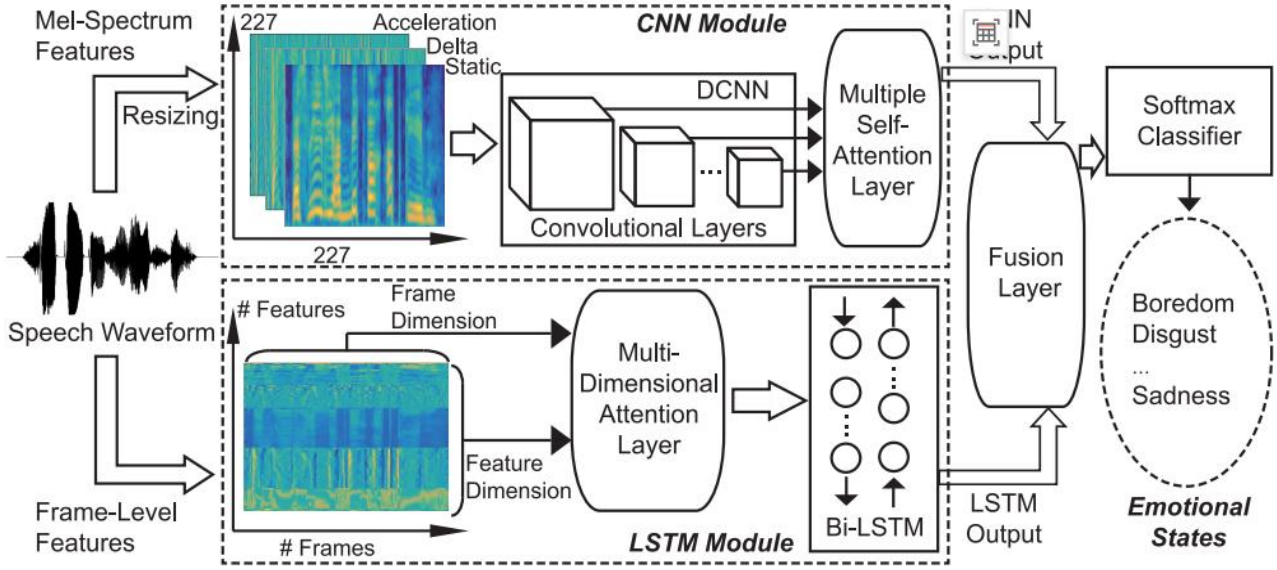

Fig. 1 CRNN-MA model architecture

In the proposed CRNN-MA, multiple self-attention layers are used in the CNN module to obtain rich emotional information by calculating the weights of different frames. Then, a multi-dimensional attention layer is added as input to the LSTM module, taking into account the feature and frame dimensions. Finally, a fusion layer is added to connect the two modules, taking multiple features as input to better fuse advanced features.

Table 2. Unweighted accuracy (%) for the most advanced method, CRNN, and proposed CRNN-MA on the IEMOCAP data-set

Setups	Methods	UA
5-fold CV	ACNN	59.5
	CRNN	59.0
	CENN-MA	60.2
Last Session	STC	60.4
	CRNN	59.2
	CRNN-MA	60.6

To evaluate the performance of CRNN-MA on a large data-set, the experiment used an IEMOCAP data-set containing five sessions (20,243 sentences in total). In this study, 5,531 statements were used, covering four emotional states: happy 1,636, angry 1,103, sad 1,084 and neutral 1,708. The data-set was set up in two session-independent ways: one with a session-independent 5x cross-validation (CV) setup using a focus Convolutional neural network (ACNN), and the other with a spectral time channel attention (STC) approach, where the last session serves as the test set. The experimental results are shown in Table 2, showing that the proposed CRNN-MA method has achieved significant improvement in the performance of emotion recognition (SER) compared with existing large-scale affective data studies

The model includes a parallel structure of CNN and long short-term memory (LSTM) modules, in which the CRNN-MA model proposes multiple attention mechanisms, using multiple self-attention layers, multi-dimensional layers and fusion layers. Compared with the current optimal standard model on the IEMOCAP data set, this model has a small improvement in accuracy, which can effectively improve the performance of SER systems.

3. Multi-modal Emotion Recognition Fusion Based on Deep Learning

3.1. Multi-modal Emotion Recognition Fusion of Dynamic Convolution and Residual Gated

Multi-modal data contains rich emotional information. When human beings communicate through multi-modal information, emotional information is inconsistent in time due to the asynchronicity of language and facial expression. Therefore, Yu et al. proposed a multi-modal emotion recognition model based on dynamic convolution and residual gating [9].

First, the model extracts the low-level features, high-level local features, and context dependencies of each mode. The interaction information within and between modes of text, images, and audio is then modeled. Finally, the weight of each interaction information in the final emotion classification is automatically learned through residual gating, and the multi-modal fusion features are input to the classifier for emotion prediction. In order to automatically learn the weight of each set of interaction representations in the final emotion classification, a residual gated fusion method is designed to effectively merge multiple interaction representations through competition or cooperation.

Table 3. Experimental results on the CMU-MOSEI data-set

Models	ACC-2	ACC-7	F1-value	MAE	Corr
LF-LSTM	80.6	48.8	80.6	61.9	65.9
MCTN	79.8	49.6	80.6	60.9	67.0
MulT	82.5	51.8	82.3	58.0	70.3
RAVEN	79.1	51.0	79.5	61.4	66.2
MFRM	82.4	50.9	82.6	59.8	69.0
Multilogue-Net	82.1		80.3	59.0	
MAG-BERT	82.2		82.6	54.3	76.4
PMR	83.3	52.5	82.6		
MI	83.5	52.8	83.1	56.8	71.2
ΔSOTA	0.2↑	0.3↑	0.5↑	2.5↑	5.2↑

By comparing the proposed model (MI) with the benchmark model, the accuracy of the proposed model in the CMU-MOSEI and IEMOCAP data sets reaches 83.5% and 83.9%, respectively, which is better than Mult, MFRM and other benchmark models. In addition, Table 3 shows that the proposed model also exceeds the current best model PMR in the classification index. Among them, ACC-2 increased by 0.2 percentage points, ACC-7 increased by 0.3 percentage points, and F1-value increased by 0.5 percentage points. At the same time, the mean absolute error (MAE) is 2.5 percentage points lower than MAG-BERT, and the correlation coefficient (Corr) is 5.2 percentage points higher, finally achieving the best-known result.

The multi-modal emotion recognition model based on dynamic convolution and residual gating can fuse the three modes of text, image and audio. Then, the model can extract the low-level features, high-level features and contextual dependencies of different modes, and model the model through cross-modal dynamic convolution. Finally, the model can integrate effectively through residual gating module. This model can avoid the inaccuracy of recognition caused by the inundated information, and improve the accuracy and robustness of the multi-modal emotion fusion analysis system. In the future, this model can be extended to more modes, such as body language, brain wave signals, skin electrical signals, etc., to further improve its scope of application.

3.2. Multi-modal Emotion Recognition Fusion of Dynamic Convolution and Residual Gated

The multi-modal emotion fusion recognition network represented by three Spaces is proposed by S.Zhang. It is a kind of multi-modal emotion recognition technology for the game scene. By recognizing the players' emotions, the multi-modal emotion recognition technology can analyze the

influence of the game plot on the players, so as to further optimize the plot setting of the game and make the plot more scientific [10].

In the model, for the three extracted single-mode feature sequences, different networks are used for vector representation, and a double-layer long short-term memory network is used for vector representation of audio feature sequences and visual feature sequences. The final state hiding representation of the network is combined with a FC Layer. In the process, four kinds of loss functions are used to calculate the loss. Finally, these loss functions are weighted and combined as the loss function of the whole system, and the model is learned by minimizing the overall loss function.

The fusion method proposed by the model is TSRA method, which inputs the fused data into the emotion recognition network. In this model, a stacked long short-term memory network is used for emotion recognition, and the training efficiency and accuracy are improved by adding the depth of the network.

Table 4. Comparison of experimental results of GMSA data-set

Method	Mult_acc_2	Mult_acc_3	Mult_acc_5	F1_score	MAE	Corr
LMF	69.37	54.27	21.23	56.82	59.07	5.45
TFN	79.78	65.62	40.93	78.62	43.62	59.1
TSRA	**80.5**	**66.3**	**40.98**	**80.86**	**43.28**	**59.82**

The experimental results of the multi-modal data fusion method, namely TSRA method, on GMSA data sets are shown in Table 4. By comparing the existing fusion methods (TFN and LMF), the superior performance of TSRA method is proved from the results of various data sets and evaluation indicators.

This method improves the multi-modal emotion recognition network, uses the long and short term memory network to stack, increases the network depth, and improves the accuracy and robustness of the system recognition. However, the current recognition performance is far from human cognition, and the accuracy of recognition is still unsatisfactory. Therefore, it is necessary to continue to strengthen the learning of the algorithm, overcome the problems encountered, and constantly improve the applicability of the method. This method can then add more modes for analysis, such as brain wave modes. Or to find out which modal fusions lead to faster, more accurate results.

3.3. Emotion Recognition and Fusion Based on Multi-granularity and Multi-modal Interaction

In the task of multi-modal emotion recognition in dialogue, a key research problem is how to make full use of multi-modal data flow to realize inter-modal and intro-modal interaction. Therefore, Y. Liu proposed a dialogue emotion recognition model named MGMNet, which adopts a multi-granularity and multi-modal context approach [11].

First of all, a Bridge Transformer is constructed through the cross-modal attention mechanism, which uses text modes as a bridge to capture cross-modal speech and expression information. The model extracts coarse-grained audio and video features by measuring learning for emotion transfer detection.

Next, MGMNet builds a bi-linear fusion network using non-linguistic shift vectors to construct multi-modal fusion features. At the same time, the model introduces the multi-modal feature fusion target TMIB based on the information bottleneck to solve the possible single-modal bias problem, which ensures the effective integration of different modal information.

Finally, MGMNet combines two kinds of different granularity features to form a multi-granularity and multi-modal context information dialogue emotion recognition model, and finally achieve efficient and accurate emotion recognition.

Table 5. Comparison of F1 scores for each category on the IEMOCAP and MELD datasets

Models	IEMOCAP							MELD
	Happy	Sad	Neutral	Angry	Excited	Frustrated	W-Avg F1	W-Avg F1
DialogueTRM	48.7	77.52	**74.12**	66.27	70.24	67.23	69.23	63.55
MMGCN	42.34	78.67	61.73	69.00	74.33	62.32	66.22	58.31
M2FNet	60.00	82.11	65.88	68.21	72.60	**68.31**	69.86	66.71
MM-DFN	42.22	78.98	66.42	**69.77**	75.56	66.33	68.18	59.46
EmoCaps	**71.91**	**85.06**	64.48	68.99	78.41	66.76	71.77	64.00
UniMSE							70.66	65.51
MGMNet	65.22	79.75	71.20	66.07	80.25	67.96	72.33	66.95

Taking the experimental results in Table 5 on IEMOCAP as an example, the proposed MGMNet achieves 72.33% performance on the IEMOCAP data-set, which improves the performance by 0.56% compared with the previous optimal model. In addition, MGMNet's seven classification accuracy on MELD also reached the best level. Although in the six IEMOCAP categories, only the Excited emotion achieves the best performance, in comparison, the accuracy of this model remains high in other emotion categories, and only the Anger emotion performs poorly. In addition, the performance of Happy emotion in the model is much different from that of EmoCaps. According to the analysis, this is due to the small number of data samples in this category, which makes it difficult for the model to summarize this kind of emotion.

This approach focuses on context modeling and multi-modal fusion in dialogue. The model incorporates coarse-grained and fine-grained features and uses cross-modal attention mechanism and bi-linear fusion to improve emotion recognition results. However, due to the sample problem in the data set, the model training accuracy is reduced, so the model needs to adapt to the training data set of short samples to improve the accuracy of recognition. Due to the ambiguity of human emotion, the performance of the algorithm should be taken into account in the follow-up research. Therefore, researchers should model the subjective emotional tendency of speakers in the data set from the perspective of personalized modeling. Thus the system can be applied to human-machine interaction only

4. Conclusion

In this paper, we introduce the current high-accuracy and stable system of text, face, and speech single-mode emotion recognition. For multi-modal emotion recognition fusion, this paper introduces three methods with high fusion rate and recognition rates. Single-mode emotion recognition based on deep learning is the key to the fusion of multi-mode emotion recognition. The accuracy and robustness of these deep learning methods are almost higher than the current standard model. However, there are research on body gesture recognition and its fusion with other modes. Moreover, the three-modal fusion of body gesture with text, speech and image lacks accurate data support. In addition, research articles on four-mode signal fusion such as text, speech, image, and EEG are also scarce. For the future, the collaborative recognition of faces and limbs is a direction worth exploring, and there is a shortage of large-scale, high-quality data sets in the process of three-mode fusion. For four-mode fusion, how to improve its accuracy and robustness is an important topic for future research.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Liu Y., Ai H., Zhang W. Multiple modal emotion recognition based on deep learning review. Journal of Xi'an University of Posts and Telecommunications, 2022, 27 (01): 60-71, 95.

[2] Wu J., Li W., Zhang Q., et al. Multimodal affective dialogue technology: Research review and development trend. Artificial Intelligence, 2024 (03): 45-56.

[3] Liu W., Qiu J.-L., Zheng W.-L., Lu B.-L. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14 (2): 715-729.

[4] Cheng D., Zhang D., Chen Y. Multimodal emotion recognition. Journal of Southwest University for Nationalities (Natural Science Edition), 2022, 48 (04): 440-447.

[5] Goyal R., Chaudhry N., Singh M. Personalized emotion detection from text using machine learning. In: 2022 3rd International Conference on Computing, Analytics and Networks (ICAN), Rajpura, Punjab, India, 2022, pp. 1-6.

[6] Dwijayanti S., Iqbal M., Suprapto B. Y. Real-time implementation of face recognition and emotion recognition in a humanoid robot using a convolutional neural network. IEEE Access, 2022, 10: 89876-89886.

[7] Savchenko A. V., Savchenko L. V., Makarov I. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. IEEE Transactions on Affective Computing, 2022, 13 (4): 2132-2143.

[8] Jiang P., Xu X., Tao H., Zhao L., Zou C. Convolutional-recurrent neural networks with multiple attention mechanisms for speech emotion recognition. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14 (4): 1564-1573.

[9] Liu W., Qiu J.-L., Zheng W.-L., Lu B.-L. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14 (2): 715-729.

[10] Suk. Game user oriented analysis of modal emotion recognition method research. University of Electronic Science and Technology, 2022.

[11] Liu Y. Multimodal scenario dialogue emotion recognition research. Huazhong University of Science and Technology, 2023.