A Survey of Multimodal Emotion Recognition

Liu Xinyu
Huazhong University of Science and Technology, Wuhan City, Hubei Province, China
2714976804@qq.com

ABSTRACT

Multi-modal emotion recognition refers to the method of recognizing human emotions in a more accurate way through the fusion of multiple signals, such as speech, text, expression, etc., which contributes more to the field of human-computer interaction and is also a more advanced direction in the field of artificial intelligence. The research of multimodal emotion recognition starts with the development process of speech emotion identification, combs and introduces the research situation of multimodal emotion recognition, introduces two emotion expression theories, and then introduces the relevant databases of emotion identification according to this classification. Then, the common methods of multimodal emotion recognition are analyzed from three dimensions: feature extraction, motif fusion and emotion classification. Finally, this paper summarizes the applications, challenges, and future directions of multimodal emotion recognition reference for other researchers in this field.

Keywords: Multimodal; Emotion Recognition; Intelligent Interaction; Modal Fusion.

1. TOWARD EMOTION-AWARE INTELLIGENT INTERACTION

The importance of recognition technology to judge human emotion through speech, text, expression and other information will gradually increase. Emotion is the core component of human psychological activities and plays a decisive role in individual activities, interpersonal communication and social life. Speech has always been the main way to express information and convey feelings, and is one of the most widely used modes of communication in people's daily life [1]. It has more auxiliary language information and is also an important material for emotion recognition technology [2].

In 1995, Professor Picard of MIT established "emotion computing" as a new interdisciplinary subject in computer science with a pioneering perspective, laying a theoretical foundation for machine understanding of human emotions. Back in the 1980s, scholars represented by Bezooijen pioneered the exploration of speech emotion features, and built an early emotion classification framework through traditional machine learning methods such as Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM). Even so, limited by the surface mining of feature engineering, these methods struggle to capture deep semantic associations of emotion expressions. With exponential increases in computational power, deep learning injects new momentum into emotion recognition. In 2013, Li's team first introduced deep neural networks (DNNs) into the field, enabling automatic feature extraction through multi-layer nonlinear transformations. Convolutional neural network (CNN) and long-short term memory network (LSTM) are used to improve recognition accuracy by leaps and bounds. Input features are gradually upgraded from traditional acoustic parameters to more representative Mel Frequency Cepstrum Coefficient (MFCC) and spectrogram. These frequency domain features can more accurately describe the formant distribution and energy change of speech. In 2017, Mirsamadi et al. proposed a recurrent neural network (RNN) architecture based on the attention mechanism, marking the new stage of emotion computing. This mechanism enables the model to favor the key frames of emotion expression through dynamic weight allocation, effectively solving the information attenuation problem in long sequence modeling. From rule-driven to data-driven, from manual features to automatic representation, speech emotion recognition technology is evolving along the path of "perception-understanding-empathy," opening up broad prospects for building human-computer interaction systems with true emotional intelligence.

Even so, when emotion recognition is carried out, the emotional information carried by the single mode of speech is limited, and the accuracy is limited. In the modes of posture, expression, text, etc., more emotions are contained when speaking, among which the relationship between text and sound is the closest.

2. EMOTIONAL REPRESENTATION THEORY

Emotion has the characteristics of diversity, subjectivity and timeliness, which affects people's behavior and choice all the time, and is one of the unique characteristics of human thought.

2.1 Discrete Emotion Theory

Discrete-emotion theory treats each emotion as a separate entity, arguing that complex emotions are simply a combination of basic emotions, which vary in number from theory to theory. In this case, most people agree with Ekman and Friesen's six basic emotions, "happiness, sadness, anger, fear, disgust, and surprise," because they behave much the same across cultures.

2.2 Dimensional Emotion Theory

The theory of dimensionality is a new way to visualize abstract emotions as vectors in multidimensional space. Under this theoretical framework, every point accurately corresponds to a specific emotion, and the subtle differences between different emotions are skillfully represented by the distance and angle between vectors. In the study of emotion space, two-dimensional and three-dimensional space have attracted much attention. For example, PAD emotion space proposed by Russell and Mehrabian constructs a three-dimensional system based on "pleasure degree-arousal degree-dominance degree", Schlosberg's inverted conical three-dimensional emotion space adopts "pleasure degree-attention degree-activation level", Russell's improved annular emotion space focuses on "pleasure degree-intensity", Plutchik's parabolic cone emotion space is divided according to "intensity-similarity-polarity". In addition, some researchers boldly explore higher dimensions, such as Krech constructed a four-dimensional emotional space, Frijda even proposed a six-dimensional emotional space, opening up a broader world for emotional research.

3. EMOTION RECOGNITION DATABASE

Deep learning is greatly affected by its dependency, so speech emotion recognition database plays an important role in the process. The accuracy of emotional materials often directly determines the accuracy of the final conclusion. According to the emotion theory above, this paper introduces the commonly used emotion recognition database into two categories, in which discrete emotion database is labeled with individual basic emotions, while dimensional emotion database is labeled with emotion coordinates in dimensional space.

3.1 Discrete Emotion Database

CASIA [4] is a Chinese speech emotion database built by the Institute of Automation of Chinese Academy of Sciences. 9600 voices output by 2 professional actors and 2 professional blogs in the database contain 6 emotions, namely anger, fear, joy, sadness and surprise.

RSSDESS [5] is an audio-visual emotional database created by Anthony Emotional Voice and Song Database Center of Ryerson University in Canada. The database includes 12 professional actors and 12 professional spokesmen. According to the expression of two words, they are expressed in the form of spoken and singing, using neutral North American sounds.

CHESSD [6] Chinese multimodal emotion database is constructed by Institute of Automation, Chinese Academy of Sciences. The database consists of emotion fragments extracted from movies, TV plays and talk shows, and contains 6 basic emotions of 238 people.

ENETFACE '05' [7] is an audio-visual emotion database created by the Enterface project team. In this database, 24 subjects of different nationalities listened to 6 consecutive short stories, and 81% of men and 19% of women elicited specific emotional expressions.

3.2 Dimension Affection Database

IEMOCAP [8] is a multimodal emotion database created by the Speech Analysis and Interpretation Laboratory at the University of Southern California. In this database, five professional actors and five professional actresses contributed about 12 hours of audio-visual data through improvisation or script interpretation. In particular, each sentence of each conversation was strictly labeled, covering nine discrete emotion categories and three dimensions of valence, arousal, and dominance.

VAM [9] is a German multi-modal emotion database extracted from German TV talk show Veraam Mag. The database contains 947 voices from 12H recordings, output by 11 men and 36 women. Audiovisual signals and faces are also labeled according to valence, arousal, dominance, etc.

RECOCPOK [10] is a French database of multimodal spontaneous collaboration and emotional interaction, built by the University of Verdeburg, Switzerland. During video conference, images, voices, electrocardiograms and skin electrical signals of 46 people in the database were recorded through 2 emotional dimensions and 5 social behavior dimensions.

CMU-ERSEI [11] Carnegie Mellon University's emotional database of multiple patterns extracted from Youtuic video reviews, which contains 65H gender-labeled videos with 6 discrete tags and 7 dimensional tags.

4. MULTIMODAL EMOTION RECOGNITION METHODS

Single modality cannot form a comprehensive scene representation, can only intercept the information in the complex reality of the side, the expression of incomplete emotion. Multi-modal information can form complementary, enhance the use of multi-modal emotion recognition rather than single mode, through context comparison, resolve ambiguity in a single template, improve the accuracy of understanding.

4.1 Speech Feature Extraction

In the current research of speech emotion recognition, the commonly used speech emotion features are mainly classified into three categories: prosody features, speech quality features and spectral features. Speech quality features use formants, harmonic noise ratio (HNR), jitter and low light to carefully evaluate the purity and clarity of speech. Spectral features generally use Mel Frequency Cepstrum Coefficient (MFCC) to capture the power spectrum of speech on a nonlinear Mel scale to simulate the response pattern of the human auditory system. Tools such as COVAREP, OpenSMILE, pyAudioAnalysis, Librosa, OpenEAR are widely used to manually label these features. However, manual feature extraction requires a high level of expertise from researchers and covers a relatively limited range of emotions. This is why deep learning techniques that have emerged in recent years have become increasingly popular among researchers. Schneider et al. proposed wav2 vec in 2019, which obtains a wider range of speech features by performing unsupervised training on unlabeled speech features. By 2020, Baevski et al. introduced the Transformer model on this basis and further introduced wave2vec2.0.

4.1.1 Text Feature Extraction

Traditional methods usually use the bag-of-words model (BAGGSBow) [12] to extract text emotion features, in which case the connections between words cannot be well captured in context. Later, a word embedding model is introduced to solve this problem, such as Ship-Grams or Word2uc with continuous bag-of-words [13] to consider more complex semantics and syntax. To further break the limitations of previous models, Peters et al. [14] proposed ELMO using deep bi-directional LsTMBILnTM networks

(Embeddings From LuageModels) a more in-depth consideration of context to solve problems such as polysemy.

4.2 Multimodal Fusion

The complementarity of different modalities can be more comprehensive, so the multi-modal recognition model can have better advantages than the single-modal emotion recognition model. The structure, time and semantics should be aligned in the fusion.

4.2.1 Early Fusion

Early fusion usually concatenates the characteristics of each modality into a relatively unified form after data preprocessing, and then classifies emotions according to this, so it is also called feature-level fusion. This method enables models rather than humans to process the association and alignment between different modalities, so that the self-learning ability of the patterns is better utilized, thus improving the relationship between transmission efficiency and capturing the basic features of multiple modalities in depth. Poria et al. [15] used this approach to study the eigenvectors of multiple modalities, laying the foundation for future research. With the introduction of deep learning and neural networks, bidirectional long-term memory model (BILSTM) networks were added to multimodal fusion systems, such as Williams et al. [16]. After that, the self-determination mechanism of TranSFormer was adopted by Tsai et al. Mult TFN is proposed for more detailed modeling Models such as TensorFusion Network, e.g., [17], can model multimodal interactions, but in this case the computational complexity is increased. Although the early fusion has some flaws relative to pure fusion, for example,

redundant processing during fusion may remove too little to bring about a decline in vehicle performance, and key information may be omitted because too much is removed.

4.2.2 Late Fusion

Late fusion generally identifies and processes various modal information first, and then integrates them together for emotion classification according to weighted judgment, so it is also called decision-level fusion. (TSignclassFusion). Each modality in this fusion method is not affected by each other, and the robustness is stronger. Even so, the interaction correlation degree between modalities is less likely to affect each other. Poria et al. established an emotion recognition model as early as 2016 [18], which treats multiple modalities separately and finally identifies them by weighted integral integration. After that, recurrent neural network was introduced by Huang et al. [19], LSTM was applied to emotion prediction model, and support vector regression (SVR) was used for integration at decision level. In order to improve Sun et al. [20] to add self-attention mechanism, a two-layer LSTM model was used to better capture temporal dependence (Reartrix).

4.2.3 Blended Mix

Hybrid fusion, as a comprehensive multi-modal fusion strategy, refers to the fusion operation in both feature extraction stage and final decision stage, which skillfully integrates the advantages of early fusion and late fusion, and can not only extract multi-modal information comprehensively, but also give full consideration to the interaction and internal relations among various modalities. Take Wöllmer et al. [23] for example, they skillfully fuse the information of vision and speech modes with the help of bidirectional long-term memory network (BLSTM) in the feature extraction stage, and then fuse with the information obtained from text mode in the decision-making stage to complete emotion classification. In recent years, with the vigorous rise of large models, some people have introduced them into the field of hybrid fusion. Lea et al. [24] use modal alignment at feature level fusion to accurately map feature vectors of each modality to text feature space of large language model (LLM) for subsequent processing. At decision level, LLM fine-tuned on multitask emotion instruction (M^2SE) is used to deeply fuse and comprehensively consider information to finally generate accurate answers. Even though hybrid fusion is not perfect, it inherits the disadvantages of both early fusion and late fusion. Its complexity and high requirements for models are undoubtedly a challenge, and there is still much room for improvement in the future.

4.3 Sentiment Classification

Choosing the appropriate emotion classification algorithm to judge and make decisions about the emotions transmitted to us is the final processing in emotion recognition. The sizes displayed in different datasets are different, and the environment is also different. For example, traditional machine learning methods such as simple Bayes, support vector machines, and logistic regression usually use feature engineering combined with classifier models, which are better at small data volumes. Recurrent neural network (RNN) and its variant (LSTMGRU) in deep learning avoid manual feature extraction for scenes with large data volume and can better capture context. Random forest in ensemble method [21] has strong performance and low requirement for features, even so, the training speed is slow. K-Nearest NeighborskNN [25] Lazy learning K nearest neighbor (K-NoueligeysKnn) does not need training, the principle is simple but the prediction calculation cost is too high. Kastraconajan et al. [22] adopt double-layer graph scroll network (GCN) in application, and combine fully connected layer and Soft-Ox classifier, which has significantly improved performance than traditional models.

5. EMOTION RECOGNITION APPLICATIONS

Emotion recognition has been instrumental in areas such as medical care, safety, and everyday interactions. For example, emotional labeling systems can more intuitively determine a patient's current mood when treating mental disorders such as depression, helping doctors better formulate treatment plans. Technology can identify a driver's current mental state and then alert or dissuade those who should not continue driving to reduce traffic accidents. In addition, emotion recognition makes intelligent items such as voice assistants more accurate for users, while unmanned service systems such as intelligent customer service can also provide faster and more convenient help for people's daily life.

6. CHALLENGES AND PROSPECTS

Although multimodal emotion recognition has made certain progress at this stage, and deep learning technology has been applied to human-computer interaction, natural language processing, etc., even so there are still many shortcomings and deficiencies. For example, emotion data is now collected in multimodal fields. The integration method is not sufficient, and the diversity itself also has many imperfections.

6.1 Multimodal Emotion Dataset

Emotion is a complex and rich single modality to define emotion, which inevitably leads to losses and errors. Even so, emotion datasets containing multiple motifs have fewer deadlier template datasets, such as physiological signals.

6.2 Multimodal Fusion Method

Now there is no perfect fusion model, which can perfectly transform different models into the same feature dimension for decision-making. The different methods and ideas of each mode in this model will inevitably affect transfer learning when the mode is fused. The processing and judgment of various modal features by Model in transfer learning is precisely the most important part of MOdL effect. In this case, the mainstream deep learning interpretability is more applicable, which is also a big problem of MordEL fusion.

7. CONCLUDING REMARKS

In this paper, the multimodal emotion recognition is discussed systematically. Firstly, the differences of different emotion expression models are clarified, and the corresponding multimodal emotion database is sorted out. Then, the specific process and method of multimodal emotion recognition are explained in detail, which are analyzed in three stages: feature extraction, modal fusion and emotion recognition. Finally, the practical application of multimodal emotion recognition in various fields of life is introduced. Finally, the study analyzes the challenges faced by multimodal emotion recognition, and informs the prospects of the future development of multimodal emotion recognition.

REFERENCES

[1] Crystal, D. Non-segmental phonology in language acquisition: A review of the issues. Lingua, 1973,32(1-2):1−45.
[2] Sun Xiaohu, Li Hongjun. A Review of Speech Emotion Recognition [J]. Computer Engineering and Application, 2020, 56 (11): 1-9.
[3] EKMAN P. Facial expressions of emotion: new findings, new questions[J]. Psychological Science,1992, 3(1): 34-38.
[4] Tao JH, Liu FZ, Zhang M, et al. Design of speech corpus for mandarin text to speech. Proceedings of the Blizzard Challenge 2008 Workshop. Hyderabad, 2008. 1-4.
[5] Livingstone SR, Russo FA. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS One, 2018, 13(5): e0196391.
[6] LI Ya, TAO Jianhua, CHAO Linlin, et al. CHEAVD: a Chinese natural emotional audio–visual database[J]. Journal of Ambient Intelligence and Humanized Computing,2017,8(6):913-924.
[7] Martin O, Kotsia I, Macq B, et al. The eNTERFACE'05 audio-visual emotion database. Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06). Atlanta: IEEE, 2006. 8.
[8] Busso C, Bulut M, Lee CC, et al. IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2008, 42(4): 335-359.
[9] GRIMM M, KROSCHEL K, NARAYANAN S. The Vera am Mittag German audio-visual emotional speech database[C]Proceedings of the 2008 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE,2008:865-868.
[10] RINGEVAL F, SONDEREGGER A, SAUER J, et al. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions[C]Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. Piscataway: IEEE, 2013: 1-8.
[11] BAGHER ZADEH A, LIANG P P, PORIA S, et al. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph[C]Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers. Stroudsburg: ACL, 2018: 2236-2246.
[12] Soumya George K., Joseph S. Text classification by augmenting bag of words (BOW) representation with co-occurrence feature. IOSR J. Comput. Eng. 2014; 16:34-38.
[13] LIU Y, AI H, ZHANG W D. A survey of multimodal emotion recognition based on deep learning[J]. Journal of Xi'an University of Posts and Telecommunications, 2022, 27(1): 60-71.
[14] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1(Long Papers). Stroudsburg: ACL, 2018: 2227-2237.
[15] Poria S., Cambria E., Hussain A., Huang G.B. Towards an intelligent framework for multimodal affective data analysis.

Neural Netw. 2015; 63:104-116.
[16] Williams J., Kleinegesse S., Comanescu R., Radu O. Recognizing emotions in video using multimodal dnn feature fusion; Proceedings of the Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML); Melbourne, Australia. 20 July 2018; pp. 11-19.
[17] Zadeh A, Chen M H, Poria S, Cambria E and Morency L P. 2017. Tensor fusion network for multimodal sentiment analysis [EB/OL]. [2025-04-08]
[18] Poria S., Cambria E., Howard N., Huang G.B., Hussain A. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing. 2016; 174:50-59.
[19] Huang J., Li Y., Tao J., Lian Z., Wen Z., Yang M., Yi J. Continuous multimodal emotion prediction based on long short-term memory recurrent neural network; Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge; Mountain View, CA, USA. 23 October 2017; pp. 11-18.
[20] Sun L., Lian Z., Tao J., Liu B., Niu M. Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism; Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop; Seattle, VA, USA. 16 October 2020; pp. 27-34.
[21] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45 (1): 5-32.
[22] Devarajan, K., Sivan, R., & Poopalaratnam, T. (2025). Enhancing emotion recognition through multi-modal data fusion and graph neural networks. Intelligence-Based Medicine, *9*, 100138.
[23] Wöllmer M., Weninger F., Knaup T., Schuller B., Sun C., Sagae K., Morency L.P. Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intell. Syst. 2013; 28:46-53. DOI: 10.1109/MIS.2013.34.
[24] Lea, L., et al. EmoVerse: Enhancing Multimodal Large Language Models for Affective Computing via Multitask Learning[J]. Neurocomputing, 2025, 650: 130810.
[25] COVER T, HART P. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.