Research Advanced in Multimodal Emotion Recognition Based on Deep Learning

Weiqi Kong

School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, Jiangxi Province, 330022, China

* Corresponding Author Email: lideng323@tzc.edu.cn

Abstract. In summary, the field of computer science has long been intrigued by emotion recognition, which seeks to decode the emotional content hidden within data. Initial approaches to sentiment analysis were predominantly based on single-mode data sources like textual sentiment analysis, speech-based emotion detection, or the study of facial expressions. In recent years, with the increasingly abundant data representations, Many people has gradually pay attention on the multimodal emotion recognition. Multimodal emotion recognition involves not only text, but also audio, image, and video, which is of great significance for enhancing human-computer interaction, improving user experience, and improving emotion-aware applications. This paper thoroughly discusses the research advancements and primary techniques of multimodal emotion recognition tasks, with an emphasis on the aforementioned tasks. Specifically, this paper first introduces representative methods for single-modal emotion recognition based on graphic data, including its basic process, advantages and disadvantages, etc. Secondly, this article introduces pertinent studies on multimodal emotion recognition and offers a quantitative comparison of how different approaches perform on standard multimodal data sets. Lastly, it addresses the complexities inherent in multimodal emotion recognition research and suggests potential areas for future study.

Keywords: Emotion recognition; deep learning; multimodal.

1. Introduction

Multimodal emotion recognition is a fascinating research direction in the field of computer science and natural language processing, holding significant relevance in the era of digital information exchange and intelligent systems. Traditional emotion recognition research primarily relies on unimodal data sources, such as text sentiment analysis, speech emotion recognition, or facial expression analysis. However, in today's multimedia-rich social media platforms and smart devices, people communicate not only through text but also express emotions through various modalities like audio, images, and videos. Hence, understanding and analyzing emotions in multimodal data has become a crucial task, profoundly impacting the enhancement of human-computer interaction, improved user experiences, and the refinement of emotion-aware applications.

Human communication and decision-making are significantly influenced by emotions. They not only convey information, but also influence people's decisions, behaviors, and social interactions. For example, systems that are capable of recognizing emotions can be utilized in contexts such as sentiment analysis. They enable the monitoring of user's feelings on social media, thereby offering a deeper insight into the public's emotions regarding certain events or products. Additionally, emotion recognition can find applications in the medical field by assisting doctors in gaining insights into the psychological states of patients through the monitoring of their voice, expressions, and text. Early research on emotion recognition focused on emotion prediction using three single-modal information such as text, speech, or vision (face). When using unimodal information for emotion recognition, the source of information comes from a single modality, so there are deficiencies in some cases. For example, when the amount of unimodal data is small, the training of the network may appear overfitting. Not only that, but sometimes unimodal data can also even provide wrong information, which can affect the final prediction results. Therefore, it is particularly necessary to carry out research on multimodal emotion recognition.

Although multimodal emotion recognition helps to mine the potential relationship between different data to improve recognition accuracy, it also presents a series of challenges. Firstly, the diversity and complexity of data across different modalities make emotion recognition a more demanding task. For example, text can contain rich semantic information, while images convey abundant visual features such as facial expressions and body language. Secondly, inconsistencies may be present among various information patterns, thus it's essential to develop suitable methods for integrating this information to enhance the accuracy and sturdiness of emotion recognition. Lastly, multimodal emotion recognition requires deep learning techniques to efficiently process data from multiple sources and fuse them to produce comprehensive emotional analysis results.

The primary objective of this study is to investigate the obstacles encountered in multimodal emotion recognition and offer solutions based on deep learning methods. In the subsequent sections, we will examine in depth the specific issues related to multimodal emotion recognition and present the pioneering deep learning strategies we developed to overcome these issues. By integrating text, audio, and image data, our research aims to advance emotion recognition, making it more intelligent and adaptable to the complexity of multimodal data.

2. Unimodal Emotion Recognition Research

2.1. Facial Expression-Based Emotion Analysis

Significant advancements have been achieved in the field of unimodal emotion analysis, within the broader scope of emotion recognition research. These studies primarily focus on three main modalities: facial expressions, text, and audio. These investigations serve as crucial foundations for multimodal emotion recognition. The goal of emotion analysis is to understand and delineate human feelings by scrutinizing methods through which people express these emotions, like facial expressions. People differ in how they communicate their emotions - some primarily use their language, enhancing the emotional tone of their speech, others predominantly use their facial expressions, incorporating more emotional data into their facial features. This diversity adds complexity and diversity to unimodal emotion analysis.

In daily interactions, facial expressions play an important role in conveying emotions, hence their crucial importance in emotional analysis. Facial expression recognition (FER) systems, depending on the various feature representations, can be broadly split into two types: those that rely on static images and those that deal with dynamic sequences. Dynamic sequence-based FER encompasses temporal dynamics, including spatial-temporal characteristics and saliency of expressions. In the dynamic sequence-based Facial Expression Recognition (FER), accurately capturing emotional changes necessitates the consideration of the facial expressions' temporal sequence. To address these challenges, researchers have proposed methods based on spatial-temporal attention networks. By integrating corresponding attention modules into spatial and temporal sub-networks, these methods boost the efficiency of feature extraction for convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

The typical process of recognizing facial expressions includes three main steps: identifying the face, selecting and extracting features, and classification. The techniques used can be classified into traditional methods and methods based on deep learning, depending on the type of feature representation used. Various features are deployed by traditional methods of facial expression recognition, such as geometric, appearance, statistical, and motion features. The processes of geometric feature methods revolve around the identification and interpretation of facial landmark shapes and positions, collecting geometric details of facial expressions. On the other hand, the appearance feature method primarily concentrates on identifying facial texture attributes, like the local binary mode (LBP) and Gabor wavelet features in the frequency domain. Conversely, statistical feature methods aim to capture the comprehensive statistical data in facial imagery, incorporating techniques such as Principal Component Analysis (PCA) and the Independent Component

Correlation Algorithm (ICA). Motion feature methods involve capturing motion information in dynamic image sequences, with optical flow methods being commonly used.

Deep learning techniques have been at the forefront of recent progress in facial expression recognition. Deep learning systems like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been extensively utilized to enhance the precision of recognizing facial expressions.For example, CNNs can automatically learn expressive features from images, while RNNs can handle temporal information in dynamic sequences, making them highly effective in sequence-based emotion analysis.

Deep learning models not only can automatically learn features but can also optimize model performance to the fullest extent through end-to-end training. By minimizing the necessity for manual feature engineering, this method makes the system more adaptable. As a result, significant advancements in facial expression recognition have been achieved through deep learning.

2.2. Page Numbers

The two main strategies primarily utilized in traditional text sentiment analysis methods are dictionary-based methods and supervised machine learning models. While these methods were widely used in the early days, they have certain limitations, including the heavy reliance on human effort and time, as well as suboptimal performance when dealing with large-scale data. Here's a brief overview of these two approaches: Lexicon-based methods: These methods rely on manually constructed sentiment lexicons that associate words or phrases with emotions. In this method, words from a given text are cross-referenced with a lexicon, and a sentiment polarity (e.g., positive, negative, or neutral) is assigned based on the results of this correlation. But there is a prevalent requirement for continuous updates in light of ever-changing language and cultural dynamics, leading to a labor-intensive process of lexicon upkeep. An alternative approach is through supervised machine learning models, where classification models are built using training data which already possess sentiment labels. Naive Bayes, support vector machines (SVMs), and decision trees are among the typical algorithms used in machine learning. These models can automatically learn the relationship between text features and emotions. Nonetheless, they necessitate a significant quantity of labeled data and intricate feature engineering tasks, which in large-scale data situations can be complicated and lengthy.

In the field of text sentiment analysis tasks, deep learning methods have gained substantial success. Deep learning models can automatically learn text features without the need for manual sentiment lexicon construction or complex feature engineering. Here are some aspects of deep learning's application in text sentiment analysis: Word Embedding Models: In deep learning, text data is often initially encoded into word vectors using word embedding models such as Word2Vec and BERT. These models map each word to continuous vectors in a high-dimensional space, capturing semantic relationships between words. These word vectors can serve as inputs to subsequent deep learning models. Deep learning models known as Recurrent Neural Networks (RNNs) are frequently utilized for managing sequential data. In text sentiment analysis, RNNs can capture temporal information in text sequences. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) variants have demonstrated superior results in tasks related to text sentiment analysis. However, RNN models may encounter issues like gradient disappearance or gradient explosion when processing long sequences, requiring careful handling in practice.

The strengths of deep learning lie in its end-to-end training capability and its ability to automatically learn high-level features, making it a powerful tool in text sentiment analysis tasks. However, when applying deep learning methods, attention to details such as data preprocessing, model selection, and hyperparameter tuning is crucial to achieve optimal performance.

3. Literature References

Two significant research areas in multimodal emotion recognition are feature representation and multimodal emotion fusion. A good trait representation should capture rich emotional cues that can be generalized across different speakers, contextual, semantic content, etc. A good fusion mechanism should be able to effectively integrate various modal data. This section will present notable methods for multimodal emotion recognition, with an emphasis on the two aforementioned directions.

3.1. Multimodal Feature Representation

Commonly used multimodal representation learning strategies include joint representation and collaborative representation. Incorporating single-modal signals into a unified representation space is a principal function of joint representation, with the most emblematic method being the direct joining of single-modal attributes. Existing research, influenced by the impressive attribute representation capabilities of neural networks, employs these networks to represent the characteristics of text, visual, and auditory data. This approach is extensively utilized in multimodal emotion recognition. Mai et al. introduced a novel adversarial encoder-decoder-classifier structure to establish a joint modality-invariant embedding space [1]. In light of the distinct distribution characteristics across different modalities, this research additionally used adversarial training to diminish the modality gap, converting the source modality distributions to match the target modality distributions. Furthermore, additional constraints are imposed on the embedding space by introducing reconstruction loss and classification loss. Zhao et al. [2] proposed a pre-trained model MEmoBERT for multimodal emotion recognition, which learns multimodal joint representations from a large amount of unlabeled video data through self-supervised learning.

Instead of merging different modalities into a shared space, collaborative representations acquire separate representations for each modality, guided by a specific constraint. Cooperative representations are mainly classified into similarity and structured coordination space models. In the collaborative space, the similarity model reduces the gap between modalities. Fu et al. [3] utilized an improved method known as sparse local discriminant canonical correlation analysis to learn shared representation of multiple modalities. The En-SLDCCA method was used to extract correlation coefficients from audio and video modalities, and these coefficients were then employed to create a fused representation of audio and video features. A structured collaborative representation model enforces additional constraints between modal representations. Zhang et al. demonstrate that the Canonical Correlation Analysis (CCA) method [4], for instance, calculates a linear projection that amplifies the correlation between two random variables while maintaining the orthogonal relationship within the newly created space. Zhang et al. [5] introduced deep canonical correlation analysis (DCCA) into multimodal emotion recognition. The basic idea of DCCA is to transform each modality separately and coordinate different modalities into A multidimensional space.

3.2. Multimodal Feature Fusion

Combining data from various modalities is a crucial phase in multimodal assignments. The three primary tactics are fusion at the feature level, decision level, and model level. In feature-level fusion, the characteristics of different modalities are integrated using concatenation and other techniques, as demonstrated by Zadeh and his team.[6] proposed a Tensor Fusion Network (TFN) to fuse multimodal information using the product of multimodal features. But this will greatly increase the dimension of features, making the model too large to train. Zeng et al. [7] proposed a data-driven multiplicative fusion technique to fuse multiple modalities. During the training process, it detects the modalities and filters out invalid emotional features, thus learning more reliable emotional cues.

Zadeh et al. found that decision-level fusion independently extracts, classifies, and fuses the features of each modality after having obtained local decision results, creating a final decision vector. This is much more straightforward and flexible than feature-level fusion, as decision results from each modality typically share the same definition.[8] proposed a dynamic fusion graph (DFG)

technique to fuse multiple modalities. DFG can learn the interactions between n-modalities and the effective number of parameters, which is highly interpretable. In contrast to feature-level fusion and decision-level fusion, model-level fusion is superior in comprehending the multimodal interaction within the model and optimizing the benefits of deep neural networks. A conditional attention fusion model was proposed by Chen et al. [9], utilizing Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) as the primary unimodal model for capturing long-term dependencies. The weights assigned to different modalities are automatically determined by current input features and recent historical information, improving traditional fusion strategies by dynamically focusing on different modalities at each time step. Huang et al. [10] used the Transformer model to learn the semantic association between the two modalities of voice and video and achieved model-level fusion for continuous emotion recognition.

4. Common Multimodal Dataset

In the realm of multimodal emotion recognition, several datasets have gained prominence, facilitating comprehensive research in both bimodal (text and images) and trimodal (text, images, and audio) settings. Among the bimodal emotion datasets, the Yelp dataset stands out, comprising user-generated restaurant reviews with associated textual content and images, serving as a valuable resource for studies in bimodal emotion analysis. The Twitters dataset, on the other hand, features tweets collected from Twitter, encompassing textual content and related images, making it a popular choice for analyzing sentiment expressions in social media. Additionally, the multi-ZOL dataset includes posts gathered from various ZOL technology product forums, providing both textual content and corresponding images, thus catering to sentiment analysis requirements, particularly within the technology product domain.

The main features of each dataset, including the number of chapters, sentiment labels, emotion labels, modality labels, and sentiment scoring, are condensed in Table 1. Here, "T" stands for text, "I" signifies images, and a "-" means that the specific characteristic is either not applicable or not present in the dataset.

The CMU-MOSEI (CMU Multimodal Opinion Sentiment and Emotion Intensity) dataset stands out in the field of trimodal emotion datasets. It is made up of video clips from YouTube which contain not only textual content but also audio and facial expression video data. Researchers widely employ this dataset for diverse multimodal analyses, encompassing emotion recognition, emotion intensity estimation, and emotion polarity assessment. The CMU-MOSI dataset mirrors CMU-MOSEI in terms of content, offering video clips from YouTube with textual, audio, and video data. It is frequently utilized for tasks such as multimodal sentiment recognition and emotion intensity prediction. The YouTube dataset, which includes a significant amount of comment text, along with accompanying facial expressions and audio data, is often preferred for the research in sentiment analysis, emotion detection, and multimodal emotion study. On the other hand, the ICT-MMMO (ICT-Multimodal Multimodal Movie Corpus) dataset, containing subtitles, audio, and video information pulled from film clips, is often utilized for emotion recognition and sentiment analysis specifically in the context of films. The IEMOCAP, an abbreviation for Interactive Emotional Dyadic Motion Capture, is a dataset consisting of text, audio recordings, and video data collected from scripted dialogues in theatrical scenarios. It is commonly utilized for tasks related to recognizing emotions and analyzing emotional expressions. The Multimodal EmotionLines Dataset (MELD) is a popular choice for sentiment analysis tasks and multimodal emotion recognition research. It consists of text data, audio recordings, and video data obtained from film dialogues. These datasets collectively provide the necessary resources for researchers to explore bimodal and trimodal emotion analysis, emotion intensity prediction, and emotion classification across diverse domains.

Detailed data regarding the multimodal emotion datasets is provided in Table 2, which includes information such as the count of video clips and speakers, the presence of sentiment and emotion

labels, modality labels, and the origin of the data. These datasets are valuable resources for conducting various emotion-related tasks in multimodal sentiment analysis.

5. Discussion

Despite the major advancements in multimodal emotion recognition research due to the abovementioned studies, numerous emerging challenges pose a threat to multimodal emotion recognition techniques. These include: (1) The uncertainty surrounding the kind and quantity of patterns required for effective emotion classification; (2) The struggle to presicely capture and depict the inherent traits of various expressions; and (3) The complexities involved in achieving multi-modal fusion drawing from single modality. (4) Multi-source information real-time acquisition problem.

Based on the above problems, the future development trend of emotion recognition technology may include:(1) Cross-domain multimodal emotion recognition. Multimodal emotion recognition is too domain-dependent and not generalizable enough. Designing a domain-independent multi-modal emotion recognition system is a problem that needs to be solved, such as analyzing Douyu comments with a model trained on the car review dataset.Few-shot learning-based multimodal recognition of emotions. Multi-modal emotion recognition requires a high amount of data. The lack of any single-modal data will affect the final recognition results. The introduction of small-sample learning without reducing the accuracy is also an urgent problem to be solved. (4) Lightweight multi-modal emotion recognition. Model is too complex. At present, the multimodal emotion recognition method based on deep learning has too many model parameters, and the training time of the model is too long. How to simplify the network structure is also a problem that needs attention. (5) Feature extraction and optimization of multimodal data. Feature extraction is the most important part of emotion recognition, which directly affects the final recognition result. Examining ways to further refine the extracted features to enhance the model's robustness is also valuable. For example, how to efficiently remove redundant and repeated emotional features.

6. Conclusion

The latest advancements in multimodal emotion recognition research are presented in this paper. It initially outlines the primary tasks and techniques related to unimodal emotion recognition. The paper then methodically examines multimodal emotion recognition, focusing on areas such as data sets, multimodal feature representation, and multimodal feature fusion.This paper delves deeper into the pivotal aspect of multimodal emotion fusion, offering detailed insights into three primary fusion strategies: feature-level fusion, decision-level fusion, and model-level fusion. It wraps up with a comprehensive discussion on the prevalent issues in the realm of emotion recognition, while also anticipating the potential future trajectory of multimodal emotion recognition

References

[1] Mai S, Hu H, Xing S. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(01): 164-172.

[2] Zhao J, Li R, Jin Q, et al. Memobert: Pre-training model with prompt-based learning for multimodal emotion recognition [C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 4703-4707.

[3] Fu J, Mao Q, Tu J, et al. Multimodal shared features learning for emotion recognition by enhanced sparse local discriminative canonical correlation analysis[J]. Multimedia Systems, 2019, 25: 451-461.

[4] Hotelling H. Relations between two sets of variates[M]//Breakthroughs in statistics: methodology and distribution. New York, NY: Springer New York, 1992: 162-190.

[5] Zhang K, Li Y, Wang J, et al. Feature fusion for multimodal emotion recognition based on deep canonical correlation analysis [J]. IEEE Signal Processing Letters, 2021, 28: 1898-1902.

[6] Zadeh A, Chen M, Poria S, et al. Tensor Fusion Network for Multimodal Sentiment Analysis [C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 1103-1114.

[7] Zeng Z, Tu J, Pianfetti B, et al. Audio-visual affect recognition through multi-stream fused HMM for HCI[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE, 2005, 2: 967-972.

[8] Zadeh A A B, 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). 2018: 2236-2246.

[9] Chen S, Jin Q. Multi-modal conditional attention fusion for dimensional emotion prediction [C]//Proceedings of the 24th ACM international conference on Multimedia. 2016: 571-575.

[10] Huang J, Tao J, Liu B, et al. Multimodal transformer fusion for continuous emotion recognition [C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 3507-3511.