

Research And Application Analysis of Multimodal Emotion Recognition Methods Based on Speech, Text, And Facial Expressions

Jiaqi Sun

Information Engineering College, Changchun College of Electronic Technology, Changchun,
130061, China

dburkev84913@student.napavalley.edu

Abstract. In this study, the focus is primarily on the diverse methods for recognizing human emotions through language, text, and facial expressions via computational technology. Emphasizing the real-world applicability of these techniques, the paper underscores the significance of multimodal emotion recognition in areas such as human-computer interaction, psychology, and emotion analytics. Multimodal methods, which combine data from various sources like voice tone, facial cues, and textual context, offer a robust approach for discerning nuanced emotional states. Compared to single-mode analysis, these multimodal techniques tend to produce more accurate and comprehensive results, bridging the gaps left by any one mode in isolation. As technology increasingly integrates with daily human activity, the importance of nuanced, reliable emotion recognition is becoming paramount for fostering more natural and empathic machine-human interactions. Moreover, in the realm of psychology, these methods offer groundbreaking possibilities for diagnosis and treatment. By discussing the future applications and methodologies of multimodal emotion recognition, this paper aims to provide a comprehensive roadmap for both academic research and practical applications in the evolving landscape of emotion-aware computing.

Keywords: Multimodal Emotion Recognition, Human-Computer Interaction, Psychological Analysis, Machine Learning Methods.

1. Introduction

The advent of advancements in information technology and computer science has led to increasing interest in multi-modal systems, particularly in the realm of emotion recognition. Traditional emotion recognition models often rely on single data streams, such as facial expressions or speech patterns, and are relatively simple in their analyses. However, human emotional expression is a complex interplay of multiple factors including language, text, and facial or bodily expressions. As such, multi-modal emotion recognition—utilizing data from varied sources like verbal cues, text analytics, and facial expressions—offers a far more nuanced and comprehensive understanding of human emotions.

The promise of multi-modal emotion recognition lies in its ability to synthesize multiple types of data for more accurate results. In real-world applications, such as customer service, healthcare, and even autonomous vehicles, understanding the emotional state of humans can be crucial for decision-making and interaction quality. However, integrating disparate data streams presents significant technical challenges, including but not limited to, data fusion techniques, real-time processing, and resolving conflicting emotional indicators from different modalities. Moreover, ethical considerations cannot be overlooked. Issues surrounding data privacy and the psychological impact of continuous emotional monitoring are still open questions that require careful scrutiny. There is also the challenge of ensuring that these systems are free from bias and are culturally sensitive, given that emotional expression can differ widely among different populations [1]. In conclusion, while multi-modal emotion recognition opens up avenues for more effective and nuanced understanding of human emotions, it also introduces a host of technical and ethical challenges that must be rigorously addressed. Nevertheless, the burgeoning interest in this multidisciplinary field suggests that it holds the potential for transformative applications across various sectors.

2. Single modal emotion recognition methods

2.1. Text emotion recognition

This recognition can be divided into four feature fusion methods. The features of different modes are fused to obtain a comprehensive feature vector, and then the classifier is used to classify the emotion; the deep learning method: the feature extraction and emotion classification of multi-modal text is used by deep neural network [2]. Transfer learning method: The existing single-modal emotion classification model is transferred to the multi-modal emotion classification task, and the model parameters are fine-tuned to adapt to the multi-modal data [3]. The above methods can be used alone or in combination to improve the accuracy and robustness of multi-modal text emotion recognition.

2.2. Speech emotion recognition Speech emotion recognition

This recognition can be divided into three categories: 1. Acoustic features-based classifiers: Acoustic features are modeled using traditional computer algorithms, and then emotion classification is performed using these models. 2. Convolutional neural networks: Features such as spectrograms are automatically extracted using convolutional layers, and then classified through fully connected layers. 3. Recurrent neural networks: Ability to process sequential data, suitable for time series of speech signals. Short term Memory networks (LSTMs) and gated cycle units (GRUs) are commonly used variants of RNNs. Acoustic features-based classifiers utilize traditional computer algorithms to model acoustic features from speech signals, which are then used to perform emotion classification. These features can include parameters such as pitch, intensity, and spectral characteristics of the voice. By analyzing these features, it is possible to differentiate emotions such as happiness, sadness, anger, and surprise. Convolutional neural networks (CNNs) are effective in automatically extracting features such as spectrograms from input data. The convolutional layers extract local features from the input, followed by the fully connected layers that perform classification. CNNs have been successful in emotion recognition tasks as they can capture complex patterns and relationships in the data. Recurrent neural networks (RNNs) are particularly suitable for processing sequential data, making them suitable for analyzing time series of speech signals. RNNs have an internal memory that allows them to maintain short-term memory of previous inputs, enabling them to capture contextual information in the data. Short term Memory networks (LSTMs) and gated cycle units (GRUs) are commonly used variants of RNNs that have shown excellent performance in emotion recognition tasks. Overall, these techniques have significantly advanced the field of emotion recognition, enabling accurate assessment of emotions from speech signals. They have found applications in various domains, including human-computer interaction, social research, speech analysis, and more [4].

2.3. Facial emotion recognition Emotion recognition

The fundamental principle of emotion recognition is to extract emotion-related features through image analysis techniques, such as color, texture, shape, and others, from human faces. Then, through the use of convolutional neural networks or recurrent neural networks, the meaning of facial sequences is understood. This allows for the accurate assessment of emotions in real-time and can be applied across a wide range of scenarios, including human-computer interaction, social research, and more [5].

3. Multimodal emotion recognition methods

Multi-modal emotion recognition method is a method to identify and analyze emotion states from multi-modal data, mainly through the following multi-modal recognition methods: 1 Model level fusion: Use different models to model different models, and then perform output fusion; 2 Knowledge transfer transfer the model that has been trained on single-modal data to the multi-modal emotion recognition task; 3 multi-task learning: Multi-task learning: Multiple tasks (such as text emotion

classification, image emotion classification) are included in the training process at the same time, and the performance of multi-modal emotion recognition is improved by sharing model parameters [6].

4. Application Analysis

4.1. Inquiry application

Enter multi-modal inquiry applications, which stand at the intersection of technology and human psychology, offering a more comprehensive approach to customer interactions. By integrating capabilities such as voice recognition, text analysis, and even facial or emotional expression recognition, multi-modal systems provide a far more nuanced understanding of customer needs and behaviors. These systems do not just capture what is being said, they interpret how it is being said. This involves analyzing the tone, pitch, and pace of speech, text semantics, and even non-verbal cues like facial expressions to gain insights into the customer's emotional state and intent [7,8]. The advantages are manifold. For instance, if a customer expresses frustration through text but uses a calm tone in voice interaction, the system could reconcile these differing signals to offer a more empathetic and effective response. Similarly, multi-modal applications could prioritize urgent inquiries based on vocal stress indicators or text semantics, thereby allocating resources more efficiently. The market prospects for multi-modal inquiry applications are incredibly promising. As businesses increasingly aim to offer more personalized and emotionally resonant customer experiences, the demand for systems that can understand and interpret multi-faceted human interactions is likely to soar. Additionally, the rise in the use of artificial intelligence and machine learning algorithms in customer service amplifies the capabilities of these applications, enabling more accurate and timely responses. In a world that is ever more interconnected yet craving personalized interactions, multi-modal inquiry applications are not just a technological advancement; they are a paradigm shift in how businesses understand and engage with their customers. With their capability to deliver what traditional systems simply cannot, multi-modal inquiry applications are poised to redefine the future of customer service.

4.2. Intelligent customer service

With the rise of the Internet era, people have higher and higher requirements for the Internet. Therefore, ordinary intelligent customer service can no longer meet the needs of current people, and multi-modal intelligent customer service has become a new industry [8]. Multi-modal customer service can analyze user input, identify user's intention and demand, and conduct appropriate questions and answers and interactions according to user's intention and context. By analyzing the user's tone, emotion and emotion, it can identify the user's emotional state and provide corresponding emotional responses and solutions [9]. Therefore, compared with the traditional intelligent customer service, the advantage of multi-modality is that it can extract more useful information to provide more accurate and personalized services.

4.3. Application in the field of mental health

Multimodal mental health technology can record the user's words and convert them into text, and then interpret and analyze the text to analyze the user's mental state [10]. Meanwhile, video analysis and video understanding technology can analyze the user's mental state, such as facial expression, tone and posture, to analyze the user's mental state, and record the customer's language. Analyze the user's psychological state.

5. Challenges

Managing multi-modal data is a complex task that often involves a myriad of challenges, the foremost being the inherent ambiguity and diversity present in information sourced from different origins. These data streams can include text, images, audio, and video, among others. The ambiguity arises when similar data points from different sources seem to suggest conflicting interpretations. For

instance, text-based sentiment analysis might indicate customer satisfaction, while vocal tone analysis from the same customer could suggest dissatisfaction. This disparity complicates the analytical process, as it becomes difficult to ascertain which source provides the most accurate representation of the reality being studied. A secondary, albeit equally critical, challenge is the fusion of these diverse data types into a cohesive, meaningful whole. This is a non-trivial problem that extends beyond mere data concatenation or basic algorithmic manipulation. One must decide how to weight each type of data for the final interpretation and what computational models are most suited for integrating these disparate forms of information. Should greater importance be assigned to visual cues over textual content in understanding consumer behavior. Is an ensemble method more appropriate than a single, more complex model for data integration? These are intricate questions that require robust methodologies to answer. In essence, multi-modality presents both an opportunity and an obstacle. On one hand, leveraging multiple sources of data can potentially offer richer insights and more accurate predictions. On the other, the complexity and diversity inherent in multi-modal data require sophisticated techniques for handling ambiguity and facilitating effective data fusion. The objective, then, is to navigate these challenges adeptly enough to harness the full analytical potential of multi-modal data sets.

6. Conclusion

In conclusion, multimodal technology represents a significant advancement with the potential to outperform human capabilities in both accuracy and convenience across a wide array of tasks. While it offers transformative potential, it is important to acknowledge that the technology is still in its developmental phase and thus has its limitations. Despite these shortcomings, there is strong reason to be optimistic about the future. As technological innovation continues to advance, it is highly likely that these issues will be addressed and resolved, leading to even more robust and reliable multimodal systems. This technology not only stands to revolutionize the way tasks are completed but could also introduce new methods and paradigms that have yet to be imagined. The promise lies not just in perfecting existing models but also in the untapped possibilities that this technology could unveil. In that sense, the ongoing evolution of multimodal systems could well be a cornerstone for future technological landscapes, offering unprecedented capabilities and benefits. Therefore, as the technical level continues to improve, it is reasonable to expect that this technology will resolve its current limitations and emerge as an indispensable tool in various fields.

References

- [1] Abdullah, S. M. S. A., Ameen, S. Y. A., Sadeeq, M. A., & Zeebaree, S. (2021). Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2(02), 52-58.
- [2] Chen, J., Wang, C., Wang, K., Yin, C., Zhao, C., Xu, T., ... & Yang, T. (2021). HEU Emotion: a large-scale database for multimodal emotion recognition in the wild. *Neural Computing and Applications*, 33, 8669-8685.
- [3] Tan, Y., Sun, Z., Duan, F., Solé-Casals, J., & Caiafa, C. F. (2021). A multimodal emotion recognition method based on facial expressions and electroencephalography. *Biomedical Signal Processing and Control*, 70, 103029.
- [4] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020, April). M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 02, pp. 1359-1367).
- [5] Chuang, Z. J., & Wu, C. H. (2004, August). Multi-modal emotion recognition from speech and text. In *International Journal of Computational Linguistics & Chinese Language Processing*, Volume 9, Number 2, August 2004: Special Issue on New Trends of Speech and Language Processing (pp. 45-62).
- [6] Cai, L., Dong, J., & Wei, M. (2020, November). Multi-modal emotion recognition from speech and facial expression based on deep learning. In *2020 Chinese automation congress (CAC)* (pp. 5726-5729). IEEE.

- [7] Xu, H., Zhang, H., Han, K., Wang, Y., Peng, Y., & Li, X. (2019). Learning alignment for multimodal emotion recognition from speech. arXiv preprint arXiv:1909.05645.
- [8] Siriwardhana, S., Kaluarachchi, T., Billingham, M., & Nanayakkara, S. (2020). Multimodal emotion recognition with transformer-based self-supervised feature fusion. IEEE Access, 8, 176274-176285.
- [9] Tripathi, S., Tripathi, S., & Beigi, H. (2018). Multi-modal emotion recognition on iemocap dataset using deep learning. arXiv preprint arXiv:1804.05788.
- [10] Sharma, G., & Dhall, A. (2021). A survey on automatic multimodal emotion recognition in the wild. Advances in data science: Methodologies and applications, 35-64.