A survey on multimodal emotion recognition

摘要：

**修改前：**Humans are emotional creatures. Multiple modalities are often involved when we express emotions, whether we do so explicitly or implicitly. Therefore，enabling machines to have emotional intelligence, i.e., recognizing and interpreting emotions, is becoming increasingly vital.

In recent years, significant research efforts have been devoted to emotion recognition of multimodal data. With advancements in AI technologies and the diversification of application domains, emotion recognition finds utility in smart driving, education classroom, and business promotion.

In this tutorial, we discuss several key aspects of multimodal emotion recognition (MER). We begin with a brief introduce on widely used multimodal emotion recognition datasets. We then summarize existing representative methods, followed by the description of main challenges in MER. Finally, we outline several practical applications and discuss some future directions.

**修改后：**

Humans are emotional creatures. Multiple modalities are often involved when we express emotions, whether we do so explicitly or implicitly. Therefore, enabling machines to have emotional intelligence, i.e., recognizing and interpreting emotions, is becoming increasingly vital.

Recent research has focused on multimodal emotion recognition. With the development of deep learning and computer vision technology, the application areas of multimodal emotion recognition (MER) are diversified. MER has been applied in smart driving, education classroom, and business promotion.

This tutorial delves into several key aspects of MER. It begins with an overview of representative methods, followed by a summary of commonly used datasets for MER. We then summarize the practical applications in the field. Finally, primary challenges and future directions are outlined.

1. Introduction

Emotions significantly shape human experiences, impacting interactions, decisions, and overall well-being. The recognition of emotions is essential, not just for improving human-computer interaction, but also for various applications, including mental health support and enhancing customer satisfaction[1]. Therefore, there is an urgent demand for the creation of emotion recognition method that can precisely interpret and respond to human emotions, enabling seamless and effective human-computer interaction.

Emotionally intelligent machines can offer personalized services, benefiting vulnerable groups like the elderly, disabled, and children. Additionally, by recognizing human emotions in real-time, they can identify abnormal behaviors, issue timely reminders, and prevent societal disruptions.

This article provides a comprehensive overview of various aspects of Multimodal Emotion Recognition (MER), encompassing computational methodologies, data source, challenges, applications, and future prospects.

1. Methods

In this section, according to the development of MER, the relevant methods are divided into simple approaches, sequential approaches, contextual approaches and graph-based approaches. The research status, advantages and disadvantages of these methods will be introduced respectively.

* 1. 1. Simple approaches

Simplified architectures, blending handcrafted features with basic frameworks like CNNs, efficiently extract fundamental and spatial features. Utilizing layered processing, CNNs initially analyze specific image segments, gradually integrating this information to grasp emotional nuances comprehensively.

In the realm of facial expression recognition, Ly et al.[2] leveraged CNN model variants like Inception-Resnet and 3D point CNN to extract deep features from facial regions obtained using MTCNN. Moreover, Kumar et al.[3] conducted a comprehensive comparative study, analyzing various facial expression recognition techniques, and highlighting their contrasts and similarities. Xu et al.[4] leverage ECG and textual modality processed by a convolutional autoencoder, alongside feature-level fusion.

To summarize, simple approaches offer accessibility and practicality, effectively capturing basic emotional cues across data modalities. However, their reliance on basic features may restrict their ability to capture intricate nuances within complex emotional states.

2.1.2 Sequential approaches

Sequential methodologies cater to MER scenarios where data is inherently sequential, Sequential methodologies in MER heavily rely on RNN, LSTM, and GRU. This ability to analyze emotional changes within a sequence makes them well-suited for capturing nuanced emotional shifts.

Do et al.[5] put forward a model that incorporates video information as input. They employed three CNN-based architectures to extract high-level features and utilized weighted decision fusion for emotion recognition. Nguyen et al.[6] devised a methodology employing a two-stream auto-encoder and an LSTM network, integrating visual and audio signal streams in an end-to-end manner.

However, the sequential processing nature of these methods can lead to increased computational requirements and slower processing speeds, limiting their real-time applicability.

2.1.3 Contextual approaches

Contextual methodologies represent a sophisticated approach to MER, harnessing the power of context to enhance emotional understanding across various modalities. By considering the broader context in which data is presented, these methodologies offer an insightful way to capture complex emotional nuances.

Liu et al.[7] introduced a novel attention network-based approach for MER, assigning weights to different modalities based on their importance. Wei et al.[8] designed a Fully Multimodal Video-To-Emotion System that utilized a hierarchical attention approach, considering sound spectra, multi-scale visual extraction, transformer, and aligned multimodal learning. Yoon et al.[9] proposed a cross-modal translator-based approach for MER, utilizing heterogeneous datasets without requiring modal alignment.

Contextual MER has focused on attention-based networks, hierarchical fusion, transformers, and neural network models for extracting contextual information. The intricate processing demanded by contextual analysis often necessitates substantial computational resources, resulting in extended processing durations.

2.1.4 Graph-based approaches

Graph-based approaches represent a cutting-edge paradigm in MER, especially suited for tasks involving intricate relationships and dependencies among diverse data modalities.

Miao et al.[10] proposed an end-to-end multi-output DL model for aesthetic and emotion conjoint analysis based on a multimodal GCN and co-attention. DeepLabv3+ network with Xception and GloVe word embeddings was used for feature processing. The model employs a stacked multimodal GCN network and a co-attention module to enable interactive learning between aesthetics and emotional feature representations.

Graph-based approaches offer significant capabilities but also pose certain challenges. Creating meaningful graph structures from various data sources demands domain expertise and may consume considerable time.

1. Data Sources

The most common multimodal combinations in the MER field involve integrating Audio, Text, and Visual (ATV) modalities. This section offers a thorough analysis of these datasets, focusing on various important aspects, as outlined in Table 1 Dataset. The ‘sent’ and ‘emo’ in the table represent the sentiment and emotion labels.

Table 1 Dataset

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| dataset | Year | samples | speaker | Model | sent | emo |
| CMU-MOSEI | 2018 | 23453 | 1000 | ATV | 🗸 | 🗴 |
| CMU-MOSI | 2016 | 2199 | 98 | ATV | 🗸 | 🗴 |
| MSED | 2022 | 9190 | - | TV | 🗸 | 🗸 |
| SEWA | 2019 | 538 | 408 | AV | 🗴 | 🗸 |
| OMG | 2018 | 2444 | 1 | ATV | 🗴 | 🗸 |
| MELD | 2018 | 13708 | 407 | ATV | 🗴 | 🗸 |

These datasets serve as crucial benchmarks, providing a standardized platform for evaluating different MER methodologies objectively.

1. Applications

MER has gained significant traction as a subfield of affective computing, showing rapid advancements and growing prominence in recent times.

Education: By analyzing the emotional states of their students, teachers can quickly identify those who may be struggling with the learning material or require extra support.

Human–Robot Interaction: With the help of MER, robots can detect emotions such as sadness, frustration, or happiness and can offer appropriate responses to improve the interactions.

Security and Surveillance: The smart home framework can utilize facial emotion recognition to enhance security measures in smart homes.

1. Challenges and Future

This section delves into the difficulties faced by MER systems and provides an in-depth discussion of potential avenues for future research.

* 1. Heterogeneity of modalities

MER encounters a significant challenge rooted in the inherent heterogeneity of modalities. This heterogeneity arises from the diverse nature of information conveyed through different modalities. Another dimension of this challenge lies in the varying temporal resolutions of modalities.

The future will require the advancement of fusion techniques, including attention mechanisms and transformers, capable of accommodating diverse feature representations and temporal resolutions across modalities.

* 1. Data security

When implementing MER systems, importance must be placed on data security. Preserving data integrity and security is of critical concern. As future trends, prominent strategies play pivotal roles.

1. Conclusion

MER seeks to improve emotion recognition accuracy by analyzing various modalities, including facial expressions, vocal modulations, and physiological changes. In addition, the review meticulously categorizes existing state of art methods.

While MER presents exciting opportunities, it also poses some challenges. These challenges encompass the need to handle diverse modalities effectively and cope with noisy and unstructured data. Moreover, addressing privacy concerns and ethical considerations remains an essential concern. In addition, the integration of MER with affective computing applications.

参考文献：

1. Zhao, Sicheng, et al. "Emotion recognition from multiple modalities: Fundamentals and methodologies." IEEE Signal Processing Magazine 38.6 (2021): 59-73.
2. Ly, S. T., Do, N. T., Lee, G., Kim, S. H., & Yang, H. J. (2019, June). Multimodal 2D and 3D for In-The-Wild Facial Expression Recognition. In Cvpr workshops (pp. 2927-2934).
3. Kumar, S., Sagar, V., & Punetha, D. (2023). A comparative study on facial expression recognition using local binary patterns, convolutional neural network and frequency neural network. Multimedia Tools and Applications, 82(16), 24369-24385.
4. Xu, J., Hu, Z., Zou, J., & Bi, A. (2019). Intelligent emotion detection method based on deep learning in medical and health data. IEEE Access, 8, 3802-3811.
5. Do, Luu-Ngoc, et al. "Deep neural network-based fusion model for emotion recognition using visual data." The Journal of Supercomputing (2021): 1-18.
6. Nguyen, Dung, et al. "Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition." IEEE Transactions on Multimedia 24 (2021): 1313-1324.
7. Liu, Shuai, et al. "Multi-modal fusion network with complementarity and importance for emotion recognition." Information Sciences 619 (2023): 679-694.
8. Wei, Qinglan, Xuling Huang, and Yuan Zhang. "Fv2es: A fully end2end multimodal system for fast yet effective video emotion recognition inference." IEEE Transactions on Broadcasting 69.1 (2022): 10-20.
9. Yoon, Yeo Chan. "Can we exploit all datasets? Multimodal emotion recognition using cross-modal translation." IEEE Access 10 (2022): 64516-64524.
10. Miao, Haotian, et al. "Multi-output learning based on multimodal GCN and co-attention for image aesthetics and emotion analysis." Mathematics 9.12 (2021): 1437.