The Application of Artificial Intelligence-based Multimodal Emotion Analysis

Hongji Zhou

Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, United Kingdom
pshzho15@liverpool.ac.uk

Abstract. Nowadays, Artificial Intelligence (AI) and Multimodal Emotion Analysis represent cutting-edge advancements in the realm of computational intelligence. AI, the emulation of human cognitive processes by machines, has revolutionized various fields, including emotion analysis. Multimodal Emotion Analysis refers to the integration of multiple sensory inputs, such as images, speech, and text, to understand and interpret human emotions comprehensively. As opposed to single-modal approaches, the multimodal approach provides a more comprehensive and precise analysis of affective states. By combining machine learning algorithms with sophisticated data processing techniques, AI systems can now recognize and analyze emotional cues from diverse modalities, providing deeper insights into human affective states. This interdisciplinary approach has significant implications across numerous domains, from human-computer interaction and social robotics to mental health diagnostics and marketing research. With the potential to enhance the understanding of human emotions and behaviors, AI-driven multimodal emotion analysis stands at the forefront of innovation, promising to reshape how people interact with technology and interpret human experiences.

Keywords: Artificial Intelligence, Multimodal Emotion Analysis, Machine Learning.

1 Introduction

With further improvements in computing power, increasing data volumes, and continuous algorithmic advancements, it is foreseeable that multimodal applications will demonstrate significant potential in more fields, bringing greater convenience and innovation to people's lives [1]. Especially notable this year are the significant advancements and achievements within the field of multimodal emotion recognition. This article aims to consolidate the current stage of Artificial Intelligence (AI) accomplishments in multimodal emotion recognition, while also analyzing and summarizing methodologies and datasets pertinent to this domain. Multimodal emotion recognition pertains to the ability of artificial intelligence systems to detect and interpret emotions from multiple sources of input, such as facial expressions, vocal intonations, gestures, and textual content. This interdisciplinary field

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_32

amalgamates principles from computer vision, natural language processing, signal processing, and affective computing. Recent breakthroughs have been witnessed in various aspects of multimodal emotion recognition. These include advancements in deep learning architectures tailored for multimodal fusion, the development of novel feature extraction techniques, and the integration of context-aware models for improved emotion understanding. Furthermore, the proliferation of annotated multimodal datasets has played a pivotal role in driving progress within this domain. These datasets encompass diverse modalities and emotional contexts, facilitating the training and evaluation of multimodal emotion recognition systems.

Aruna Gladys and her collaborator, V. Vetriselvi, provide an in-depth evaluation of the current advancements and future directions in Multimodal Emotion Recognition (MER) [2]. They recognize the advancements in leveraging a fusion of physiological, audio-visual, and linguistic modalities to improve the accuracy of emotion detection. The deployment of deep learning methodologies, such as attention-based models, Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs), is acknowledged for its contribution to the field's progress. Despite these strides, Aruna Gladys identifies several challenges that must be addressed to further the development of MER. The necessity for more interactive and multilingual datasets is emphasized, as is the creation of models that are invariant to language and culture. The utilization of unsupervised and semi-supervised learning techniques is proposed to capitalize on the abundance of unlabeled data, which could significantly boost the training of MER models. Additionally, Aruna Gladys calls for an exploration into Multimodal Neural Architecture Search (M-NAS) frameworks, which hold the potential to construct optimal fusion models within a domain-specific context. The paper also underscores the imperative for model interpretability, which is increasingly critical for gaining trust and ensuring the ethical deployment of affective computing solutions.

In addition, affective computing is also developing very rapidly. Emotional computing is an emerging interdisciplinary field focused on equipping intelligent systems with the ability to recognize, infer, predict, and interpret human emotions. This domain integrates various disciplines, including cognitive science, computer science, neuropsychology, artificial intelligence, social science, and neuroscience. The primary objective of emotional computing is to detect emotional cues during human-computer interactions and generate suitable affective responses. This field employs techniques for identifying emotions from data of diverse modalities and granularities.

The research in emotional computing primarily focuses on sentiment analysis and emotion recognition. Sentiment analysis typically involves a broad categorization of emotions, classifying data into binary positive vs. negative or ternary positive, negative, and neutral sentiments. In contrast, emotion recognition entails a more detailed analysis, categorizing data into numerous emotion labels, often exceeding four categories. Over the past two decades, artificial intelligence researchers have strived to endow machines with the cognitive capabilities to discern, interpret, and express emotions and moods. These endeavors fall under the wider umbrella of emotional computing research.

Emotion is a complex psychophysiological phenomenon, and psychologists have yet to reach a consensus on a unified definition of emotion or mood. The diverse array of theories surrounding emotions and moods highlights the multifaceted nature of this phenomenon [3].

2 Retrospective Introduction of AI

Timeline about development of AI and multimodal emotion recognition is demonstrated in Fig. 1.

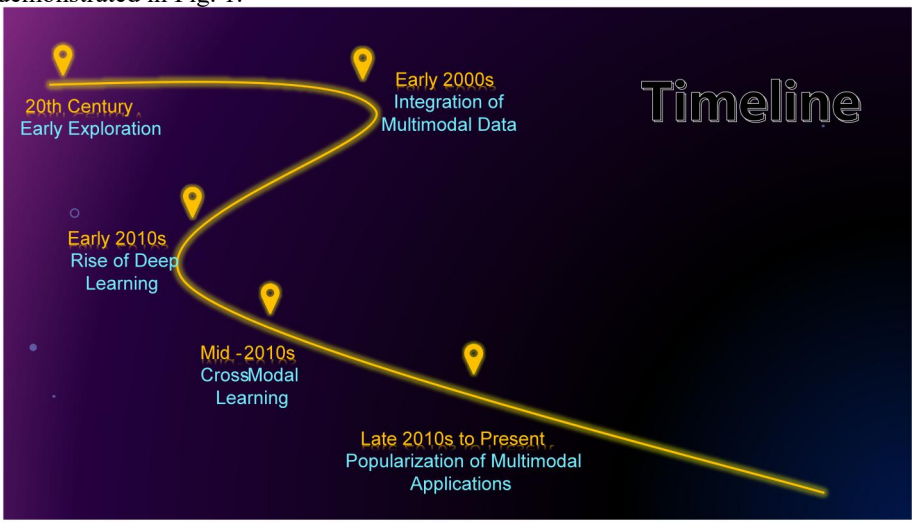

Fig. 1. Timeline of AI development (Figure Credit: Original).

2.1 Stage 1: Early Exploration (20th Century).

In the 1950s to the early 1960s, the field of artificial intelligence began to emerge, mainly focusing on symbolic reasoning and expert systems, with no involvement of multimodal applications. In the 1980s, multimodal technologies such as computer vision and speech recognition began to develop gradually but remained in their early stages.

2.2 Stage 2: Integration of Multimodal Data (Early 2000s)

With the massive generation of digital images and speech data, multimodal data processing became a research hotspot. Researchers began to explore how to integrate data from different modalities to improve the performance of various applications.

2.3 Stage 3: Rise of Deep Learning (Early 2010s)

The rise of deep learning technology injected new vitality into the development of multimodal applications. Through end-to-end learning of deep neural networks, multimodal data such as images, text, and speech could be effectively processed. Significant progress was made in many fields, which were also applied in multimodal scenarios.

2.4 Stage 4: Cross-Modal Learning (Mid-2010s)

Researchers have started investigating methods to transfer and share knowledge across different modalities to enhance the performance of multimodal tasks. Cross-modal learning techniques allow models to acquire knowledge from one modality and apply it to another. For instance, jointly training on image and text data can improve outcomes in image captioning and text generation.

2.5 Stage 5: Popularization of Multimodal Applications (Late 2010s to Present)

Multimodal applications began to be widely used in various fields, including intelligent transportation, medical diagnosis, and smart assistance. Products such as smart speakers and smartphones integrated multimodal technologies, providing users with more intelligent interaction experiences. In the field of autonomous driving, the fusion and processing of multimodal sensors (such as cameras, radars, lidars, etc.) became key technologies for vehicle perception, decision-making, and control.

2.6 Summary

In the scholarly discourse encapsulated by Tom B. Brown, et al they present an analytical compendium of global trends in AI research and development [4]. Their report meticulously documents a twofold increase in scholarly publications within the AI sphere from 2010 to 2021. Prominent nations like China and the United States are accentuated as prolific contributors to the AI literature across a spectrum of platforms, including academic journals, conferences, and repositories. Brown's work accentuates the inherently collaborative and open nature of AI research, with a marked emphasis on the robust acceleration in Sino-American cooperative research endeavors. Additionally, the report chronicles a dramatic upsurge in AI patent filings, indicative of an innovation surge within the field. It also provides a nuanced view of the distribution of scholarly outputs, ranging from peer-reviewed journal articles to the contributions made through pre-peer-reviewed electronic repositories. In the development of Multimodal Sentiment Analysis (MSA). Ankita Gandhi has emphasized that with the progress in AI and Natural Language Processing (NLP), MSA is becoming an increasingly important technique for analyzing user emotions towards products and services [5]. They have noted that due to the widespread adoption of social media, there is a greater propensity for people to share their

opinions through multimedia formats such as videos. This trend has propelled the development of MSA to automate the analysis and comprehension of complex emotional expressions that blend text, audio, and visual data. MSA integrates knowledge from psychology, computer science, cognitive science, and social science cognitive science to create intelligent systems capable of recognizing, analyzing, and expressing emotions. Furthermore, they have highlighted the potential of MSA in creating commercial value and supporting everyday decision-making.

3 Representative Emotion Recognition Methods

3.1 The Structure of General MER System

Zhang, S outlines the workflow of a Multimodal Emotion Recognition (MER) system as comprising three main steps: (1) Multimodal Feature Extraction, (2) Multimodal Information Fusion, and (3) Design of Emotion Classifiers [6].

Step 1: Multimodal Feature Extraction involves obtaining effective feature representations that capture human emotional expressions from various modalities, such as audio, visual, and text. Step 2: Multimodal Information Fusion entails integrating different emotion recognition modalities through various fusion strategies. Common approaches include feature-level fusion, decision-level fusion, and model-level fusion. This step demonstrates that multimodal information fusion significantly outperforms bimodal or trimodal emotion recognition in terms of accuracy. Step 3: Emotion Classifier Design involves employing suitable classifiers to map the extracted feature representations to target emotions, resulting in the final emotion recognition labels, which can be either discrete or dimensional categories. For emotion classification or prediction, most existing machine learning methods are applicable. Representative classifiers include Bayesian networks, Multilayer Perceptron (MLP), k-Nearest Neighbors (KNN), and Support Vector Machine (SVM).

3.2 Facial Expression-based Emotion Recognition

Chen, L thinks the essence of capturing and comprehending human emotional fluctuations within human-robot interaction by observing facial expressions, body language, and speech [7]. Its core lies in simulating the human process of perceiving and understanding emotions, facilitated through robots.

The specifics encompass several aspects: (1) Multimodal Emotion Recognition. This entails integrating various sensory cues, including facial cues, vocal signals, and bodily gestures, for emotion identification. These modalities typically co-occur and are employed for real-time analysis and inference of emotional states and intentions, guiding the robot's corresponding responses. (2) Facial Expression Feature Extraction. Facial expressions represent one of the most intuitive forms of emotional expression. Extraction methods primarily involve deformation-based and motion-based techniques. Deformation methods include Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA), while motion methods focus on facial

expression variations. (3) Speech Emotion Feature Extraction. Speech Emotion Feature Extraction involves categorizing emotion features into acoustic and linguistic features. Acoustic features are further divided into prosodic, spectral, and voice quality features. These are typically extracted at the frame level and contribute to emotion recognition through global feature statistical analysis. (4) Body Gesture Feature Extraction. Body gestures also play a significant role in emotional communication. Feature extraction methods encompass global and local approaches, such as Motion Energy Images (MEI) and Motion History Images (MHI). (5) Emotion Databases and Classification Issues. Current emotion databases typically represent emotions in adjective label forms (e.g., anger, happiness, sadness, surprise, and neutral), leading to standard pattern classification problems based on discrete emotion description models. (6) Emotion Recognition Algorithms. Research primarily concentrates on algorithm development based on different modal emotion data. Deep learning algorithms, such as CNNs and Deep Belief Networks (DBNs), are also applied for learning and recognizing emotion features. (7) Multimodal Information Fusion. Information fusion serves as the theoretical foundation for multimodal emotion recognition, spanning multiple domains. In emotion recognition, multimodal information fusion can be categorized into decision-level fusion and feature-level fusion, both aiming to fully leverage multimodal emotion information to enhance emotion recognition performance.

Through these techniques, robots can better comprehend and respond to human emotional states, thereby providing a more natural and human-like interaction experience.

3.3 K-Means Clustering-Based Kernel Canonical Correlation Analysis (KMKCCA)

Chen, L developed the KMKCCA algorithm, a sophisticated method specifically designed for multimodal emotion recognition in human-robot interaction scenarios [8]. It operates by initially extracting multimodal features from facial expressions and speech, which are then preprocessed through K-means clustering to enhance feature homogeneity and reduce dimensionality. This clustering step is crucial for improving the algorithm's ability to differentiate between various emotional states. Subsequently, the algorithm employs kernel canonical correlation analysis to fuse the selected features, leveraging kernel methods to handle nonlinear relationships within the data. The fused features are then classified using an SVM, which serves as an effective classifier for recognizing emotional states. KMKCCA's advantages include higher recognition rates, improved handling of feature heterogeneity, and the ability to solve nonlinear problems inherent in emotional feature representation. Additionally, it reduces the feature space's dimensionality, which streamlines the computational process and focuses on relevant features. Empirical validation has demonstrated KMKCCA's effectiveness, showcasing its superiority over methods lacking a K-means clustering step. This algorithm is particularly adept at capturing the complementarity between different modalities, thereby enhancing the overall performance in emotion recognition tasks within interactive systems.

3.4 Imbalance in Multimodal Emotion Recognition in Conversations (IMBA-MMERC)

IMBA-MMERC is an innovative system developed to address the common problem of class imbalance in MMERC. This system is particularly significant given the escalating role of Intelligent Personal Assistants (IPAs) in various interactive applications, where the accurate detection of user emotions is crucial for enhancing service quality and personalization. The cornerstone of IMBA-MMERC lies in its strategic approach to sample generation, where it employs an enhanced synthetic minority oversampling technique algorithm tailored to the nuances of multimodal conversational data. This approach is complemented by the utilization of mutual information maximization to ensure the alignment of different modalities, thereby preserving the conversational coherence and modality-specific feature integrity. IMBA-MMERC further distinguishes itself through the integration of Graph Convolutional Networks (GCN), which adeptly model the intricate interactions and dependencies among modalities within a conversational context. The framework's architecture is fortified by a well-classified encouraging loss function, which serves as a regularization technique. This function is pivotal in maintaining the classifier's robust performance on the majority classes, despite the influx of synthetic minority class samples. Empirical validation is thoroughly established through extensive experiments on two English benchmark datasets and one Chinese dataset, consistently demonstrating IMBA-MMERC's superiority over state-of-the-art methods. The system's efficacy is further underscored by ablation studies, which analyze the individual contributions of sample generation and the well-classified encouraging loss, as well as by case studies that showcase the system's nuanced handling of complex emotional expressions. Despite its empirical successes, Qianer acknowledges the need for further validation in more intricate real-world scenarios, where the variability and complexity of human emotions are heightened [9]. The challenge of class imbalance, while algorithmically mitigated, may benefit from additional research focused on developing more sophisticated strategies that are sensitive to the dynamic and context-specific nature of conversational emotions.

In conclusion, the IMBA-MMERC system represents a substantial advancement in the field of affective computing, offering a promising avenue for research and development aimed at achieving more natural and empathetic human-computer interactions. The framework's potential impact extends beyond the current scope, suggesting possible applications in diverse domains, where emotional intelligence can markedly improve user experience and satisfaction.

4 Discussion and Prospects for Multimodal Emotion Analysis

With the swift progress of AI technologies, especially in the domains of machine learning and deep learning, people have witnessed groundbreaking applications across various domains. One such domain is Multimodal Sentiment Analysis (MSA), an integral branch of AI that has emerged as a pivotal technology for understanding and analyzing human emotional states. MSA amalgamates data from multiple sensory

channels, such as speech, text, facial expressions, and physiological signals, to provide a more comprehensive and accurate assessment of an individual's emotional state and psychological well-being. The advancement of this technology provides new perspectives and tools for the early detection and intervention of mental health issues. As an example, by analyzing text and images on social media platforms, AI can assist in identifying users who may be experiencing mental health challenges. In clinical settings, MSA can aid physicians in more accurately assessing patients' emotional states, thereby facilitating the provision of more personalized treatment plans. In 2024, NVIDIA demonstrated its advanced Avatar Cloud Engine (ACE) technology, enabling non-player characters (NPCs) in games to engage in smooth, natural dialogues with players, supporting multi-turn interactions that advance the game's storyline. The technology integrates services like Riva for speech tasks and Audio2Face for real-time lip-syncing and facial expressions, combined with third-party large models to create intelligent in-game characters, thereby enhancing the immersive experience and player engagement [10].

Integrating these technological advancements with the British General Practitioner (GP) system, a system could be conceptualized for the pre-emptive diagnosis and adjunctive treatment of psychological disorders.

The operational flow of this system is as follows: (1) Initial Screening: The system collects multimodal data from the user through a brief video chat, capturing speech, facial expressions, and linguistic expressions. (2) Data Analysis and Scoring: The AI analyzes the collected data, employing Multimodal Sentiment Analysis techniques to score the user's psychological health. (3) Tiered Treatment Recommendations: Tier 1: Psychologically healthy, no treatment required. The system provides suggestions for maintaining good psychological health. Tier 2: Mild psychological issues present. The system offers a two-week self-guided treatment plan, including psychological health education and emotional management skills, with a reassessment after 15 days. Tier 3: Severe psychological issues. The system immediately assists the user in contacting a professional psychologist for further diagnosis and treatment.

The integration of AI and MSA into a system that mirrors the British GP model for pre-emptive diagnosis and adjunctive treatment of psychological disorders presents several notable advantages: Firstly, enhanced accuracy: MSA integrates multiple data sources for a precise assessment of psychological health. Additionally, in terms of efficiency and accessibility, this system is faster and more convenient than traditional consultations, thus making mental health assessments more accessible to a broader population. Furthermore, regarding resource optimization, it ensures a more equitable distribution of mental health professionals, providing timely care for those in need. Lastly, its preventive significance: MSA identifies early signs of psychological disorders, potentially preventing more severe conditions.

However, the implementation of such a system is not without its challenges: Firstly, privacy concerns: The necessity for data collection poses risks of personal information leakage, which must be mitigated through stringent privacy protection measures. Secondly, industry disruption: The introduction of this system could disrupt the traditional mental health profession, potentially leading to social movements, such as strikes, if not managed with careful consideration of existing professionals' roles

and concerns. Lastly, professional recognition: Gaining recognition from international psychological associations is another challenge, as the system must meet the rigorous standards and ethical guidelines set forth by these organizations.

Addressing these challenges is crucial for the successful integration of AI into mental health care. It requires a balanced approach that respects privacy, considers the impact on existing professionals, and aligns with international standards to ensure the system's effectiveness and acceptance. By doing so, the power of AI could be harnessed to enhance mental health services while minimizing potential negative consequences.

The design of such a system not only enhances the accessibility and efficiency of mental health services but also enables the early identification and intervention of potential psychological issues, thereby reducing the pressure on the healthcare system and providing more personalized and timely mental health support for individuals. Additionally, the development and implementation of the system must strictly adhere to privacy protection and data security regulations to ensure the proper safeguarding of users' personal information. This approach can contribute to the construction of a healthier and more humane mental health service system for society.

5 Conclusion

Recent years have witnessed significant advancements in AI for multimodal sentiment analysis. Deep learning methods, notably CNNs, RNNs, and attention mechanisms, have been extensively employed to process diverse modalities such as images, text, and speech, thereby achieving more accurate and robust sentiment analysis. Multimodal data fusion techniques play a crucial role, with novel fusion strategies, including feature-level, decision-level, and model-level fusion being proposed to exploit correlations between different modalities and enhance sentiment analysis performance. Applications in real-time and natural language interaction, such as virtual assistants, intelligent meeting systems, and affect-aware educational systems, leverage multimodal sentiment analysis to facilitate smarter and more personalized interactions. Additionally, attention is increasingly directed towards cross-cultural and cross-lingual sentiment analysis, aiming to overcome cultural and linguistic biases in emotional expression for broader and more universal applications. Furthermore, privacy and ethical considerations have emerged as important issues, prompting discussions on safeguarding user privacy and ensuring fairness and transparency in sentiment analysis systems. In summary, recent developments in AI for multimodal sentiment analysis exhibit a trend towards diversification, intelligence, and human-centricity. With ongoing technological advancements and expanding application scenarios, multimodal sentiment analysis promises to deliver greater convenience and intelligence to people's lives.

References

1. Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. F., & Pantic, M.: A survey of multimodal sentiment analysis. Image and Vision Computing, **65**, 3-14 (2017).
2. Gladys, A. A., & Vetriselvi, V.: Survey on Multimodal Approaches to Emotion Recognition. Neurocomputing, **556**, 1-21 (2023).
3. Zhang, J., Yin, Z., Chen, P., & Nichele, S.: Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. Information Fusion, **59**, 103-126 (2020).
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., et al.: Language models are few-shot learners. Advances in neural information processing systems, **33**, 1877-1901 (2020).
5. Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A.: Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, **91**, 424-444 (2023).
6. Zhang, S., Yang, Y., Chen, C., Zhang, X., Leng, Q., & Zhao, X.: Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. Expert Systems with Applications, **237**, 121692 (2023).
7. Chen, L., Wu, M., Pedrycz, W., & Hirota, K.: Emotion recognition and understanding for emotional human-robot interaction systems. 1st edition. Springer Nature (2020).
8. Chen, L., Wang, K., Li, M., Wu, M., Pedrycz, W., & Hirota, K.: K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human–robot interaction. IEEE Transactions on Industrial Electronics, **70**(1), 1016-1024 (2022).
9. Li, Q., Huang, P., Xu, Y., Chen, J., Deng, Y., & Yin, S.: Generating and encouraging: An effective framework for solving class imbalance in multimodal emotion recognition conversation. Engineering Applications of Artificial Intelligence, **133**, 108523 (2024).
10. NVIDIA's AI-Powered NPCs Bring New Life to Gaming Experiences. URL: https://www.thepaper.cn/newsDetail_forward_26775164. Last Accessed 2024/05/25