Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

A Review of Multimodal Interaction Technologies in Virtual Meetings

LIN, Weikun [1*]

[1] Shandong University of Science and Technology, China

**** LIN, Weikun is the corresponding author, E-mail: Welton.lin2233@gmail.com***

Abstract: Multimodal interaction technologies enhance the planning of human-to-human virtual meetings that involve Call Centers located in one part of the world and customers in other activities through the use of interacting with people regardless of their languages thus addressing problems of culture, equity and more interaction and participation from humans is sought after. Hence, language translation and transcription in real-time is an important aspect of multimodal technologies. In addition to these functions, emotion analysis, natural language processing and sentiment analysis, in their turn, allow the host to understand how participants feel and the main ideas expressed in the audio or video minutes. Also, participants have the opportunity to record their notes, which will in turn be generated in the form of minutes using natural language processing. Thus, turning audio and videos into texts significantly improves the efficiency and quality of meetings with natural human engagement.

Keywords: Virtual Meetings, Multimodal Interaction, User Experience, Natural Language Processing, Technological Development.

Disciplines: Artificial Intelligence. **Subjects:** Natural Language Processing.

DOI: https://doi.org/10.5281/zenodo.13988124 **ARK:** https://n2t.net/ark:/40704/JCTAM.v1n4a08

1 INTRODUCTION

In the digital age, virtual meetings have become a crucial means of communication and collaboration. According to statistics, the global virtual meeting market reached billions of dollars in 2020, and it is expected to continue growing in the coming years. The COVID-19 pandemic has driven many businesses and educational institutions to shift to remote work and online learning, leading to a surge in demand for virtual meetings. These meetings not only enhance work efficiency but also break geographical barriers, enabling more flexible communication and collaboration among individuals [1]. However, as the number of participants increases and communication needs diversify, traditional, single-mode interaction methods are insufficient to meet user demands. Therefore, improving user experience and interaction quality in virtual meetings is of paramount importance [2].

In this context, multimodal interaction technology emerges as a novel form of interaction that enhances the interactivity and sense of participation in virtual meetings by integrating multiple information delivery methods, such as voice, video, and text. The implementation of this technology relies on advanced speech recognition, natural language processing (NLP), and computer vision technologies, allowing users to achieve more natural and fluid communication during meetings. Moreover, multimodal interaction technology not only facilitates information sharing in meetings but also enhances the quality of interaction among users through features such as emotion analysis [3]. This paper systematically reviews the application of multimodal interaction technology in virtual meetings, analyzes its technical architecture and research progress, explores the needs and challenges in different scenarios, and anticipates future development directions.

2 OVERVIEW OF MULTIMODAL INTERACTION TECHNOLOGY

2.1 Basic Concepts of Multimodal Interaction

Multimodal interaction is a kind of technology that can introduce into different perceptual modalities (like visual, auditory or tactile etc.) to support human computer interaction in order to gain more intuitive community and diversity for the users [4]. Compared with the general one-mode interaction manner, multimodal interface technology can provide humans not only differently artificial information channel alternation and another part of several independent complementary channels in parallel use it to significantly improve system/event would affect on the user-friendly sex.

This technology applied in many areas such as virtual meetings, intelligent assistants, education and health by seeking to allow users interact with machines more continuously.

Multimodal interaction has the characteristics of diverse information and easy to interact with users [5]. The essence of it is to deliver data through the means you communicate (face-to-face, video, in gestures etc) that enriches your communication. During virtual meetings, in addition to speaking with their voices, users can communicate using text input or facial expressions and gestures. This multidimensional way of information presentation is effective to address the communication limitations and compensates for it in traditional singl-mode, which also can improves user's perception ad immersion that are vital especially remote meetings & virtual collaboration scenarios [6].

Moreover, multimodal interaction technology emphasizes emotional expression, enabling the capture and transmission of emotions during the communication process [7]. Through video analysis and speech emotion recognition technologies, the system can assess users' emotional states in real time, such as joy, surprise, and anxiety, thereby enriching the emotional dimension of communication. Research indicates that in face-to-face interactions, non-verbal cues such as facial expressions, tone of voice, and body language play a significant role; the combination of these factors helps to convey users' true emotions more accurately, making virtual meetings or remote communications more vivid and warm. By incorporating these features, multimodal interaction technology not only enhances user experience but also opens new possibilities for human-computer interaction, facilitating more efficient and emotionally rich modes of communication.

2.2 Technological Foundations of Multimodal Interaction

The realization of multimodal interaction relies on several advanced technologies that work synergistically to provide users with a smooth interactive experience. The primary technological foundations include:

Speech Recognition: Speech recognition technology is one of the key components of multimodal interaction [8]. It converts users' spoken language into text, enabling machines to understand and respond to voice commands. In virtual meetings, speech recognition technology facilitates real-time subtitles and automatic meeting minutes, enhancing the efficiency of meetings and the user experience. Modern speech recognition technology can not only recognize multiple languages but also handle different accents and background noise, significantly improving its accuracy and applicability in complex scenarios. Furthermore, advancements in fine-tuning large language models have further enhanced the capabilities of speech recognition systems, particularly in specialized fields such as medical documentation [9]. As demonstrated in recent research, efficient fine-tuning methods enable these models to adapt to specific terminologies and contexts, resulting in improved performance and usability [10]. Beyond their impact on speech recognition, large language models have transformed the broader field of natural language processing (NLP). With their vast parameter sizes and deep contextual understanding, these models excel in generating human-like text, translating between languages, and performing complex tasks like summarization and question-answering with remarkable accuracy [11].

Natural Language Processing (NLP): Natural language processing technology plays an essential role in multimodal interaction by analyzing and understanding text to extract key information and emotional states. NLP is used not only to transform speech recognition results into analyzable text but also to provide deeper understanding through semantic and sentiment analysis. For instance, NLP technology can assist virtual meeting systems in extracting key topics from participants' discussions and assessing the discussion atmosphere through sentiment analysis, thereby automatically generating meeting summaries or offering targeted feedback [12]. With the application of deep learning technologies, the capabilities of NLP in sentiment recognition, topic summarization, and automated question answering have significantly improved.

Computer Vision: Computer vision technology enables systems to capture and interpret users' facial expressions, gestures, and other non-verbal behaviors through video stream analysis. By employing advanced facial recognition, posture estimation, and emotion analysis algorithms, computer vision technology can identify users' emotional states and body movements, providing the system with more contextual information. In virtual meetings, computer vision can help the system recognize which participants are engaged and which may be losing attention, thereby assisting the host in guiding discussions more effectively. Additionally, computer vision technology can be utilized for tasks such as automatically focusing on important speakers and identifying key content, further enhancing the intelligent experience of meetings [13].

The organic integration of these technologies allows multimodal interaction systems to respond to user needs in a more intelligent and intuitive manner. In the future, multimodal interaction technology is expected to further integrate with other technologies (such as Augmented Reality (AR) and Virtual Reality (VR)), driving comprehensive innovation and development in application scenarios like virtual meetings.

2.3 Application Scenarios of Multimodal Interaction in Virtual Meetings

Multimodal interaction technology has a wide range of applications in virtual meetings, covering aspects such as meeting organization, participation, feedback, and

summarization. The goal is to enhance the efficiency, quality, and engagement of meetings. Here are some specific application scenarios:

Real-Time Translation and Transcription: In multilingual environments, real-time translation and transcription functions can instantaneously convert spoken content into text and provide translation support between different languages[14]. This allows participants from diverse countries and cultural backgrounds to communicate seamlessly, increasing global participation in meetings. By integrating speech recognition and natural language processing technologies, participants can easily understand meeting content, preventing information loss due to language barriers.

Sentiment Analysis and Feedback: Multimodal interaction technology can identify participants' emotional states (such as happiness, anger, or confusion) through video analysis techniques, providing real-time feedback to the meeting host. This sentiment analysis capability helps hosts understand the emotional fluctuations of participants, enabling them to adjust the pace, content, or interaction format of the meeting to ensure participants' engagement and satisfaction.

Key Point Extraction and Meeting Summarization: During meetings, the system can automatically extract key points and generate meeting minutes. Utilizing natural language processing technology, the system can identify discussion topics, important decisions, and action items. This not only alleviates the burden on participants but also ensures accurate recording of critical information, aiding in follow-up and implementation by the team.

Enhanced Engagement and Interactivity: By combining video, audio, and text in multimodal interaction, participants can more authentically express their emotions and attitudes in virtual meetings. This enhanced sense of participation can foster deeper discussions and exchanges, allowing participants to experience higher levels of interactivity. Additionally, participants can interact with others through gestures, facial expressions, and other non-verbal cues, enhancing the collaborative atmosphere within the team.

Personalized Learning and Feedback: In specialized virtual meetings, particularly in educational or training settings, the system can offer personalized learning suggestions and content adjustments based on participants' emotional responses and feedback. This personalized support not only improves learning effectiveness but also ensures participants remain engaged throughout the learning process.

Security and Privacy Protection in Remote Meetings: Protecting the privacy and data security of participants is crucial in virtual meetings[15]. Multimodal interaction technology needs to ensure user data encryption and protection while facilitating interaction, preventing information leaks. By implementing secure user authorization and data management measures, trust in virtual meetings among participants can be enhanced.

2.4 Applications of Deep Learning in Virtual Meetings

In recent years, deep learning has played a crucial role in enhancing multimodal interactions in virtual meetings. Liu et al. (2024) introduced a lightweight information split network for infrared image super-resolution, providing a solution to improve visual clarity in low-light or complex environments during virtual meetings[16]. Additionally, Cao et al. (2023) developed an adaptive receptive field U-shaped temporal convolutional network, which excels at recognizing behaviors, especially in detecting participant actions and emotional responses, enhancing real-time interaction within meetings[17].

For security and identity verification in virtual meetings, Li and Zhang (2022) proposed a Siamese network-based cat face recognition model, which can be adapted for biometric recognition in virtual meetings to ensure accurate participant verification[18]. Meanwhile, Zhang et al. (2024) developed an efficient traffic sign detection system based on YOLO-PPA, which has high precision in real-time object detection, making it applicable for real-time monitoring and analysis in virtual meeting environments[19]. Furthermore, Zhang et al. (2024) integrated the Mamba and Transformer-MAT models for long-short range time series forecasting, showcasing its potential in analyzing complex data. This technology can be leveraged to predict participant behavior and process real-time data in virtual meetings[20].

3 TECHNOLOGY ARCHITECTURE AND RESEARCH PROGRESS

3.1 Acquisition and Processing of Multimodal Data

In multimodal interaction systems, the acquisition of multimodal data is the first step toward achieving its functionality, involving the collection and processing of various information sources such as audio, video, and text. Each data modality has its unique collection methods and processing requirements. The accurate acquisition of sensor data is crucial for ensuring the real-time performance and reliability of multimodal systems. For example, the study by Bačić, Feng, & Li (2024) demonstrates that through personalized algorithm optimization, the JY61 IMU sensor can provide more precise motion detection results. This personalized approach can enhance the interaction efficiency of sensor-based multimodal systems[21].

Audio Data Collection: User audio signals are collected through microphones. These audio signals typically contain rich information, such as linguistic content, tone, and emotional cues. To ensure the accuracy of subsequent analyses, audio data must undergo preprocessing after

collection, including background noise reduction and volume normalization. Following this, Automatic Speech Recognition (ASR) technology converts the audio into text, making its content analyzable by natural language processing (NLP) modules[22]. Additionally, emotional recognition can also be derived from audio data by analyzing vocal attributes such as pitch, speaking speed, and pauses to infer the user's emotional state.

Video Data Collection: Users' facial expressions, eye movements, and body language are captured through cameras. These non-verbal cues are critical for understanding user attitudes and emotions. The processing of video data includes action detection, expression analysis, and attention estimation. In meeting scenarios, computer vision technology based on video can real-time identify participants' response states, such as attention levels and emotional changes, aiding the system in providing adaptive feedback.

Text Data Collection: User text information is obtained through input devices such as keyboards, touchscreens, or speech-to-text applications. Text data not only conveys direct linguistic information but can also undergo semantic analysis and sentiment analysis using NLP techniques, extracting more valuable insights such as key points of the meeting, user emotions, and questions.

Data Preprocessing: During the processing of multimodal data, data preprocessing is a critical step, typically involving noise removal, data normalization, and other operations to ensure data quality. Furthermore, feature extraction is the core of multimodal processing. The system needs to extract representative features from a vast amount of raw data for subsequent pattern recognition and modeling. Effective feature extraction can significantly enhance the system's ability to recognize and understand different modalities and lay the foundation for subsequent multimodal fusion[23].

This integrated approach allows multimodal systems to create a richer, more nuanced understanding of user interactions, facilitating more effective communication and collaboration in virtual meetings.

3.2 Intermodal Collaborative Processing Mechanisms

A core issue in multimodal interaction is how to coordinate and integrate data from different modalities, which refers to the collaborative processing between modalities. Data from different modalities possess complementarity, and by merging them, the system can obtain diverse and useful information from each modality. Existing research has proposed several main fusion strategies: early fusion, late fusion, and intermediate fusion. Early fusion involves integrating data at the data level, meaning that multimodal data is directly combined before feature extraction to form a unified data representation. This method can fully utilize the complementary information from each modality, thereby enhancing interaction performance. However, early fusion requires high demands for temporal alignment and quality consistency of the data, making it susceptible to the differences between modalities. For example, differing sampling rates between speech and video can lead to information loss or distortion during fusion.

Compared to early fusion, late fusion involves the independent processing and feature extraction of data from each modality, followed by fusion at the decision level. This approach allows each modality module to operate independently, maintaining processing flexibility. While this method has advantages in modularity and scalability, it may result in inadequate utilization of the complementary information between different modalities, thereby affecting overall performance. Mid-level fusion, on the other hand, combines modalities at the feature level, where preliminary processing and feature extraction of data from each modality occur first, followed by fusion in the feature space. This approach retains the benefits of independent processing for each modality while enabling information sharing between modalities, thus enhancing the overall understanding capability of the system. With the advancement of deep learning technologies, researchers have begun to explore adaptive modality selection methods based on deep learning, which can automatically select the most representative modality for processing and fusion according to the context, making the system more flexible and efficient in complex interaction scenarios[24].

3.3 Deep Learning-Based Multimodal Models

Deep learning applications in multimodal interaction have driven rapid advancements in the entire field. Thanks to deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers, the capability to process multimodal data has been significantly enhanced. These models reduce the complexity of manually designing features by automatically learning high-level representations of the data[25].

In multimodal interaction, CNNs are widely used for video processing. By capturing local image features from video frames, CNNs can effectively recognize details such as facial expressions and body movements. For example, in virtual meetings, CNNs can be employed to recognize participants' facial expression changes in real-time, helping the meeting system better understand the emotional states of attendees[26]. Additionally, 3D Convolutional Networks (3D-CNNs) can capture the spatiotemporal features of videos, enhancing the ability to recognize dynamic behaviors. RNNs and their variants, such as Long Short-Term Memory networks (LSTMs) and Gated Recurrent Units (GRUs), are extensively used for processing speech and text data[27]. They can capture dependencies in time series, enabling the system to better understand contextual information when analyzing continuous speech or text[28]. For example, RNNs

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

can assist speech recognition systems in accurately capturing semantic and emotional shifts during long conversations. The attention mechanism is also widely applied in multimodal interaction, especially when dealing with data that combines audio and visual information. For instance, Transformer models utilize attention mechanisms to automatically select the most relevant features for processing, thereby improving the efficiency and accuracy of data handling[29]. During the multimodal fusion process, attention mechanisms help the system extract important features from different modalities, facilitating efficient information integration.

4 APPLICATIONS OF MULTIMODAL INTERACTION TECHNOLOGY IN VIRTUAL MEETINGS

4.1 Real-time Translation and Transcription Functions

Such an on-demand translation service requires effective speech recognition technologies which are quite advanced and usually comprise three main phases, feature extraction, acoustic modeling and language modeling. Feature extraction is the earliest stage in the speech recognition process where techniques like Mel Frequency Cepstral Coefficients (MFCCs) and the rest turn the segment of audio collected to a format the machine learning models can handle. This step makes sure that the system would be able to effectively extract relevant contents from the speech.

The next stage calls for the acoustic model where there is a relationship established between the phone waves and phonemes and vocabulary mapping. In contemporary times modern acoustic models include LSTMs (Long Short Term Memory) networks and Transformer structures that are deep learning based. LSTMs are effective in models where there are time-based sequences since they can memorize time dependant sequences of the data. LSTM is shown to be a very decent choice in classifying fake news; however, the long training time may pull it back some time[30]. Through self-attention mechanisms that are present in the transformer system, features that are related to each other on a global scale can be captured which greatly enhances the system's accuracy and stability of recognition.

Building on this foundation, language models further optimize the fluency and contextual consistency of generated text. Traditional N-gram models assist text generation by calculating the probabilities of word sequences, but modern neural language models (such as BERT and the GPT series) have gradually become mainstream, generating more natural and coherent text based on complex contextual relationships. By combining these technologies, real-time translation systems can achieve efficient multilingual communication, greatly reducing the challenges posed by language barriers.

Natural Language Processing (NLP) technologies also play a crucial role in transcribing meeting content. During the transcription process, tokenization and syntactic analysis form the basis for understanding the text structure and semantic relationships. By using part-of-speech tagging and dependency parsing, the system can identify subject-verb-object structures within the text, aiding subsequent topic modeling and information extraction. Additionally, topic modeling techniques (such as Latent Dirichlet Allocation (LDA)) can automatically identify the main themes from meeting discussions, enhancing the efficiency of information extraction. To ensure the adequate retention of important information, automatically generating meeting minutes requires the application of extractive and abstractive summarization techniques. These technologies can effectively extract key information and generate accurate, clear meeting records, providing a reliable basis for subsequent discussions and decision-making.

4.2 Application of Sentiment Analysis

Sentiment analysis monitors participants' emotional states in real time by analyzing non-verbal signals such as facial expressions and vocal characteristics. This involves technologies like facial expression recognition, voice feature analysis, and emotional feedback mechanisms. Facial expression recognition technology leverages Convolutional Neural Networks (CNN) and Regional Convolutional Neural Networks (R-CNN) to analyze participants' facial expressions in real-time. These models, trained on large datasets, can accurately identify various emotional states (such as anger, happiness, sadness, and confusion), providing feedback to the meeting host about the participants' emotional states, which helps in adjusting the flow of the meeting.

Voice feature analysis is another critical component of sentiment analysis. By analyzing vocal features such as pitch, volume, and speech rate, combined with algorithms like Support Vector Machines (SVM) and Long Short-Term Memory networks (LSTM), the system can classify the emotional states of participants. Changes in pitch, variations in volume, and differences in speech speed reflect the emotional state of the participants, and the system provides real-time feedback based on these features. Additionally, the emotional feedback mechanism can offer the analysis results to the host in real-time, allowing them to adjust the meeting flow according to the participants' emotional states. For instance, if the system detects confusion or dissatisfaction from most participants, it can automatically prompt the host to clarify a concept or shift the direction of the discussion, thereby enhancing the interactivity of the meeting and improving participant satisfaction.

4.3 Automatic Meeting Minutes Generation

The process of automatically generating meeting minutes relies on natural language processing (NLP)

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

technology, primarily involving three steps: speech transcription, key information extraction, and structured summary generation[31]. First, advanced speech recognition models such as DeepSpeech and Google Speech-to-Text are used to convert spoken content in the meeting into text. These models, trained on large datasets using deep learning techniques, can achieve high accuracy in speech-to-text conversion across various languages and accents, ensuring reliable transcription.

After obtaining the text, key information extraction techniques identify important decisions, action items, and discussion points from the meeting. This process typically involves Named Entity Recognition (NER) and relation extraction techniques to ensure that the system efficiently extracts crucial information mentioned by the participants, laying a solid foundation for the meeting minutes. Lastly, structured summary generation techniques apply sequence-to-sequence (Seq2Seq) models to create structured meeting summaries that include titles, themes, decisions, and action items. The Seq2Seq model's advantage lies in its ability to handle variable-length inputs and outputs, generating more coherent and natural summaries. Additionally, Seq2Seq models incorporating attention mechanisms can focus on key statements made by participants during the generation process, ensuring that the produced minutes are more representative and accurate.

5 TECHNICAL CHALLENGES AND FUTURE OUTLOOK

5.1 Data Privacy and Security

Multimodal interaction systems raise the need for defending user's voice, video, text, and other data due to the increased use of cloud-based systems. To counter those instances, strong and enhanced encryption methods like homomorphic encryption and differential privacy are necessary so the data is secure during the ever evolving processing phases. Homomorphic encryption helps in computing over the encrypted data rather than decrypting it hence protecting users' data privacy while differential privacy consists of adding noise to the data in order to conceal users' identities and avert data leakage[32]. Therefore, in the upcoming studies, the provision of such demands will provide a detailed amount of attention on how to blend these privacy protection solutions and tools within the systems being utilized without affecting real-time interactivity of the system.

In addition, federated learning as an emerging trend is advantageous because there is no need to send the actual data since the training of the model can still be carried out. Therefore, by training the model at each participant device and sharing only the trained model parameters to the central server, the chances of data loss and/or leakage are very minimal. This approach is particularly important for managing multilevel data during virtual meetings.

5.2 Processing and Integration of Multimodal Data

The efficient processing and integration of data from different modalities remains one of the key technical issues in multimodal interaction systems. Every modality like voice, image, text is quite specific in its characteristics and its temporal nature, and traditional approaches to these problems do not adequately address the inter-domain problem. Deep learning based adaptive fusion models, have become quite important in this regard.

Cross-modal attention mechanisms are very critical in the integration of the different modalities in the system. These mechanisms use context information to determine the importance of different modalities and thus control the behavior of the system to employ that most applicable modality. For instance, in a meeting, if a visual signal such as a facial expression differs with what the person is saying, then the attention head decides which is the primary focus and this increases the accuracy of the decision making.

A further technique that may also be applied is the mid-fusion. This consists of first extracting features from each modality and then integrating them at the feature space. This technique adds the disadvantages of collating beneficial resources from the different modalities and merging them all at once. It preserves the pros of processing the signals independently of one another and enhances the system's comprehension of the interaction. Future works could be focused on

5.3 Model Optimization and Efficiency Improvements

The effectiveness of the real-time multimodal interaction systems must be enhanced while dealing with large-scale online meetings[33]. However, deep learning's complex model architecture creates a disadvantage in resource use through costly computational excesses when working with video, audio and text data at the same time. To resolve this issue, future work can focus on pruning and quantization techniques. Model pruning involves the finding of less contributing model parameters, thereby lowering model complexity and increasing efficiency. Quantization is the conversion of floating point computation to low bit mathematics reducing the use of hardware resources. Through Note these two processes are useful in allowing the systems improve their realtime capability especially with high concurrency user data handling so as to meet the extent of the demands for real-time interaction.

At the same time, the technique that allows to integrate information embedded in large and complicated models into smaller compact forms namely knowledge distillation can achieve great accuracy as well as save on computation

resources. In multimodal interaction scenarios, such models can also be trained through knowledge distillation, which in turn speeds up the process of dealing with real-time data thereby easing the pressure off the server. Distributed computing and edge computing will also enhance the efficiency of the system further. With the increase in the number of participants in virtual meetings, alone server will not be sufficient to gather and process all the data. Distributed computing architectures can distribute the processing tasks involving multimodal data over several servers or devices, thereby increasing the processing speed of the tasks significantly. Concurrently, some of the operations can be shifted to edge devices closer to where the data is generated such as personal user terminals or network edge nodes thereby lowering the congestion on network data transfer and improving real time responsiveness of the system. The integration of these technologies offers a robust technical assistance to ensure seamless execution of extensive virtual meetings.

6 CONCLUSION

Multimodal interaction technology has been gradually adopted into virtual meetings, which has improved the effectiveness of communication in terms of participation and time management, especially now that the trend towards remote work and global teamwork is on the rise. Enabling different technologies such as speech recognition, NLP, computer vision and sentiment analysis inside the virtual meetings is expected to bring a better experience by making the communication more natural rather than official. The addition of features such as real-time interpretation and transcription, sentiment analysis and feedback, as well as auto drafting of meeting minutes aids in communication by minimizing differences in languages and omission of vital information, making the participants more engaged and satisfied. At the same time, the adoption of multimodal interaction technology brings out technical issues such as data privacy and security, multimodal data processing and fusion, as well as model optimization and efficiency enhancement. However, with the increase of deep learning, distributed computing among other advanced technologies, more flexibility and efficiency are expected in virtual meetings, especially in large scale concurrent complicated interaction situations. By optimizing algorithms and applying techniques such as federated learning, model pruning, and quantization, the real-time and interactive nature of virtual meetings will be further enhanced. Overall, multimodal interaction technology will drive the development of virtual meetings in the future, becoming an important tool for improving the efficiency of global collaboration and communication in enterprises and organizations. As relevant technologies continue to improve and innovate, the application scenarios of virtual meetings will become more widespread and in-depth, further promoting efficient and immersive remote collaboration and communication.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

LIN, Weikun

Shandong University of Science and Technology, China.

REFERENCES

[1] Brennan, J. N., Hall, A. J., & Baird, E. J. (2023). Online case-based educational meetings can increase knowledge,

skills, and widen access to surgical training: The nationwide Virtual Trauma & Orthopaedic Meeting series. The Surgeon, 21(5), e263-e270.

[2] Bohao, D., & Desheng, L. (2021, May). User visual attention behavior analysis and experience improvement in virtual meeting. In 2021 IEEE 7th International Conference on Virtual Reality (ICVR) (pp. 269-278). IEEE.

[3] Bahreini, K., Nadolski, R., & Westera, W. (2016). Data fusion for real-time multimodal emotion recognition through webcams and microphones in e-learning. International Journal of Human–Computer Interaction, 32(5), 415-430.

[4] Multimodal interaction is a technology that achieves human-computer interaction by integrating multiple sensory modalities, such as vision, hearing, and touch.

[5] The notable features of multimodal interaction include the diversity of information and the sense of user engagement.

[6] Wang, D. (Ed.). (2016). Information Science and Electronic Engineering: Proceedings of the 3rd International Conference of Electronic Engineering and Information Science (ICEEIS 2016), January 4-5, 2016, Harbin, China. CRC Press.

[7] Rakkolainen, I., Farooq, A., Kangas, J., Hakulinen, J., Rantala, J., Turunen, M., & Raisamo, R. (2021). Technologies for multimodal interaction in extended reality—a scoping review. Multimodal Technologies and Interaction, 5(12), 81.

[8] Deng, L., Wang, Y. Y., Wang, K., Acero, A., Hon, H. W., Droppo, J. G., ... & Huang, X. D. (2004). Speech and language processing for multimodal human-computer interaction. Real World Speech Processing, 87-113.

[9] Y. Chen, J. Zhao, Z. Wen, Z. Li, and Y. Xiao, Temporalmed: Advancing medical dialogues with time-aware responses in large language models, in Proceedings of the 17th ACM International Conference on Web Search and Data Mining, 2024, pp. 116–124.

[10] Leong, H. Y., Gao, Y. F., Shuai, J., Zhang, Y., & Pamuksuz, U. (2024). Efficient Fine-Tuning of Large Language Models for Automated Medical Documentation. arXiv preprint arXiv:2409.09324.

[11] Y. Chen, Q. Fu, Y. Yuan, Z. Wen, G. Fan, D. Liu, D. Zhang, Z. Li, and Y. Xiao, Hallucination detection: Robustly discerning reliable answers in large language models, in Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 245–255.

[12] Kachhoria, R., Daga, N., Ramteke, H., Akotkar, Y., & Ghule, S. (2024, March). Minutes of Meeting Generation for Online Meetings Using NLP & ML Techniques. In 2024 International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 1-6). IEEE.

[13] Qu, M. (2024). High Precision Measurement Technology of Geometric Parameters Based on Binocular Stereo Vision Application and Development Prospect of The System in Metrology and Detection. Journal of Computer Technology and Applied Mathematics, 1(3), 23-29.

[14] Lavanya, R., Gautam, A. K., & Anand, A. (2024, April). Real Time Translator with Added Features for Cross Language Communiciation. In 2024 10th International Conference on Communication and Signal Processing (ICCSP) (pp. 854-857). IEEE.

[15] Atawneh, S., Alshammari, Z., Mousa, A. L., & Shawar, B. A. A Security Framework for Addressing Privacy Issues in the Zoom Conference System.

[16] Liu, S., Yan, K., Qin, F., Wang, C., Ge, R., Zhang, K., ... & Cao, J. (2024, August). Infrared image super-resolution via lightweight information split network. In International Conference on Intelligent Computing (pp. 293-304). Singapore: Springer Nature Singapore.

[17] Cao, J., Xu, R., Lin, X., Qin, F., Peng, Y., & Shao, Y. (2023). Adaptive receptive field U-shaped temporal convolutional network for vulgar action segmentation. Neural Computing and Applications, 35(13), 9593-9606.

[18] Li, H., & Zhang, W. (2022, November). Cat face recognition using Siamese network. In International Conference on Artificial Intelligence and Intelligent Information Processing (AIIIP 2022) (Vol. 12456, pp. 640-645). SPIE.

[19] Zhang, J., Zhang, W., Tan, C., Li, X., & Sun, Q. (2024). YOLO-PPA based efficient traffic sign detection for cruise control in autonomous driving. arXiv preprint arXiv:2409.03320. https://arxiv.org/abs/2409.03320

[20] Zhang, W., Huang, J., Wang, R., Wei, C., Huang, W., & Qiao, Y. (2024). Integration of Mamba and Transformer--MAT for Long-Short Range Time Series Forecasting with Application to Weather Dynamics. arXiv preprint arXiv:2409.08530.

[21] Bačić, B., Feng, C., & Li, W. (2024). JY61 IMU SENSOR EXTERNAL VALIDITY: A FRAMEWORK FOR ADVANCED PEDOMETER ALGORITHM PERSONALISATION. ISBS Proceedings Archive, 42(1), 60.

[22] Kheddar, H., Hemis, M., & Himeur, Y. (2024). Automatic speech recognition using advanced deep learning approaches: A survey. Information Fusion, 102422.

[23] Liu, W., Zhou, L., Zeng, D., Xiao, Y., Cheng, S., Zhang, C., ... & Chen, W. (2024). Beyond Single-Event Extraction: Towards Efficient Document-Level Multi-Event Argument Extraction. arXiv preprint arXiv:2405.01884.

[24] Zhang, Y., Wang, F., Huang, X., Li, X., Liu, S., & Zhang,

H. (2024). Optimization and application of cloud-based deep learning architecture for multi-source data prediction. arXiv. https://arxiv.org/abs/2410.12642

[25] Jiang, L., Yang, X., Yu, C., Wu, Z., & Wang, Y. (2024, July). Advanced AI framework for enhanced detection and assessment of abdominal trauma: Integrating 3D segmentation with 2D CNN and RNN models. In 2024 3rd International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC) (pp. 337-340). IEEE.

[26] Kord, S., Taghikhany, T., & Akbari, M. (2024). A novel spatiotemporal 3D CNN framework with multi-task learning for efficient structural damage detection. Structural Health Monitoring, 23(4), 2270-2287.

[27] Jiang, H., Qin, F., Cao, J., Peng, Y., & Shao, Y. (2021). Recurrent neural network from adder's perspective: Carry-lookahead RNN. Neural Networks, 144, 297-306.

[28] Liu, W., Cheng, S., Zeng, D., & Qu, H. (2023). Enhancing document-level event argument extraction with contextual clues and role relevance. arXiv preprint arXiv:2310.05991.

[29] Ge, Q., Li, J., Wang, X., Deng, Y., Zhang, K., & Sun, H. (2024). LiteTransNet: An interpretable approach for landslide displacement prediction using transformer model with attention mechanism. Engineering Geology, 331, 107446.

[30] Yu, P., Cui, V. Y., & Guan, J. (2021, March). Text classification by using natural language processing. In Journal of Physics: Conference Series (Vol. 1802, No. 4, p. 042010). IOP Publishing.

[31] Jin, Y., Zhou, W., Wang, M., Li, M., Li, X., & Hu, T. (2024, June). Online learning of multiple tasks and their relationships: Testing on spam email data and eeg signals recorded in construction fields. In 2024 5th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA) (pp. 463-467). IEEE.

[32] Cao, Y., Weng, Y., Li, M., & Yang, X. The Application of Big Data and AI in Risk Control Models: Safeguarding User Security.

[33] Yuyan Chen, Yichen Yuan, Panjun Liu, Dayiheng Liu, Qinghao Guan, Mengfei Guo, Haiming Peng, Bang Liu, Zhixu Li, and Yanghua Xiao. 2024e. Talk funny! a large-scale humor response dataset with chain-ofhumor interpretation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 17826–17834.

[34] Yu, P., Cui, V. Y., & Guan, J. (2021, March). Text classification by using natural language processing. In Journal of Physics: Conference Series (Vol. 1802, No. 4, p. 042010). IOP Publishing.