Multi-Modal Perception and Understanding: Application of Large Model in Real-Time Video Analysis

Chen Xi[1,a,*], Ma Baijun[2,b], Li Haoxuan[3,c]

[1]Beijing University of Posts and Telecommunications, Beijing, Chnia

[2]Zhengzhou Fengyang Foreign Language School, Zhengzhou, China

[3]Renchao Branch Experimental School of Renmin University Affiliated Middle School, Beijing, China

[a]2021213046@bupt.cn, [b]1637098361@qq.com, [c]13311058108@163.com

Keywords: Multimodal Perception; Real-Time Video Analysis; Large Model; Applications

Abstract: With the advent of artificial intelligence technology into a new era, research and application of multi-modal perception and understanding have reached a stage of high-quality development. This paper focuses on multi-modal perception and understanding in real-time video analysis. It introduces scientific propositions to enhance multi-modal perception and understanding in this context. Based on the dynamic evolution of multi-modal perception and understanding development, a theoretical analysis framework for developing multi-modal perception and understanding is constructed according to the inherent logic of real-time video analysis. This framework can explain the mechanism of multi-modal perception and understanding development, jointly generated by the real-time video analysis mechanism and the cyclic mechanism involving multi-modal perception, understanding, and large models. The potential for advancing the goal of high-quality real-time video analysis is further explored from technical challenges and practical implications related to developing multi-modal perception and understanding. The purpose of developing modal perception and understanding is to provide multi-modal perception and understanding that meet the expected real-time video analysis standards, continuously improve the quality of real-time video analysis, and enhance user satisfaction. To achieve high-quality development of real-time video analysis measures such as strengthening data control based on the internal circulation of data quality, constructing a multi-modal perception and understanding model, establishing a mechanism for interaction and feedback between real-time video analysis quality perception, and setting up an evaluation system for real-time video analysis efficiency and accuracy should be implemented. These actions will promote the application of multi-modal perception and understanding and effectively meet the requirements of real-time video analysis.

1. Introduction

Multi-modal perception and understanding are two of the primary responsibilities in artificial intelligence and are general terms for intelligent system applications [1]. It can be divided into basic perception and non-basic perception, which consist of sensors and algorithms. Since the rise of deep learning technology, multi-modal perception and understanding have become the key to intelligent system applications, with the efficiency and accuracy of real-time video analysis having become evaluation indicators. Unlike traditional video analysis, multi-modal perception and understanding emphasize real-time, accuracy, and robustness more. Therefore, the issue of applying multi-modal perception and understanding in real-time video analysis has been proposed, and large models provide new solutions for real-time video analysis.

Multi-modal perception and understanding originate from deep learning with data as its core, and its theoretical foundation contains the ability to process heterogeneous data from multiple sources. It is also an essential tool for intelligent systems to understand the environment. From a technical perspective, multi-modal perception and understanding aim to understand the real world comprehensively. However, it is done by modernizing intelligent systems by integrating large models

DOI: 10.25236/icamfss.2024.008

and multi-source data. However, this is only at the theoretical research level. To this day, multi-modal perception and understanding have embarked on a unique path of combining theory and application. The comprehensive promotion of large models not only rewrites the technological landscape of real-time video analysis and reflects the progress of artificial intelligence technology but also rewrites the development path of intelligent systems, exerting a profound impact on multi-modal perception and understanding. Therefore, discussing the application of multi-modal perception and understanding in real-time video analysis must have a forward-looking and global perspective. Thus, the application proposition of multi-modal perception and understanding has been proposed in intelligent system applications.

In short, multi-modal perception and understanding are the necessary conditions and guarantees for achieving real-time video analysis in intelligent systems. From the perspective of technological development, significant progress has been made in multi-modal perception and understanding, but there are still shortcomings in data fusion, model training, and real-time performance. Multi-modal perception and understanding have not yet fully identified a practical path for real-time video analysis, and efforts are still being made. Therefore, further research is needed on multi-modal perception and understanding, which is not only a requirement for technological development but also a practical requirement for intelligent system applications.

Based on the above background analysis, this paper proposes a multi-modal perception and understanding method that combines large models to improve the efficiency and accuracy of real-time video analysis. The problem of multi-modal data fusion and processing is solved through deep learning theory and big data methods. Its main content is the theoretical foundation, technical challenges, and multi-modal perception and understanding solutions. It effectively addresses the technical risks in real-time video analysis, which has theoretical and practical significance.

2. Theoretical Basis of Multi-Modal Perception and Technical Requirements of Real-Time Video Analysis

2.1. Definition and Development of Multi-Modal Perception

Multi-modal perception is a concept that has developed in parallel with artificial intelligence technology. It is imbued with the concept of interdisciplinary cooperation, highlighting the practical orientation of technological integration and reflecting the comprehensive strategy of information technology since the 21st century. However, it is still hard to obtain a unified understanding when we attempt to construct the definition and essence of multi-modal perception using certain traditional perception standards.

2.2. Application of Real-Time Video Analysis in Multi-Modal Perception

Real-time video analysis is an essential standard in intelligent monitoring, representing the real-time performance of multi-modal perception applications. The academic and industrial communities have discussed different definitions of real-time video analysis from various perspectives. Some scholars believe that real-time video analysis refers to how multi-modal data is processed and interpreted in real-time requirements or the particular demonstration of multi-modal perception and understanding under real-time requirements. Real-time video analysis is particularly challenging because it falls under applied science, aiming for real-time performance. The practical history of real-time video analysis can be traced back to the early development of video surveillance systems, whose main activities include acquiring, processing, and analyzing video streams. The concept and technology of real-time video analysis are closely related to the progress of artificial intelligence, big data, and cloud computing. Real-time video analysis is essential for intelligent systems through multi-modal perception and understanding. The main contribution of real-time video analysis theory in the era of deep learning is to improve the efficiency and accuracy of analysis. Therefore, the concept of real-time video analysis initially focused on multi-modal data processing and interpretation based on real-time standard attributes [2].

3. The Critical Role of the Large Model in Real-Time Video Analysis

3.1. Construction and Optimization of Large Model

Compared with large models, traditional models emphasize the unidirectional relationship between algorithms and data more, while large models have the characteristics of combining deep learning and data-driven approaches. Although some scholars argue a direct relationship exists between model size and performance, most scholars advocate that larger models can provide more rational evaluations of complex tasks. Researchers proposed a deep neural network model with many parameters and levels, which became a typical tool for handling complex tasks, thus developing the concept of large models. These scholars believe that large models have a high degree of nonlinearity and are "smart amplifiers". Large models can unleash their enormous potential only when data and computing power are sufficient. Therefore, large model optimization results from improving real-time video analysis performance. Some scholars have also summarized large models as dual path models, namely data-based pre-training models, and task-based fine-tuning models. The former focuses on learning the general features of the data, while the latter focuses on learning the task's specific features, namely the model's adaptability. Although large models have experienced some failures in practice, they can significantly improve the efficiency and accuracy of real-time video analysis in the long run. As a result, the application concept of large models has gradually become a consensus in artificial intelligence research and practice.

3.2. The Specific Application of the Large Model in Real-Time Video Analysis

Figure 1 explained the application of the large model in real-time video analysis.

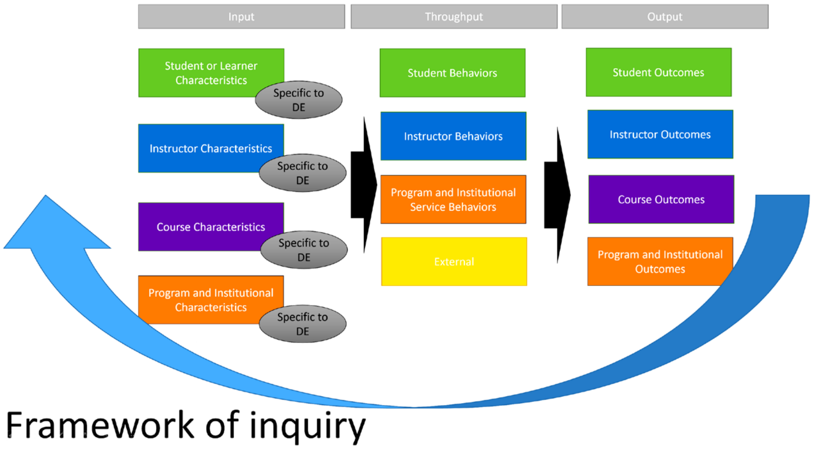

Figure 1: Application of the large model in real-time video analysis

The essence of the concept of large models focuses on complex problems in real-time video analysis. The big model is the application of deep learning thinking in intelligent monitoring. To overcome the shortcomings of traditional algorithms in processing large-scale video data, it has entered the research field as a new alternative mode – the deep learning framework. The basic concept of this framework is that large models should ensure effective implementation of real-time video analysis, set professional standards for video analysis output, capture critical information in videos through techniques such as convolutional neural networks, and use deep learning methods to measure the complexity of video content. The large model framework has restructured real-time video analysis,emphasizing enhancing the model's generalization ability and building video data's real-time accuracy, stability, and reliability [3].

4. Technical Challenges and Solutions of Multi-Modal Perception and Understanding

4.1. Data Fusion and Processing

Data fusion and processing are the main technical challenges of multi-modal perception and understanding, emphasizing the complexity of integrating multi-source heterogeneous data. It directly reflects the comprehensive status of the system's perception and understanding of the environment through the algorithm. Some essential elements of the development of data fusion technology are gradually taking shape, and data processing efficiency and various performance evaluation systems are also receiving progressive attention. However, from a practical application perspective, some data fusion and processing practices are still in their early stages, contradicting the theoretical logic framework and generation mechanism of multi-modal perception and understanding, leading to accuracy and real-time performance issues [4].

4.2. Model Training and Optimization

From a technical perspective, model training and optimization are the fundamental steps of deep learning and the core manifestation of large model performance. Therefore, multi-modal perception and understanding are primarily data-driven in generating logic. Large models are the primary technical means for real-time video analysis and the performance mainstay of intelligent monitoring systems. Currently, model training is divided into three main forms to strengthen quality control from an algorithmic perspective: The first is pre-training. Pre-training balances model universality and task-specific adaptability is clarified. The second is fine-tuning. By establishing dataset standards and evaluation standards, and making these standards publicly available to the research community, standardized control of model performance is achieved. The third is the continuous optimization of the model's internal processes. Researchers and others have recently used deep learning methods to improve model training efficiency and performance. However, compared to the ideal state, the generalization of current large models still needs further improvement.

4.3. Efficiency and Accuracy of Real-Time Video Analysis

The fundamental difference between real-time and traditional video analysis lies in their efficiency and accuracy attributes. The efficiency standards and accuracy criteria for real-time video analysis aim to process video data quickly and accurately. The development of real-time video analysis mainly reflects the progress of technology and the needs of applications. In the technical framework of real-time video analysis, accurate recognition, rapid response, stable operation, and efficient processing are the core values and highest principles for developing real-time video analysis. The diversity of current video data types and scenario differences lead to a complex situation in real-time video analysis. Although the application of large models significantly improves the efficiency and accuracy of real-time video analysis, due to the limitations of model structure and algorithms, real-time video analysis lacks a perfect self-optimization mechanism. Therefore, the "weakness" in real-time video analysis affects its performance in practical applications [5].

5. Application Cases and Analysis of Multi-Modal Perception and Understanding

5.1. Case 1: Intelligent Monitoring System

From an application perspective, traditional monitoring systems cannot accurately provide the multi-modal data required for real-time video analysis. Users' performance evaluation of intelligent monitoring systems takes satisfaction evaluation as the primary form, but traditional systems lack relevant information and processing mechanisms for multi-modal data fusion. The core of this problem may be technical limitations. In intelligent monitoring, monitoring systems are often described as "smart eyes", whose video content analysis directly reflects the system's level of intelligence. However, traditional systems mainly focus on visual information and have limited perception of other modal data, such as sound and temperature. Usually, the fusion of multi-modal data is challenging to obtain or measure. Asymmetric information and imperfect technology directly lead to obstacles in improving the performance of intelligent monitoring systems [6].

5.2. Case 2: Unmanned Vehicles

From a technical perspective, the limitations of sensor technology have long constrained the ability to perceive the environmental perception of autonomous vehicles. Since the 21st century, unmanned vehicles integrating multiple sensors have reshaped environmental perception through multi-modal perception and understanding [7]. However, traditional sensor systems' drawbacks still constrain unmanned vehicles' performance. Not only due to the accuracy and response speed of sensors but also the complexity and diversity of the environment, the environmental perception of autonomous vehicles still needs to be improved. Under the premise of multi-modal perception and understanding, data fusion is seen as a direct way to enhance perception ability. However, the actual effect of a perception system dominated by a single sensor on autonomous vehicles remains discussed. Meanwhile, the difficulties in data processing lead to a lack of real-time capability and accuracy in unmanned vehicles. Therefore, multi-modal perception and understanding do not always to achieve the goals of autonomous vehicles. Multi-modal perception and understanding is not only a technical challenge but also faces practical application challenges.

5.3. Case 3: Telemedicine Diagnosis

Telemedicine diagnosis cannot be avoided as a critical application of multi-modal perception and understanding in data acquisition and processing. In telemedicine diagnosis mechanisms, multi-modal perception is a standard and effective diagnostic tool that plays an essential role in patient health analysis, making telemedicine both a technical and an applied concept. As a result, telemedicine diagnosis, primarily based on multi-modal data, has become a key mechanism for medical diagnosis. The practical interpretation of telemedicine is generally a diagnostic path gradually formed based on multi-modal perception and understanding, although this path involves attempts at technological innovation. Telemedicine has always revolved closely around multi-modal perception and understanding, from data collection to analysis. Telemedicine should strive to improve diagnostic accuracy to meet medical quality requirements [8]. However, the amplification of data brings a dilemma: data processing efficiency. Overall, there is still room for improvement in data processing and diagnostic accuracy in telemedicine, and its technology needs further improvement, which is also an essential task of telemedicine diagnosis.

6. Conclusion

Multi-modal perception and understanding have entered a new era of intelligent system applications, which poses new challenges and requirements for real-time video analysis. Multi-modal perception is not only a symbol of the ' intelligence ' of intelligent systems, an essential means of real-time video analysis, but also an urgent need to achieve efficient video analysis and maintain the performance of intelligent systems, reflecting the inherent requirements of developing intelligent systems. This paper constructs the theoretical analysis framework and practical mechanism of multi-modal perception and understanding in this context. In recent years, modern information technologies such as large models have driven the progress of real-time video analysis. By empowering real-time video analysis through multi-modal perception and understanding and improving the accuracy and scientificity of analysis, its value aligns with the inherent logic of intelligent system development. Therefore, methods based on large models also provide a new path for real-time video analysis. In summary, the continuous improvement and development of multi-modal perception and understanding contribute to better addressing real-time video analysis challenges and promoting high-quality intelligent systems development.

References

[1] Zhu W, Wang X, Gao W. Multimedia intelligence: When multimedia meets artificial intelligence[J]. IEEE Transactions on Multimedia, 2020, 22(7): 1823-1835.

[2] Amer A, Dubois E, Mitiche A. A real-time system for high-level video representation: application to video surveillance[C]//Image and Video Communications and Processing 2003. SPIE, 2003, 5022: 530-541.

[3] Feng W, Ji D, Wang Y, et al. Challenges on large scale surveillance video analysis[C]//Proceedings of the IEEE conference on computer vision and pattern recognition

workshops. 2018: 69-76.

[4] Nousias S, Pikoulis E V, Mavrokefalidis C, et al. Accelerating deep neural networks for efficient scene understanding in multi-modal automotive applications[J]. IEEE Access, 2023, 11: 28208-28221.

[5] Li C, Yang B, Ding H, et al. Real-time video-based smoke detection with high accuracy and efficiency[J]. Fire Safety Journal, 2020, 117: 103184.

[6] Fujino Y, Kitagawa K, Furukawa T, et al. Development of vehicle intelligent monitoring system (VIMS)[C]//Smart Structures and Materials 2005: Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems. SPIE, 2005, 5765: 148-157.

[7] Chen Q, Xie Y, Guo S, et al. Sensing system of environmental perception technologies for driverless vehicle: A review of state of the art and challenges[J]. Sensors and Actuators A: Physical, 2021, 319: 112566.

[8] Pappas Y, Vseteckova J, Mastellos N, et al. Diagnosis and decision-making in telemedicine[J]. Journal of patient experience, 2019, 6(4): 296-304.