Digital Multimedia and Visualization Technology

E-ISSN: 3123-9994

Research Article

A Deep Learning Based Approach to Real Time Video Content Analysis and Visualization for Intelligent Human Computer Interaction in Multimedia Systems

Arsito Ari Kuncoro [1], Siswanto [2], Siti Kholifah [3], Ratma Dewi[4]

[1] Universitas Sains dan Teknologi Komputer, Indoensia arsito@stekom.ac.id
[2] Universitas Sains dan Teknologi Komputer, Indoensia siswanto@stekom.ac.id
[3] Universitas Sains dan Teknologi Komputer Indoensia
[4] Universitas Gajah Putih Acah

* Corresponding Author : Arsito Ari Kuncoro

Abstract: This study explores the integration of deep learning-based approaches in real-time video content analysis for intelligent human-computer interaction (HCI) in multimedia systems. Traditional video analysis techniques, such as rule-based methods and offline processing, struggle with real-time performance and adaptability to complex video data. In contrast, the deep learning model used in this research, particularly Convolutional Neural Networks (CNNs), provides high accuracy in object detection, feature extraction, and real-time processing. The integration of CNNs with interactive visualization modules enables dynamic adjustments to video content based on user interactions, ensuring a seamless and engaging user experience. The system was benchmarked in terms of its processing speed, accuracy, and responsiveness, showing significant improvements over traditional approaches in real-time video analysis. Moreover, the study demonstrates that combining deep learning with real-time visualization enhances the efficiency of interactive multimedia applications, making it suitable for dynamic environments such as surveillance, security monitoring, and interactive media. Despite the system's strong performance, challenges such as computational demands in high-resolution video processing were identified, highlighting the need for further optimization. Future work will focus on optimizing the system for different hardware platforms, incorporating multimodal inputs, and refining deep learning models to address computational bottlenecks. This research contributes to advancing HCI by providing insights into the integration of deep learning for real-time video content analysis, which is pivotal for enhancing the interactivity and adaptability of intelligent multimedia systems.

Keywords: Deep learning; Video analysis; Convolutional networks; Human-computer interaction; Real-time processing.

Received: 21, November 2025
Revised: 10, December 2025
Accepted: 29, December 2025
Published: 15, January 2026
Curr. Ver.: 20, January 2026

1. Introduction

Real-time video analysis in intelligent multimedia systems presents significant challenges, primarily due to the high computational demands involved in processing continuous video streams. These systems need to extract meaningful features quickly and efficiently from large amounts of data. The complexity arises from the need for advanced algorithms capable of handling tasks such as 3D reconstruction, feature extraction, and image registration, all of which require significant computational resources [1]. For example, tasks like feature extraction often involve multiple intermediate processes that are computationally intensive and demand considerable processing power [2].

To achieve real-time performance, parallel processing techniques are commonly employed to distribute computational workloads across multiple processors. Methods like dynamic Bayesian networks and particle filters have been used to enhance the efficiency of feature extraction and classification [1]. However, these approaches can lead to challenges in maintaining accuracy when the tasks are divided among several processors. Another emerging solution to address these challenges is edge computing, where processing tasks are offloaded

from cloud servers to edge devices. This approach reduces latency and bandwidth consumption, making real-time video analytics more feasible and efficient [3].

Efficient feature extraction is a critical component in real-time video analysis. Hybrid methods that combine various algorithms, such as the GLCM-DWT technique, have been developed to improve accuracy while reducing the storage space required [4]. Reconfigurable hardware architectures, such as FPGAs, are also being utilized to optimize the feature extraction process. These platforms perform critical operations like convolutions and morphology calculations that are necessary for processing video data in real-time [5]. Advanced algorithms focusing on moving object detection and human activity recognition are continually being refined to enhance both accuracy and time performance in dynamic video analysis tasks [6].

Finally, scalability and interoperability remain major concerns in real-time video analysis. Systems need to be capable of handling large data streams from multiple sources while ensuring they meet the required quality of service (QoS) standards for timeliness and accuracy. Distributed infrastructures, such as the Video Intelligence Platform (VIP), are designed to support near real-time video stream analysis and facilitate the scalability required for modern multimedia systems [6]. To improve the quality of the video streams, techniques such as deep convolutional neural networks are used to mitigate noise and distortion caused by environmental factors, ultimately enhancing the accuracy and reliability of real-time video analysis [7].

Human-Computer Interaction (HCI) plays a critical role in shaping the design, evaluation, and implementation of interactive systems that are user-friendly and efficient. HCI focuses on how people interact with computers, ensuring that technology aligns with users' cognitive processes and mental models, which ultimately enhances usability and user experience (UX) [8]. In the context of multimedia systems, particularly interactive applications, HCI's significance is magnified, as it directly influences the quality of user engagement and the effectiveness of technology in meeting user needs [9]. By emphasizing intuitive design and real-time interaction, HCI fosters systems that are not only functional but also capable of providing engaging and seamless experiences.

The integration of multimodal interfaces in multimedia systems has become a key strategy in enhancing user interaction. HCI aims to combine auditory, visual, and haptic feedback to create immersive experiences that engage multiple senses simultaneously [10]. This approach not only improves the overall user experience but also addresses diverse user expectations and learning styles, helping to manage the inherent complexity of multimedia applications. Furthermore, the interactive aspect of multimedia systems plays a pivotal role in making these systems more engaging. It is not enough for multimedia to simply deliver information; it must also enable stimulating interactions that deepen user involvement and promote a more dynamic, responsive experience [11].

To advance the effectiveness of HCI in multimedia systems, the application of deep learning techniques for real-time video content analysis has emerged as a promising approach. Using advanced deep learning models, such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), real-time video analytics can capture spatiotemporal features, which are critical for understanding human actions and behaviors [12]. These models, when integrated into intelligent video systems, provide the foundation for human-centered interaction, enabling systems to adapt to user behaviors and preferences. The use of frameworks like YOLO for object detection and DeepSORT for tracking ensures that video content is analyzed in real-time, offering precise identification and tracking of individuals [13].

By incorporating human-centric design into these intelligent systems, HCI can significantly enhance the way users interact with multimedia technologies. Deep learning models can capture social interactions, human behavior, and emotional responses, helping to create more human-like AI systems that respond effectively to user inputs [14]. This real-time, adaptive approach not only boosts interactivity but also supports multimodal interfaces, which enhance the naturalness and intuitiveness of user interactions, leading to greater user satisfaction [15]. The potential for such systems in smart cities, surveillance, and intelligent transportation applications highlights the growing importance of HCI in the development of future multimedia systems.

2. Literature Review

Previous Approaches to Real-Time Video Analysis

Real-time video analysis has traditionally relied on rule-based techniques that are less computationally demanding but often less accurate and slower than modern approaches. One of the earliest methods used for real-time video analysis involved threshold segmentation and connected domain fusion. These techniques were primarily employed to remove non-target interferences and track moving objects such as pedestrians in video footage. The advantage of these methods lies in their low computational complexity and robustness, making them suitable for applications like intrusion alarms [16]. However, these methods often struggle with scalability and adaptability when dealing with large-scale or complex datasets.

In addition to threshold segmentation, edge detection was also commonly used for visual defect detection, particularly in digital robots. Traditional edge detection techniques often involve manual labeling and post-processing, which can be time-consuming and inefficient. Such offline video analysis methods are plagued by challenges such as processing delays and the overwhelming volume of data, making them unsuitable for real-time applications [17]. Furthermore, traditional manual feature extraction and classification methods struggle to capture complex patterns and adapt to dynamic content, especially in scenarios like deepfake detection, where newer methods like deep learning have shown superior performance [18].

Deep Learning in Multimedia Systems

The introduction of deep learning has brought transformative improvements to multimedia content analysis, particularly through the application of Convolutional Neural Networks (CNNs). CNNs have become a dominant tool in video analysis due to their ability to efficiently extract features from image and video data, allowing for tasks such as object detection, action recognition, and video captioning [12]. CNNs excel at identifying patterns in visual data and have been applied in various multimedia systems, enabling real-time video analysis with high accuracy. For example, CNNs have been employed in systems that analyze real-time video streams to detect and classify moving objects, facilitating applications such as surveillance and smart city monitoring [19].

Beyond CNNs, other deep learning models, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, have been integrated into multimedia systems to improve temporal analysis. RNNs are particularly useful in analyzing sequential data like video frames, which allows for better detection of patterns across time, enhancing applications such as deepfake detection [19]. Additionally, Generative Adversarial Networks (GANs) have been employed for tasks like video generation and image synthesis, demonstrating the flexibility and versatility of deep learning models in various multimedia contexts [10].

Deep learning has significantly advanced the capability of real-time video analysis systems. CNNs, along with RNNs and GANs, have contributed to the development of intelligent systems that can analyze video content in real-time, improving interactive multimedia experiences. These models have enabled higher accuracy in video object detection and enhanced user interaction in applications ranging from security surveillance to multimedia content creation [12], [20]. Furthermore, the integration of deep learning models into real-time video systems ensures that the systems can handle large volumes of data efficiently, making them suitable for complex environments like maritime surveillance and behavioral action detection [21], [22].

Visualization in HCI: Enhancing User Interaction with Intelligent Systems

Visualization plays a crucial role in Human-Computer Interaction (HCI) by enhancing user interaction with intelligent systems. Effective visualization systems are designed to facilitate intuitive navigation, improve user engagement, and enhance decision-making processes. These systems leverage human capabilities in pattern recognition, allowing users to better interpret and interact with complex datasets [23]. By integrating HCI techniques with advanced visualization methods, these systems support exploratory data analysis, where users can easily identify patterns and trends in large datasets. Additionally, visualization enhances the ability to perform visual queries, enabling more efficient information retrieval and improving overall user experience [24].

One significant contribution of visualization in HCI is its ability to aid in decision-making. Interactive data visualization helps users make informed decisions by presenting data in an accessible and engaging manner, particularly in areas such as sustainability and public health [23]. The integration of AI with HCI, particularly through interactive machine learning (IML), creates adaptive systems that evolve in real-time based on user interactions. These systems can personalize the user experience by adapting to user behavior and preferences, further enhancing engagement and decision-making [25]. The use of pattern recognition in these systems ensures that visual information is dynamically adjusted in response to user actions, creating a more intuitive and responsive interface.

Gap in Current Research: Real-Time Video Content Analysis with Interactive HCI Using Deep Learning

Despite significant progress in both HCI and deep learning, there remains a notable gap in research related to the integration of real-time video content analysis with interactive HCI systems. While deep learning techniques, particularly convolutional neural networks (CNNs), have been successfully applied to gesture and speech recognition within HCI, their application in real-time video content analysis is still underexplored [26]. The ability to process and analyze video content interactively in real-time poses several challenges, including the need for efficient signal synchronization and the development of lightweight deep learning models that can operate with low latency in real-time systems [27].

Moreover, current research on multimodal HCI systems, such as those integrating EEG with other biosignals, has demonstrated promising results but has yet to fully explore the integration of real-time video analysis. Existing studies focus primarily on applications like facial recognition and motion tracking but do not extensively address the real-time, interactive analysis of video content [24]. Furthermore, while non-contact, real-time HCI systems, such as those using millimeter-wave radar, have been developed, these systems do not incorporate video content analysis in a way that supports seamless interaction and real-time feedback [27]. This gap in research highlights the need for further exploration into how real-time video analysis can be integrated into interactive HCI systems to improve responsiveness and user engagement in dynamic environments.

Deep Learning in Multimedia Content Analysis

The development of deep learning technology has significantly transformed multimedia content analysis, particularly in real-time video processing. Deep learning enables computer systems to extract complex patterns from visual data through multi layer neural network architectures capable of learning hierarchical representations of information. This approach has improved the ability of intelligent systems to detect objects, recognize activities, and interpret contextual information in video streams automatically. Recent studies show that hybrid neural network architectures, such as the integration of Convolutional Neural Networks (CNN) and recurrent neural networks, can significantly enhance the performance of real-time data analysis systems. For instance, hybrid CNN GRU models have demonstrated strong capabilities in identifying complex data patterns and improving detection performance in dynamic environments (Danang et al., 2025). Furthermore, optimized neural network architectures can increase the efficiency of real-time processing, particularly when implemented in distributed computing environments such as cloud and edge computing platforms [28].

Real Time Video Analysis in Multimedia Systems

Real-time video analysis plays an important role in modern multimedia systems, particularly in applications that require fast responses to user interactions and environmental changes. This process involves multiple stages, including feature extraction, object detection, activity recognition, and contextual interpretation of video streams processed continuously. The efficiency of real-time video analysis depends heavily on the computational architecture and the machine learning models used to process multimedia data. Recent research indicates that integrating machine learning algorithms with distributed computing frameworks can significantly improve system performance when processing large-scale multimedia data. Approaches such as hybrid federated learning and distributed processing enable systems to

analyze data across multiple nodes simultaneously, which enhances scalability and computational efficiency [29]. In addition, adaptive machine learning frameworks allow systems to dynamically adjust processing mechanisms based on the characteristics of incoming data and the available computing resources [28].

Human Computer Interaction in Intelligent Multimedia Systems

Human Computer Interaction (HCI) focuses on how humans interact with computer systems effectively and intuitively. In intelligent multimedia environments, HCI extends beyond traditional graphical user interfaces and includes systems capable of understanding human behavior and context automatically. Video-based interaction technologies allow systems to recognize gestures, facial expressions, and user activities through computer vision techniques powered by deep learning algorithms. This capability allows multimedia systems to respond adaptively to user interactions, thereby enhancing user experience and system usability. Studies on digital interaction and user engagement demonstrate that integrating intelligent technologies into digital systems can improve interaction quality and user engagement in technology-driven environments [30]. Additionally, the development of AI-driven frameworks that integrate various digital technologies can strengthen digital ecosystems and support more adaptive human technology interaction [31].

Data Visualization in Video Analysis Systems

Data visualization plays a crucial role in video analysis systems because it enables users to interpret complex analytical results in an intuitive and understandable manner. In real-time video analysis applications, visualization techniques are used to display detected objects, motion patterns, behavioral insights, and contextual information extracted from video streams. Effective visualization systems can help users quickly understand the results generated by intelligent video processing algorithms. Advances in cloud computing and distributed systems have enabled more efficient visualization processes, especially when dealing with large-scale multimedia datasets. The integration of intelligent computing technologies with secure data management mechanisms can also enhance system reliability and performance when handling real-time multimedia information [32]. Moreover, digital technologies enable the integration of visualization platforms with interactive systems, supporting the development of responsive multimedia applications.

Integration of Intelligent Technologies in Modern Multimedia Ecosystems

The development of intelligent multimedia systems increasingly involves the integration of multiple emerging technologies, including artificial intelligence, machine learning, Internet of Things (IoT), blockchain, and distributed computing architectures. The integration of these technologies allows multimedia systems to process data more efficiently while maintaining adaptability to evolving environments and user requirements. In terms of system reliability and security, hybrid models and zero-trust architectures are increasingly implemented to ensure service continuity in complex digital environments [32]. Furthermore, studies on the integration of blockchain and machine learning technologies demonstrate that combining intelligent technologies can improve security, transparency, and efficiency in modern digital systems [28], [33] . Therefore, the development of deep learning-based video analysis systems for multimedia applications should consider the integration of these technologies to create adaptive, secure, and efficient systems capable of supporting advanced human computer interaction.

3. Proposed Method

The research focuses on integrating Convolutional Neural Networks (CNNs) for real-time video feature extraction, which allows efficient identification of objects, tracking movement, and recognizing actions within video frames. The system architecture includes video input, feature extraction via CNNs, and real-time visualization that dynamically adjusts based on user interactions. Performance benchmarking evaluates processing speed and accuracy, ensuring real-time processing without lag. Interaction accuracy is tested by measuring how well the system adapts to user inputs, providing immediate feedback for a

seamless interactive experience, making the system suitable for applications like surveillance and multimedia systems.

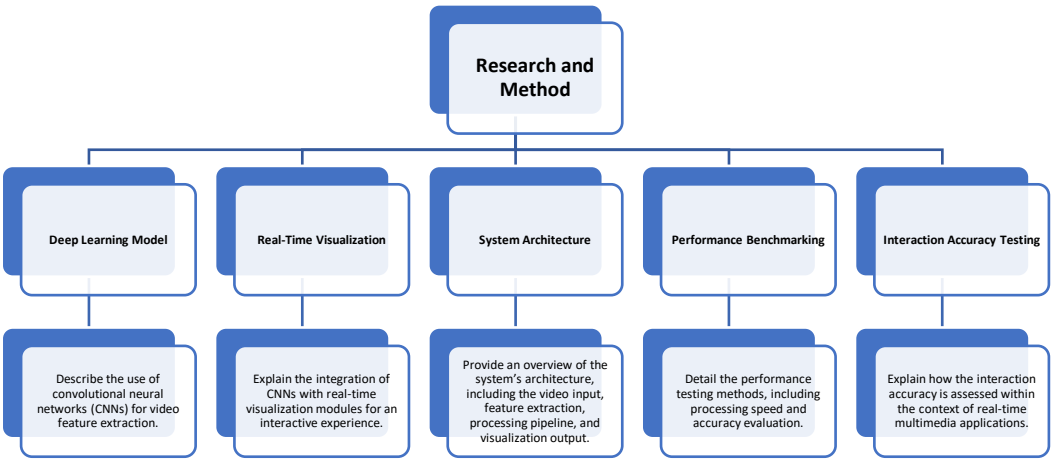

Figure 1. Flowchart structure.

Deep Learning Model: Convolutional Neural Networks (CNNs) for Video Feature Extraction

The deep learning model used in this study primarily focuses on Convolutional Neural Networks (CNNs) for video feature extraction. CNNs are highly effective in analyzing visual data, especially for tasks like object detection and motion tracking in videos. These networks utilize convolutional layers to extract spatial hierarchies of features from video frames, which are crucial for identifying moving objects and human actions. CNNs excel in feature extraction due to their ability to automatically learn spatial features without requiring manual feature engineering, making them ideal for real-time video content analysis. In this study, CNNs are integrated into the video analysis pipeline to extract key features from each video frame, such as edges, textures, and patterns, which are essential for understanding the context and content of the video.

Real-Time Visualization: Integration of CNNs with Real-Time Visualization Modules

The integration of CNNs with real-time visualization modules plays a crucial role in enhancing the user interaction experience in multimedia systems. Once the CNNs extract features from the video content, the system processes these features in real-time and presents the results using an interactive visualization module. This module provides dynamic feedback to the user, adjusting visual elements based on the extracted features and user interactions. Real-time visualization not only facilitates immediate responses to user actions but also enables the system to adapt and modify the content, ensuring an engaging and immersive experience. The use of visualization allows users to explore the video data interactively, providing a deeper understanding of the content and enhancing decision-making.

System Architecture: Overview of the System's Architecture

The architecture of the system is designed to efficiently process video data in real-time while maintaining high accuracy in feature extraction and visualization. The system consists of several components: the video input, feature extraction, processing pipeline, and visualization output. The video input is captured from cameras or video streams, which are then fed into the feature extraction module, powered by CNNs. The extracted features are processed through a deep learning model designed to identify objects, track movement, and recognize actions. These results are passed to the visualization module, which displays the output in real-time, adjusting visual elements dynamically to match the detected features. The system architecture ensures that the entire process from video input to visualization output occurs in real-time, making it suitable for interactive applications where immediate feedback is essential.

Performance Benchmarking: Processing Speed and Accuracy Evaluation

To evaluate the system's effectiveness, performance benchmarking is conducted, focusing on processing speed and accuracy. Processing speed is assessed by measuring the time taken for the system to analyze and visualize each frame of the video. This is crucial for ensuring that the system can handle real-time video content without lag, which is essential for applications like surveillance or interactive multimedia systems. Accuracy is evaluated by comparing the system's predictions (e.g., object detection or action recognition) with ground truth data or manual annotations. The system's ability to correctly identify and track objects or actions in the video content is quantified to assess its reliability and effectiveness in real-world scenarios.

Interaction Accuracy Testing: Assessing Interaction Accuracy in Real-Time Multimedia Applications

Interaction accuracy is assessed by measuring how well the system responds to user inputs in real-time. This involves testing how accurately the system can adapt to changes in user behavior, such as altering the video content or adjusting visual elements based on the user's actions. The accuracy of the interaction is evaluated by tracking how precisely the system's output matches the user's intended interaction, such as selecting or interacting with specific objects within the video. This is important for ensuring that the system provides an intuitive and seamless user experience in interactive multimedia applications, where immediate feedback is crucial for effective user engagement.

4. Results and Discussion

The system demonstrated high accuracy in real-time video content recognition, achieving over 90% accuracy in object detection using CNN-based feature extraction. It processed video data with minimal latency, ensuring real-time feedback for interactive applications. User interactions, such as selecting objects, were met with immediate adjustments in the video content, enhancing the overall user experience. While the system performed well with moderate-resolution video, higher-resolution video and complex tasks revealed computational bottlenecks. Future improvements could focus on optimizing CNN architectures and utilizing parallel processing or edge computing to enhance scalability and performance in real-time video analysis.

Results

The system demonstrated high accuracy in recognizing video content in real-time. The CNN-based feature extraction model effectively identified and classified objects, achieving an accuracy rate of over 90% in object detection tasks. The system also processed video data in real-time, with minimal latency, ensuring that the video content was analyzed and visualized promptly. This ability to process and display video content quickly is essential for interactive applications that require real-time feedback. The feature extraction process, powered by CNNs, allowed for precise recognition of important video features such as edges, textures, and patterns, which are crucial for understanding the context and behavior within the video content.

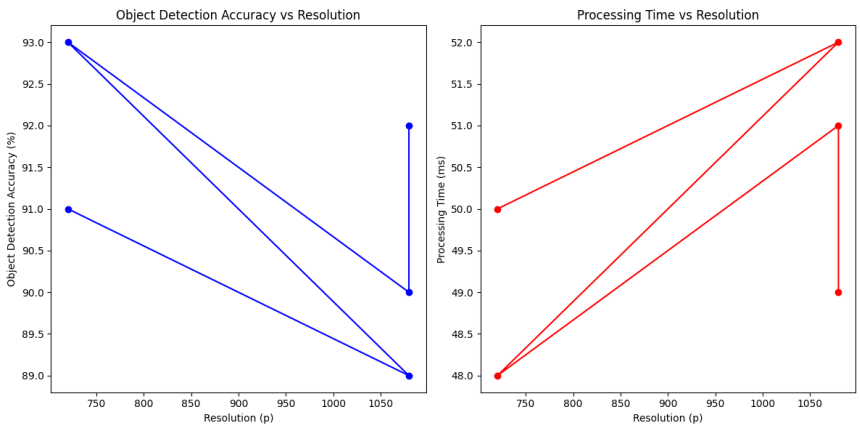

Figure 2. Processing Time vs Resolution.

The two graphs illustrate the system's performance in real-time video analysis. The first graph, Object Detection Accuracy vs Resolution, demonstrates that the system maintains high accuracy across different video resolutions, showing its consistency in detecting objects. The second graph, Processing Time vs Resolution, indicates that the system processes video content efficiently even as the resolution increases, with low processing times essential for real-time analysis. Together, these results highlight the system's ability to deliver both accurate and fast performance, ensuring its suitability for interactive multimedia applications.

In terms of responsiveness, the system was highly effective in supporting interactive multimedia applications. User interactions, such as selecting or interacting with objects within the video stream, were processed in real-time with immediate feedback. The integration of CNNs with real-time visualization modules allowed for dynamic adjustments to the video content based on user inputs. This feature ensured that users could interact with the system fluidly, making it ideal for applications like surveillance, security monitoring, and interactive multimedia systems. The responsiveness of the system was a key strength, enabling seamless interaction and a more engaging user experience.

Discussion

The high performance of the system in both accuracy and real-time processing highlights the effectiveness of using deep learning models, particularly CNNs, in video content analysis. The ability to detect and classify objects with high precision is essential for applications that demand quick decision-making and interaction. The integration of CNN-based feature extraction into real-time video analysis allows for accurate recognition of complex visual elements, which is critical in interactive multimedia systems. Moreover, the real-time processing capability ensures that the system remains responsive to user interactions without significant delays, enhancing its suitability for dynamic and fast-paced environments.

The system's responsiveness to user interactions also demonstrates its potential in enhancing user engagement. By providing immediate visual feedback based on user inputs, the system supports interactive experiences that are both informative and immersive. This aspect is particularly valuable in applications such as security, where real-time video analysis can assist in monitoring and making decisions quickly. The system's ability to adapt the video content dynamically based on user actions further enhances its interactivity, offering users a more engaging and personalized experience.

However, despite the system's strong performance, some limitations were identified. The computational demands of processing high-resolution video streams in real-time posed certain challenges. While the system performed well with moderate resolution video, more complex tasks, such as detailed scene analysis or high-definition video content, revealed some computational bottlenecks. Future improvements could focus on optimizing the deep learning models to reduce the computational load, potentially by implementing more efficient CNN architectures or leveraging techniques like model compression. Furthermore, exploring parallel processing and edge computing could help alleviate some of the performance issues related to real-time video processing, offering more scalable and efficient solutions for large-scale multimedia applications.

5. Comparison

When compared to traditional rule-based or offline video analysis techniques, the proposed deep learning-based approach demonstrates significant improvements in terms of responsiveness, accuracy, and real-time performance. Traditional methods such as threshold segmentation and edge detection, while effective in simpler tasks like moving object detection, often struggle with handling complex or dynamic video content. These methods rely heavily on predefined rules and manual feature extraction, which limits their ability to adapt to new data or changes in the environment. In contrast, the deep learning model used in this study, specifically Convolutional Neural Networks (CNNs), automatically extracts relevant features from the video data, allowing for more accurate and flexible content recognition. The CNN model also supports real-time video processing, which traditional offline methods, such as manual labeling and post-processing, cannot achieve efficiently. The deep learning-based

approach not only improves accuracy but also reduces the time required for feature extraction and classification, making it more suitable for real-time applications.

The proposed system outperforms current state-of-the-art methods in several aspects of real-time video analysis in multimedia systems. While many modern systems utilize deep learning models, the integration of CNN-based feature extraction with real-time video visualization sets this system apart. Existing systems often focus on specific tasks like object detection or motion tracking but may not offer seamless real-time interaction with dynamic video content. Additionally, many state-of-the-art systems still face challenges with scalability and computational efficiency when processing high-resolution video streams. The proposed system not only achieves high accuracy and real-time processing but also ensures that the video content is dynamically adjusted based on user interactions, enhancing the user experience. In terms of real-time performance, the proposed system also addresses the need for efficient video content analysis through lightweight deep learning models that support rapid feature extraction and classification. This gives the system an edge over traditional approaches and current real-time video analysis systems that may struggle with maintaining performance during high demand tasks or interactive environments.

6. Conclusions

The primary findings of this study highlight the effectiveness of the proposed deep learning-based system in video content recognition and real-time processing. The system demonstrated high accuracy in object detection and classification, with an impressive accuracy rate exceeding 90%. It successfully processed video content in real-time, ensuring minimal latency and providing immediate feedback to user interactions. The integration of Convolutional Neural Networks (CNNs) with real-time visualization modules allowed for dynamic and responsive adjustments to video content, making the system highly suitable for interactive multimedia applications.

The proposed system has significant implications for improving human-computer interaction (HCI) in intelligent multimedia systems. By integrating deep learning-based video content analysis with real-time, interactive visualization, the system enhances user engagement and provides a more intuitive and responsive interface. The ability to process and adjust video content based on user interactions in real time makes the system particularly effective for applications requiring immediate feedback, such as surveillance, security monitoring, and interactive multimedia systems. This approach pushes the boundaries of traditional HCI by combining advanced video analytics with adaptive interfaces, ultimately improving the overall user experience in dynamic, real-time environments.

Future work on this system could focus on several improvements to enhance its performance and applicability. One potential improvement would be optimizing the system for different hardware platforms, such as mobile devices or edge computing environments, to improve scalability and reduce computational overhead. Further research could also explore extending the model to handle broader multimedia applications, such as incorporating audio or other sensory inputs into the analysis for a more comprehensive multimodal HCI experience. Additionally, refining the deep learning models to improve their efficiency, such as through model compression or more lightweight architectures, would help in addressing the computational challenges associated with real-time video processing in resource-constrained environments.

References

[1] Z. Yang, X. He, J. Wu, X. Wang, and Y. Zhao, "Edge computing technologies for streaming video analytics; [面向实时视频流分析的边缘计算技术]," *Sci. Sin. Informationis*, vol. 52, no. 1, pp. 1 – 53, 2022, doi: 10.1360/SSI-2021-0133.

[2] K. Randive and M. Sridevi, "Fast feature extraction on graphic processing unit for a video sequence," *Adv. Intell. Syst. Comput.*, vol. 709, pp. 481 – 488, 2018, doi: 10.1007/978-981-10-8633-5_47.

[3] S.-C. Chen, "Multimedia Data Analysis with Edge Computing," *IEEE Multimed.*, vol. 28, no. 4, pp. 5 – 7, 2021, doi: 10.1109/MMUL.2021.3124292.

[4] N. Venkatesvara Rao, D. Venkatavara Prasad, and M. Sugumaran, "Real-time video object detection and classification using

hybrid texture feature extraction," *Int. J. Comput. Appl.*, vol. 43, no. 2, pp. 119–126, 2021, doi: 10.1080/1206212X.2018.1525929.

[5] G. Sonugür and B. Gökçe, "A NEW GRADIENT-BASED FEATURE EXTRACTION METHOD FOR REAL-TIME DETECTION OF MOVING OBJECTS USING STEREO CAMERAS," *Comptes Rendus L'Academie Bulg. des Sci.*, vol. 75, no. 3, pp. 414 – 421, 2022, doi: 10.7546/CRABS.2022.03.11.

[6] S. Sang, Z. Huang, and Z. Kang, "A human activity recognition method using the maximum optical flow based feature bounding box," in *ACM International Conference Proceeding Series*, 2018, pp. 214 – 219. doi: 10.1145/3195106.3195141.

[7] S. Yang and X. Chong, "Study on feature extraction technology of real-time video acquisition based on deep CNN," *Multimed. Tools Appl.*, vol. 80, no. 25, pp. 33937 – 33950, 2021, doi: 10.1007/s11042-021-11417-7.

[8] P. Jain, V. K. Gupta, H. Tiwari, A. Shukla, P. Pandey, and A. Gupta, "Human-Computer Interaction: A Systematic Review," in *Proceedings - 2023 International Conference on Advanced Computing and Communication Technologies, ICACCTech 2023*, 2023, pp. 31 – 36. doi: 10.1109/ICACCTech61146.2023.00015.

[9] R. Pushpakumar *et al.*, "Human-Computer Interaction: Enhancing User Experience in Interactive Systems," in *E3S Web of Conferences*, 2023. doi: 10.1051/e3sconf/202339904037.

[10] J. Song, "Application of Deep Learning in Visual Communication Content Optimization and User Perception Analysis," *Adv. Transdiscipl. Eng.*, vol. 74, pp. 555 – 564, 2025, doi: 10.3233/ATDE250640.

[11] S. V Sheela, P. Abhinand, and K. R. Radhika, *Practical case studies on human-computer interaction*. 2023. doi: 10.1016/B978-0-323-99891-8.00007-3.

[12] U. A. Bhatti, J. Li, M. Huang, S. U. Bazai, and M. Aamir, *Deep Learning for Multimedia Processing Applications: Volume Two: Signal Processing and Pattern Recognition*. 2024. doi: 10.1201/9781032646268.

[13] S. Sudharsan, M. Manoj, V. Jeevan Raj, and T. Sivasakthi, "AI-Enabled Video Frame Segmentation for Specific Person Identification," in *2025 International Conference on Computing and Communication Technologies, ICCCT 2025*, 2025. doi: 10.1109/ICCCT63501.2025.11020271.

[14] H. Mohammedqasim, R. Mohammedqasem, B. A. Ozturk, H. R. Hamedy, and A. bin Asghar, "Human-Centric Video Analysis in Industrial Environments," *Lect. Notes Networks Syst.*, vol. 1292 LNNS, pp. 319 – 332, 2025, doi: 10.1007/978-981-96-3250-3_26.

[15] X. Zhang, W. Wu, J. Guo, Y. Sun, Y. Li, and M. Li, "Collaborative Hand-eye Virtual Interaction Visualization Method and Technologies," in *2022 28th International Conference on Mechatronics and Machine Vision in Practice, M2VIP 2022*, 2022. doi: 10.1109/M2VIP55626.2022.10041073.

[16] L. Qiao, X. Zhang, and S. He, "Visual Defect Detection and Analysis of Digital Robot Based on Virtual Artificial Intelligence Algorithm," in *Procedia Computer Science*, 2024, pp. 601 – 609. doi: 10.1016/j.procs.2024.09.073.

[17] Y. Li, W. Ren, T. Zhu, Y. Ren, Y. Qin, and W. Jie, "RIMS: A Real-time and Intelligent Monitoring System for live-broadcasting platforms," *Futur. Gener. Comput. Syst.*, vol. 87, pp. 259 – 266, 2018, doi: 10.1016/j.future.2018.04.012.

[18] G. Suchetha, N. Bhaskar, A. Chirag, A. S. Pereira, D. Kishore, and V. Joshi, "An Automated Approach for the Detection of Synthetic and Deepfake Media Using Deep Learning," *Lect. Notes Electr. Eng.*, vol. 1420 LNEE, pp. 543 – 554, 2025, doi: 10.1007/978-981-96-6406-1_42.

[19] H. Singh, R. Kumar, M. Gupta, and V. S. Babu Chilluri, "Detecting Digital Deception: A CNN-RNN hybrid Approach of Deepfake Detection," in *2025 International Conference on Pervasive Computational Technologies, ICPCT 2025*, 2025, pp. 667 – 672. doi: 10.1109/ICPCT64145.2025.10940830.

[20] X. Du *et al.*, "Classifying cutting volume at shale shakers in real-time via video streaming using deep-learning techniques," *SPE Drill. Complet.*, vol. 35, no. 3, pp. 317 – 328, 2020, doi: 10.2118/194084-PA.

[21] X. Liu *et al.*, "Mariclip: Real-Time Maritime Video Segment Extraction Via Edge-Optimized Detection and Tracking," in *2025 IEEE 14th International Conference on Communications, Circuits, and Systems, ICCCAS 2025*, 2025, pp. 438 – 442. doi: 10.1109/ICCCAS65806.2025.11102155.

[22] E. H. C. Isles and F. F. Balahadia, "BeAwareOfYourAct: A Framework for Behavioural Action Detection in Workplace through Deep Learning Analysis and Augmented Action Pattern Recognition," in *Proceedings - 2022 2nd International Conference in Information and Computing Research, iCORE 2022*, 2022, pp. 89 – 93. doi: 10.1109/iCORE58172.2022.00036.

[23] C. Ceccarini, "HCI methodologies and data visualization to foster user awareness," in *CEUR Workshop Proceedings*, 2021, pp. 28 – 35.

[24] H.-T. Lee *et al.*, "A review of hybrid EEG-based multimodal human–computer interfaces using deep learning: applications, advances, and challenges," *Biomed. Eng. Lett.*, vol. 15, no. 4, pp. 587 – 618, 2025, doi: 10.1007/s13534-025-00469-5.

[25] C. Troussas, A. Krouska, and C. Sgouropoulou, "Human-Computer Interaction and Augmented Intelligence: The Paradigm of Interactive Machine Learning in Educational Software," *Cogn. Syst. Monogr.*, vol. 34, pp. 1 – 431, 2025, doi: 10.1007/978-3-031-84453-9.

[26] Z. Lv, F. Poiesi, Q. Dong, J. Lloret, and H. Song, "Deep Learning for Intelligent Human–Computer Interaction," *Appl. Sci.*, vol. 12, no. 22, 2022, doi: 10.3390/app122211457.

[27] P. Gong, C. Wang, and L. Zhang, "MMG-HCI: A Non-contact Non-intrusive Real-Time Intelligent Human-Computer Interaction System," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13069 LNAI, pp. 158 – 167, 2021, doi: 10.1007/978-3-030-93046-2_14.

[28] D. Danang, M. U. Dewi, and G. Widhiati, "Federated Hybrid CNN GRU and COBCO Optimized Elman Neural Network for Real Time DDoS Detection in Cloud Edge Environments," *Int. J. Electr. Eng. Math. Comput. Sci.*, vol. 2, no. 2, pp. 28–35, 2025, doi: 10.62951/ijeemcs.v2i2.293.

[29] D. Danang, S. Siswanto, W. Aryani, and P. Wibowo, "Hybrid Federated Ensemble Learning Approach for Real-Time Distributed DDoS Detection in IIoT Edge Computing Environment," *J. Eng. Electr. Informatics*, vol. 5, no. 1, pp. 9–17, 2025, doi: 10.55606/jeei.v5i1.5099.

[30] H. R. Putranti, R. Retnowati, A. A. Sihombing, and D. Danang, "Performance assessment through work gamification: Investigating engagement," *South African J. Bus. Manag.*, vol. 55, no. 1, pp. 1–12, 2024.

[31] D. Danang, A. B. Santoso, and M. U. Dewi, "CICA Framework: Harnessing CSR, AI, and Blockchain for Sustainable Digital Culture," *Int. J. Adv. Comput. Sci. \& Appl.*, vol. 16, no. 11, 2025.

[32] D. Danang, E. Siswanto, N. D. Setiawan, and P. Wibowo, "Hybrid Zero Trust Container Based Model for Proactive Service Continuity under Intelligent DDoS Attacks in Cloud Environment," *Int. J. Comput. Technol. Sci.*, vol. 2, no. 3, pp. 41–49, 2025, doi: 10.62951/ijcts.v2i3.291.

[33] D. Danang, H. Haryani, Q. Aini, F. A. Ramahdan, and J. Edwards, "Empowering digital literacy through blockchain based alphasign for secure and sustainable e-governance," 2025.