Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

Design and Implementation of an AI-Enhanced Video-Based Document Creation and Collaboration Platform

Sejin Jang
Dept. of Computer Engineering,Sun Moon Univ
Chungnam, Asansi, South Korea

Myeonsung Jung
Dept. of Computer Engineering,Sun Moon Univ
Chungnam, Asansi, South Korea

Bohyeon Baek
Dept. of Computer Engineering,Sun Moon Univ
Chungnam, Asansi, South Korea

Minju Park
Dept. of Computer Engineering,Sun Moon Univ
Chungnam, Asansi, South Korea

Bonggeun Song
Dept. of Computer Engineering,Sun Moon Univ
Chungnam, Asansi, South Korea

Seungjae Lee (corresponding author)
Dept. of Computer Engineering,Sun Moon Univ
Chungnam, Asansi, South Korea

***Abstract*— This paper introduces an intelligent web-based platform that transforms video and audio data into interactive documents through automatic summarization, translation, and collaborative editing. By integrating modern APIs such as OpenAI, Clerk, and Liveblocks, the system enables users to generate timestamped notes, bookmarks, and multilingual summaries linked to specific video moments. It supports real-time co-editing, OAuth-based authentication, and dark/light modes for an enhanced user experience. This system provides an efficient and accessible way to analyze, organize, and collaborate on video-based information for educational, corporate, and research applications.**

***Keywords*— Video summarization, AI collaboration, Multilingual translation, Real-time editing, React, Clerk API, Liveblocks, Whisper STT**

I. INTRODUCTION (HEADING 1)

With the exponential growth of video-based content across education, business meetings, and online platforms, users increasingly face challenges in efficiently capturing, analyzing, and sharing valuable information embedded in videos. Traditional note-taking and documentation tools—such as Google Docs or OneNote—are not optimized for handling time-dependent media, resulting in fragmented workflows and decreased productivity. Moreover, existing video platforms like YouTube provide only limited annotation features and no collaborative editing support.

Recent advancements in artificial intelligence (AI) and cloud computing have created opportunities to build intelligent systems capable of understanding, summarizing, and translating multimedia content automatically [1,2]. Tools such as OpenAI's GPT and Whisper engines now enable high-quality speech-to-text transcription, semantic summarization, and natural multilingual translation, significantly enhancing the usability of video materials in global contexts.

In this study, we propose and develop an AI-enhanced video-based document creation and collaboration system. Unlike traditional documentation platforms, this system directly integrates with video and audio sources—such as YouTube URLs or uploaded media files—to automatically transcribe and summarize content. Users can annotate key moments with timestamped notes, create searchable bookmarks, and collaborate in real-time within the same workspace.

The distinctiveness of this system lies in its seamless combination of AI-driven summarization, multilingual translation, and synchronous document editing [3]. By integrating Clerk API for authentication and Liveblocks API for real-time synchronization, the system achieves smooth and secure multi-user collaboration. Built using React and TypeScript, it also provides an intuitive, responsive, and dark-mode-compatible interface optimized for continuous engagement.

This research contributes to the growing field of human–AI interaction by offering a practical, scalable, and user-friendly platform for converting multimedia content into structured, shareable knowledge artifacts [4].

II. METHODOLOGY

The platform adopts a modular web-based architecture using React for the frontend and Firebase as the backend service. Core components include user authentication through Clerk API, real-time synchronization with Liveblocks, and OpenAI-based AI processing.

The workflow begins when users upload or link a video or audio file (e.g., YouTube URL, MP4, or MP3). The system extracts the audio track, which is then transcribed using Whisper, an advanced speech-to-text model. This transcript is summarized by a GPT-based model, producing concise content that captures essential points. The summary is then automatically translated into four languages—Korean, English, Japanese, and Chinese—enabling global accessibility [5,6].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

Timestamped note-taking and bookmarking allow users to record key insights linked to specific moments in the video. All data are synchronized via Firebase, ensuring that notes and summaries remain updated in real time. The Liveblocks API manages collaborative editing, enabling multiple users to simultaneously modify documents with minimal latency. Each user's cursor position and text changes are synchronized through WebSocket communication, preserving editing consistency and preventing data loss.

Security and scalability are ensured by the Clerk authentication layer, which verifies every user session using OAuth 2.0 protocols. All content, including media and summaries, is stored in Firebase Cloud Storage, ensuring both high availability and low access latency [7].

III. EXPERIMENTS AND RESULTS

To validate the proposed system, a series of experimental evaluations were conducted focusing on transcription accuracy, summarization quality, translation fluency, real-time collaboration stability, and overall user satisfaction. The development and testing environment consisted of React 18, TypeScript, Node.js, Firebase, and Vercel Cloud deployment. The system was accessed using modern browsers such as Chrome, Edge, and Firefox.

During testing, multiple video and audio samples—ranging from 3-minute lectures to 40-minute seminars—were uploaded to evaluate system robustness under diverse content lengths. The Whisper STT model demonstrated strong transcription performance, achieving an average word accuracy rate of 97.8% across all test samples. This result aligns with findings from similar AI-driven transcription systems [8]. Summarization using GPT-based models showed consistent contextual understanding, with human evaluators rating the summaries as capturing over 92% of key content relevance compared to the original transcript.

For multilingual translation, the system's performance was benchmarked against professional translation outputs using BLEU scoring. The results indicated an average BLEU score exceeding 85%, confirming that the translations were fluent and semantically faithful. In user surveys, 83% of participants agreed that the translated summaries significantly reduced manual translation workload.

Real-time collaboration was another critical aspect of the evaluation. The system was tested with up to 20 concurrent users editing the same document. The latency for update synchronization averaged below 150 milliseconds, even under moderate network traffic. Cursor tracking, user presence indicators, and simultaneous editing behaved smoothly, with no instances of data conflict or text overwriting. This demonstrated the high reliability of the Liveblocks API in maintaining collaborative coherence [9].

A usability study was conducted among 30 participants, including university students, educators, and IT professionals. They rated their experience using a five-point Likert scale. The overall satisfaction score reached 4.7/5, particularly highlighting the clarity of AI-generated summaries, timestamp-based note features, and the stability of collaborative editing. Users also emphasized that dark mode and multilingual translation made the system accessible and visually comfortable during long sessions.

Collectively, these results confirm that the developed system not only performs efficiently in technical benchmarks but also enhances the overall productivity of users who work with video-based content in academic and professional settings [10,11].

IV. CONCLUSIONS

This research presented an integrated AI-driven system that transforms video and audio materials into structured, multilingual, and collaboratively editable documents. The proposed platform successfully bridges the gap between static note-taking and dynamic video interaction, offering users a new way to manage complex information environments.

Through comprehensive testing, the system demonstrated high performance in automatic transcription, summarization, and translation. The use of the Whisper model provided reliable speech recognition across various accents and audio qualities, while the GPT-based summarization produced coherent and contextually relevant summaries. The multilingual translation feature further expanded accessibility, allowing content to be shared seamlessly across linguistic boundaries.

From a collaboration perspective, the integration of Liveblocks and Clerk APIs enabled real-time editing among multiple users without latency issues or data conflicts. The synchronization accuracy and low network delay proved that this architecture can be scaled for team-based projects, educational institutions, and corporate meeting environments. Additionally, user feedback indicated a clear improvement in workflow efficiency, as the system allowed simultaneous content review, summarization, and discussion within a single platform.

Future work will focus on several directions. First, the AI summarization module will be enhanced through domain-specific fine-tuning, enabling more accurate comprehension of specialized topics such as medical or engineering content. Second, visual scene understanding using computer vision models will be incorporated to identify key video frames automatically, thereby improving contextual summarization [12]. Third, the system will be optimized for mobile devices and touch interfaces, extending accessibility beyond desktop environments. Lastly, advanced data analytics and access control will be introduced to support enterprise-grade document management and data security compliance.

In conclusion, the proposed platform provides a robust foundation for next-generation video-based collaboration systems, integrating AI, cloud computing, and human-centered design principles. By automating tedious transcription and translation tasks while enabling collaborative creativity, this system lays the groundwork for a new era of intelligent multimedia knowledge management.

ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science ICT), Korea, under the National Program for Excellence in SW, supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation) in 2025"(No. 2024-0-00023).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

REFERENCES

[1] Kim, T. S., et al. “Papeos: Augmenting Research Papers with Talk Videos,” CHI ’23.
[2] Tang, C., et al. “CWcollab: A Context-Aware Web-Based Collaborative Multimedia System,” arXiv, 2022.
[3] Chen, Y.-T., et al. “iVRNote: Interactive Note-Taking Interface in VR Environments,” arXiv, 2019.
[4] Schroeter, R., Hunter, J., & Kosovic, D. “Vannotea – A Collaborative Video Indexing and Discussion System,” K-CAP Workshop, 2003.
[5] Sun, C., et al. “Real-Time Cooperative Editing Systems,” ACM TOCHI, 1998.
[6] Novikov, B., & Proskurnin, O. “Towards Collaborative Video Authoring,” LNCS, 2003.
[7] Goldman, D. B., et al. “Designing for Collaborative Video Editing,” SIGGRAPH/ICCV, 2006.
[8] Radford, A. et al. “Whisper: Robust Speech Recognition via Large-Scale Weak Supervision,” OpenAI Technical Report, 2022.
[9] Zhang, X. et al. “Real-Time Document Collaboration Architecture,” Applied Sciences, 2022.
[10] Bødker, S., & Klokmose, C. N. “Collaborative Video Editing,” Human–Computer Interaction, 2011.
[11] Liu, Y., et al. “Enhancing Video Collaboration through AI-Based Summarization,” IEEE Access, 2023.
[12] Park, M. & Jang, S. “Vision-Driven Event Detection in Educational Videos,” Sensors, 2024.