Using Video Recognition AI x LLM to Automate the Creation of Reports

LIU Jianquan, YAMAZAKI Satoshi, MIYANO Hiroyoshi

Abstract

The rapid progress of large language models (LLMs) has generated great excitement for their use across various sectors such as transportation, finance, logistics, manufacturing, construction, retail, and healthcare. These sectors often handle complex types of data, including text, speech, images, and videos, creating a strong demand for LLMs capable of processing such varied inputs effectively. Currently, there is significant advancement in LLMs for still images, with a growing focus on applying this technology to videos, which is rich in valuable information. In this paper, we explore NEC's latest advancements in using LLM for processing long videos and examine how this technology is used in industries to streamline tasks like creating reports. We also outline plans for future developments in this exciting domain.

Keywords

LLM, image recognition, video summarization, shortened video, explanatory text, report

1. Introduction

In November 2022, the launch of ChatGPT by OpenAI in the United States marked a pivotal moment for generative AI—a field of AI focused on creating new and unique content—leading many to claim we have entered the "Fourth AI Boom."[1] ChatGPT is a type of generative AI service that generates a variety of content, including text responses to a wide range of prompts[2]. This capability is made possible thanks to large language models (LLMs).

LLMs are sophisticated AI models trained on vast amounts of text data. They power general-purpose AI capabilities in natural language processing tasks such as translation and text summarization. These advanced models contain anywhere from millions to billions of parameters, allowing them to exhibit rich, human-like language intelligence. LLMs gained prominence around 2018 with the introduction of models like BERT[3] and GPT[4], showcasing their ability to excel in tasks ranging from translation, text classification, sentence summarization, and contextual understanding. The advent of ChatGPT has further accelerated LLM technology development, with potential applications across various industries including transportation, finance, logistics, manufacturing, construction, retail, and healthcare. These industries handle not just textual data, but also speech, images, and videos. Therefore, there is growing demand for developing models capable of processing such multifaceted data, as well as leveraging the capabilities of LLMs.

In today's digital era, an enormous volume of video content is created every day. While there is demand for analyzing these videos to understand on-site situations and generate explanatory texts or accident reports to improve efficiency, most recordings remain underutilized. As such, methods leveraging LLMs to identify specific video scenes and explain them are increasingly needed.

While LLMs have advanced significantly in image processing, creating AI models that can thoroughly comprehend the wealth of information contained in videos is still a significant challenge. In response to this, NEC is leading the way in developing methods that employ LLMs specifically for video analysis. This initiative makes use of the company's extensive experience in AI-driven

video recognition technology, a key area of expertise that NEC has developed over the years.

This paper presents NEC's latest advancements in utilizing LLM for analyzing long videos, as detailed in section 2. It also explores the practical applications of this technology in industries, including how it can automate the creation of reports, a topic covered in section 3. Section 4 discusses the potential future developments and directions of this technology.

2. Research & Development on the Technology of Video Recognition AI x LLM

This section discusses the latest technology[5)6)] related to Video Recognition AI x LLM technology developed by NEC. This technology, which we call "descriptive video summarization," combines video recognition AI with LLM to extract scenes from long videos that meet the user's criteria and then generates a summary that explains these scenes. This technology, developed with the concept of summarizing videos based on a user's narratives,[7)] enables users to efficiently access desired information within a short period, without needing to watch an entire long video.

In developing this technology, we focused on implementing a system that is "user-centric" and incorporates a "narrative" element.

Realizing a "user-centric" approach involves users providing explicit questions or instructions. Through leveraging LLM, the system skillfully interprets the users' intent and extracts only the scenes from the video that they want to see. For example, consider drive recorder videos. An insurance claims adjuster would find scenes important for determining accident causes and assessing damages. Meanwhile, traffic police may specifically want clips showing traffic violations and safe driving practices for enforcement purposes. Applying conventional video analysis has limitations in extracting desired meaning or value from video footage. Different users can seek vastly different insights from the same video. Thus, incorporating the user perspective is essential.

Furthermore, when it comes to generating summaries with a "narrative" element, making complex matters easily understandable requires more than just listing linguistic expressions and information. It is necessary to clearly define the cause-and-effect relationships involved. To achieve this, we go beyond merely relying on LLM. Instead, we apply a variety of recognition engines to the video to identify real-world elements like people, objects, actions, environments, and events. Based on this information, summaries are generated that clearly illustrate what is happening in the video, in a storytelling-like manner, making it easy to understand.

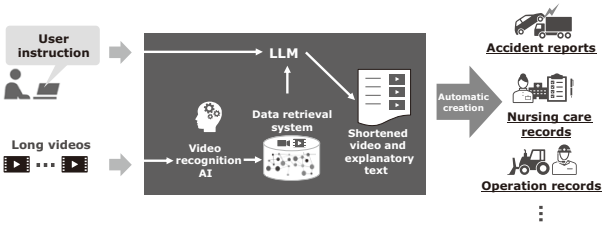


Fig. 1 Process flow within the technical implementation framework.

In the next section, we will introduce the basic framework that enabled this technology, highlight the features of the technology, and discuss the initial experiment results.

2.1 Basic Framework

The process flow of the technology developed by NEC is illustrated in **Fig. 1**. The basic framework for technical implementation consists of video recognition AI, data retrieval system, LLM, and module for generating shortened videos and explanatory text.

First, as a preprocessing step, multiple long videos are fed into several video recognition AI engines. These engines then individually detect various objects and environments within the video footage, such as people, vehicles, and buildings, as well as any changes they undergo. The recognition results are then combined serially to create a unique graph structure that compactly represents video scenes. This graph structure is stored in a video database within the data retrieval system.

The next step is processing online queries. The user inputs the desired information or search criteria for the preprocessed long videos into the system as a natural language text query (prompt text). Inside the system, the LLM semantically comprehends the user prompt text, breaks it down into multiple search conditions, and accurately captures the user's intentions and requirements.

The system then performs high-speed, accurate matching between the user's text query expressing their needs and interests, and the video scene graph structure in the data retrieval system. Only specific partial graphs consistent with the user's intent are extracted. Scenes connected to those partial graphs are pulled from the full videos to create condensed summary videos. Concurrently, using the recognition results for objects like people, cars, and buildings generated earlier by the video AI engines, text summaries explaining the story of these shortened videos are automatically created by the LLM.

Finally, by appropriately embedding relevant images, videos and texts to match established report and record formats used in various industries, the system can automatically generate documentation such as accident reports, nursing care records, and construction work logs.

2.2 Features of the Technology

Next, we will outline the three main features of the newly developed technology.

Feature 1: Find scenes efficiently and create reports faster

The combination of video recognition AI and LLM makes it possible to understand each scene in a video. Specifically, more than 100 video recognition AI engines are applied to recognize the various objects and environments that make up a scene, such as people, cars, buildings, animals, trees and other natural objects, and the weather, as well as their changes, individually. By using LLM to analyze only the recognition results, users can find the scene they are looking for more efficiently than when analyzing an entire video, eliminating the need to repeatedly check a video.

Feature 2: Accurate interpretation of video context to generate expert-quality reports

To improve the quality of the generated text, the LLM is pre-finetuned using sample videos from a specific domain. For example, when applied to drive recorder videos, road traffic-related videos are analyzed in advance. This gives the LLM the expertise to correctly understand what happened in the video. As a result, it is possible to create highly reliable reports while addressing hallucination, which has been an issue with the accuracy of generative AI.

Feature 3: Generate reports in seconds without large-scale computing resources

This technology can create a video of a desired scene and explanatory text in a few seconds from a video that is over an hour long. To achieve this, NEC integrated a compact, high-performance LLM and a high-speed data retrieval system developed by NEC.

Based on the basic framework described earlier, we developed a system that can operate in either cloud or on-premises environments and is accessible to users via a web browser. As an example, **Fig. 2** shows a demonstration of analyzing drive recorder videos. In the demo, the screen displays the moment a large truck runs a red light and collides with a black passenger car at an intersection. Simultaneously, the technology automatically generates text analyzing and explaining the cause of the collision. Experiments were conducted to verify the effectiveness of this technology using this system.

2.3 Experiment results

NEC verified this technology in a use case of creating accident investigation reports from drive recorder videos. By automating the search for accidents, causative scenes, and report drafting—which had previously required manual effort—the time required to create reports was cut in half.

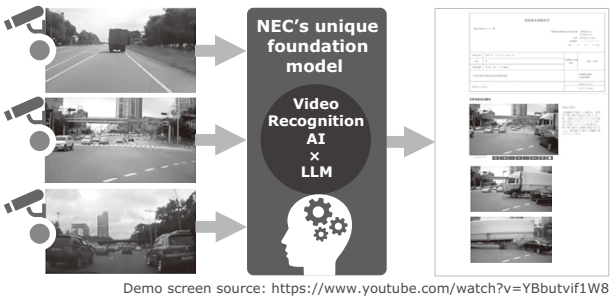
3. Industrial Applications of Video Recognition AI x LLM Technology

In this section we discuss the industries and use cases where the Video Recognition AI x LLM technology can be applied.

Firstly, as shown in **Fig. 3**, this technology can be utilized in the industries of transportation and finance. By analyzing drive recorder videos with this technology, it is possible to automatically generate text and shortened videos explaining the circumstances of an accident and how it occurred. Based on the text and video, an accident investigation report can be automatically created in

Demo screen source: https://www.youtube.com/watch?v=YBbutvif1W8

Fig. 2 Demo screen of the technology.



Demo screen source: https://www.youtube.com/watch?v=YBbutvif1W8

Fig. 3 Application example: Automatic creation of traffic accident investigation reports.

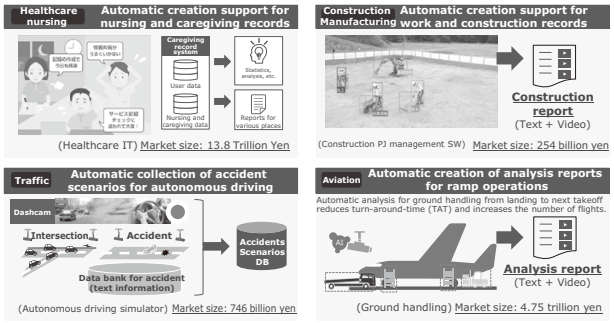


Fig. 4 Application scenarios in healthcare, caregiving, manufacturing, and other sectors.

a format that is appropriate for non-life insurance claims and traffic safety instructions. As demonstrated by the experiment results, the time required to create report drafts, which was previously done manually, can be cut in half.

Beyond the realms of transportation and finance, video has been increasingly utilized for the purpose of safety management and operational efficiency in a variety of other industries, including healthcare, caregiving, manufacturing, construction, aviation, and retail. However, it takes an enormous number of hours to manually check long videos and create reports on near-misses and areas for improvement. By utilizing the Video Recognition AI x LLM technology, it is possible to automatically generate explanatory texts and reports from videos containing complex scenes that consist of various objects and environments and that change over time.

The application of this technology to these industrial sectors is illustrated in **Fig. 4**. For example, by applying this technology to camera images of a factory's production line, you can streamline work checks at key points such as finished product inspections. You will no longer need to check the 24 hours worth of video images for the day, but merely need to read through the report generated by this technology. Other possible applications include journaling by nurses and caregivers, shift check at shops, and aircraft ground handling at airports. This technology can be widely applied to improve the efficiency of video checking. It can also be applied to the B2C area. An example would be efficiently creating a digest video that follows a specific player in a sport game video.

4. Conclusion

In this paper, we have outlined our research and development initiatives related to Video Recognition AI x LLM technology, focusing on both the technical aspects and practical applications. As digitalization becomes more widespread globally, Video Recognition AI x LLM technology will emerge as a key player in the industrial application of generative AI. This technology is indispensable for tasks such as analyzing recorded video, understanding real-time situations on the ground, generating explanatory texts, and creating detailed accident reports, all of which contribute to greater efficiency in operations.

Moving forward, we aim to continually improve this technology to meet the specific performance and cost requirements demanded by various industrial sectors. Our goal is to develop a robust technology that can be reliably implemented in diverse settings. By leveraging Video Recognition AI x LLM, we are dedicated to pursuing research and development efforts that contribute to creating greater efficiency in society.

* ChatGPT is a trademark of OpenAI in the United States.
* All other company names and product names that appear in this paper are trademarks or registered trademarks of their respective companies.

References

1) Hitoshi Matsubara et al.: The Stirrings of the 4th AI Boom - How ChatGPT Will Change the World, Nikkei Business, June 2023 (Japanese)
2) Sho Shimazu: A Hundred Schools of Thought Contend, The Emergence of 'Generative AI', Nikkei Business, March 2023 (Japanese)
3) Jacob Devlin et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018 https://arxiv.org/abs/1810.04805
4) OpenAI: Improving language understanding with unsupervised learning, June 2018 https://openai.com/research/language-unsupervised
5) NEC Press Release: NEC uses generative AI (LLM) and video recognition AI to automatically generate explanatory text from video, December 2023 https://www.nec.com/en/press/202312/global_20231205_01.html
6) NEC R&D: Summarizing long videos into shorter videos with text according to user instructions: Video Recognition AI × LLM, December 2023 https://www.nec.com/en/global/rd/technologies/202314/index.html
7) Yongkang Wong et al: Compute to Tell the Tale: Goal-Driven Narrative Generation, MM '22: The 30th ACM International Conference on Multimedia, pp. 6875-6682, October 2022 https://doi.org/10.1145/3503161.3549202

Authors' Profiles

LIU Jianquan
Director & Senior Principal Researcher
Visual Intelligence Research Laboratories

YAMAZAKI Satoshi
Senior Researcher
Visual Intelligence Research Laboratories

MIYANO Hiroyoshi
Senior Director
Visual Intelligence Research Laboratories

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese

English

Vol.17 No.2 Special Issue on Revolutionizing Business Practices with Generative AI
– Advancing the Societal Adoption of AI with the Support of Generative AI Technologies

Remarks for Special Issue on Revolutionizing Business Practices with Generative AI
Approaches to Generative AI Technology: From Foundational Technologies to Application Development and Guideline Creation

Vol.17 No.2
June 2024

Special Issue TOP

Papers for Special Issue

Market Application of Rapidly Spreading Generative AI
NEC Innovation Day 2023: NEC's Generative AI Initiatives
Streamlining Doctors' Work by Assisting with Medical Recording and Documentation
Using Video Recognition AI x LLM to Automate the Creation of Reports
Understanding of Behaviors in Real World through Video Analysis and Generative AI
Automated Generation of Cyber Threat Intelligence
NEC Generative AI Service (NGS) Promoting Internal Use of Generative AI
Utilization of Generative AI for Software and System Development
LLMs and MI Bring Innovation to Material Development Platforms
Disaster Damage Assessment Using LLMs and Image Analysis

Fundamental Technologies that Enhance the Potential of Generative AI
NEC's LLM with Superior Japanese Language Proficiency
NEC's AI Supercomputer: One of the Largest in Japan to Support Generative AI
Towards Safer Large Language Models (LLMs)
Federated Learning Technology that Enables Collaboration While Keeping Data Confidential and its Applicability to LLMs
Large Language Models (LLMs) Enable Few-Shot Clustering
Knowledge-enhanced Prompt Learning for Open-domain Commonsense Reasoning
Foundational Vision-LLM for AI Linkage and Orchestration
Optimizing LLM API usage costs with novel query-aware reduction of relevant enterprise data

For AI Technology to Penetrate Society
Movements in AI Standardization and Rule Making and NEC Initiatives
NEC's Initiatives on AI Governance toward Respecting Human Rights
Case Study of Human Resources Development for AI Risk Management Using RCModel

NEC Information

2023 C&C Prize Ceremony