ATINER CONFERENCE PRESENTATION SERIES No: DAT2024-0341

ATINER's Conference Paper Proceedings Series
DAT2024-0341
Athens, 29 October 2024

Implementing Image Analysis with Azure AI Vision and Open AI for Predictive Analysis

Mihail Mateev

Athens Institute for Education and Research
9 Chalkokondili Street, 10677 Athens, Greece

ATINER's Conference Paper Proceedings Series
DAT2024-0341
Athens, 29 October 2024
ISSN: 2529-167X

Mihail Mateev, Assistant Professor, Faculty of Structural Engineering, CAE Department, UACEG, Bulgaria

Implementing Image Analysis with Azure AI Vision and Open AI for Predictive Analysis

One of the most used predictive analytics applications involves extracting necessary metadata from images and videos to evaluate the condition of real-world systems and recommend measures to sustain these systems. Microsoft's Azure AI Vision service offers access to sophisticated algorithms that analyze images and generate insights based on the visual aspects of interest to the user. Additionally, Azure OpenAI Service has introduced an image analysis feature that leverages large language models (LLMs) to comprehend the content of images. GPT-4 Turbo with Vision, developed by OpenAI, is a significant multimodal model (LMM) capable of interpreting images and providing text-based answers to queries regarding those images, combining capabilities in both natural language processing and visual comprehension. This research proposes an efficient approach to implementing image analysis and automated metadata generation for images. The article compares Azure AI Vision and GPT-4 Turbo with Vision, exploring how these technologies can be utilized together for enhanced predictive analysis.

Keywords: *ChatGPT, AI, Generative AI, automation, Open AI, Azure AI vision, Microsoft Azure, cloud computing, predictive analysis, IoT, power platform*

Acknowledgments: *Special thanks to my family for the support and Microsoft for the Azure resources, which made this research possible.*

Introduction

ChatGPT (Chat Generative Pre-Trained Transformer) has been one of the game changers in AI and the whole AI industry during the last two years.

Different business domains in the modern industry can benefit from Generative AI and Ghat GPT, not just for creating human-like content effortlessly but also for specific areas, such as logistics, manufacturing, and the construction industry.

This study explores image analysis with Generative AI and GPT models via OpenAI API. OpenAI is a relatively new organization and laboratory for artificial intelligence research. It was established in 2015 by several tech leaders, such as Elon Musk and Sam Altman. Artificial Intelligence is the research field that OpenAI works in across different areas such as natural language processing (NLP), computer vision, reinforcement learning, and robotics.

OpenAI provides a service via API and UI, offering analysis with models from GPT (Generative Pre-trained Transformer) series like GPT-3 and GPT-4.

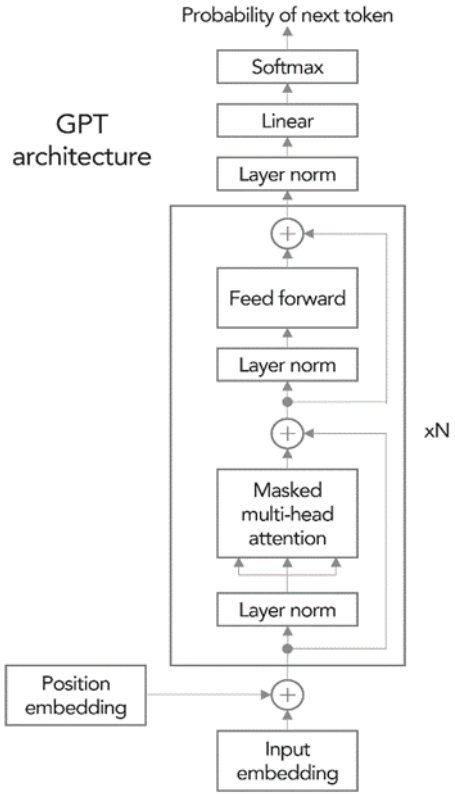
The current paper focuses on implementing different types of analysis using digital content as input (video and images) based on OpenAI with Generative pre-trained transformers (GPT) models.

"Generative pre-trained transformers (GPT) are a type of large language model (LLM) and a prominent framework for generative artificial intelligence. The first GPT was introduced in 2018 by OpenAI." [1]

This concept uses Neural Networks and Reinforcement Learning, but deep-level details are never unveiled from OpenAI. Demonstration of the high-level design of the original GPT model is explained in [1]. A simplified high-level schema of the GPT model is demonstrated in Figure 1.

"The most notable GPT foundation models have been from OpenAI's GPT-n series. The most recent is GPT-4, for which OpenAI declined to publish the size or training details. "The competitive landscape and the safety implications of large-scale models." [2]

Figure 1. *High-level GPT Model Diagram [3]*



This research relies on an analysis done with OpenAI, especially ChatGPT 4 Turbo with Vision, which runs on Microsoft Azure as SaaS using its implementation. The prototype also uses other OpenAI models to improve the efficiency and quality of different analytics solutions.

OpenAI developed a large multimodal model (LMM) called GPT-4 Turbo with Vision that can look at images and generate text answers to queries about them. It combines both visual understanding and natural language processing.

Computer vision is a specific subfield of Artificial Intelligence (AI) that applies machine learning, neural networks, and other AI-related techniques to extract specific, relevant information from images, videos, and other sources. The computer vision approach can be applied to various kinds of analysis:

- Descriptive analysis
- Diagnostic analysis
- Predictive analysis
- Prescriptive analysis

Descriptive analysis is a way to analyze data that reveals the current state of a specific case. It can show, illustrate, and condense data and help with other types of analysis. Descriptive analysis is good for Computer Vision, which interprets what an image or video shows. It can use different technologies, but AI-related technologies make it simple to obtain metadata from different kinds of digital content.

The diagnostic analysis is finding out the main reasons why something (for example, construction failure) happens. Labels from images and videos can help explain what caused something (events, behaviors, outcomes) to happen. It uses descriptive analysis to get information about the case and identify the key cause of a specific situation.

Predictive analysis focuses on forecasting future outcomes based on different data sources. This type of analysis can leverage AI to have a pure analytic solution (without artificial intelligence components) or to use a hybrid approach as part of the solution. Predictive analysis can use various LLMs – GPT, DALL-E, DAVINCI, and WHISPER. Computer Vision is usually essential to extract the information required as input for predictive analysis, and the prediction is made with some of the suitable LLMs analyzing the metadata generated from the Computer Vision.

Prescriptive analysis offers suggestions for future actions based on outcomes from predictive analysis. It uses statistical algorithms, machine learning technologies, and other AI-related solutions to give the best advice for the next steps in a specific situation.

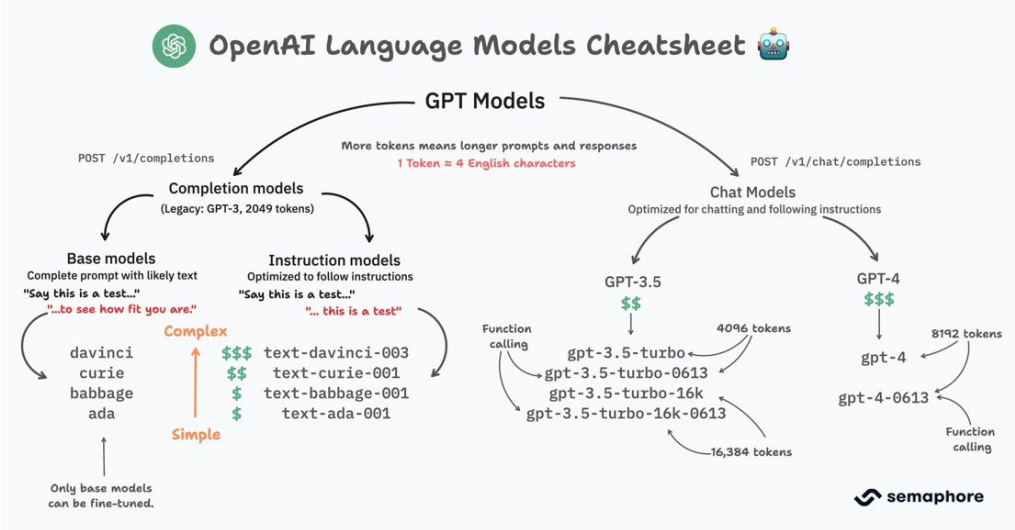
Chat models are designed for conversational interactions. They work as completion models, but they can also distinguish between different speakers, resulting in a smoother dialogue.

Both modes help with analysis. The conversational mode lets users chat and change requirements, and the completion mode gives a full analysis in one go with a clear explanation. Completions make it simpler to automate business processes. ChatGPT-4 Turbo with Vision also uses completions for image analysis.

Some of the LLM models from OpenAI can operate with both chat and completion modes through the respective API, but some models are limited to one mode only.

Figure 2 shows the high-level taxonomy of OpenAI LLM models:

Figure 2. *Cheatsheet with the Models Offered by OpenAI [4]*



In May 2024, the first official version of GPT-4-Turbo with Vision was launched. This is a GPT model that can process text and images and operate in chat and completion modes.

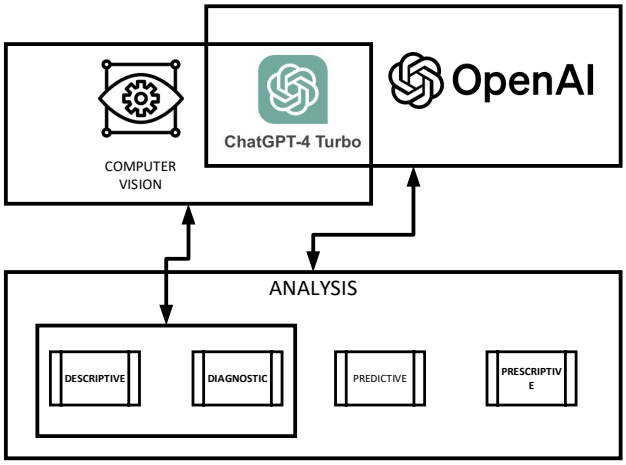
ChatGPT-4 Turbo with Vision (ChatGPT 4 TV) offers fast and precise image and video processing, enabling the analysis of digital content (descriptive and diagnostic analysis) as the initial stage of more advanced predictive and prescriptive analysis. OpenAI LLMs enable various deep and complex queries related to potential system failures.

LLM models like ChatGPT 4 TV allow all stages to be performed using the same model (suitable for prototyping). However, for production, the best LLM for each stage is usually used for cost and performance optimization.

The current research is focused on creating a reference architecture for the main types of analysis using as an input image and video analysis with OpenAI GPT models, especially GPT-4 Turbo with Vision (GPT4-TV) to implement with a short time to market solutions, able to be implemented fast, to provide accurate results and to be available also for low code/no code technologies. Extracting information from images and videos is important to provide correct metadata for predictive analysis to implement predictive maintenance and decrease the overall maintenance cost for different solutions.

The overall schema explaining relations between different types of analysis using Computer Vision and Open AI is demonstrated in Figure 3.

Figure 3. *Implementation of Different Types of Analysis Using Computer Vision, Open AI – General Dependencies*



One interesting case is analyzing GPT4-TV video streams and extracting frames from video using Azure AI Vision.

A few weeks ago, OpenAI's latest iteration, GPT-4o, introduced in May 2024, integrates multimodal capabilities that process text, images, and audio within a single framework. This integration facilitates a more holistic understanding of multimedia content, making GPT-4o particularly adept at tasks such as video summarization.

Integration of GPT-4o for digital content analysis is in the plans for the author's next publications.

The images that you show to the GPT-4 Turbo with Vision model can be used to answer general questions about what they contain. It is also possible to use Vision enhancement to analyze video content.

GPT-4 Turbo with Vision provides access to exclusive enhancements from Azure AI Services. The video prompt integration uses Azure AI Vision to select some frames from a video and make a transcript of the video's speech. This allows the AI model to provide summaries and answers about the video content.

Video analysis combined with cutting-edge natural language processing (NLP) technologies have created new opportunities for innovation in today's fast-changing digital environment. It is possible to integrate Azure AI Video Indexer (an application based on Azure AI Vision) and GPT-4 Turbo with Vision. These options are aimed at enhancing video analysis capabilities with the power of Large Language Models. They enable business value distillation from the video content, offering insights and automation opportunities through ad-hoc chat interfaces for content discovery or automated custom tasks. Here are some compelling use cases demonstrating this potential for a variety of industries.

Methodology

Concept, Technologies, and Methodology Description

The study uses a flow in two scenarios for image analysis and two for video description:

- Image analysis

A. Analysis of Images using ChatGPT-TV

LLM model extracts metadata from images

B. Image analysis based on Azure AI Vision and GPT-4-TV in 2 steps.
    1. Objects are recognized from Azure AI Vision
    2. GPT-4-TV makes detailed descriptions of recognized objects.

The second approach gives better accuracy on the overall images on account of additional steps and a more complex schema of the solution, providing image recognition.

- Video analysis

A. Analysis of videos using GPT-4-TV using AI Vision to get frames from video for detailed image recognition in two steps:
    1. Getting frames from the video: AI Vision
    2. Image Analysis: GPT-4-TV

B. Using GPT-4-TV and AI Vision in 3 steps:
    1. Getting frames from the video: AI Vision
    2. Objects are recognized from Azure AI Vision
    3. 2. GPT-4-TV makes a detailed description of recognized objects

This research also demonstrates a reference architecture on how to build copilot-style apps that use both GPT-4 Turbo with Vision and Azure AI Vision and

Search in Microsoft's Azure AI Studio. They allow image inputs to query your organizational data directly and generate AI responses based on them. This significantly improves natural language processing and image recognition tasks and enables new generative AI scenarios. You can also use video inputs when you combine GPT-4 Turbo with Vision and Azure AI Vision.

The Theoretical and Technical Framework

The experimental environment uses ChatGPT 4-Turbo with Vision (ChatGPT-4-TV) for image analysis. This model has the following specifics:

- Open AI-based
- LLM based.
- Good at solving not defined requirements in both conversational and completion modes
- Option to use Retrieval Augmented Generation (RAG)

In the case related to analyzing the video stream (input stream from drones and robots' cameras during construction observation), Azure AI Vision empowers ChatGPT-4-TV.

Azure AI Vision is a different solution for computer vision analysis, which uses cognitive analysis to understand videos and images. It has the following characteristics:

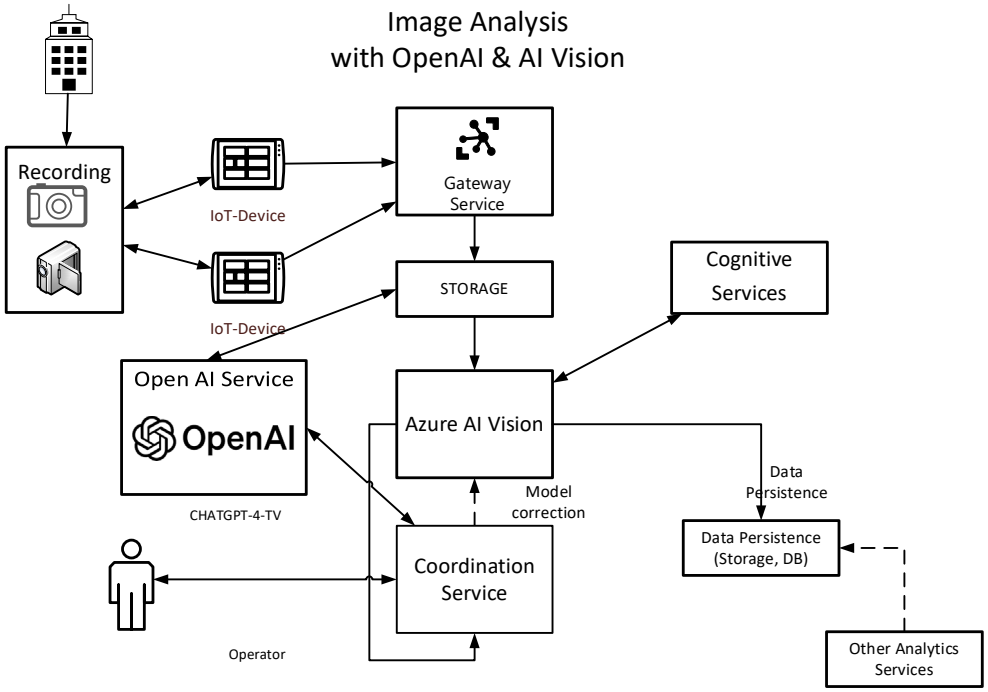
- It's not LLM based.
- Good in clearly defined tasks, object recognition, etc.
- Working with video streaming.

AI Vision extracts frames from the stream and sends them to the OpenAI service using a ChatGPT-4-TV deployment.

The research is focused on the realization of a common effective approach to the analysis of images and video using OpenAI GPT models (especially GPT-4-TV) and optimizing the effectiveness of the solution with Computer Vision based on Azure AI Vision services. Results from different solutions are recorded and compared to recommend the set options with specific precision for different cases.

Figure 5 shows the contextual schema of a solution using OpenAI and AI Vision for digital content analysis.

Figure 5. *Contextual Schema for Image and Video Analysis, Based on OpenAI and AI Vision*



The solution covers scenarios in which digital content (images, videos) is generated using different platforms (drones, robots, technicians). This content is analyzed in different ways (with and without object detection to increase the precision of the analysis).

Image analysis can be considered an essential part of descriptive and diagnostic analysis and is the main input for predictive and prescriptive analysis.

GPT-4 Turbo with Vision goes beyond these limitations by using visual data, allowing a higher level of image comprehension. This model is not just about identifying objects in an image; it's about grasping the situation and specifics—such as generating detailed image captions, giving abundant contextual explanations, answering questions about visual content, or applying smart tags. GPT-4 Turbo with Vision enhances data analysis to new levels, understanding the visual world in ways that surpass simple pixels.

- GPT-4 Turbo with Vision + Azure AI Services

GPT-4 Turbo with Vision on Azure OpenAI Service offers cutting-edge AI capabilities, enterprise-grade security, and responsible AI governance. In addition, it provides exclusive access to Azure AI Services tailored enhancements. When combined with Azure AI Services, it enhances your experience by introducing an array of advanced functionalities, including:

- Getting frames from video for post-action image analysis
- Video Retrieval

Video Retrieval helps GPT-4 Turbo with Vision respond to video prompts using a selected set of images from the video as reference data. This means that when you query specific scenes, objects or events in a video, the system gives more precise answers without transferring all the frames to the large multimodal model (LMM). This reduces time and cost, especially for long videos that might otherwise go over the token context 128k window of GPT-4 Turbo with Vision. Azure AI Vision Video Retrieval will supply the relevant frames to the AI model to produce an accurate answer.

Research

Implementation of Image Analysis Based on Computer Vision and OpenAI/ ChatGPT

The input data is based on 1000+ images and 100+ videos of rusty steel construction elements taken from real structures in the USA (Figure 6).

Figure 6. *Sample Image of Steel Construction with Corrosion*

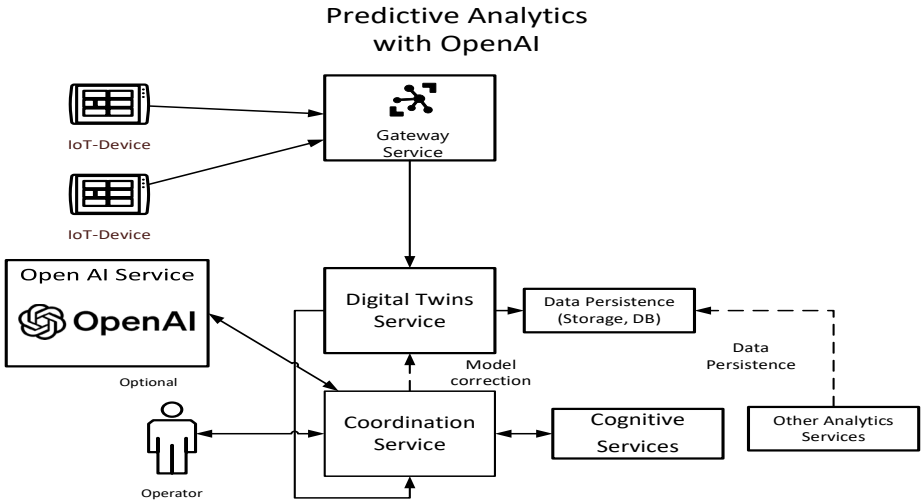
- Implementation of a Computer Vision Concept:

The experimental environment uses ChatGPT 4-Turbo with Vision (ChatGPT-4-TV) for image analysis.

The experimental setup includes a Digital Twins (DT) module based on Azure (Azure Digital Twins/ADT). ADT is used to extract the context from the observed solution, decompose the case, and unify the analysis case, converting it from domain-specific to domain-agnostic subcases suitable for LLM analysis.

One high-level schema of the solution is presented in Figure 7 [5].

Figure 7. *Experimental Setup: Predictive Analysis Using Computer Vision Open AI Based on Microsoft Azure [5]*



The coordination service helps create and ask questions that clarify the search and allow the OpenAI service to provide the right solution.

Digital Twins (DT) is a part of the solution that can be skipped, used for two things:

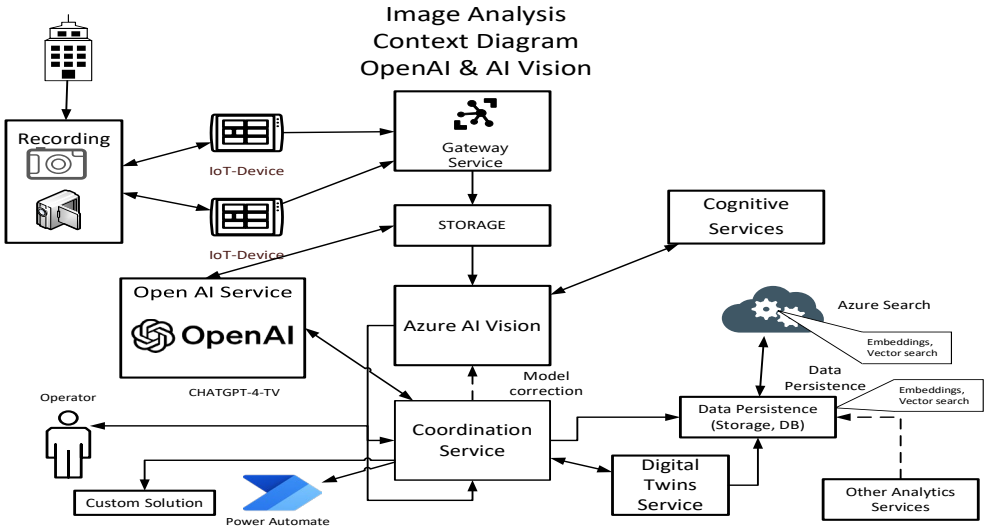
- Breaking down complex cases
- Simulations in Predictive Analysis

The coordination service has two implementations:

1. Power Platform (Power Automate) for PoC
2. Custom implementation as a cloud service for high-load solutions

A full-detail schema is demonstrated in Figure 8.

Figure 8. *A Full, Detailed Context Diagram of the Experimental Setup*



Based on the research, were recorded two KPIs

- Correctness
- Time for analysis

Results are recorded for both images using ChatGPT-4-TV and videos with and without Azure AI Vision.

The experiment is done with and without an RAG solution (Retrieval Augmented Generation).

Results

In this paper, there are added metrics related to the experimental project PoC using the following parameters:

1. Coordination Service
    a. Custom solution
    b. Power Platform (Low-Code/No-Code)
2. Digital Content
    a. Video
    b. Images
3. Analysis
    a. Without Azure AI Vision for object detection
    b. With Azure AI Vision for object detection
4. RAG
    a. Without RAG
    b. Using RAG

The implementation uses SaaS Components, where a Digital Twins Instance is constantly updated by a coordination service. This service gets data from an Azure Digital Twins instance and sends it to the OpenAI service for subcase analysis, custom-built analysis services, and Cognitive Services for overall model enhancement.

The input is based on images and videos of rusty steel structures in the USA, which are available on the public Internet as free content.

Tables 1-4 demonstrate experimental results with all combinations of parameters.

Table 1. *Experimental Results Accuracy – no RAG*

Coordination Service	**#images**	**#vidoes**	**#success rate**	**#time image analysis**	**#time video analysis**	**#time for implementation [day]**
Custom Solution	1000	100	99.1%	15 min	20 min	1
Low-Code/ No-Code	1000	100	99.1%	49 min	29 min	10

Table 2. *Experimental Results Accuracy + AI Vision, no RAG*

Coordination Service	**#images**	**#vidoes**	**#success rate**	**#time image analysis**	**#time video analysis**	**#time for implementation[day]**
Custom Solution	1000	100	99.9%	15 min	20 min	1
Low-Code/ No-Code	1000	100	99.9%	49 min	29 min	10

Table 3. *Experimental Results Accuracy with RAG*

Coordination Service	**#images**	**#vidoes**	**#success rate**	**#time image analysis**	**#time video analysis**	**#time for implementation [day]**
Custom Solution	1000	100	99.1%	9 min	12 min	5
Low-Code/ No-Code	1000	100	91.1%	19 min	20 min	20

Table 4. *Experimental Results Accuracy + AI Vision, with RAG*

Coordination Service	**#images**	**#vidoes**	**#success rate**	**#time image analysis**	**#time video analysis**	**#time for implementation[day]**
Custom Solution	1000	100	99.9%	9 min	12 min	1
Low-Code/ No-Code	1000	100	99.9%	19 min	20 min	10

Conclusions

- Modern Image Analysis can rely on OpenAI ChatGPT-4-TV.
- Azure AI Vision for object detection provides additional precision, increasing the correctness of the description by around 1%.
- The overall analysis offered by automation solutions can effectively solve complex cases even on the PoC level using both custom and Low-Code/No-Code solutions. The approach is domain-agnostic and can be used in different business domains.
- Cost and performance optimization can be improved up to two times using Retrieval Augmented Generation (RAG), especially for the analysis of complex images and videos.

Abbreviations

- CS Cognitive Services
- IoT Internet of Things
- CDT Cognitive Digital Twins
- ChatGPT Generative pre-trained transformers
- LLM – Large Language Models
- SLM – Small Language Models
- DT Digital Twins
- ADT Azure Digital Twins

- DT Cognitive Digital Twins
- ChatGPT Generative pre-trained transformers
- ChatGPT-4-TV-> ChatGPT 4-Turbo with Vision.
- RAG -> use Retrieval Augmented Generation

References

[1] "Generative Pre-Trained Transformer." Wikipedia, May 26 2024, en.wikipedia.org/wiki/Generative_pre-trained_transformer#Foundational_models.

[2] "GPT-4 Technical Report". Arxiv.Org, 2023, https://arxiv.org/abs/2303.08774. Accessed May 8 2024.

[3] [14] Hestness, Joel. "Cerebras Sets Record For Largest AI Models Ever Trained On Single Device - Cerebras". Cerebras, 2022, https://www.cerebras.net/blog/cerebras-sets-record-for-largest-ai-models-ever-trained-on-single-device#ml-impacts. Accessed August 13 2023.

[4] T. Fernandez, "How to Choose the Best OpenAI Model for Your AI Application," Semaphore CI, August 10, 2023. [Online]. Available: https://semaphoreci.com/blog/openai-models

[5] Mateev M. Predictive Analytics Based on Digital Twins, Generative AI, and ChatGPT, Proceedings of the 27th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2023, pp. 168-174 (2023); https://doi.org/10.54808/WMSCI2023.01.168

[6] Contextualizing Large Language Models (LLMs) with Enterprise Data. (2023). Retrieved July 16 2023, from https://www.linkedin.com/pulse/contextualizing-lar ge-language-models-llms-enterprise-debmalya-biswas/

[7]https://www.sablono.com/de/blog/bim-and-digital-twin-technology/, accessed: 25-Nov-2019

[8] https://learn.microsoft.com/en-us/azure/digital-twins/tutorial-end-to-end, accessed: 26-Dec-2022 .

[9] Aydın, Ömer and Karaarslan, Enis, OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare (December 21, 2022). Aydın, Ö., Karaarslan, E. (2022). OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare. In Ö. Aydın (Ed.), Emerging Computer Technologies 2 (pp. 22-31). İzmir Akademi Dernegi., Available at SSRN: https://ssrn.com/abstract=4308687 or http://dx.doi.org/10.2139/ssrn.4308687

[10] "Introducing Chatgpt". 2023. Openai.Com. https://openai.com/blog/chatgpt.

[11] El Mokhtari, Karim, Ivan Panushev, and J. J. McArthur. 2022. "Development Of A Cognitive Digital Twin For Building Management And Operations". Frontiers In Built Environment 8. doi:10.3389/fbuil.2022.856873.

[12] D'Amico, Rosario Davide, John Ahmet Erkoyuncu, Sri Addepalli, and Steve Penver. 2022. "Cognitive Digital Twin: An Approach To Improve The Maintenance Management". CIRP Journal Of

[13] Ünal, Perin. "Cognitive Digital Twins: Digital Twins That Learn by Themselves, Foresee the Future, and Act Accordingly." Digital Twin Consortium, June 9, 2023. https://www.digitaltwinconsortium.org/2022/09/cognitive-digital-twins-digital-twins-that-learn-by-themselves-foresee-the-future-and-act-accordingly/.

[14] Generative pre-trained transformer - Wikipedia. (2023). Retrieved July 16 2023, from https://en.wikipedia.org/wiki/Generative_pre-trained_transformer#Foundational_models

[15] OpenAI (2023). "GPT-4 Technical Report" (PDF). Archived (PDF) from the original on 2023-03-14. Retrieved 2023-03-16.
[16] Mateev M., Design and Implementation of Cognitive Digital Twins with Generative AI and ChatGPT, 4th Annual International Conference on Computer & Software Engineering 17-20 July 2023, Athens, Greece, Abstract Book, pp. 43
[17] Kuo-Liang Lin and Jyh-Bin Suen, 2019, "A Vision-Based Method for Determining Degradation Level of a Road Marking", "ATINER CONFERENCE PRESENTATION SERIES No: CIV2019-0138, Athens, 2019