AI-VaS: Empowering Edge Video Analytics with Cloud-Based Services

Cristian Cerasuolo[1,†], Andrea Vincenzo Ricciardi[1,†], Domenico Rocco[1,†], Stefano Saldutti[1,†], Bruno Vento[1,*,†] and Antonio Vitale[2,†]

[1] *A.I. Tech srl - www.aitech.vision*
[2] *A.I. Ready srl*

Abstract

A.I. Tech was born as an academic spin-off of the University of Salerno and is specialized in the design and development of advanced video analytics solutions based on deep learning techniques. The solutions developed by A.I. Tech are designed to address the specific needs of various vertical markets, including retail, business intelligence, security, safety, smart parking, smart cities and smart roads. In particular, in this paper we present the innovative AI-VaS platform, which enables A.I. Tech products to be powered with cloud-based services. These services can be accessed at any time from any device capable of transmitting images via the standard HTTPS protocol.

Keywords

Artificial Vision, Cloud Services, Deep Learning, Edge Computing, Smart Surveillance

1. Company presentation

A.I. Tech develops solutions for video analysis built upon state-of-the-art artificial intelligence and deep learning methodologies. The company has established strategic partnerships with some of the most important global players in their respective industries including, but not limited to, NVIDIA, Panasonic, Samsung, Hanwha Techwin, Mobotix, Axis, Hikvision and Dahua.

Over the years, A.I. Tech has earned significant international recognition. In 2017, the company was named one of the Top 25 AI companies globally by CIO Applications Magazine. The following year, it was ranked among the Top 10 Most Innovative AI Solution Providers. Its technologies were selected as finalists at the Benchmark Innovation Awards consecutively from 2018 through 2022. In 2018, A.I. Tech claimed the Business Intelligence category award with its AI-RETAIL video analytics solution. In 2020, the company received the Corporate LiveWire award for Most Innovative Video Analytics Solution. That same year, A.I. Tech was a finalist at the Security and Fire Excellence Awards, competing with its AI-CROWD-DEEP product in the Security Software Product Innovation of the Year category and with the WOW project for Security Project of the Year. Additionally, its AI-TRAFFIC system for traffic analysis earned the IoMOBILITY AWARD 2020 in the Mobility Analytics category. The company's consistent drive for innovation was further recognized by Corporate LiveWire, which presented A.I. Tech with its Innovation and Excellence Award for 2022, later renewing the title in 2023, 2024 and 2025, affirming the company as a leader in AI technology innovation.

Given the highly technical and scientific nature of its work, which requires deep expertise in artificial intelligence, computer vision and embedded systems, A.I. Tech maintains a close partnership with the Department of Information and Electrical Engineering and Applied Mathematics (DIEM) at the University of Salerno. This collaboration includes agreements for student internships and ongoing research projects over the coming years. Through this partnership, the advanced academic expertise in

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

[*] Corresponding author.

[†] These authors contributed equally.

cristian.cerasuolo@aitech.vision (C. Cerasuolo); andrea.ricciardi@aitech.vision (A. V. Ricciardi); domenico.rocco@aitech.vision (D. Rocco); stefano.saldutti@aitech.vision (S. Saldutti); br1.vento@gmail.com (B. Vento); vitale.antonio@gmail.com (A. Vitale)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-07

AI and computer vision developed by the DIEM research group is transferred into real-world applications, resulting in a portfolio of cutting-edge AI products available on the international market.

2. Overview of the AI-VaS solution

AI-VaS is an innovative and versatile platform specifically designed to provide image analysis and image understanding as-a-service. This architecture enables on-demand access to advanced analytics models from any authorized device, at any time. As a result, the system offers exceptional flexibility and scalability, making it highly adaptable to a wide range of deployment scenarios. AI-VaS is fully integrated within the video analysis solutions developed by A.I. Tech and is primarily employed for two key purposes: to confirm events detected by edge devices or to conduct a preliminary analysis of those events.

A typical example of its usage is illustrated in Figure 1. From left to right, the diagram shows an edge device, such as a smart surveillance camera, running a video analysis application. When an event is detected, the application sends a data package containing the event metadata and the associated video frame to AI-Dash, A.I. Tech's proprietary Event Management System. AI-Dash's role is to collect, aggregate, analyze and visualize events from all intelligent devices installed at the customer's premises. In this workflow, AI-Dash acts as a client for the AI-VaS platform: it forwards the event metadata and images to the AI-VaS service, which processes the request and returns the analysis results. These outcomes can then be used to enrich the original event information or to support automated decision-making processes downstream.

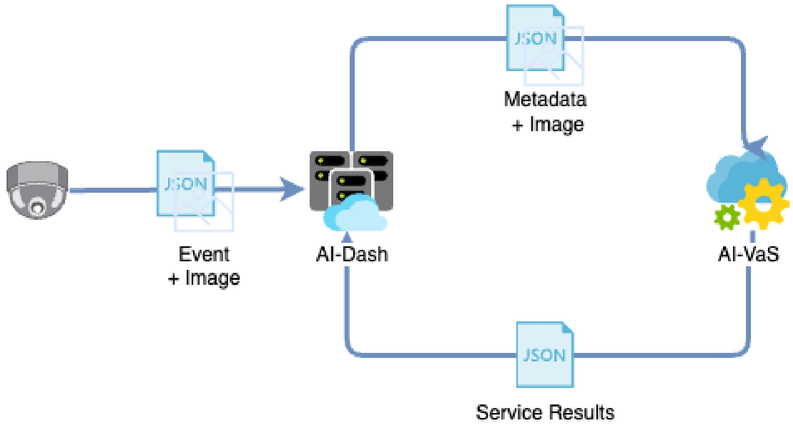

Figure 1: AI-Dash communicating with AI-VaS.

The overall architecture of the AI-VaS platform is detailed in Figure 2. Any compatible client, such as the previously mentioned AI-Dash, can request one of five available services: AI-VaS BIO, AI-VaS CARGO-LOSS, AI-VaS CROWD, AI-VaS IMAGE CAPTIONING, or AI-VaS LPR. Communication is handled via a lightweight protocol based on HTTPS message exchange. Through this interface, the client submits licensing credentials, request metadata and any required resources (such as images or other media) to the AI-VaS Reception server.

The AI-VaS Reception component is responsible for formalizing the incoming requests, temporarily storing the associated media files and routing each task to the appropriate service module. This is done by placing the task in one of several processing queues, each tied to a specific service. Every AI-VaS service module autonomously retrieves tasks from its respective queue and processes them accordingly.

Once processing is complete, the results are placed into a shared output queue, which serves as a central buffer for outgoing responses. The AI-VaS Dismisser module continuously monitors this queue and ensures that the analysis results are promptly delivered back to the originating clients, thereby closing the request-response loop.

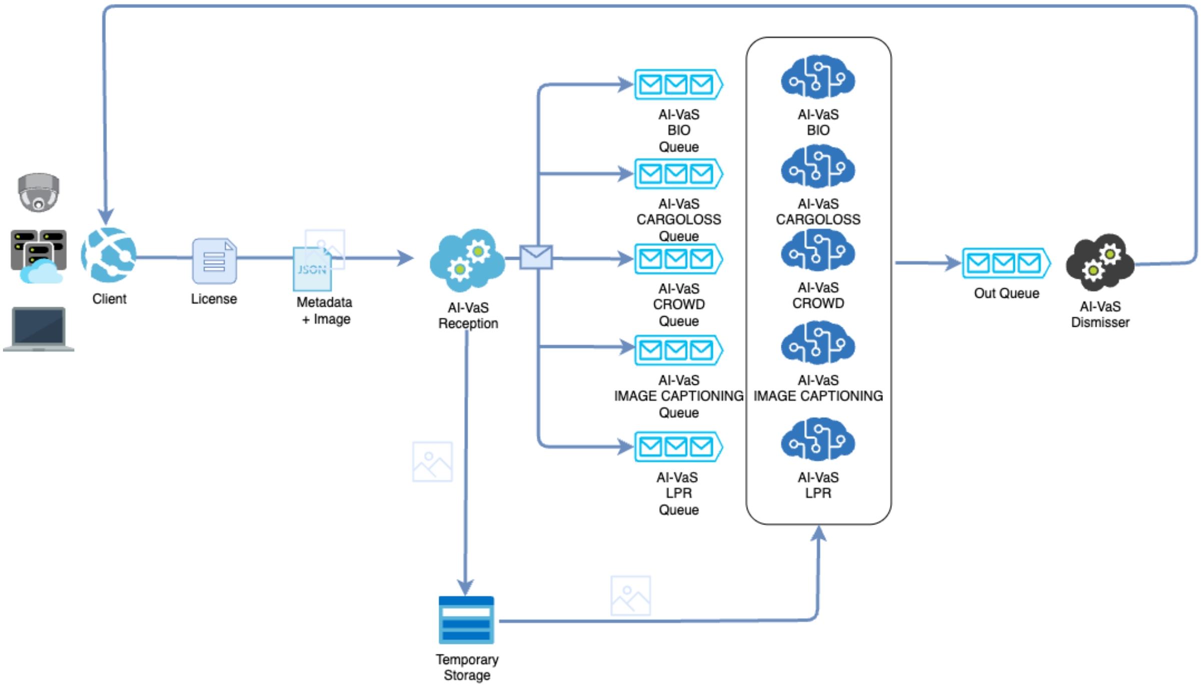

Figure 2: Architecture of AI-VaS platform.

3. Cloud-based services

In this section, we describe the five analytics service solutions currently available on the market.

AI-VaS BIO is a video analytics service for face detection and analysis, designed to extract attributes such as age, gender, and emotion from facial imagery [1, 2, 3]. Compared to traditional on-edge deployments, AI-VaS BIO offers significant advantages in terms of scalability, flexibility, and maintainability. By offloading the processing to centralized infrastructure, it becomes possible to deploy more powerful and computationally demanding multi-task architecture that would not be feasible on resource-constrained edge devices. This enables improved accuracy and the use of more sophisticated models for face analysis, ultimately enhancing the quality of the extracted soft-biometric information. In the context of digital signage [4], AI-VaS BIO enables the dynamic personalization of advertising content displayed on public monitors. By analyzing the soft-biometric features, such as age group, gender, and emotional state of individuals in front of the screen, the system can tailor the visual content in real time to better match the viewer's profile. An example is shown in Figure 3a.

AI-VaS CARGO-LOSS is a specialized vertical solution developed to identify cargo losses from vehicles traveling on highways—an issue of critical importance for road safety. When cargo falls or is unintentionally abandoned, these objects become hazardous obstacles on the road, especially dangerous for high-speed vehicles that may not have enough time to react. At the core of the system lies AI-LOST, a video analysis application specifically designed to detect abandoned or removed objects within surveilled areas. AI-LOST continuously monitors road footage and identifies potential threats by highlighting anomalies—suspected as road-abandoned objects. When AI-LOST highlights a suspicious object, the information is passed to the CARGO-LOSS module. This is where state-of-the-art zero-shot object detection algorithms take over. These algorithms are particularly innovative because they can recognize previously unseen or undefined objects based on textual descriptions, rather than relying on pre-labeled categories. This makes them exceptionally versatile in dynamic, real-world scenarios like highway cargo loss, where the types of objects can vary widely and unpredictably. In short, this two-step pipeline ensures high accuracy, significantly reducing false alarms while maximizing detection reliability. An example is shown in Figure 3b.

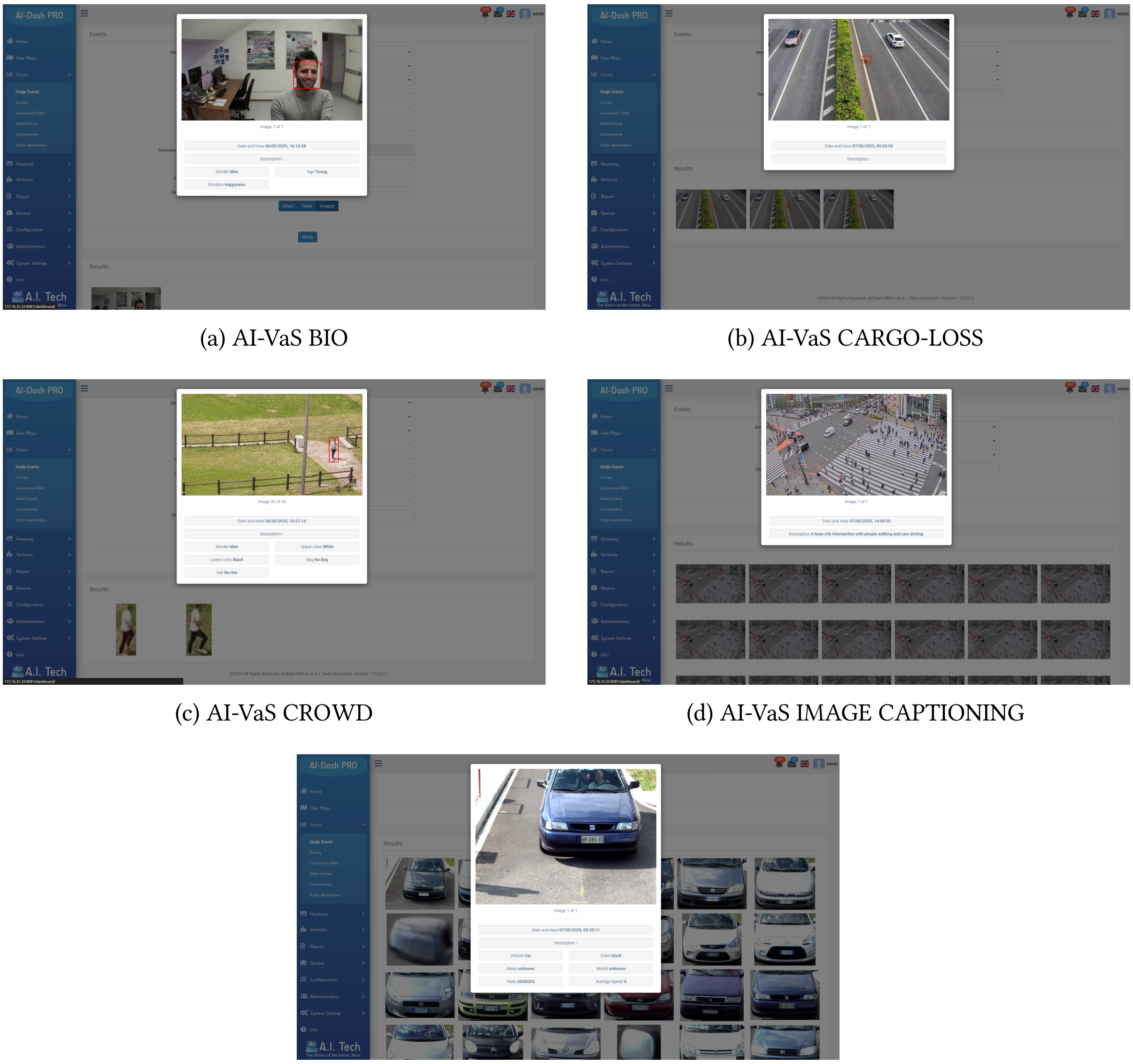

Image 1 of 1

Date and Hour **06/05/2025, 16:12:38**

Description -

Gender **Man**

Age **Young**

Emotion **Happyness**

(a) AI-VaS BIO

Image 1 of 1

Date and Hour **07/05/2025, 09:33:53**

Description -

(b) AI-VaS CARGO-LOSS

Image 32 of 35

Date and Hour **06/05/2025, 16:57:14**

Description -

Gender **Man**

Upper color **White**

Lower color **Black**

Bag **No Bag**

Hat **No Hat**

©2025 All Rights Reserved. AI-Dash PRO is an A.I. Tech srl product. Version 1.15.3312

(c) AI-VaS CROWD

Image 1 of 1

Date and Hour **07/05/2025, 14:09:25**

Description **A busy city intersection with people walking and cars driving.**

(d) AI-VaS IMAGE CAPTIONING

Image 1 of 1

Date and Hour **07/05/2025, 09:25:11**

Description -

Vehicle **Car**

Color **black**

Make **unknown**

Model **unknown**

Plate **AR285ES**

Average Speed **8**

(e) AI-VaS LPR

Figure 3: Some examples of A.I. Tech video analytic plugins in action. Fig. 3a AI-VaS BIO: page of AI-DASH that shows an image of a person with the corresponding soft-biometric features extracted by the service. Fig. 3b AI-VaS CARGO-LOSS: a cardboard box, detected by AI-LOST as a potentially abandoned object, is confirmed by the service as a waste item discarded on the highway Fig. 3c AI-VaS CROWD: page of AI-DASH that shows a frame sent by and event from the surveillance plugin. Below the frame, the AI-VaS CROWD service outputs are shown, including: upper and lower clothing colors, presence of a bag or hat, and the gender of the person highlighted in the red box. Fig. 3d AI-VaS IMAGE CAPTIONING: page of AI-DASH showing input frame of the AI-VaS IMAGE CAPTIONING service with the output caption that describes what is depicted in it. Fig. 3e AI-VaS LPR: page of AI-DASH that shows one of the images of the vehicle and the plate value returned by the service.

AI-VaS CROWD is an analytic service specifically designed for Pedestrian Attribute Recognition (PAR). Leveraging a proprietary deep learning-based classifier, it processes image patches containing individuals and is capable of: (i) classifying gender; (ii) estimating the color of both upper and lower garments from a predefined set of colors; (iii) detecting the presence of a hat; and (iv) determining whether the person is carrying a bag. Through the dashboard interface, as illustrated in Figure 3c, users can perform forensic analyses by filtering individuals based on specific attribute configurations. This functionality proves particularly valuable for distinguishing individuals in crowded scenes or for searching for a person matching a specific identikit.

AI-VaS IMAGE CAPTIONING is a general-purpose service designed to automatically generate descriptive captions for images. It leverages state-of-the-art techniques in image understanding to produce accurate and meaningful textual descriptions. Within the dashboard, this service is applied to images associated with events received from video surveillance systems, enhancing the interpretability of visual content. Additionally, it could be used to contribute to accessibility by generating alternative text descriptions, thereby supporting users with visual impairments. An example is shown in Figure 3d.

AI-VaS LPR is the video analysis service for the license plates detection and recognition, starting from one or more images of the vehicle. The service has various usage scenarios such as, for example, in the management of parking lots, for the management of black and white lists or in the association of the license plate to the parking entrance ticket or to detect access to limited traffic areas and transits in preferential lanes. Both the sensitivity for the detection of the license plate and that for the detection of individual characters are configurable. Furthermore, thanks to the use of an engine based on semantic technologies, the service is able to automatically correct license plates based on the specific nationality. The supported nations are: Italy, France, Spain and Greece. For each vehicle image, the service detects the plate and, if the results are different, it returns the one with the highest confidence. An example is shown in Figure 3e.

4. Performance evaluation

In this section, we describe the performance of the five analytics services.

The data were collected under high-load conditions, meaning that, for each service, a queue of 200 events was pre-filled to ensure a continuous and consistent workload. This approach allowed the system to process events sequentially, thereby maintaining sustained resource utilization and avoiding idle periods that could affect measurement accuracy. Furthermore, the services were monitored individually, i.e., one at a time, to prevent resource contention and ensure that the collected metrics reflected the isolated behavior of each service under load. The recorded data include CPU usage (in percentage), GPU usage (in percentage), RAM consumption (in gigabytes), and VRAM consumption (in gigabytes).

The various services were monitored on a machine equipped with two Intel Xeon E5-2660 v2 processors, providing a total of 20 physical cores and 40 threads, 125.82 GiB of DDR3 RAM running at 1333 MHz, and an NVIDIA GeForce GTX 1060 with 6 GB of VRAM.

Table 1
Summary of resource usage for each AI-VaS service, including initialization time, average GPU memory usage, system RAM usage, GPU utilization, and CPU utilization during high-load operation. All values are based on empirical observations collected in a controlled test environment.

Service	Init Time (s)	GPU Mem (GB)	RAM (GB)	GPU Util (%)	CPU Util (%)
BIO	80	2.5	3.2	8	3.3
CARGO-LOSS	20	1.6	2.9	60	2.9
CROWD	60	1.0	2.9	10	2.9
IMAGE CAPTIONING	35	3.6	3.0	80	3.0
LPR	15	0.7	2.7	3.5	2.7

AI-VaS BIO. The system took approximately 80 seconds to reach full operational capacity, utilizing 2.5GB of GPU memory and 3.2GB of system RAM on average. CPU usage was minimal, with an average utilization of just 3.3%. As shown in Table 1, the GPU utilization consistently hovered around 8%.

AI-VaS CARGO-LOSS. After initializing in 20 seconds, the system maintained an average GPU utilization of 60%, with CPU usage staying low at 2.9%. It required 1.6GB of GPU memory and 2.9GB of RAM. The GPU utilization, as reported in Table 1, is increased due to the use of larger models that are specifically designed to run efficiently on GPUs.

AI-VaS CROWD. This service took about 60 seconds to stabilize, during which it used approximately 1GB of GPU memory and 2.9GB of system RAM. The GPU utilization averaged around 10%. The CPU maintained a steady load of 2.9%. As indicated in Table 1, the system's GPU usage remained relatively stable throughout the operation.

AI-VaS IMAGE CAPTIONING. The system reached a stable state within 35 seconds. The GPU utilization averaged 80%, while CPU usage was consistently low at 3%. The system consumed 3.6GB of GPU memory and 3GB of system RAM. The high GPU utilization is due to the use of larger, more computationally demanding models, similar to AI-VaS CARGO-LOSS, which are optimized for GPU execution. Table 1 provides a summary of the usage metrics.

AI-VaS LPR. Achieving stable performance within just 15 seconds, the system used 0.7GB of GPU memory and 2.7GB of RAM. GPU utilization averaged around 3.5%. The CPU load was also low, maintaining an average usage of 2.7%. As shown in Table 1, the GPU utilization remained consistent during the license plate recognition task.

5. Conclusions

A.I. Tech's AI-VaS platform stands out as a robust, scalable, and flexible solution for advanced video analysis, leveraging cutting-edge AI and deep learning technologies. Its diverse service modules, tailored for applications ranging from face biometrics to cargo loss detection and license plate recognition, demonstrate a well-rounded approach to real-world challenges. The platform's architecture efficiently balances computational load, offering powerful cloud-based analytics while maintaining low system resource usage, ensuring both performance and reliability. Through strategic partnerships and ongoing research collaborations, A.I. Tech continues to push the boundaries of innovation in AI-driven video analytics, positioning itself as a leader in this fast-evolving field.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT solely for grammar and spelling checks. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Greco, A. Saggese, M. Vento, V. Vigilante, A convolutional neural network for gender recognition optimizing the accuracy/speed tradeoff, IEEE Access 8 (2020) 130771–130781. doi:10.1109/ACCESS.2020.3008793.

[2] A. Greco, A. Saggese, M. Vento, V. Vigilante, Gender recognition in the wild: a robustness evaluation over corrupted images 12 (2021).

[3] A. Greco, A. Saggese, M. Vento, V. Vigilante, Effective training of convolutional neural networks for age estimation based on knowledge distillation, Neural Comput. Appl. (2021).

[4] A. Greco, A. Saggese, M. Vento, Digital signage by real-time gender recognition from face images, in: 2020 IEEE International Workshop on Metrology for Industry 4.0 IoT, 2020, pp. 309–313.