WHITE PAPER

# Free-text search in AXIS Camera Station Pro

January 2025

# Summary

AXIS Camera Station Pro comes with several video forensic search tools preinstalled. These include timeline scrubbing, data search, and smart search with pre-classified objects and free-text search.

Free-text search allows you to search for any moving objects by describing them in your own words. The freedom to create detailed search filters with a wide range of descriptive attributes makes it possible to find relevant footage quicker.

The free-text search function is based on text-image matching provided by a pre-trained open-source foundation model that's been optimized by Axis for surveillance use cases. The search can be applied to one camera or several cameras at the same time.

A numerical representation of your free-text query is compared with numerical representations of images of detected moving objects. The results are displayed as thumbnails, including camera name, time, and date, sorted by relevance to your search query.

With free-text search, we use AI to increase the accuracy and efficiency of our forensic search solutions and ultimately enhance human decision-making. To comply with legal and ethical standards, the search function includes a separate Axis-developed moderation function that restricts the use of offensive words in search queries. All searches are also logged and visible to administrators, making it possible to follow up and take corrective action in case of misuse.

# Table of Contents

# 1 Introduction

The free-text search tool in AXIS Camera Station Pro allows you to search video recordings using your own words instead of predefined filters.

This white paper outlines how the search method works and presents some guidelines on how to use it. We also describe the moderation function and query logging that are in place to ensure compliance with legal and ethical standards.

# 2 Background: forensic search in AXIS Camera Station Pro

AXIS Camera Station Pro comes with several video forensic search tools preinstalled, including timeline scrubbing, data search, as well as smart search with pre-classified objects and free-text search.

The smart search function uses scene metadata generated by the Axis device. The metadata includes object type (person, vehicle type, or unknown object) for moving objects, along with attributes like clothing and vehicle color, license plates, speed, location, and timestamp.

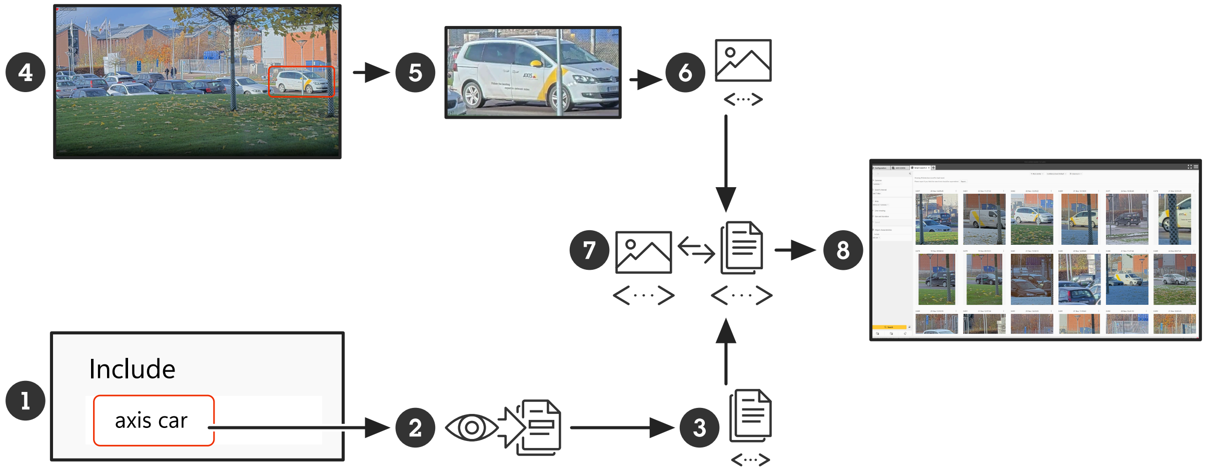
For devices with limited analytics capabilities, the search function is based on motion detection in the device combined with object classification performed on the AXIS Camera Station Pro server. Forensic search in AXIS Camera Station Pro is thus a hybrid solution where the capabilities of the edge devices are used as far as possible but supplemented with data from the server where necessary.

Traditionally, searches using scene metadata had to be conducted using predefined search filters. With these, you choose fixed object descriptors from a list, including object type (such as "vehicle"), vehicle type (if applicable, such as "car"), color (such as "blue"), and more. The new free-text search method instead allows you to create your own search filter.

While pre-classified search delivers high-precision results, it can't detect novel object types that aren't pre-defined. To address this limitation, free-text search gives you the freedom and flexibility to search using your own words. You can describe any moving object in greater detail with natural language and associations to fine-tune your search and get more results.

# 3 How does the free-text search work?

A numerical representation of your free-text query is compared with numerical representations of images of detected moving objects. The result of this text-image matching is presented and sorted by the best match. The results are displayed as thumbnails, including camera name, time, and date, sorted by relevance to your search query.



*Simplified overview of free-text search process. Note that steps 4-6 take place continuously, even when you're not searching, to create feature vectors of all detected moving objects.*

1 *You type your free-text search query.*
2 *A moderation module prevents the use of toxic and unethical words.*
3 *The foundation model creates a numerical representation (a feature vector) of the search query.*
4 *A camera detects movement in a scene.*
5 *The camera selects one cropped image to represent the moving object.*
6 *The foundation model creates a feature vector of the object after analyzing its shape, patterns, color, and more.*
7 *The two feature vectors are compared.*
8 *The result of the comparison is sorted by best match and presented as thumbnails.*

Free-text search can be applied to one camera or several cameras at the same time.

To narrow the scope of your free-text search you can combine it with other smart search functions, such as similarity search or time-based search, by using one search type after another.

# 4 Constructing search queries

You can search for any moving object and any type of vehicle. Follow the guidelines for best results.

Note that you should only search for moving objects. Searching for stationary objects will in most cases not work.

Search phrases are moderated and logged to prevent unethical search behaviors.

## 4.1 Prompt guidelines

- Describe situations as you would describe an image. The model is fed with still images so searching for actions (such as falling, running, or stealing) can be difficult since it would require more context.

- Describe objects using a few key descriptors: "a person in a red sweater" or “a yellow pickup truck”. Like other multimodal models, the free-text search model performs well with descriptors such as objects and colors, but is less suitable with counting ("three persons"), slang, or emotional cues ("angry-looking man"). The object description shouldn't be subjective, too vague, or include too specific details.

- Combine multiple object attributes using *and*: "person with red hat and backpack".
- Describe text, text logos, or brand names: "van with text Axis".
- Don't focus on describing environments. The processing is done on cropped images of objects, which means that the model might not see the objects' surroundings. Broad scene or environment descriptors (such as "city", "urban", "park", "garden", "lake", or "beach") might therefore not give good results.
- Experiment with alternate phrasing if you're not happy with a result.
- The free-text prompt supports only English.

### 4.2 Query moderation

We've implemented query moderation based on common practice to ensure effective filtering. The moderation model is a natural language processing model that checks the query to restrict offensive wordings. It checks whole text strings for harmful, inappropriate, or toxic content. Additionally, we've enhanced these capabilities with proprietary measures, including a custom list of prohibited search categories and words. When a query contains words or phrases from this list, we reject the search to maintain a safe search environment and ensure ethical results. You can provide anonymous user feedback to Axis if you disagree with a word being blocked or want to suggest blocking a word.

### 4.3 Logging of search queries

AXIS Camera Station Pro maintains an audit trail of user operations. Audit trails not only keep track of the specific operations and user identity, but also retain any data used in the operation. This means that all user searches, including search prompts, are logged. Administrators can use the logs to identify inappropriate search behavior among users, flag unethical search prompts, and take corrective action.

Note that video data is not shared with Axis. Your data remains on your server.

## 5 Text-image matching

The possibility to search video metadata using free-text queries significantly expands search capabilities from a predefined list of attributes to almost limitless search criteria. In AXIS Camera Station Pro this function is based on an open-source foundation model, trained on billions of image-text pairs and fine-tuned by Axis for surveillance use cases to improve performance.

### 5.1 Foundation model optimized for surveillance

The foundation model is a text-to-image model trained on large datasets of text-image pairs. It's a zero-shot model that matches text with relevant images. A zero-shot model is a type of artificial intelligence (AI) model that can recognize and classify objects or concepts without prior training data. In other words, the model can perform tasks without having seen any examples of the task before. This ability is crucial for making sure we provide optimal performance in matching natural language with images.

The model was trained on a large amount of text-image combinations and operates on a neural network of more than 2.5 billion parameters. At Axis, we've used our own unique training material to further tune this model, improving its ability to interpret images with typical surveillance camera views and objects. This means that we've optimized the model for surveillance use cases.

### 5.2 Feature vectors

When you make a free-text search, the foundation model creates a feature vector of the search query.

The foundation model also continuously produces descriptions of every object tracked in the scenes and creates feature vectors to represent them. Each object is represented by only one feature vector, which is stored in our database. This makes searching fast since feature vectors are already precomputed and readily available in the database.

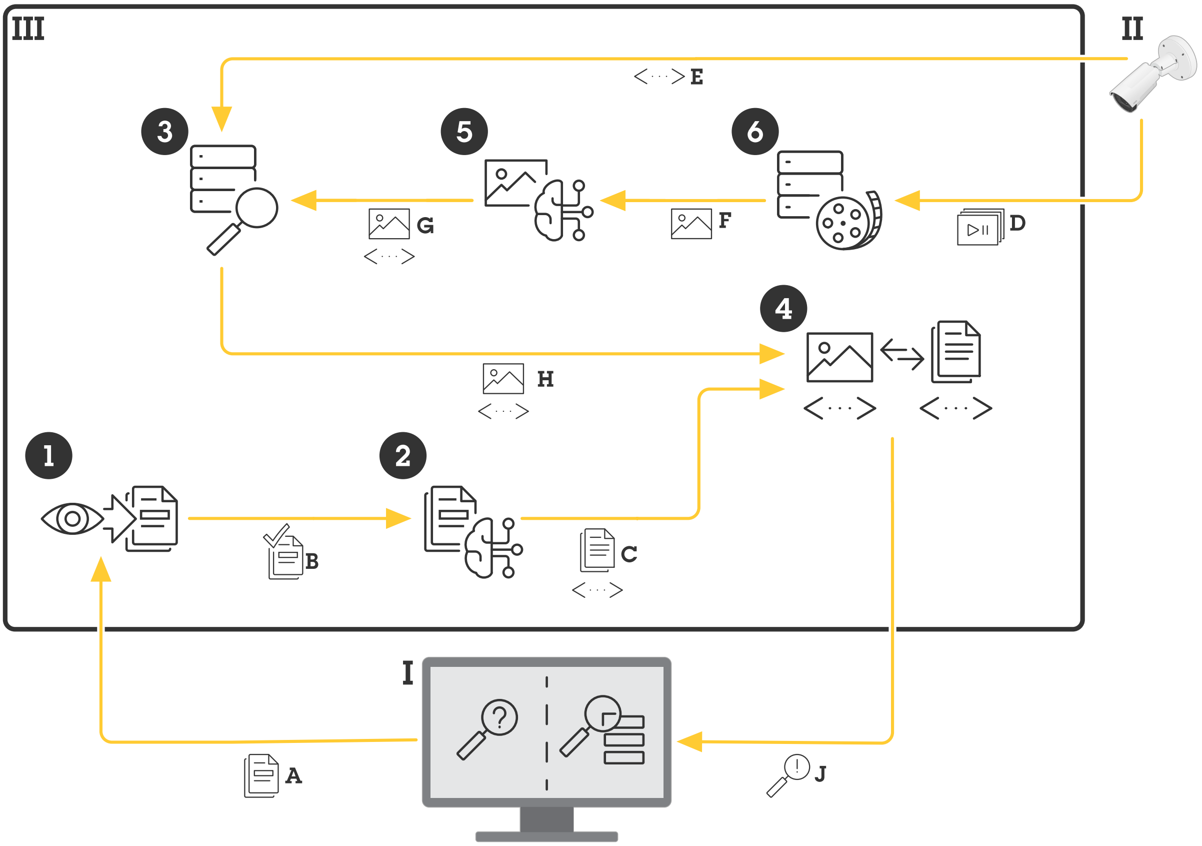
Both types of feature vectors are fed into the vector comparison engine to determine the similarity distance between your search query and all available feature vectors that represent detected objects.

A feature vector is a numerical representation of text or images. The feature vectors of persons or objects are thus only abstract representations of the person's or object's appearance. Feature vectors don't contain any human-interpretable information about features, such as hair or clothing color, that can be explicitly mapped to a specific person or used for identification. The feature vectors can only be used for comparisons with other feature vectors.

## 6 Process overview

The process overview diagram shows the main process steps, including which locations the steps take place in and which type of data each step produces.

Note that the upper loop in the diagram, including camera (II), recordings storage (6), foundation model (5), and search database (3) is a process that takes place continuously to create feature vectors of all detected moving objects, and not only when you make a search.



*Main locations (I–III) for the free-text search process*

- *I AXIS Camera Station Pro client*
- *II Camera(s)*
- *III AXIS Camera Station Pro server*

*Main process steps (1–6)*

- *1 Search query moderation*
- *2 Foundation model (text)*
- *3 Search database*
- *4 Vector comparison*
- *5 Foundation model (image)*
- *6 Recordings storage*

*Data type or outcome (A–J)*

- *A Text string*
- *B Text string*
- *C Feature vector (text)*
- *D Video*
- *E Metadata*
- *F Images*
- *G Feature vectors (image)*
- *H Feature vectors (image)*
- *J Results from search*

(I) **AXIS Camera Station Pro client**: Here you type your search query and receive sorted search results

(II) **Camera(s)**: Free-text search works on Axis cameras with AXIS OS 5.51 or later, but the better the camera, the better the results you get. Older devices provide less granular metadata based on motion detection only. Newer devices produce AXIS Scene Metadata, which includes object classification. The camera's moving object detection and tracking is used to find one representative image of each detected object, thus reducing the number of images to analyze on the server.

(III) **AXIS Camera Station Pro server**: Here, all metadata and video data from the cameras is processed and stored. Before you make a free-text search, the server must (for each detected moving object) decode the video and extract an image of the detected object. The foundation model then processes this image to create the feature vector. These operations are quite costly in terms of processing capacity so to improve performance the feature vectors are saved to a database enabling quick searching in the future. If your server has spare capacity we strongly recommend you to enable background processing of the video from your most important cameras because this will make searching significantly faster.

(1) **Search query moderation**: The moderation model checks the query to restrict offensive content.

(2) **Foundation model (text)**: The foundation model creates a numerical representation (feature vector) of the moderated search query text string.

(3) **Search database**: The search database holds complete metadata from AXIS Scene Metadata or metadata created by the server, including object classification data with attributes, time, position, and feature vectors.

(4) **Vector comparison**: The feature vector representation of the search query text string is compared with the feature vector representations of object images detected in video.

(5) **Foundation model (image):** The foundation model creates numerical representations (feature vectors) of each object track in the recorded video. This is a continuous process that takes place also when you're not searching.

(6) **Recordings storage**: This is where video from the camera is stored and where the foundation model gets its images.

# 7 Responsible use of AI

With free-text search, we use AI to increase the accuracy and efficiency of our forensic search solutions and ultimately enhance human decision-making.

Responsibility and accountability are fundamental to Axis AI approach. This involves ensuring that the AI systems we create adhere to ethical principles, comply with laws, and effectively manage risks. Axis provides tools that allow our customers to be confident in the integrity of their operations. The free-text search feature in AXIS Camera Station Pro includes a fine-tuned text classification model for text prompt moderation. We developed this model to moderate search queries in order to help you prevent unethical use.

Free-text search connects to Axis cloud services once a week to check whether the AI models require updating in order to comply with new regulations or requirements. If connection fails, free-text search operations will be unavailable until connection has been reestablished.

To further comply with legal and ethical standards in the application of AI, our products provide access controls based on user authentication credentials and access permissions. This allows our customers to enforce user compliance with operational policies.

T10218907/EN/M1.9/2501

# About Axis Communications

Axis enables a smarter and safer world by creating solutions for improving security and business performance. As a network technology company and industry leader, Axis offers solutions in video surveillance, access control, intercom, and audio systems. They are enhanced by intelligent analytics applications and supported by high-quality training.

Axis has around 4,000 dedicated employees in over 50 countries and collaborates with technology and system integration partners worldwide to deliver customer solutions. Axis was founded in 1984, and the headquarters are in Lund, Sweden

©2025 Axis Communications AB. AXIS COMMUNICATIONS, AXIS, ARTPEC and VAPIX are registered trademarks of Axis AB in various jurisdictions. All other trademarks are the property of their respective owners. We reserve the right to introduce modifications without notice.