

MAGNAS: A CCTV Searching Platform for Digital Investigation

Mahra Alameri, Maitha Alnoaman, Noora Alhashmi and Richard Ikuesan

Computing and Applied Intelligence, College of Technological Innovation, Zayed University, Abu Dhabi, UAE

Mahraalameri8@gmail.com

Maithaalnoaman@outlook.com

Nooraalhashmi74@gmail.com

Richard.ikuesan@zu.ac.ae

Abstract: Closed-circuit television (CCTV) systems generate huge amounts of video material, making manual inspection slow, inefficient, and prone to missing evidence. As surveillance scenarios grow more complex, investigators require automated tools capable of understanding extensive, unstructured recordings and supporting natural-language search, which most existing systems cannot provide. Current video-analysis systems usually focus on isolated tasks such as detection, action recognition, or captioning, producing fragmented outputs that lack the consistency and explainability needed for forensic work. This research presents MAGNAS, an intelligent CCTV retrieval platform that converts raw footage into a searchable metadata index using a multi-phase pipeline integrating person detection, multi-object tracking, visual attribute extraction, and action recognition. It generates structured representations including bounding boxes, timestamps, appearance attributes, and actions stored in an SQLite database. Users describe people or events via a natural-language interface, and the system translates these into structured filters for accurate set-theoretic retrieval. MAGNAS was evaluated using mAP for detection, Recall@K and Precision@K for retrieval, and temporal IoU for action alignment. Results show strong person-detection performance and high retrieval accuracy, especially for queries with multiple appearance attributes. Attribute extraction was largely precise, despite challenges with delayed VLM processing, inconsistent action detection under difficult camera angles, and limited temporal accuracy (tIoU = 0.2649). Overall, MAGNAS significantly reduces search time and improves identification of people of interest in large-scale CCTV archives, highlighting the value of organized, explainable video indexing for future investigative technologies and law enforcement applications.

Keywords: CCTV, Video retrieval, Natural language processing, Computer vision, Digital forensics

1. Introduction

The CCTV systems have evolved into one of the most prolifically used security technologies, and video surveillance has been implemented in airports, banks, retail spaces, and smart cities and brings about rapidly growing amounts of unstructured footage as the number of cameras in the world has exceeded a billion by 2021 (Subha et al., 2023). This information is useful to prevent crime, rebuild the events, and assist investigations, and human operators can only observe a few screens and their performance drops in case of fatigue, which is why manual inspection is inconsistent and inefficient as stated by Subha et al. (2023). Artificial intelligence and computer vision studies have thus prioritized the automation of video interpretation with models like YOLO to detect objects and CNN/RNN to recognize behaviors, and video-to-text systems based on natural language search, as Khurana and Deshpande (2021) explain. Initial techniques based on background subtraction, semantic tagging, and motion detection developed into semantic indexing and multimodal systems such as TagBook that showed that it was possible to unsupervisedly characterize events, as stated by Mazloom et al. (2016). Recent developments now involve VLAB (He et al., 2025) and PRVR (Dong et al., 2022), which have been trained to better the video-language association, but most of them are evaluated on controlled data sets and have issues related to low-resolution, occlusion, and low-light conditions found in real CCTV images, as pointed out by Wan et al. (2021). Although technologies have made progress, existing systems tend to focus on single functions like anomaly detection, captioning, or retrieval without providing solutions that can be used in the real-world surveillance setting that can be scalable, which is why improved unified solutions are required.

The fast growth of CCTV systems has led to huge amounts of unstructured video that need to be brought to bear on intelligent actionable information as highlighted by Subha et al. (2023) but current manual reviewing systems are not keeping up with the volume and complexity. Other forms of early indexing and tagging techniques enhanced organization but were limited to low-level features enabling low scalability like Saravanan (2016) and TagBook by Mazloom et al. (2016), multimodal hypergraph retrieval by Li et al. (2024), and hierarchical graph search models by Song et al. (2022) enhanced the semantic alignment of video and text. The minimal effect of NLP-based technologies such as video captioning by Wu et al. (2024), multi-sentence summarization by Masoomah Nabati and Alireza Behrad (2020), and Video-QA by Khurana and Deshpande (2021) have also helped reduce the semantic gap, although their results are still underwhelming on continuous CCTV footage, where the computational cost, data capacity, and inefficiency become critical issues. Although supervised contrastive

learning and multimodal captioning models have enhanced robustness and similar accuracy, according to Hirata et al. (2021) and Wu et al. (2024), real CCTV conditions of low light, blur, occlusion, and varying dynamics of a scene still remain a problem to model reliability, an issue brought up by Wan et al. (2021) and Li et al. (2024). The existing literature is currently divided with anomaly detecting, captioning and retrieving being studied separately without any combination of identification, indexing, and video-to-text translation into a single, scalable system that guarantees the timely and efficient CCTV processing, which is why the system that would help alleviate these limitations and provide the means of timely efficacious CCTV processing is necessary. Three major research questions will guide the study:

RQ1: What approaches can combine multiple pretrained models to generate descriptive representations of CCTV footage that support faster and more intuitive search?

RQ2: What methods enable a single, unified query to retrieve relevant moments from long CCTV recordings based on actions, objects, or events?

RQ3: How can CCTV footage be transformed into structured outputs such as timelines, labels, or event summaries that allow investigators to navigate videos more efficiently?

The primary aim of this project is to develop a unified CCTV retrieval system capable of single-query natural-language search, enabling investigators to rapidly locate individuals, actions, and events across extensive surveillance footage by converting visual content into structured metadata and reducing reliance on lengthy manual review. To achieve this aim, the system integrates multiple pretrained models that detect objects, actions, and relevant visual cues in CCTV video, transforms these outputs into searchable representations such as labels, timestamps, and contextual descriptions, and implements a complete end-to-end pipeline that unifies detection, indexing, and retrieval into one efficient platform. Overall, the objective is to enhance investigative workflows by shortening video-analysis time, improving the accuracy and explainability of retrieval results, and providing a scalable metadata-driven solution suitable for real-world forensic and security applications.

2. Problem Background

Early video indexing relied on low-level features with limited scalability (Saravanan, 2016). TagBook introduced unsupervised semantic characterization (Mazloom et al., 2016), while multimodal approaches like hypergraph retrieval (Li et al., 2024) and hierarchical graph models (Song et al., 2022) improved video-text alignment. Video captioning systems (Wu et al., 2024; Nabati & Behrad, 2020) and Video-QA frameworks (Khurana & Deshpande, 2021) reduced the semantic gap but underperform on continuous CCTV. Recent models like VLAB (He et al., 2025) and PRVR (Dong et al., 2022) enhance video-language pretraining, yet real-world conditions—occlusion, low light, blur—remain problematic (Wan et al., 2021). Supervised contrastive learning (Hirata et al., 2021) improves robustness, but no existing system fully integrates detection, tracking, attribute extraction, action recognition, and natural-language search into one scalable forensic platform suitable for real CCTV conditions.

3. Methodology

The three-phase approach is used in this project. At the initial stage, a set of pretrained models were experimented on sample CCTV video to find out which models could be reliably used to assist in detection, attribute extraction and action recognition in the environment of real surveillance. During the second phase, the chosen components have been integrated into an entire platform with user interface, Flask backend, and a five-step automated pipeline that takes uploaded videos and allows search through natural languages. During the last stage, the system was tested with the help of the new CCTV videos in order to measure the accuracy of retrieval, the speed of the response, and the overall stability of the system in the conditions of the real investigation.

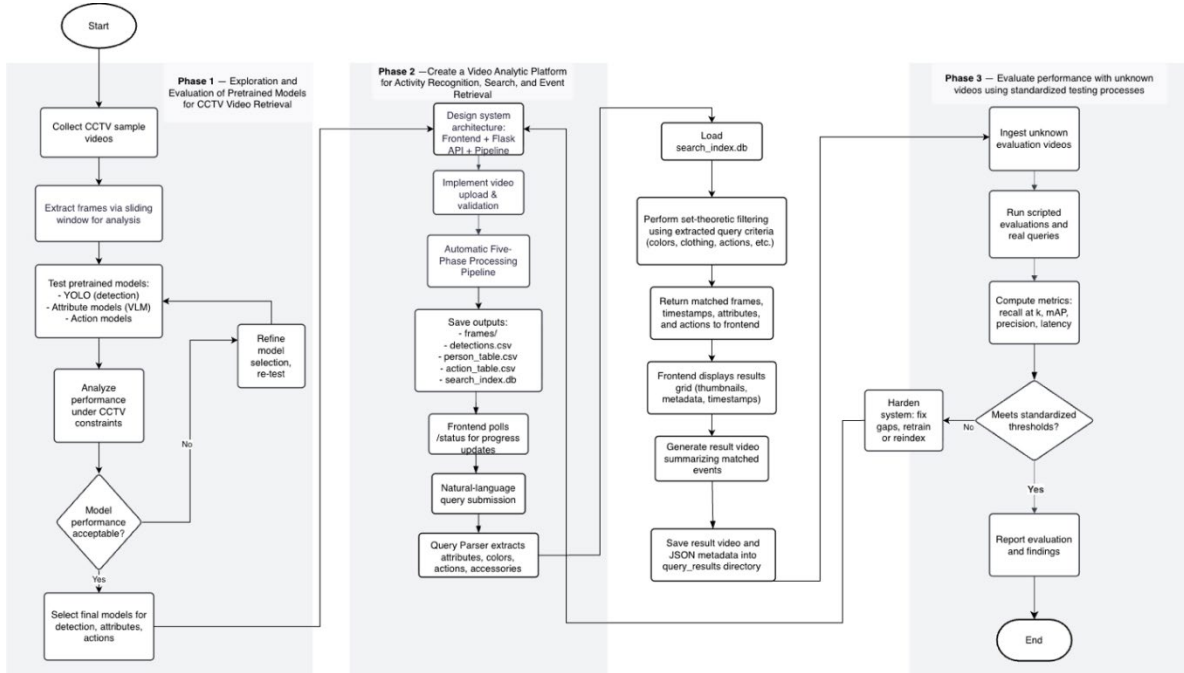


Figure 1: Operational Framework of the Study

3.1 Phase 1: Exploration and Evaluation of Pretrained Models for CCTV Video Retrieval

Phase 1 set MAGNAS baseline with a large variety of pretrained models, reviewed, and experimentally tested to be able to detect objects, extract attributes, recognize actions, track objects, and align texts with images. Due to the low resolution of CCTV footage, compressed artifacts, blur, poor illumination, occlusion, and low effective frame rate, this stage aimed at identifying models that would execute in the most favorable conditions when subjected to real surveillance. The analysis focused on the stability of detecting, the consistency of detecting over time, resistance to blur, fidelity of clothing colors and generalization to real recordings.

3.1.1 Review of pretrained models evaluated

Some models were experimented in terms of their functions in the analysis chain. In order to look at object-detection and tracking, YOLOv8 (m, l, x), YOLOv10, YOLO11, and RT-DETR were evaluated. YOLOv8m and YOLOv8x using ByteTrack were shown to be much more successful on degraded CCTV clips, particularly in low-light, compressed, and crowded scenes, and provides good small-person detection and dependable track IDs with minimal costly ReID modules.

Action recognition models such as SlowFast, TimeSformer, VideoMAE (base and large), MViTv2, Uniformer and InternVideo2 have also been tested. Most of the models performed poorly on noisy CCTV clips, and low FPS, though they were strong in Kinetics. VideoMAE-large was chosen due to its 16 frames input and powerful temporal representation and a custom HMDB51-finetuned checkpoint achieved much better surveillance-action performance than Kinetics-only pretrained models.

Person-attribute classifiers included SOLIDER, PromptPAR and OpenPAR, which were found to have an issue with poor crops, changes in lighting and compression noise. Then, vision-language models were tested based on zero-shot and instruction-tuning (CLIP, BLIP-2, LLaVA-1.5, Qwen-VL, Qwen2-VL, Qwen3-VL). Qwen3-VL-2B-Instruct that was deployed locally and quantized provided the highest accuracy, high-blur image, and structured outputs, particularly when the prompts contained context of the detected objects.

3.1.2 Experimental findings

Key insights were made in phase 1 experiments. Action models trained on Kinetics tended to get similar movements confused e.g. walking vs. running, when squeezed and at low frame rates. VideoMAE-large finetuning on HMDB51 enhanced significantly the recognition of surveillance-defensive actions, including walk, run, wave, pick, throw, punch, kick, smoke, drink, and clap. Dedicated attribute classifiers could not process blurred or low-resolution person crops and vision-language models, particularly Qwen3-VL-2B-Instruct with contextual hints, such as person holding a phone, were far more accurate on held objects, accessories, and

clothing. The results of YOLOv8 with Byte Track provided the most consistent person track across lighting variations and occlusion, and removed individual ReID models. Quantized Qwen3-VL-2B-Instruct deployed locally was able to operate offline with a tolerable latency of approximately 1.5 to 2 seconds per person crop and was therefore feasible in practice. These outcomes noted the most consistent factors during the noise of CCTV and informed the final MAGNAS architecture.

3.1.3 Final adopted model and pipeline

Phase 1 was successful as it ended with a strong and an effective model setup. YOLOv8m and YOLOv8x were used in object detection with low-resolution, compressed, low-light CCTV and in tracking with the assistance of Byte Track since they possess high ID continuity without ReID. Qwen3-VL-2B-Instruct utilized the attributes of people based on the context of the items that were detected by YOLO to enhance recognition of objects and clothes. VideoMAE-large finetuned on HMDB51 provided action classification achieving 83% validation accuracy on surveillance-style actions. This combination formed the foundation for Phase 2 integration.

3.2 Phase 2: System Implementation and Integration

3.2.1 Operational architecture and ingestion workflow

MAGNAS is based on the modular system with interface, backend, pipeline, and storage to transform raw CCTV video into searchable metadata. The MagnasFinal.html interface allows the investigator to upload the footage, monitor the progress, and search by natural language, and the Flask backend allows managing authentication, uploads, pipeline execution, and logging. Each video is processed by a five-stage background pipeline, the first stage of which is to extract frame, followed by detecting and tracking people with the help of YOLO, attribute generations with a vision-language model, short term action recognition and, finally, all data are summarized into a SQLite search index. All the intermediate and final output are stored in the table directory so that they can be easily accessed.

3.2.2 Query processing, search logic, and retrieval mechanism

After the ingestion pipeline completes and search_index.db is created, the natural-language queries can then be typed in by the investigators through the interface. Once queries are typed in A query such as: a man in a dark green hoodie with a backpack walking will be logged in the magnas_activity.log and the query parser used to process it in the system. The parser acquires formal search attributes such as colors, types, accessories, gender and actions of clothes. MAGNAS retrieves the results according to set-theoretic metadata filtering of metadata in search_index.db. The backend loads detections.csv, person_table.csv and action_table.csv, and finds all the entries that meet all the requirements and counts their intersection. This provides results to the investigation that are deterministic, explainable and reliable. Organized information that is sent by the backend includes frame numbers, timestamps, bounding boxes, attributes of persons, and inferred actions. Each individual can view the findings in frame thumbnails on the front-end. Bounding boxes on an MP4 video are created on command by generating a summarized video file using create_result_video.py and stored in the query results directory.

3.3 Phase 3: Evaluate the Performance of the Platform Using Unknown Videos Based on Standardized Testing Processes

The main aim of Phase 3 is to thoroughly test the platform's reliability, scalability, and accuracy by using originally unseen CCTV video. This guarantees that the created platform applies beyond the training environment and works well in real-world surveillance scenarios. The study uses established evaluation practices to assess algorithmic correctness, system responsiveness, and user experience while maintaining forensic transparency and security.

3.3.1 Evaluation metrics

On the platform, standardized retrieval metrics are computed after testing. Recall@K and Precision@K are used to determine the performance of the top-K relevant clips (Radlinski and Craswell, 2010). Mean Average Precision (MAP) measures the ranking quality across queries (Bellogin et al., 2017). Temporal accuracy of localized events is assessed by temporal IoU and mean Temporal Average Precision (mTAP) (Gao et al., 2017). Query latency (P50/P95) and indexing throughput measure responsiveness and scalability (Mackenzie et al., 2017).

4. Results and Analysis

In this section, MAGNAS performance will be of all components including interface, processing, retrieval accuracy, temporal localization, and overall functionality, in which quantitative and qualitative results are presented.

4.1 Ingestion and Indexing Results

The ingestion pipeline consists of five primary steps, each of which is designed to enable high-quality information extraction and reliable indexing. The system extracts timestamped frames at 3 frames per second (adjustable), balancing temporal precision with storage and processing performance. YOLOv8 is put to use for accurate person recognition, followed by Byte Track for reliable identity tracking. For each identified individual, the system saves a narrow crop for attribute and action analysis and a context crop with 150 px padding, which proved useful for catching surrounding objects and hand-held items that frequently fall outside tight bounding boxes. Each crop is processed using a locally hosted Qwen3-VL-2B model, which extracts precise human characteristics. VideoMAE-Large is implemented with movable temporal windows and post-processing smoothing achieving 83% accuracy. All Metadata has been compiled and saved in search_index.db for faster retrieval using combination indexing (frame, track_id) and better attribute-action fusion.

4.2 Quantitative Evaluation

4.2.1 Retrieval accuracy metrics (Recall@K, Precision@K, mAP)

The retrieval-stage evaluation demonstrates that the system achieves strong ranking and retrieval quality across diverse query types. The overall Mean Average Precision (mAP) reached 0.9597, indicating excellent ranking consistency and the system's ability to place relevant clips at higher positions. Recall@K: Recall@1 = 0.4450 (44.5%), Recall@5 = 0.9800 (98.0%), Recall@10 = 1.0000 (100%), Recall@20 = 1.0000 (100%), Average Recall = 0.8563 (85.63%). Figure 2 shows Recall@K results showing the system's ability to retrieve relevant clips within the top-K returned items. High Recall@5/10/20 demonstrates strong retrieval coverage.

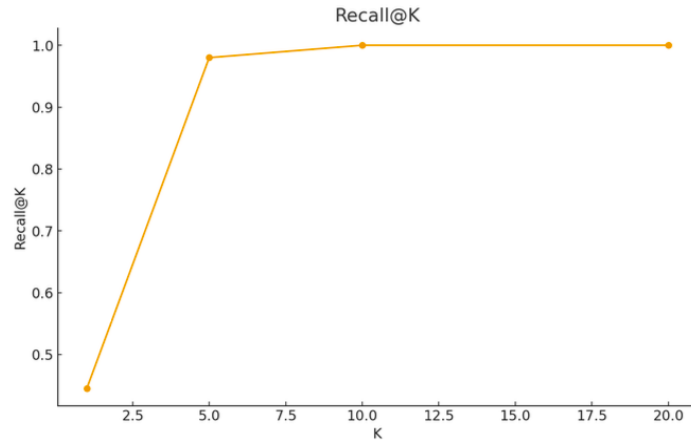


Figure 2: Recall@K results showing strong retrieval coverage

Precision@K: Precision@1 = 0.9000 (90%), Precision@5 = 0.5200 (52%), Precision@10 = 0.2700 (27%), Precision@20 = 0.1350 (13.5%), Average Precision = 0.4563 (45.63%). Precision@1 remains high, gradually decreasing as K increases. Per-Query AP Performance: Hijab/headscarf: AP = 1.0, Walking/stairs: AP = 1.0, Shaking hands: AP = 1.0, Green jacket: AP = 1.0, Group standing/talking: AP = 0.8875, Baseball swing: AP = 1.0, Riding horse: AP = 1.0, Helmet/hat: AP = 0.71, Jumping/diving: AP = 1.0, Wearing glasses: AP = 1.0. Overall, the system exhibits excellent retrieval accuracy. High mAP and Recall@5/10/20 values show that relevant clips are consistently retrieved. Precision decreases at larger K values, as expected in ranked retrieval systems.

4.2.2 Temporal localization (Temporal IoU and mTAP)

Temporal localization evaluates how accurately the system identifies the start and end boundaries of actions. Overall performance shows partial overlap with ground truth. Overall Metrics: Average Temporal IoU = 0.2649, Overall mTAP = 0.2391. mTAP Across IoU Thresholds: mTAP@0.3 = 0.2859, mTAP@0.5 = 0.2859, mTAP@0.7 = 0.1455.

Per-Video Performance: 4PEOPLE: Avg tIoU = 0.2327, mTAP = 0.1570; Riding: Avg tIoU = 1.0000, mTAP = 1.0000; Baseball: Avg tIoU = 0.3185, mTAP = 1.0000; Green Jacket: Avg tIoU = 0.2857, mTAP = 0.4545; Hijab: Avg tIoU = 0.3183, mTAP = 0.3788; Running: Avg tIoU = 0.1182, mTAP = 0.2020; Shakehand: Avg tIoU = 0.2926, mTAP = 0.3636. Figure 3 shows Average Temporal IoU (tIoU) per video. Simple, continuous actions such as Riding show perfect boundary alignment, while multi-person or fast-motion actions result in lower tIoU.

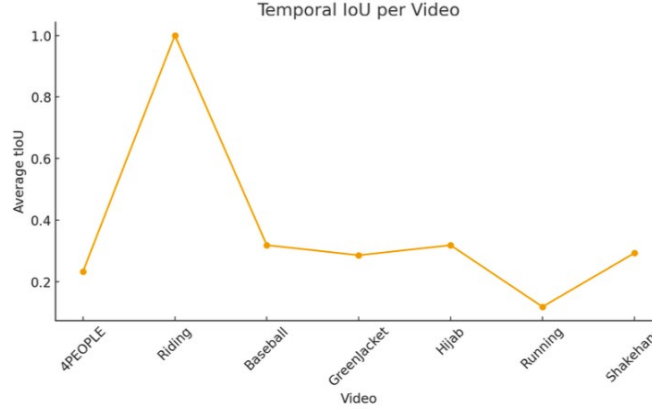


Figure 3: Average Temporal IoU per video

The system performs well on simple, continuous actions such as riding and baseball. Performance decreases in multi-person scenes, fast transitions, subtle actions, and scenes with occlusion.

4.2.3 System performance metrics

System performance was evaluated across all seven videos, each processed through the full pipeline. Overall Throughput: 100% pipeline completion, Total processing time: 659.25 seconds, Average processing time per video: 94.18 seconds, Fastest video: 39.79 seconds; slowest: 233.73 seconds, Average speed: 5.21 FPS (0.20× real-time). Hardware Utilization: GPU: NVIDIA GeForce 4090, Average GPU memory: 5.45 GB, Peak GPU memory: 14.6 GB, RAM usage: ~73 MB (GPU-bound workload). Phase-Level Breakdown: Phase 3 (VLM Attributes): Slowest step—avg 56.76 s, peak 174.43 s; Phase 4 (Action Inference): Avg 26.47 s; Phase 2 (YOLO Detection): Avg ~7 s; Phases 1 & 5 (Extraction and Indexing): Fastest—avg ~3 s each. The pipeline is stable and predictable, with no failures. The system is slower than real-time due to the combination of YOLO, VLM, and VideoMAE models, but memory usage remains stable and efficient.

4.3 Qualitative Evaluation

The qualitative results show significant improvements in retrieval accuracy, user experience, and platform stability. The system now correctly returns the subject's continuous track, even with complicated multi-attribute descriptions. Specific instances include: "man in black jacket walking," "woman in pink top with backpack," "person riding bike carrying bag." The incorporation of context crops eliminated previous failures when important things, such as cups, phones, or bags, went outside the tight crop. This guarantees that all necessary attributes and moved items have been recorded.

The natural language query engine also worked effectively. Attribute and action names were accurately translated to metadata fields, including: "riding" → mapped to action ride_bike; "girl", "woman", "female" → reliably classified as female; "cap", "hat", "beanie" → matched via the accessories field; "carrying cup", "holding mug", "with bottle" → identified via objects_held.

All major issues documented in earlier versions of the system were fully resolved: No more incorrect person retrieval due to the new (frame, track_id) intersection method; No more corrupted CSV files after implementing strict response cleaning for VLM outputs; Gender detection no longer defaults to male owing to improved attribute parsing; Hand-held object detection is now consistently captured through context crops; Action verb mismatches were eliminated using refined language-to-action mapping rules. These enhancements increase system stability, confirm module robustness, and prepare the platform for real-world forensic deployments.

This section demonstrated that MAGNAS is a reliable complete CCTV search platform, completing all processing steps and generating consistent retrieval results. The results revealed great retrieval accuracy, consistent person-level tracking, and predictable system performance. While temporal localization and attribute extraction

remain difficult in complicated scenarios, the overall assessment shows that the system is effective, useful, and in line with the project's goals.

5. Discussion

The results of this research created MAGNAS, an extensive surveillance-video retrieval system that includes frame extraction, detection, attribute identification, action categorization, and natural-language search. The results reveal that the system performs well in retrieval accuracy, with a high mAP and outstanding Recall@K, making it suitable for person-based forensic searches. The system's interface and logging features also ensured that pipeline execution was reliable, information was generated consistently, and it was useful. Temporal localization was worse, although it performed well for obvious, long-duration activities.

Future work can improve MAGNAS' accuracy, scalability, and real-world application. One significant objective is to improve identity consistency by incorporating a dedicated person Re-Identification (ReID) 51model, which would decrease ID switches and allow the system to track persons more reliably over longer sequences. Furthermore, more complex temporal action models, such as VideoMAEv2, SlowFast, or TimeSformer, might be used to give clearer action boundaries and improved identification of subtle or slow-movement behaviors. Extending the dataset is also important. Evaluating the system on large-scale, diversified CCTV material containing nighttime recordings, busy locations, occlusions, low-resolution cameras, and different views would considerably improve generalization. Synthetic data augmentation and controlled data generation could help expose the system to rare or difficult conditions. The platform's forensic use might potentially be increased by incorporating new analytical modules such as posture estimation, vehicle identification, interaction analysis (e.g., "person passing an object"), or anomaly detection. More intuitive user features, such as visual timelines, heatmaps of identified activities, and multi-query filtering, would make the platform more user-friendly for investigators. Finally, performance optimization by model reduction, batching, or TensorRT acceleration may minimize processing time, bringing it closer to real-time operation. Distributed or cloud-based deployments may also enable the system to manage higher amounts of video data.

6. Future Work and Conclusion

6.1 Future Works

The accuracy, scalability, and applicability of MAGNAS can be improved in the future. The ID switches and long sequence tracking could be better by adding a dedicated person Re-Identification model. More sophisticated temporal action models, including VideoMAEv2, SlowFast, or TimeSformer, might offer better action boundaries and identify smaller-scale actions. Generalization would be enhanced by extending the evaluation datasets with various CCTV footage at night, in overcrowded areas, with occlusions, low resolution, and different viewpoints. Synthetic data augmentation may expose the system to uncommon conditions. Additional modules for forensic use can include posture estimation, vehicle identification, interaction analysis, or anomaly detection. Visual timelines, multi-query filtering, and activity heatmaps can improve user experience. Optimization through model reduction, batching, or TensorRT acceleration may reduce processing time toward real-time operation. Larger video volumes could be managed via distributed or cloud deployment.

6.2 Conclusion

MAGNAS has successfully delivered a functioning, end-to-end CCTV search engine capable of quick and accurate retrieval across a wide variety of queries. While there are still limits in temporal accuracy and in demanding visual situations, the system shows great promise for practical forensic applications with continued development and broader dataset review. MAGNAS has the potential to become a strong real-world investigative tool.

Ethics Declaration: This study did not require ethical approval because no human or animal subjects were involved. The authors affirm that no human research participants were involved in this study; hence, it is not applicable.

AI Declaration: The Authors declare that AI was not used to write the manuscript and that the content represents the authors' original writing, except where an AI tool was used to edit the manuscript to enhance clarity.

References

- Bellogín, A., Castells, P., & Cantador, I. (2017). Statistical biases in Information Retrieval metrics for recommender systems. *Information Retrieval Journal*, 20(6), 606–634.

- Dong, J., Chen, X., Zhang, M., Yang, X., Chen, S., Li, X., & Wang, X. (2022). Partially Relevant Video Retrieval. *Proceedings of the 30th ACM International Conference on Multimedia*, 246–257.
- Gao, J., Yang, Z., Sun, C., Chen, K., & Ram Nevatia. (2017). TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals. *ArXiv (Cornell University)*.
- He, X., Chen, S., Ma, F., Huang, Z., Jin, X., Liu, Z., Fu, D., Yang, Y., Liu, J., & Feng, J. (2025). VLAB: Enhancing Video Language Pretraining by Feature Adapting and Blending. *IEEE Transactions on Multimedia*, 27, 2168–2180.
- Hirata, T., Mukuta, Y., & Harada, T. (2021). Making Video Recognition Models Robust to Common Corruptions With Supervised Contrastive Learning. *ACM Multimedia Asia*, 1–6.
- Khurana, K., & Deshpande, U. (2021). Video Question-Answering Techniques, Benchmark Datasets and Evaluation Metrics Leveraging Video Captioning: A Comprehensive Survey. *IEEE Access*, 1–1.
- Li, Q., Su, L., Zhao, J., Xia, L., Cai, H., Cheng, S., Tang, H., Wang, J., & Yin, D. (2024). Text-Video Retrieval via Multi-Modal Hypergraph Networks. 369–377.
- Mackenzie, J. M., Culpepper, J. S., Blanco, R., Crane, M., Charles, & Lin, J. (2017). Efficient and Effective Tail Latency Minimization in Multi-Stage Retrieval Systems. *ArXiv (Cornell University)*.
- Nabati, M. and Behrad, A. (2020) ‘Multi-sentence video captioning using content-oriented beam searching and multi-stage refining algorithm’, *Information Processing & Management*, 57(6), p. 102302. doi:10.1016/j.ipm.2020.102302.
- Mazloom, M., Li, X. and Snoek, C.G.M. (2016) ‘TagBook: A semantic video representation without supervision for event detection’, *IEEE Transactions on Multimedia*, 18(7), pp. 1378–1388. doi:10.1109/TMM.2016.2555454.
- Radlinski, F., & Craswell, N. (2010). Comparing the sensitivity of information retrieval metrics. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 667–674.
- Saravanan, D. (2016). Segment Based Indexing Technique for Video Data File. *Procedia Computer Science*, 87, 12–17.
- Song, Y., Gao, J., Yang, X., & Xu, C. (2022). Learning Hierarchical Video Graph Networks for One-Stop Video Delivery. *ACM Transactions on Multimedia Computing Communications and Applications*, 18(1), 1–23.
- Subha, P., Brindha Devi, V., Prabavathi, R., Parvathi, N., Soundarya, S. and Varshini Priyadharshini, S. (2023) ‘Criminal activity detection and prevention using CCTV surveillance’.
- Wan, S., Xu, X., Wang, T., & Gu, Z. (2021). An Intelligent Video Analysis Method for Abnormal Event Detection in Intelligent Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems*, 22(7), 4487–4495.
- Wu, S., Gao, Y., Yang, W., Li, H., & Zhu, G. (2024). End-to-End Video Captioning Based on Multiview Semantic Alignment for Human–Machine Fusion. *IEEE Transactions on Automation Science and Engineering*, 1–9.