

Object-based image retrieval and detection for surveillance video

Swati Jagtap, Nilkanth B. Chopade

Department of Electronics and Communication Engineering, Pimpri Chinchwad College of Engineering, Pune, India

Article Info

Article history:

Received Oct 16, 2023

Revised Mar 18, 2024

Accepted Apr 30, 2024

Keywords:

Graphical user interfaces

Image retrieval

Object detection

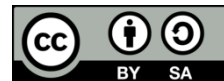
Query processing

Video surveillance

ABSTRACT

With technological advancement worldwide, the video surveillance market is growing drastically in a versatile field. Monitoring, browsing, and retrieving a specific object in a long video becomes difficult due to the enormous amount of data produced by the surveillance camera. With limitations on human resources and browsing time, there is a need for a new video analytics model to handle more complex tasks, such as object detection and query retrieval. The current approach involves techniques like unsupervised segmentation, multiscale segmentation, and feature-based descriptions. However, these methods often encounter extensive space and time consumption challenges. A solution has been developed for retrieving targeted objects from surveillance videos via user queries, employing a graphical interface for input. Extracting relevant frames based on user-entered text queries is enabled through using YOLOv8 for object detection. Users interact through a graphical user interface deployed on a Jetson Xavier Development board. The system's outcome is a time-efficient and highly accurate automated model for object detection and query retrieval, eliminating human errors associated with manually locating objects in videos upon user queries.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Swati Jagtap

Department of Electronics and Telecommunication Engineering, Pimpri Chinchwad College of Engineering
Sector 26, Pradhikaran, Nigdi, Pune-411044, India

Email: swatialjagtap@gmail.com

1. INTRODUCTION

The application of video analytics in surveillance is widespread, impacting domains like Security, Healthcare, smart cities, transportation, defense, and more in this technological era [1]. The generation of substantial data results from Surveillance cameras continuously capturing footage round the clock, making videos from these domains enormous. Efficiently managing and storing this data is vital for effective surveillance systems. Analyzing those manually would prove to be arduous due to the inherent limitations in accuracy and efficiency, along with the ample time required.

Nowadays, browsing, retrieving, and identifying any suspicious or abnormal activities in videos is labor-intensive and time-consuming. It heavily relies on dedicated human resources. However, human attention tends to decrease as time passes, increasing the likelihood of crucial incidents in the footage being overlooked. To address this challenge, it is essential to establish a framework that enables the extraction of objects or activities based on the user-described query form.

Information retrieval is efficient searching, locating, and extracting the required information from a given document, image, or video. In this process, the user enters the query into the system. Queries can be entered as statements, words, images, audio, and a video clip. In the retrieval process, the query may match the different objects with different degrees of relevancy in the given data. User queries are matched against

the database information, and the results are ranked as per the degree of matching. This ranking provides the difference between query searched and retrieved data. The topmost matched result can be shown to the user [2]. Retrieval approaches are classified based on what type of data must be extracted from the given source. The retrieval can be in the form of images, video clips, text, or region-based data. Information retrieval works based on the probabilistic model, knowledge-based model, and artificial intelligence (AI) techniques. Some researchers also use evolutionary algorithms like genetic algorithms to improve retrieval precision and recall [3]. Retrieval is also used in real-time content-based recommendation systems, utilizing deep learning methods like deep neural network (DNN) and recurrent neural network (RNN). This approach has effectively combatted data sparsity and alleviated the cold start problem within collaborative filtering [4]. Some approaches also work with analyzing the query for better understanding and getting better precision in information retrieval. The queries are in the form of speech, text, image, or some features. Query processing involves four main categories: query expansion, query optimization, query classification, and query parsing [5]. This research primarily focuses on image retrieval systems, utilizing user text queries and involving query parsing, optimization, classification, and expansion.

The system employs object detection algorithms and deep learning architecture to locate specific objects within video clips swiftly, enhancing browsing efficiency via a user-friendly interface that accommodates various query inputs such as text, images, or phrases for effective video summarization. The proposed work aims to enhance efficiency and reduce costs across various sectors, including healthcare, smart cities/transportation, retail, and security. This system can assist in patient monitoring and reduce the wait times in healthcare. Smart cities/transportation can manage traffic congestion, detect risky situations on roads, and provide data for accident analysis. For retail, it can track customer behavior, optimize product placement, and improve sales strategies. In security, the technology enables real-time recognition of people and vehicles, enhancing surveillance and crowd control in places like malls, hospitals, stadiums, and airports [6]–[8].

Contributions of this paper include: i) This paper addresses the method for retrieving the required frame from the surveillance video. It helps to reduce the searching and retrieval time of particular activities or frames from the given long video; ii) A graphical user interface is implemented to provide easy and user-friendly access to the proposed framework; iii) The object detection algorithms are used to identify the objects in each frame. The identified objects are then indexed and stored in a file frame-wise with relevant boundary coordinates containing x , y , w , and h , streamlining the computational process when comparing the desired text query with the stored objects. The string matching compares the user query with the stored label; and iv) The GPU's Intel i5, NVIDIA Jetson Nano, and NVIDIA Xavier boards are used to implement the retrieval algorithm to evaluate the retrieval time.

The paper is organized as follows: section 2 represents our research's related work, where 2.1 represents the survey on object detection, and 2.2 gives the image retrieval approaches. Section 3 describes the proposed method. Section 4 explains the results and experiments obtained. This 4.1 highlights the results for object detection, 4.2 explains the results for the retrieval framework, and 4.3 compares the results of GPU implementation. Finally, section 5 concludes the paper.

2. RELATED WORK

Advancements in computer vision, particularly through deep learning and neural networks, have greatly enhanced object detection and content-based image retrieval (CBIR). Object detection techniques like Faster R-CNN, YOLO, and SSD utilize convolutional neural networks for accurate and efficient real-time object detection in various applications. Meanwhile, CBIR systems leverage deep learning to understand the visual content of images, enabling quick and precise retrieval based on similarity metrics and indexing. These advancements have revolutionized industries, facilitating tasks such as visual search in e-commerce, medical image analysis, and video surveillance. Some of the approaches for object detection are listed under 2.1, and the techniques used to retrieve the object from video are listed under 2.2.

2.1. Object detection

Object detection is a computer vision framework that identifies and localizes the presence of an object in a frame. The localization is represented by the four coordinates, namely, the x coordinate, the y coordinate, the width, and the height, and these coordinates are called boundary boxes. During the era when image processing methods were not efficient, a majority of the initial object detection algorithms were developed relying on manually crafted features such as Viola-Jones detectors [9], histogram of oriented gradient [10], and deformable part models [11]. With the exponential technological advancement, the deep learning era has outperformed the hand-crafted approaches. Object detection in deep learning is categorized into two main genres: Two-stage detector and one-stage detector [12]. Two-stage detectors work based on

region proposal and use two models, one for feature extraction and object region extraction while the other for classification and fine localization of objects. These detectors are relatively slow but give good localization and recognition accuracy. Some of the two-stage algorithms are fast-region-based convolutional neural networks, feature pyramid networks, region-based fully convolutional networks, and cascaded region-based convolutional neural networks [13].

Single-stage detectors like YOLO streamline object detection by combining localization and classification into a single model, sacrificing some accuracy for faster computation. YOLO's innovative use of anchor boxes improves object localization, allowing it to handle objects of various sizes and shapes efficiently. Its iterative versions continually refine the algorithm, addressing previous limitations and introducing architectural enhancements for better performance [14]. single shot detector (SSD) is a single-stage detector known for effectively detecting small objects, utilizing multi-scale feature maps and a feature pyramid network. By incorporating various resolutions of feature maps, SSD efficiently handles objects of different sizes, improving accuracy in detecting large and small objects [15].

Figure 1 gives the classification of deep learning object detection models. Some of the one-stage and two-stage models are listed. Single-stage detectors are used for faster prediction of images for larger objects, while two-stage detectors can be used for good accuracy. The YOLOv8 model is used for object detection as it gives good accuracy and speed.

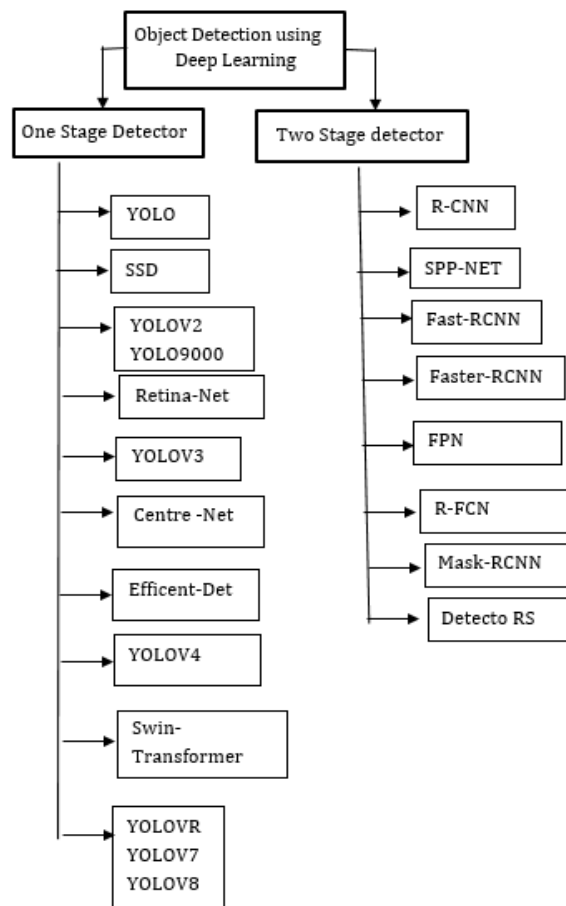


Figure 1. Deep learning models of object detection

2.2. Image retrieval

Image retrieval is a methodology employed to obtain images based on user preferences and queries from diverse sources. This process encompasses a variety of techniques. In certain methods, user queries are expressed in textual form, while others rely on extracting low-level features. Some approaches incorporate data fusion and machine learning, whereas alternative algorithms adopt semantic approaches and relevance feedback techniques [16]. Image retrieval is salvaging the relevant frames or images from the given source as per the user query. In the proposed system, we are retrieving the frames from the video as per the user query.

It is used in many applications such as medical image retrieval, fashion and design, security and surveillance, E-commerce, Elmogy *et al.* [17] suggested the three approaches to retrieve the image from the source. In one approach, image retrieval is implemented using text as an input query called text-based image retrieval (TBIR). A textual query is compared with annotated textual images in the database. In other approaches, the image is used as a user query. In this, the query image feature is extracted and compared with the features of images in the database. This method is referred to as content-based image retrieval (CBIR). In another approach, a semantic or low-level feature containing more meaningful information is extracted from the user query and compared with the image features stored in the database. This approach is called Semantic-based image retrieval.

A pair of novel techniques is introduced to address the challenges posed by imprecise indexing. These techniques pertain to both object representation and object matching. The process enhances object and event retrieval through indexing and retrieval phases. During indexing, video analysis outputs undergo feature extraction and data indexing, while in retrieval, user queries are parsed and matched with indexed data, with outcomes systematically ranked for users [18].

This research presents a video retrieval system using deep learning for person segmentation and attribute feature representation, facilitating person retrieval and summarization across surveillance cameras. It incorporates three key components: pixel-wise person segmentation, appearance-based multi-CNN, and video summarization, extracting attributes from masked images to address challenges like background clutter and appearance variations across cameras, streamlining the identification and condensation of key individuals within video sequences [19]. The study [20] helps people to retrieve their lost items from previously stored videos by giving an image as a query representing an object of interest. The system provides relevant video shots containing the object of interest as an output. The author uses the bag-of-words model in image retrieval to extract the rank list of visual instances. The experimental findings demonstrate the accuracy of identifying objects of interest solely by examining the top 10 video shots extracted from recorded video ranges from 50% to 80% across all 30 categories.

The paper presents region-based and Earth mover's distance (EMD) blob-based matching techniques for frame retrieval in surveillance videos, catering to query-by-region and query-by-example. The EMD blob-based method utilizes a combined similarity measure from dominant color, covariance matrix, and edge direction histogram. It enhances accuracy by up to 27.8% through approximate k-means clustering for computational efficiency and diverse query handling. This approach significantly improves processing time and accuracy according to experimental results [21].

The research presents a video surveillance indexing and retrieval framework comprising preprocessing, combining video indexing and compression, followed by query processing and retrieval. It supports various query types such as query-by-example, query-by-text, and query-by-region, facilitating efficient browsing and retrieval of video content [22]. The research presents a content-based image retrieval system utilizing multiscale segmentation for object extraction. Novel color and texture descriptors enhance efficiency. Results indicate superiority over traditional global histogram approaches [23].

2.3. Query-based image retrieval

Several studies have proposed video and image retrieval methods, each addressing specific objectives and employing distinct methodologies. For instance, Do-Tran *et al.* [24] focus on retrieving videos based on user-entered content, utilizing the TransNet model for semantic feature conversion and the CLIP model for comparison with user input. Che *et al.* [25] introduce a copyright detection system using CNN features and scalar quantization, incorporating a D-CNN for feature extraction and Euclidean distances for video identification. Wan *et al.* [26] develop an algorithm for event retrieval in lengthy videos through superframe segmentation, enhancing efficiency and accuracy. Guo *et al.* [27] expedite object retrieval via locality-sensitive hashing and binary code encoding of deep features, evaluating performance against other visual features. Mishra *et al.* [28] propose an image retrieval method based on object detection for separate object retrieval, improving accuracy and speed. Saikia *et al.* [29] utilize neural codes from the Faster R-CNN network for object retrieval, comparing results with confidence scores.

3. METHOD

The proposed system is used to retrieve the query given by the user from the video database with the help of OpenCV by training a model to analyze the video for detection. Three object detection models, MobileNet-SSD, YOLOv3 [30] and YOLOv8, are compared in terms of precision and recall. YOLOv8 gives better results compared to the other two. The proposed system implements object detection using YOLOv8, Keras preprocessing, Gaussian analysis, real-time detection, model training, and video analytics.

YOLOv8 is a state of art computer vision framework. This advanced model, YOLOv8, has the seamless capability for object detection tasks. It is a quicker and more precise model. It uses a new head with anchor-free and new loss features. Additionally, YOLOv8 is very effective and adaptable, supporting a wide range of output formats and running on both CPUs and GPUs [31].

Figure 2 gives the research flow of the proposed system. YOLOV8, an object detection algorithm, is utilized to detect objects in each video frame, with details indexed in an Excel sheet for efficient retrieval. This sheet contains information on object class types and boundary boxes. A graphical user interface enables users to upload videos and enter text queries for object retrieval, displaying extracted frames accordingly. The system's retrieval accuracy is evaluated across three hardware configurations to assess performance. The hardware circuit contains a Jetson development board for computation and processing. It is powered by a power supply according to its requirements. A camera is required for a high-definition video. The user can access the model using a graphical user interface (GUI) through the PC. The user uploads a video that is received from the camera and enters the query. The Jetson board acts as a processor, and the entire object detection and query retrieval model is implemented on that. It takes the input query, performs object detection, and then compares the query with the resultant frames. It returns to the frames containing the query object back to the user.

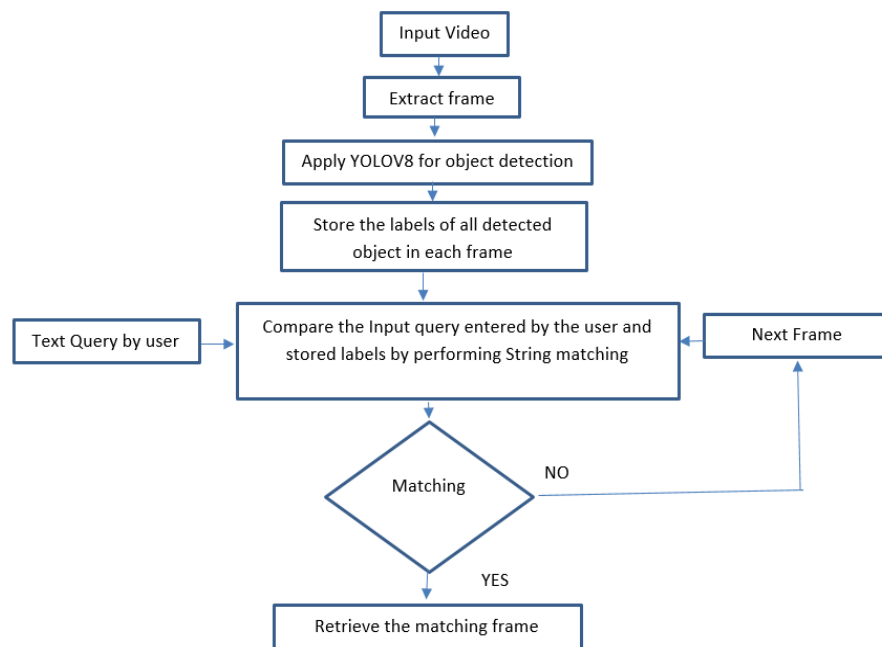


Figure 2. Research flow

4. EXPERIMENT AND RESULT ANALYSIS

The experiments performed on the video to evaluate the model are taken from Pixels, YouTube Bounding boxes [32], objections [33] dataset. Video1 is the footage of the footpath, Video2 is the traffic survey, and video3, video4, and video5 are the self-created videos generated with different objects to check the retrieval and detection of different objects as listed in the COCO dataset. Firstly, the object detection algorithm is applied to the video, and the respective frame object class ID is stored in an Excel sheet. For retrieval, a simple text query is used as the object's name is entered in the GUI, and respective frames consisting of the objects are displayed. The system is implemented on the NVIDIA Jetson Nano board, NVIDIA Xavier Board, and Intel i5. The performance is evaluated by comparing the retrieval time of all processors.

4.1. Object detection results

Evaluation parameters such as precision, recall, F1 score, and accuracy gauge object detection performance. Precision measures positive prediction values, recall identifies positive class objects, and F1 score harmonizes precision and recall. YOLO V4, SSD300, and YOLOv8 are compared, with YOLOv8 demonstrating superior scores. Detected objects, including person, buses, cars, traffic lights, and trucks, are displayed with bounding boxes and confidence scores for user visibility.

Table 1 depicts the calculation of the four metrics of the confusion matrix. The process is cycled over four videos, and an average value is calculated for all the metrics. The outcomes of the object detection model could be influenced by various parameters, including illumination, occlusion, low resolution, computational resources, and challenges with detecting small objects accurately. The detected object in each frame is indexed in an Excel sheet with details about the ID assigned to that object. This Excel sheet is used for object retrieval.

Table 1. Object detection using YOLOv8

Video	Resolution	Precision	Recall	F1 score
Video1	800×480	1	0.96	0.97
Video2	1280×720	1	0.857	0.92
Video3	1920×1040	1	0.771	0.87
Video4	640×360	0.95	0.963	0.95
Average		0.98	0.887	0.92

4.2. Retrieval framework

A GUI, created with Anvil software, offers users a visually intuitive platform to interact with digital elements such as buttons and menus without programming knowledge. Image retrieval systems are often assessed using metrics like Precision, Recall, and F1 score, which measure the system's accuracy in retrieving relevant images. Precision evaluates the ratio of relevant images retrieved to the total retrieved, providing insights into retrieval accuracy. This highlights the system's ability to minimize irrelevant results, as in (1) and (2).

$$\text{Precision} = \frac{\text{No. of Relevant Image retrieve}}{\text{Total no. of Image retrieve}} \quad (1)$$

$$\text{Recall} = \frac{\text{No of Relevant Image retrieved}}{\text{Total no. of relevant Image}} \quad (2)$$

Recall in image retrieval assesses the system's effectiveness in retrieving relevant images by calculating the ratio of retrieved relevant images to the total number of relevant images, emphasizing the system's capability to capture all relevant content from the database as given in (2). The F1 score in image retrieval provides a balanced measure by considering both precision and recall, offering a single metric to gauge the system's performance in finding relevant images amidst retrieved results. It assesses the harmonic mean of precision and recall, reflecting the overall effectiveness of the retrieval system [34].

As the YOLOv8 model is trained on the COCO dataset, users should enter their text queries to match the categories and concepts found within the COCO dataset for optimal results. The result obtained, as shown in Table 2 for query retrieval, shows that there are no false positives among the retrieved images. The average precision is one which indicates that all retrieved images are relevant to the query. The average recall score shows that the system effectively captures a significant portion of the relevant images but may miss some relevant ones in the dataset. The F1 value indicates that the system balances precision and recall well. The result is tested with a threshold value of 0.80. Comparing the outcome with established algorithms proves challenging due to the intricacies of combining text queries with retrieval using object detection.

Increasing the confidence score threshold from 0.80 to 0.90 reduces recall across all video classes, while precision remains constant at 1 for both thresholds. However, there is variation in the extent of the decrease among classes, with some experiencing more pronounced drops than others. Overall, adjusting the threshold impacts recall differently for each class, highlighting the trade-off between precision and recall in object detection tasks, as reflected in Figure 3.

Table 2. Query retrieval analysis for threshold 0.80

Query	Video	Resolution	No. of frames	Precision	Recall	F1 Score
Person	Video1	1280×720	1532	1	0.96	0.97
Bicycle	Video1	1280×720	1532	1	0.85	0.91
Truck	Video2	1280×720	2530	1	0.97	0.98
Cup	Video3	478×850	2035	1	0.73	0.84
Spoon	Video3	478×850	2035	1	0.69	0.81
bottle	Video4	1,280×720	1564	1	1	1
Average				1	0.86	0.91

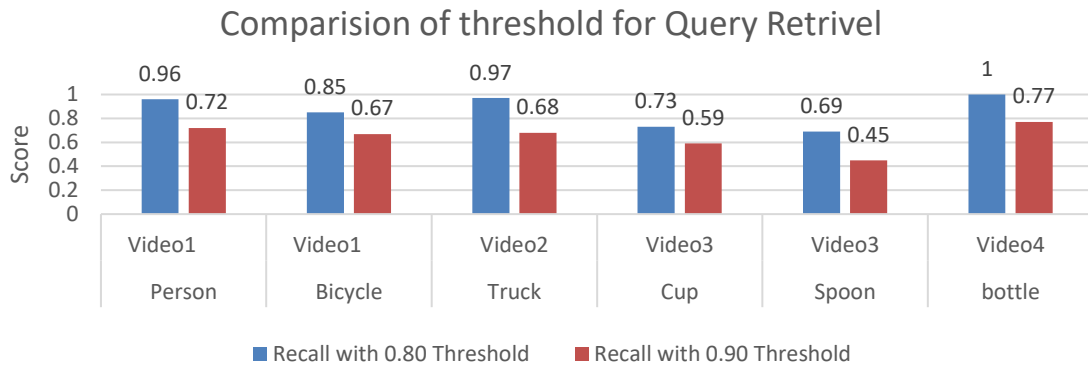


Figure 3. Comparison of the threshold of 0.80 and 0.90 for Query retrieval

4.3. Comparative results using hardware

The GPU, crucial in video processing tasks, significantly impacts the speed of text query retrieval from videos. Testing the system on three setups, Intel(R) NVIDIA GTX 1650, NVIDIA Jetson Nano, and NVIDIA Jetson Xavier, evaluates both speed and accuracy. While YOLOv8 detection results remain consistent across setups, retrieval time varies, reflecting differences in processing speed, as indicated in Figure 4. However, the influence of GPUs is just one aspect; factors like search algorithm efficiency, data volume, memory bandwidth, and system architecture also affect retrieval efficiency. While the Jetson Nano adequately handles basic AI and deep learning models, its performance may falter with larger or complex models. In contrast, the Xavier board excels in high-performance AI tasks, offering faster retrieval times, especially for demanding tasks like real-time video analysis or large-scale AI-driven retrieval.

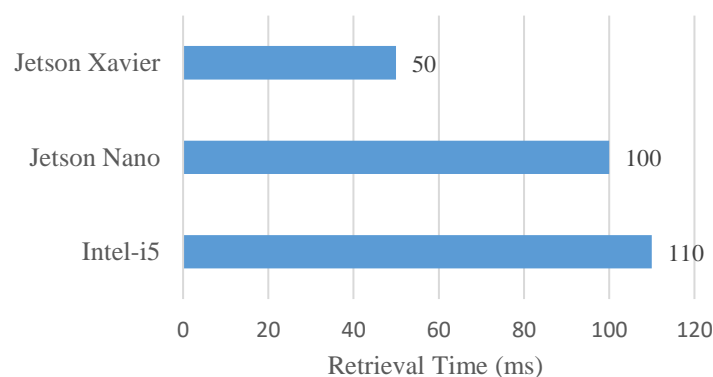


Figure 4. Average retrieval time in milliseconds for simple query retrieval on different hardware

5. CONCLUSION

The study highlights the application of object detection algorithms for image retrieval, emphasizing the crucial role of accuracy in retrieval performance. This system aims to efficiently extract desired objects from extensive video footage, thus reducing browsing time. There have been 3 algorithms tested during the object detection process, namely Mobile net SSD, YOLOv3 & YOLOv8. Mobile net SSD has an F1 score of 59%, YOLOv3 has an F1 score of 73% and YOLOv8 has an F1 score of 92%. Although all of them work quite well at detecting objects, it was observed that amongst all 3 of them, YOLOv8 is favored for its optimal balance between accuracy and speed, making it an ideal choice for detection and retrieval support.

Furthermore, the retrieval framework, presented through a graphical user interface, enhances user accessibility and simplifies retrieval. The YOLOv8 model, trained on the COCO dataset, requires users to align text queries with COCO categories for optimal results in image retrieval. Evaluation results with a threshold of 0.80 show no false positives, with balanced precision and recall. Yet, comparing existing algorithms is challenging due to the complexities of integrating text queries with object detection for retrieval. Jetson Xavier development board is used to implement the entire software, and it is concluded that this board enhances the speed by 40%.

The current retrieval process suffers from extended timeframes, necessitating the development of advanced algorithms capable of optimizing the trade-off between retrieval speed and accuracy. Moreover, there's a potential to enhance the system's functionality by training it to recognize and respond to more specific queries, such as identifying a "red car" or a "person with a black hat." This expanded training could enable the system to deliver more precise and tailored results, meeting the demands of users for detailed search capabilities.




REFERENCES

- [1] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, "A review of video surveillance systems," *Journal of Visual Communication and Image Representation*, vol. 77, May 2021, doi: 10.1016/j.jvcir.2021.103116.
- [2] K. A. Hambarde and H. Proenca, "Information retrieval: recent advances and beyond," *IEEE Access*, vol. 11, pp. 76581–76604, 2023, doi: 10.1109/ACCESS.2023.3295776.
- [3] P. Pathak, M. Gordon, and W. Fan, "Effective information retrieval using genetic algorithms based matching functions adaptation," in *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, Maui, HI, USA, 2000, doi: 10.1109/hicss.2000.926653.
- [4] C. Wan, "Research on computer information retrieval based on deep learning," *IOP Conference Series: Materials Science and Engineering*, vol. 677, no. 3, Dec. 2019, doi: 10.1088/1757-899X/677/3/032101.
- [5] P. Shivanna, S. HR, G. T. Raju, and P. Krishnan, "Performance evaluation of query processing techniques in information retrieval," in *Proceedings of International Conference on Advances in Computer Science and Application*, 2013, Lucknow. 2013, pp. 118-123.
- [6] N. Durak, A. Yazici, and R. George, "Online surveillance video archive system," in *Advances in Multimedia Modeling: 13th International Multimedia Modeling Conference*, 2006, pp. 376–385.
- [7] T. Brodsky *et al.*, "Visual surveillance in retail stores and in the home," in *Video-Based Surveillance Systems*, Springer US, 2002, pp. 51–61.
- [8] A. Hampapur *et al.*, "Searching surveillance video," in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, 2007, pp. 75–80.
- [9] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, vol. 1, doi: 10.1109/cvpr.2001.990517.
- [10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, 2005, vol. I, pp. 886–893, doi: 10.1109/CVPR.2005.177.
- [11] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2008, pp. 1–8, doi: 10.1109/CVPR.2008.4587597.
- [12] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *Digital Signal Processing: A Review Journal*, vol. 126, Jun. 2022, doi: 10.1016/j.dsp.2022.103514.
- [13] L. Du, R. Zhang, and X. Wang, "Overview of two-stage object detection algorithms," *Journal of Physics: Conference Series*, vol. 1544, no. 1, May 2020, doi: 10.1088/1742-6596/1544/1/012033.
- [14] H. Zhang and R. S. Cloutier, "Review on one-stage object detection based on deep learning," *EAI Endorsed Transactions on e-Learning*, vol. 7, no. 23, Jun. 2022, doi: 10.4108/eai.9-6-2022.174181.
- [15] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, pp. 21–37.
- [16] K. Lalitha and S. Murugavalli, "A survey on image retrieval techniques," in *Advances in Parallel Computing*, vol. 37, no. 1, IOS Press, 2020, pp. 396–400.
- [17] M. Alkhwilani, M. Elmogy, and H. El Bakry, "Text-based, content-based, and semantic-based image retrievals: a survey," *International Journal of Computer and Information Technology*, vol. 4, no. 1, pp. 58–66, 2015.
- [18] T. L. Le, M. Thonnat, A. Boucher, and F. Brémont, "Surveillance video indexing and retrieval using object features and semantic events," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 7, pp. 1439–1476, Nov. 2009, doi: 10.1142/S0218001409007648.
- [19] C. H. Tseng, C. C. Hsieh, D. J. Jwo, J. H. Wu, R. K. Sheu, and L. C. Chen, "Person retrieval in video surveillance using deep learning-based instance segmentation," *Journal of Sensors*, vol. 2021, pp. 1–12, Aug. 2021, doi: 10.1155/2021/9566628.
- [20] M.-T. Nguyen-Hoang, T.-K. Le, V.-T. Ninh, Q.-H. Che, V.-T. Nguyen, and M.-T. Tran, "Object retrieval in past video using bag-of-words model," in *2017 International Conference on Control, Automation and Information Sciences (ICCAIS)*, Oct. 2017, pp. 145–150, doi: 10.1109/ICCAIS.2017.8217565.
- [21] F. F. Chamasemani, L. S. Affendey, N. Mustapha, and F. Khalid, "Surveillance video retrieval using effective matching techniques," in *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, Mar. 2018, pp. 1–5, doi: 10.1109/INFRKM.2018.8464772.
- [22] F. F. Chamasemani, L. S. Affendey, N. Mustapha, and F. Khalid, "A framework for automatic video surveillance indexing and retrieval," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 10, no. 11, pp. 1316–1321, Aug. 2015, doi: 10.19026/rjaset.10.1828.
- [23] Kam, Ng, Kingsbury, and Fitzgerald, "Content based image retrieval through object extraction and querying," in *2000 Proceedings Workshop on Content-based Access of Image and Video Libraries*, 2000, pp. 91–95, doi: 10.1109/IVL.2000.853846.
- [24] N.-T. Do-Tran, V.-H. Tran, T.-N. Nguyen, and T.-L. Nguyen, "Efficient video retrieval method based on transition detection and video metadata information," in *2023 International Conference on System Science and Engineering (ICSSE)*, Jul. 2023, pp. 45–50, doi: 10.1109/ICSSE58758.2023.10227191.
- [25] J. Che, G. Zhang, and S. Zhang, "Video retrieval based on CNN feature and scalar quantization," in *2021 International Conference on Culture-oriented Science and Technology (ICCST)*, Nov. 2021, pp. 537–542, doi: 10.1109/ICCST53801.2021.00117.
- [26] S. Wan, X. Xu, T. Wang, and Z. Gu, "An intelligent video analysis method for abnormal event detection in intelligent transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4487–4495, Jul. 2021, doi: 10.1109/TITS.2020.3017505.
- [27] H. Guo, J. Wang, and H. Lu, "Multiple deep features learning for object retrieval in surveillance videos," *IET Computer Vision*, vol. 10, no. 4, pp. 268–272, Feb. 2016, doi: 10.1049/iet-cvi.2015.0291.




- [28] A. Mishra, K. Alahari, and C. V. Jawahar, "Image retrieval using textual cues," in *2013 IEEE International Conference on Computer Vision*, Dec. 2013, pp. 3040–3047, doi: 10.1109/ICCV.2013.378.
- [29] S. Saikia, E. Fidalgo, E. Alegre, and L. Fernández-Robles, "Query based object retrieval using neural codes," in *Advances in Intelligent Systems and Computing*, vol. 649, Springer International Publishing, 2018, pp. 513–523.
- [30] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [31] S. R. Waheed, N. M. Suaib, M. S. Mohd Rahim, M. Mundher Adnan, and A. A. Salim, "Deep learning algorithms-based object detection and localization revisited," *Journal of Physics: Conference Series*, vol. 1892, no. 1, Apr. 2021, doi: 10.1088/1742-6596/1892/1/012001.
- [32] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke, "YouTube-BoundingBoxes: a large high-precision human-annotated data set for object detection in video," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 7464–7473, doi: 10.1109/CVPR.2017.789.
- [33] A. Ahmadyan, L. Zhang, A. Ablavatski, J. Wei, and M. Grundmann, "Objectron: a large scale dataset of object-centric videos in the wild with pose annotations," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2021, pp. 7818–7827, doi: 10.1109/CVPR46437.2021.00773.
- [34] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge: Cambridge University Press, 2008. doi: 10.1017/CBO9780511809071.

BIOGRAPHIES OF AUTHORS



Swati Jagtap    received a B.E. in electronics and telecommunications from Shivaji University, India, in 2005 and an M.E. in digital systems from Pune University, India, in 2011. She is pursuing her Ph.D. in signal processing at Savitribai Phule University, Pimpri Chinchwad College of Engineering, India. She is an assistant professor at the Electronic and Telecommunication Department, Pimpri Chinchwad College of Engineering, Pune, India. Her research interests include video processing, machine learning, and deep learning. She can be contacted at email-swatialjagtap@gmail.com.



Nilkanth B. Chopade    completed his Ph.D. in engineering in the year 2009. Currently, he is working as deputy director and professor at the Electronic and Telecommunication Department, Pimpri Chinchwad College of Engineering, Pune, India. His research areas include signal transforms, signal processing, antenna arrays for mobile systems, and smart antennas. He has published more than 51 papers in Scopus-indexed journals and conferences. He has received over 20 lakhs of research Grants under different government schemes. He can be contacted at email-nbchopade@gmail.com.