Journal of Information Systems Engineering and Management
2025, 10(23s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Memory Efficient Summarization of Real-Time CCTV Surveillance System.

Dr. Amar B. Deshmukh[1], Dr. Vijaya N. Aher [2], Dr.Vidya More [3], Dr. Rahul S. Pol [4], Dr. Ajay Talele[5], Dr. Anup Ingle [6]

[1]*Associate Professor, Department of Electronics and Telecommunication, Anantrao Pawar College of Engineering & Research, Pune, India*
amarbdeshmukh@gmail.com
[2]*Assistant Professor, Department of Electronics and Telecommunication, Vishwakarma Institute of Technology, Pune, India*
vijaya.aher@vit.edu
[3]*Assistant Professor, Department of Electronics and Telecommunication, COEP Technological University, Pune, Maharashtra,*
vnm.extc@coeptech.ac.in
[4]*Associate Professor, Department of Electronics and Telecommunication, Vishwakarma Institute of Information Technology, Pune, India*
rahul.pol@viit.ac.in
[5]*Assistant Professor, Department of Electronics and Telecommunication, Vishwakarma Institute of Technology, Pune, India*
ajay.talele@vit.edu
[6]*Assistant Professor, Department of Electronics and Telecommunication, Vishwakarma Institute of Information Technology, Pune, India*
anup.ingale@viit.ac.in

ARTICLE INFO

Received: 23 Dec 2024

Revised: 10 Feb 2025

Accepted: 24 Feb 2025

ABSTRACT

Video summarization plays a crucial role in efficiently analyzing vast amounts of CCTV surveillance footage. In this research project, we propose a comprehensive approach that leverages state-of-the-art deep learning algorithms for object detection and tracking to create condensed and informative summaries of surveillance videos. The primary components of our framework include the YOLOv5 model for real-time object detection and the Deep SORT algorithm for robust object tracking. Initially, the YOLOv5 model is employed to detect objects within the CCTV footage, providing accurate bounding boxes and classifying the objects. Subsequently, the output from YOLOv5 is fed into the Deep SORT algorithm, which assigns unique IDs to each detected object and employs a Kalman filter to predict object motion and future coordinates. This predictive capability enables the system to discern moving objects within the video stream effectively. By calculating the difference between predicted and actual coordinates, the Deep SORT algorithm efficiently identifies and tracks moving objects, facilitating the extraction of relevant frames containing dynamic elements. These selected frames are then merged to generate a summarized CCTV surveillance video, highlighting key events and minimizing redundant information. The implementation of this approach has yielded promising results, demonstrating its efficacy in creating concise and meaningful video summaries. The proposed methodology not only enhances the efficiency of video analysis but also contributes to the optimization of storage and bandwidth resources in surveillance systems.

Keywords: Video Summarization, CCTV Surveillance, YOLOv5, Deep SORT, Object Detection, Object Tracking, Kalman Filter, Deep Learning.

I. INTRODUCTION

In an era characterized by the ubiquity of surveillance cameras, the volume of data generated poses a formidable challenge for efficient analysis and utilization. Video summarization emerges as a pivotal solution, seeking to distill the wealth of information within CCTV surveillance footage into concise and meaningful representations. This research endeavors to address this challenge through a comprehensive approach, integrating cutting-edge deep learning algorithms to enhance the process of videosummarization. The cornerstone of our methodology lies in the fusion of two powerful models: YOLOv5 (You Only Look Once) for real-time object detection and Deep SORT (Simple Online and Realtime Tracking) for robust object tracking with unique identifications. YOLOv5 excels in providing rapid and accurate object detection, offering precise bounding boxes and object classifications in real-time. Complementing this, Deep SORT employs a Kalman filter to predict object motion and future coordinates, ensuring the reliable tracking of objects even in dynamic environments.

The synergy between these models enables a multi-step process. Initially, YOLOv5 scans the surveillance footage, identifying and classifying objects. The output is then seamlessly integrated into Deep SORT, where each detected object is assigned a unique identifier. Leveraging the predictive capabilities of the Kalman filter, Deep SORT distinguishes moving objects and calculates the disparities between predicted and actual coordinates, a key element in the identification of dynamic elements within the video stream.

Security is a major concern around the world. Apart from security measures, video surveillance cameras have been installed in private and public buildings to overcome this challenge. In public places, houses, shops, airports, banks, etc. Several types of security surveillance cameras (ie static and moving) have been installed. These cameras play an important role in real-time surveillance and detection of suspicious activities. It also helps in investigating incidents or crime scenes, such as traffic accidents, robberies, murders and terrorism. In addition, the estimated number of cameras currently in use in the world is over 770 million. These cameras are typically active around the clock, generating more than 2,500 petabytes of video data per day. Figure 1. shows daily statistics of real-world data generated by video surveillance cameras.

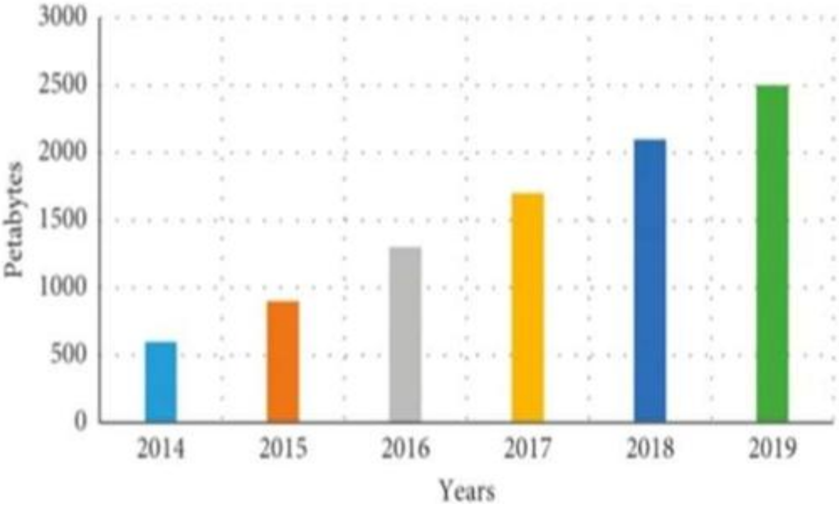

Fig.1. Daily data generated by surveillance cameras

The subsequent extraction of frames featuring moving objects paves the way for the creation of a summarized CCTV surveillance video. This condensed representation not only

facilitates a more efficient analysis of critical events but also contributes to resource optimization by minimizing redundant information. By offering a novel and integrated solution to the challenges of video summarization in the realm of CCTV surveillance, this study contributes to the advancement of efficient and meaningful video analysis in contemporary security systems.

II. LITERATURE REVIEW

This article presents a new approach to documenting video that treats it as a content analysis problem. It proposes a deep neural network to predict segment values using segment-specific and video data. This study also incorporated scene and recognition without interfering with the video to improve overall understanding. The article suggests opportunities for further development of video understanding, including integration and spatial relationships, and recommends the integration of common view strategies. It also shows that larger and more diverse datasets are needed to support deep learning in this field [1].

The video summarization plan takes an integrated approach combining audio, video and sentiment analysis. Video segmentation starts the process by breaking the content into smaller pieces so it becomes easier to analyze. In the first stage, text was extracted from these fragments using optical resolution (OCR) technology, which facilitates subsequent content. Additionally, automatic timelines are designed to facilitate faster and more accurate searches. This generalization provides a good solution for video recording and content retrieval, ensuring the main content and context [2].

Video summarization involves extracting keyframes from raw videos and aims to solve these problems by providing simple representations. The process consists of two main steps: video segmentation (dividing the video into shots) and keyframe extraction to represent visual content. This technology integration makes it easier to transfer, store and retrieve data. The two-step approach described involves training a neural network to predict changes in the video and using a coarse-to-fine strategy to filter out negatives and select keyframes. Integration of similar measures goes beyond the color histogram and motion process to ensure a good representation of the visual content of video content and an understanding of the physical continuity of video events [3].

This paper investigates the results of automatic video summarization in different applications, focusing on unsupervised video summarization as a keyframe selection problem. The goal is to minimize misrepresentation of videos and their content by using different autoencoders and different coding methods. Extract deep features using convolutional neural networks (CNN) and long-term memory networks (LSTM). The feedback loop proposes a distinctive method to enhance learning by distinguishing original videos from reconstructed videos. This paper acknowledges the difficulty of determining the distance between deep features and addresses this issue using a paradoxical method [4].

We introduce a new way to play short videos in compressed formats, using visual features and speed algorithms for instant content. Evaluation of 50 videos by the Open Video Project showed high efficiency and computational speed compared to existing methods. The structure of this article includes a basic summary, relevant activities, video organization, planning process, experimental results, and future perspectives. The focus is on the suitability of these tools for instant content. Be prepared to explore other perspectives.

Video technologies currently encounter challenges in real-time information extraction from a large volume of videos, essential for event identification, abnormal activity analysis.

Similar metrics and extensions to local feature and motion analysis. Future work includes integrating these technologies into video search and retrieval systems [5].

This paper addresses the efficiency issue of large-scale personal video collection, highlighting the need for automated and powerful video collection systems. It discusses two imaging methods: static keyframe-based and dynamic film scanning. Real-time video stream analysis focusing on dynamic scanning allows users to make flexible decisions. To develop the main content, this paper investigates the movie collection process based on visual and audio cues, change detection, keyframe detection, and text mining. In summary, this paper demonstrates the effectiveness of a real-time video summarization method suitable for portable devices [6].

This research uses the depth of object detection in computer vision, paying close attention to the problems caused by poor image quality in real situations. This article divides target detection algorithms into traditional methods and deep learning. The latter are divided into regional proposals (e.g., RCNN, SPP-net, Fast-RCNN, Faster-RCNN) and regression-based (e.g., SSD., YOLO).) offers solutions to real-world problems such as blur, occlusion, and negative, including blur, rotation, blur, and noise. This article aims to simulate various image degradation processes for analysis using mathematical models based on sample datasets. YOLO neural network was used to identify traffic signs and Darknet-53 network model was used to test them [7].

The progress in video streaming has also made progress in video analysis, but the direct processing of raw video is still not mature enough. It is computationally intensive and reduces learning accuracy. To solve this problem, it is preferred in the decision-making process and it is desirable to create a minimum number of parallel lines. Existing methods such as voxel segmentation and motion feature tracking face challenges in handling different individuals and backgrounds. To this end, this paper presents a new technique that combines YOLOv5 detectors with neural networks (RNN) for video processing. This approach includes a proposal for a functional level with RNN-based modeling of the physical body, including physical and anatomical details [8].

The analysis of visual objects is important in computer vision, including video surveillance and human-computer interaction. Even recently it is still difficult to get good viewing in real-time video. Techniques are classified based on input data: pixel, compression, and blending. Pixel space trackers such as linear filters and deep algorithms provide high accuracy but require extensive resources. Compressed space trackers use motion vectors to improve performance but may compromise accuracy. The proposed MV-YOLO framework is a hybrid tracker that combines audio vectors with semantic object detection to achieve accuracy. MV-YOLO uses a two-stage tracking method that uses motion vectors for initial estimation and optimizes the space by considering all measured objects of the frames

[9].

There are many visual materials on platforms such as YouTube and Netflix due to high-speed internet access, availability of multimedia videos and cheap storage space. This increase (YouTube alone produces more than 10 hours of video every second) requires a video compression technique. Video Summarization (VS) solves this challenge by analyzing the video, removing duplicate frames, and saving keyframes for input. This article divides VS into supervised, unsupervised, and unsupervised and explores applications in various fields. Challenges include user context, spatial dependency, and content evaluation. This research aims to fill this gap by analyzing in detail more than 40 deep learning-based video summarization algorithms developed in the last five years. Introduces tax accounting standards, training strategies and courses. The classification includes supervised and unsupervised systems, as well as systems that use central connections and monitoring systems to achieve the best results. Importantly, unsupervised methods, especially multivariate counterparts, show promise by providing similar results to supervised methods. Research highlights the importance of fully unaudited or partial/weak audit methods to reduce dependence on extensive documentation [11].

Recognizing the shortcomings of existing methods such as lower accuracy and longer training times, this article proposes a new way to improve athletic performance, especially in football. Display time and save space by removing duplicate frames. The paper concludes by emphasizing the importance of evaluating the significance of news clips in explaining the presentation and organization of the film: Section 2 presents the content of the movie video using deep learning, Section 3 discusses the experimental results, and Section 4 concludes: Future Vision [12].

This work addresses the problem of automatically analyzing videos to capture content while preserving important content and meeting certain requirements. These programs include the spatiotemporal trajectory approach, a new context method, and a new measurement method using measurement data. The process involves pixel binning for background subtraction, optical flow and trajectory subtraction, and is written using a "Tetris-like" concept. The results demonstrate the effectiveness of the proposed method in assessing homeostasis and recovery, and outperforms the fast-forward process. The paper concludes by showing the advantages of this approach and its potential in film production [13].

Decoding techniques that include image skimming, summarization and compression aim to solve these problems. Movie summarization, especially multi-view movie summarization (MVS), is a rare call. MVS processes input from multiple cameras with different perspectives, including pre-processing, feature extraction, post-processing, and content rendering. Challenges in MVS include cross-view interactions, synchronization issues, lighting changes, and visual aliasing. Despite its potential, MVS remains underutilized; This now requires expertise that can combine deep learning and standard techniques for better evaluation. This research provides an overview of the MVS approach by discussing issues, challenges, considerations, and future research directions [14].

Video technology currently faces challenges in extracting information from multiple videos in real time, which is essential for event analysis and evaluation and predict events as they occur. Researchers are actively seeking effective video summarization techniques to improve video analysis techniques and solve problems related to storage and data analysis complexity. Video summarization methods are generally divided into scene-based (static, dynamic) and content-based methods. Content-based content relies on movie content, pre-processing, and categories such as motion, event-based, and action-based approaches. This article introduces the film as a collection of films that analyze films on stage and present them at the end of the film, providing new perspectives on competition and solutions [15].

This paper addresses in the recent literature from 2023 that in addition to this research could explore incorporating Wireless Sensor Network Using Higher Security for more precise environmental control and management. [16]

Furthermore IoT based security improvement and process can be used in future. [17]

In these papers automation was discussed, it is a comprehensive and effective systematic approach to business process automation consists of 4 phases of control system: analysis, implementation, integration, and maintenance and support. [18-23]

III. METHODOLOGY

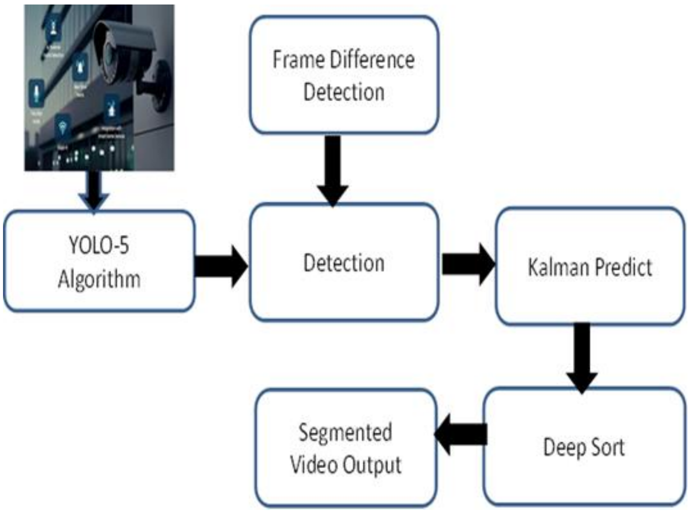

Fig.2. System Block Diagram

Data collection and preprocessing is the first step involves the collection of diverse CCTV surveillance footage to ensure the robustness and adaptability ofthe proposed approach. Ensuring the surveillance cameras are set up and configured to capture the desired area or scene. The cameras is positioned strategically to cover the areas of interest effectively. The dataset is carefully curated to encompass various scenarios, lighting conditions, and object types. Preprocessing steps include video format standardization,frame extraction, and resolution normalization to create a consistent input for subsequent analysis. Capturing video streams or images from the surveillance cameras is achieved using cameras placed in the infrastructure in place. Some options include using software provided by the camera manufacturer, open-source software like OpenCV, or custom scripts.

The YOLOv5 model is employed for real-time object detection within the surveillance footage. This deep learning model excelsin providing precise bounding boxes and object classifications. The trained YOLOv5 model is utilized to analyze each frame, identifying and classifying objects of interest, and providing accurate spatial coordinates explained in Figure 2 .

The output from YOLOv5, containing object coordinates and classifications, is seamlessly integrated into the Deep SORT algorithm. It assigns a unique identifier to each detected object, facilitating consistent tracking across frames. The Kalman filter embedded in Deep SORT predicts the motion and future coordinates of each object, enhancing the tracking robustness, especially in scenarios with occlusions and varyingspeeds.

Object tracking with Deep SORT (Simple Online and Realtime Tracking with a Deep Association Metric) is a method for tracking objects in video streams or sequences. It combines a detection algorithm, such as YOLOv5, with a tracking algorithm to follow objects across frames robustly. Deep SORT improves upon the standard SORT algorithm by incorporating deep learning features to associate detections with existing tracks efficiently.

The Deep SORT based object tracking involves object detection stage where an object detection algorithm (such as YOLOv5) to detect objects in each frame of a video sequence. These detections provide bounding boxes and corresponding confidence scores for objects of interest.

The feature extraction is used to detected the objects from each frame. It uses a machine learning algorithm, often pre-trained on large datasets, to generate feature vectors that represent the appearance of each detected object. These feature vectors help in matching objects across frames. The next step involves associating detections from consecutive frames to form tracks. Deep SORT uses the Hungarian algorithm or similar methods to solve the data association problem efficiently. It matches detections to existing tracks based on the similarity of their feature vectors and the spatial proximity of their bounding boxes.

Deep SORT maintains a set of active tracks, each representing a unique object being tracked in the video sequence. It updates the state (position, velocity, etc.) of each track based on the associated detections in each frame. Tracks are terminated if they become too old or if they exhibit suspicious behavior. Occlusions handling and ambiguities is the important step. The dynamic objects generally disappear from the scene due to movement behind the large visible object, also other challenges commonly encountered in object tracking scenarios. It employs heuristics and strategies to handle situations where objects may be temporarily hidden from view or where multiple objects overlap in the same region of the frame.

Finally, the Deep SORT algorithm provides the tracked objects along with their trajectories as output. This information can be visualized overlaid on the original video sequence or presented in a separate visualization tool. Deep SORT is widely used in various applications such as video surveillance, autonomous driving, and sports analytics, where accurate and robust object tracking is essential for understanding the dynamics of a scene. Its combination of deep learning-based feature extraction and efficient data association makes it a powerful tool for real-time object tracking tasks. Detection of motion and and frame extraction exploits the predictive capabilities of Deep SORT, the system distinguishes moving objects by calculating the difference between predicted and actual coordinates. Frames containing dynamic elements are extracted, forming a subset of the original footage. This selective extraction is crucial for the creation of a summarized video that highlights key events while minimizing redundant information.

The extracted frames are merged to create a condensed video summarization. The merging process ensures a seamless transition between frames, providing a comprehensive overview of the identified dynamic elements. The summarized video serves as an efficient representation of key events within the surveillance footage, allowing for streamlined analysis and resource optimization.

System Working Algorithm

1. Start grabbing the video frmaes from the CCTV servillence camera.
2. Read the frame from video stream.
3. Apply YOLOv5 to detect the objects
4. Extract bounding box coordinates and calculate confidence score.
5. Filter the low confidence detections by appling confidence threshould.
6. Initialize and update the Deep SORT tracker
7. Predict next position using Kalman filter.
8. Associated YOLOv5 detection with existing tracks
9. Update track information.
10. Check if frames are available then go to step 2.
11. Identify the variance between the current and background frame to recognize the moving object.
12. Use tracking algorithm Deep SORT to track the detected object over time.
13. Extract frame containing objects.
14. Generate Summerizationvideo outputcontaining significant event, such as object interactions and unusual activities.
15. End

The You Only Look Once (YOLOv5) algorithm detects objects in real-time. It is well-known for its speed and precision, which allow it to recognize and categorize objects in an image or videostream with remarkable efficiency. YOLO's central concept is todivide the input picture into a grid and forecast bounding boxes and class probabilities straight from that grid. Here's a full explanation of how YOLO works:

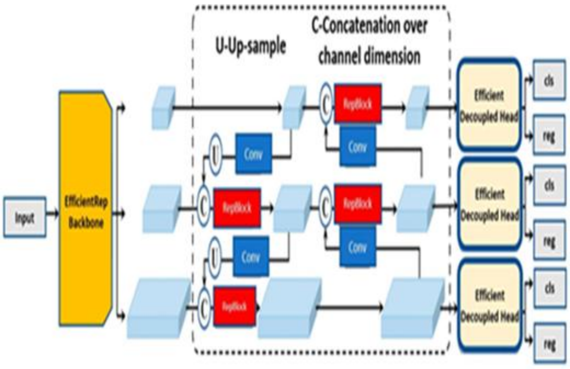

Fig. 3. YOLO Architecture

Fig. 3 presents a visual representation of the architecture of the YOLO (You Only Look Once) algorithm which divides the image frame into an S x S grid within the image itself. Each grid cell is in charge of predicting the things that fall within it. Each grid cell predicts a certain number of bounding boxes, usually two or three. Each bounding box is characterized by fivevalues: (x, y, w, h, confidence). The (x, y) coordinates denote the center of the box with respect to the grid cell's boundaries, while (w, h) represent the width and height of the bounding box,normalized to the size of the grid cell. The confidence value indicates the algorithm's level of certainty that the box encompasses an object.

Thus each produced bounding box, each grid cell estimates the probability distribution over multiple classes. A vector of class probabilitiesis used to represent this. Bounding boxes with confidence scores less than a specific level are eliminated. This stage is also called as threshold for confidence scores stage.

YOLO uses non-maximum suppression to remove duplicate and low-confidence detections. This stage eliminates superfluous bounding boxes while retaining the one with the highest confidence. The result is a list of bounding boxes, each with a class labeland a confidence score.

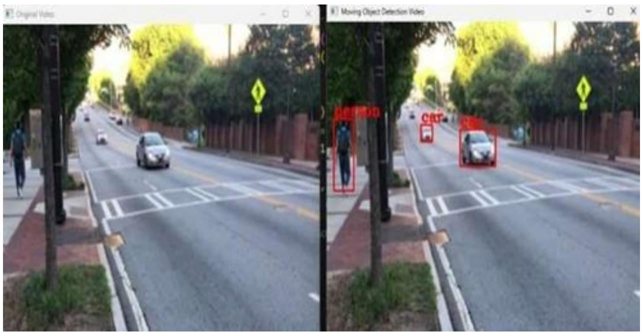

Fig.4. Output frame obtained from YOLOv5 Algorithm

Fig. 4.presents the results of object detection achieved throughthe YOLOv5 (You Only Look Once) algorithm. A)

The working of **video summaries system has number of stages such as o**bject detection, feature extraction, track initialization, Kalman filtering followed by data association, appearance features and track management. The Deep SORT assumes that you have an object detection algorithm (like YOLO) providing bounding boxes for detected objects in each frame as explained in figure 3. It uses deep learning to extract appearance features (such as embedding's) from the detected object bounding boxes. These features help in distinguishing different objects. When a new detection appears in the frame, Deep SORT initializes a new track for it. Each track is associated with a unique ID.

The Kalman filter is a recursive algorithm that estimates the state of a dynamic system observed through a series of noisy measurements. It predicts the current state based on the previous state and corrects it using the new measurement explained using equation 1.

$$X^\wedge n, n = X^\wedge n, n - 1 + Kn\ (Zn - X^\wedge n, n - 1) - - - eq\ 1$$

$$= (1 - Kn)\ X^\wedge n, n - 1 + KnZn$$

Deep SORT associates object detections with existing tracks based on appearance features. It uses a distance metric, often based on cosine similarity, to measure the similarity between appearance descriptors. The cosine similarity cos(θ) between two vectors A and B is given by equation 2:

$$(\theta)= A\cdot B/\|A\|\cdot\|B\| \text{ ----------- eq 2}$$

The similarity score is used to decide which detection is associated with which track. Deep SORT utilizes deep appearance features extracted from the object detections. The specific network architecture for feature extraction can vary. These features help distinguish between different objects. It also keeps track of the states, appearance features, and unique IDs assigned to each track.

RESULT :

From Table 1. The graph shows that the proposed YOLOv5 algorithm has the highest processing speed, at 45 fps. The next fastest algorithm is the single shot detector (SSD), at 22 fps. Theregion-based full-convolutional network (R-FCN) is the slowestalgorithm, at 6 fps.

The Table 1. also shows that the proposed YOLOv5 algorithm is more accurate than the other algorithms. For example, at a processing speed of 40 fps, YOLOv5 has an accuracy of 82%, while SSD has an accuracy of 77% and R-FCN has an accuracy of 72%. Thus, the result shows that the proposed YOLOv5 algorithm is a good choice for object detection tasks that require both high speed and accuracy.

Table 1. Processing speed of different algorithm

Different Algorithm	Processing Speed
Single Shot Detector (SSD)	22
Region Based Fully Convolutional Network (R-FCN)	6
Faster Region Based Fully Convolutional Network (F-RCNN)	17
YOLO-v5	45

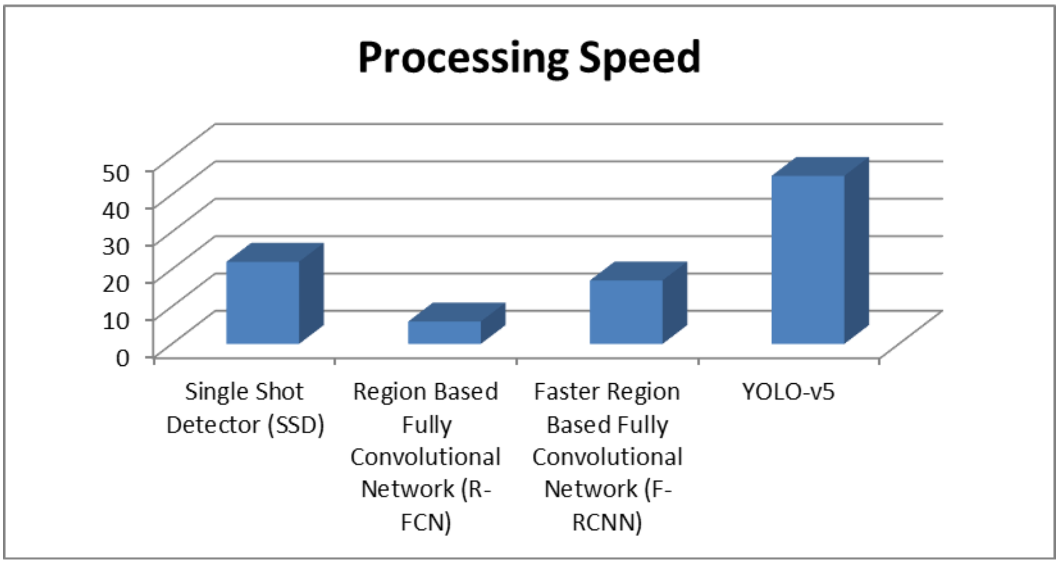

Fig.5. Comparison chart for different algorithm

Graph shows in Fig.5. Clearly describe the proposed YOLOv5 algorithm has the highest processing speed, at 45 fps. The next fastest algorithm is the single shot detector (SSD), at 22 fps. The region-based full-convolutional network (R-FCN) is the slowest algorithm, at 6 fps.

The graph also shows that the proposed YOLOv5 algorithm is more accurate than the other algorithms. For example, at a processing speed of 40 fps, YOLOv5 has an accuracy of 82%, while SSD has an accuracy of 77% and R-FCN has an accuracy of 72%. Overall, the image shows that the proposed YOLOv5 algorithm is a good choice for object detection tasks that require both high speed and accuracy.

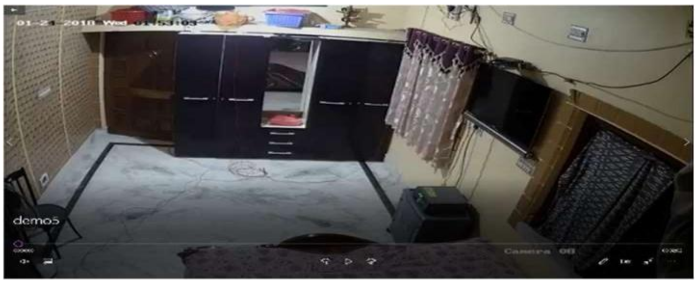

Fig.6. Original Video Frame (00:00 – 02:02)

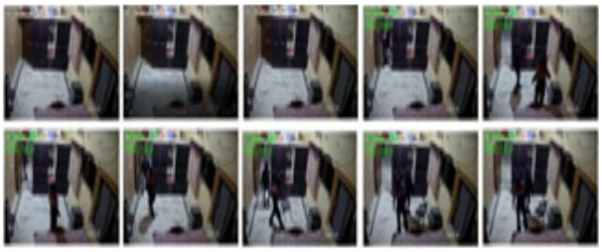

Fig.7. Selection of important frames of the sample input video

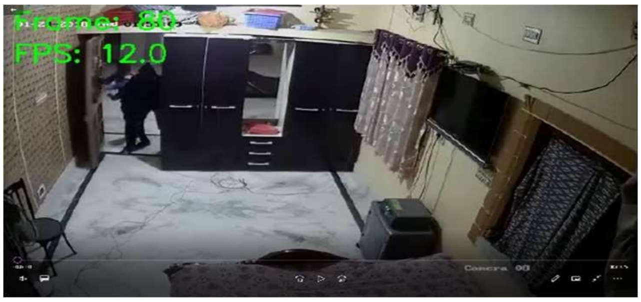

Fig.8. Summarized Output Frame (00:00 - 00:25)

From Fig.6 shows the sample video clip frame that was collected from the CCTV camera footage scattered with motion and human activity. The time duration of the given clip was 2 minutes 2 seconds. The input clip contained some stable objects kept in the room and after while it seems there is detection of human activity.

After that some of the frames were taken from the original video clip , in which some of the frames where without the presence of the human activity whereas some contains the detected object

human presence which was pertained for the detection Fig.7. and can be checked the detected object were moving or were stable, based on the criteria we extracted only the frames that contains the moving objects and the human motion. After extracting the important frames from the original video clip and removed the redundant frames which didn’t had any kind of moving activity.

As shown in Fig.8. the extracted frames were merged togther in single clip to generate the summarized video clip apart from being considerably short length compared to original video and shows only the highlighted part which user wants to use. The time duration of the summarized video was reduced to 80% of the time duration of the input image i.e. only 25 seconds.

After the extraction and merging process, the summarized video clip not only condenses the original footage but also emphasizes the dynamic elements within the scene. By focusing on frames with detected human activity and moving objects, we ensure that the viewer is presented with a concise yet comprehensive overview of relevant events.

The pre-trained object detection model played a crucial role in identifying and isolating frames featuring human presence. This allowed for a more targeted selection of frames, minimizing the inclusion of static or non-essential content. The decision to reduce the summarized video duration to 80% of the input clip's time duration further enhances its efficiency, providing a quick and informative snapshot of the key moments captured by the CCTV camera.

Additionally, the inclusion of stable objects in the room within the initial frames adds context to the video, allowing for a more holistic understanding of the environment. The careful curation of frames not only saves storage space but also streamlines the user's viewing experience, offering a streamlined narrative that centers around relevant activities.

IV. CONCLUSION

In conclusion, the future of video summarization for CCTV surveillance systems, incorporating YOLOv5 and Deep SORT, holds immense potential for advancements. Integrating advanced behavior analysis algorithms, robust anomaly detectionmechanisms, and real-time summarization can significantly enhance security measures. Human-in-the-loop systems, scalability considerations, continuous learning, and privacy enhancements further contribute to a comprehensive and cutting- edge approach. The exploration of cloud integration and multi- modal data integration ensures adaptability to evolving threats, positioning these systems at the forefront of technological advancements for heightened security and situational awareness.

V. FUTURE SCOPE

The future scope for video summarization in CCTV surveillance involves advanced behavior analysis, robust anomaly detection using unsupervised learning, real-time summarization for immediate threat response, human-in-the-loop systems for improved accuracy, scalability optimization, continuous learning for adaptive systems, privacy enhancements, cloud integration, and exploration of multi-modal data integration, ensuring cutting- edge advancements in security and situational awareness. This holistic approach ensures that the system remains at the forefrontof technological advancements, delivering heightened security and situational awareness.

REFERENCES

[1] Jiang, Yudong, Kaixu Cui, Bo Peng, and Changliang Xu. "Comprehensive video understanding: Video summarization with content-based video recommender design." In Proceedings of the IEEE/CVF international conference on computer vision workshops, pp. 0-0. 2019.

[2] Emad, Ahmed, Fady Bassel, Mark Refaat, Mohamed Abdelhamed, Nada Shorim, and Ashraf AbdelRaouf. "Automatic Video summarization with Timestamps using natural language processing text fusion." In *2021 IEEE 11th annual computing and communication workshop and conference (CCWC)*, pp. 0060-0066. IEEE, 2021.

[3] Ren, Wei, and Yuesheng Zhu. "A video summarization approach based on machine learning." In *2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 450-453. IEEE, 2008.

[4] Mahasseni, Behrooz, Michael Lam, and Sinisa Todorovic. "Unsupervised video summarization with adversarial lstm networks." In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 202-211. 2017.

[5] Almeida, Jurandy, Ricardo da S. Torres, and Neucimar J. Leite. "Rapid video summarization on compressed video." In *2010 IEEE International Symposium on Multimedia*, pp. 113-
120. IEEE, 2010

[6] Choudhary, Pradeep, Sowmya P. Munukutla, K. S. Rajesh, and Alok S. Shukla. "Real time video summarization on mobileplatform." In *2017 IEEE International Conference onMultimedia and Expo (ICME)*, pp. 1045-1050. IEEE, 2017.

[7] Liu, Chengji, Yufan Tao, Jiawei Liang, Kai Li, and YihangChen. "Object detection based on YOLO network." In

2018 IEEE 4th information technology and mechatronics engineering conference (ITOEC), pp. 799-803. IEEE, 2018.

[8] Vial, Romain, Hongyuan Zhu, Yonghong Tian, and Shijian Lu. "Search video action proposal with recurrent and static YOLO." In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 2035-2039. IEEE, 2017.

[9] Alvar, Saeed Ranjbar, and Ivan V. Bajić. "MV-YOLO: Motion vector-aided tracking by semantic object detection." In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1-5. IEEE, 2018.

[10] Saini, Parul, Krishan Kumar, Shamal Kashid, Ashray Saini, and Alok Negi. "Video summarization using deep learningtechniques: a detailed analysis and investigation." ArtificialIntelligence Review (2023): 1-39.

[11] Apostolidis, Evlampios, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. "Video summarization using deep neural networks: A survey." Proceedings of the IEEE 109, no. 11 (2021): 1838-1863.

[12] Gaikwad, D. P., S. Sarap, and D. Y. Dhande. "Video Summarization Using Deep Learning for Cricket Highlights Generation." Journal of Scientific Research 14, no. 2 (2022).

[13] Lai, Po Kong, Marc Décombas, Kelvin Moutet, and Robert Laganiere. "Video summarization of surveillance cameras." In 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 286-294. IEEE, 2016.

[14] Hussain, Tanveer, Khan Muhammad, Weiping Ding, Jaime Lloret, Sung Wook Baik, and Victor Hugo C. de Albuquerque. "A comprehensive survey of multi-view video summarization." *Pattern Recognition* 109 (2021): 107567.

[15] Elharrouss, Omar, Noor Almaadeed, Somaya Al-Maadeed, Ahmed Bouridane, and Azeddine Beghdadi. "A combined multiple action recognition and summarization for surveillance video sequences." *Applied Intelligence* 51 (2021): 690-712.

[16] Aher, Vijaya N.et.al (2024). Zig Bee-Protocol Based Dataset Acquisition of Wireless Sensor Network Using Higher Security. *SJIS-P* 36(1), 36-42.

[17] Aher Vijaya et.al. (2024). IoT Based Women Security Improvement in Methods and Process. Vol-34, (1), pp. 34-38). AWS

[18] Aher, V.N., Bhalerao, M., Pol, R., ... Chavan, P.P., Talele, A.K. , (2025) .Multichannel Data Acquisition and Monitoring System Utilizing Modbus Protocol with IIoT Automation, Communications on Applied Nonlinear Analysis,32(3) pp. 451–478.

[19] Aher, V.N., Bhalerao, M., Pol, R., ... Talele, A.K., Chavan, P.P.(2025). Simulation of Launch Pad Atomization for Military Weapon System. Communications on Applied Nonlinear Analysis, 32(3) pp. 501–511

[20] Gaikwad, S.V., Borkar, A.Y., Aher, V.N., ... Bhalke, D.G., Padwal, V.G. (2024). Quality Evaluation of an Apple using Non-Invasive Microwave Technique.. International Journal of Intelligent Systems and Applications in Engineering, 12(3s), pp. 666–671.

[21] Pol, R.S., Aher, V.N., Gaikwad, S.V., ... Borkar, A.Y., Kolte, M.T.(2024). Autonomous Differential Drive Mobile Robot Navigation with SLAM, AMCL using ROS. International Journal of Intelligent Systems and Applications in Engineering, , 12(5s), pp. 46–53.

[22] Aher, V.N., Pol, R.S., Gaikwad, S.V., ... Borkar, A.Y., Kolte, M.T. (2024).Smart Inventory System using IoT and Cloud Technology.International Journal of Intelligent Systems and Applications in Engineering, 12(4s), pp. 187–192.

[23] Gaikwad, S.V., Borkar, A.Y., Aher, V.N., Bhalke, D.G., Padwal, V.G. (2024).Design of Rectangular Wave Guide to Coaxial Line Microwave Source System for the Differential Dielectric Heating in Agriculture Using Parallel Plate Applicator. International Journal of Intelligent Systems and Applications in Engineering. 12(2s), pp. 713–717.

[24] Karambelkar, R Khaire, R Ghule, P Hiray, S Kulkarni, A Somatkar, Design and Analysis of Automatic Tripod Style Horizontal Multi Bobbin Wire Winder, 2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA)

[25] K Kolhe, AA Somatkar, MS Bhandarkar, KB Kotangale, SS Ayane, Applications and Challenges of Machine Learning Techniques for Smart Manufacturing in Industry 4.0, 2023 7th International Conference On Computing, Communication, Control And Automation