An Optimized Deep Learning Method for Video Summarization Based on the User Object of Interest

Hafiz Burhan Ul Haq, Watcharapan Suwansantisuk, Kosin Chamnongthai
Faculty of Engineering-Department of Electronics and Telecommunication Engineering,
King Mongkut's University of Technology Thonburi (KMUTT), Bangkok, 10140, Thailand

***Abstract*—Surveillance video is now able to play a vital role in maintaining security and protection thanks to the advancement of digital video technology. Businesses, both private and public, employ surveillance systems to monitor and track their daily operations. As a result, video generates a significant volume of data that needs to be further processed to satisfy security protocol requirements. Analyzing video requires a lot of effort and time, as well as quick equipment. The concept of a video summary was developed in order to overcome these limitations. To work past these limitations, the concept of video summarization has emerged. In this study, a deep learning-based method for customized video summarization is presented. This research enables users to produce a video summary in accordance with the User Object of Interest (UOoI), such as a car, airplane, person, bicycle, automobile, etc. Several experiments have been conducted on the two datasets, SumMe and self-created, to assess the efficiency of the proposed method. On SumMe and the self-created dataset, the overall accuracy is 98.7% and 97.5%, respectively, with a summarization rate of 93.5% and 67.3%. Furthermore, a comparison study is done to demonstrate that our proposed method is superior to other existing methods in terms of video summarization accuracy and robustness. Additionally, a graphic user interface is created to assist the user with summarizing the video using the UOoI.**

***Keywords*—Video summarization; deep learning; user object of interest; surveillance systems; SumMe**

I. INTRODUCTION

Globally, everyone's first priority is security. On both private and public assets, video surveillance cameras have been deployed as well as other security measures to address this difficulty. At homes, businesses, airports, banks, and other public locations, a variety of security surveillance cameras—both stationary and mobile—have been placed. These cameras are extremely important for monitoring and spotting anomalous activities. They are also useful for assisting with the investigation of incidents or crime scenes, such as car accidents, robberies, murders, and terrorist activities. Additionally, there are presently expected to be over 770 million cameras in use worldwide [1]. Over 2,500 petabytes of video data are produced each day by these cameras, which are typically always in use [2]. It is also estimated that projection growth will exceed 120 zettabytes in 2023 [3]. Every minute, 500 hours of videos are posted to YouTube [4]. Fig. 1 displays daily statistics on the actual data generated by the video surveillance cameras around the world.

Motion detection, time monitoring, facial recognition, recognition of license plates, and other content-based video analysis technologies have already made great progress in the development of video analytic technology. The issue is that manual analysis of the video recordings still needs human intervention (camera operator, security personnel, etc.). Because visual inspection requires concentration and watching the entire video, it is challenging and time-consuming to extract useful information from video footage. In the event of lengthy videos, it could potentially lead to false negatives. Therefore, it is imperative to find a solution that reduces the human time and effort required for manual analysis. To solve this issue, attempts are being made to create a video summary that quickly conveys the essence of the entire video [5]. By identifying and presenting the most interesting and up-to-date content to potential consumers, video summarizing (VS) creates a summary of substantial video content. Security surveillance systems use video surveillance to detect and analyze suspicious or anomalous activity. Individuals also use VS to share sporadic videos on social media, create highlights of different sports, create movie and television trailers, index video content to enable quick browsing of large amounts of video through video search engines, etc. [6, 7]. There have been various attempts by the researchers to propose an automated VS. The majority of VS approaches provide a summary based on choosing key frames that best depict the video during the skimming procedure. For video summarization, the shot boundary detection techniques [8–15] are widely known. Instead of concentrating on a single item, feature-based techniques [16–34] for VS provide a generalized video summary. These methods have trouble accurately recognizing the item, which prevents them from meeting the user's needs. The video is distilled using trajectory-based [6, 25] and clustering [17, 22, 35–38] algorithms that highlight related objects, actions, and events. These methods, however, do not produce a summary of any video that provides information based on the user's interests. As a result, these methods restrict the usage of retrieval tasks and do little to improve users' observing experiences. The summarizing of a video may be accomplished during the video skimming process by choosing shot portions with the use of video editing software like Filmora [38], SpenShot [39], and Davinci [40].

The aforementioned tools are expensive, need extensive storage, and require user skill. In order to capture the user's attention, it is also important to carefully choose segments that accurately portray the complete video. However, the key-frame extraction process appears appropriate for bandwidth-constrained devices and gives the video's core subject in a few frames. Similarly most of the existing techniques work on the principle of key-frame selection by eliminating the redundant

frames that may result in loss of important information related to a user's interest and create vagueness. Many surveillance cameras have been erected in public locations to monitor suspicious actions such as mobile phone snatching, terrorism, robbery, and so on, where the information contained in every single frame is critical. As a result, these strategies restrict the usage of retrieval tasks and do not contribute to improving the users' observing experience. Due to the limitation (disappearance of object and event), these techniques are unable to produce significant results. Though there are various ways that summarize the video based on the user's interest, their fundamental issue is their high processing power needs and limited accuracy.

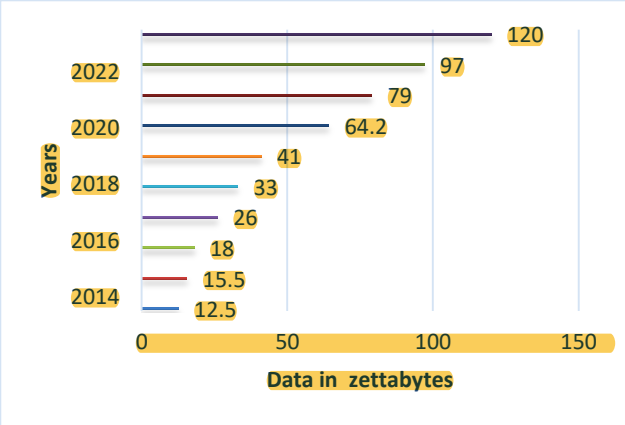


Fig. 1. Yearly production of the video's data [3].

This paper proposes a powerful VS method built on the User Object of Interest (UOoI) to address the challenges of video summarizing. The UOoI is the object that a user selects to collect all of the frames in which the selected object appears to summarize the movie. Examples of such objects are people, purses, mobile phones, motorcycles, etc. The proposed VS method has three main steps: i) the selection of the UOoI phase; ii) the detection of the object phase; and iii) the summarization of the video based on the UOoI phase. In order to exclude the unneeded noisy items (other than the OoI) that are essential for the object segmentation, the UOoI selection is first carried out from a database. Then, in order to detect an object that is thought to be a UOoI, the detector YOLOv3 is used. The VS algorithm detects the objects and then summarizes the video, relying on the UOoI. The implications of the proposed method may be summed up as follows, based on the discussion above:

- Initially selection of UOoI is done. The proposed technique chooses the object from the repository and automatically throws out any unnecessary objects; the YOLOv3 is then utilized to discover the needed object.
- The VS technique may identify a single object as well as several objects in a video clip.
- The proposed technique effectively summarizes the video and outperforms all difficulties demonstrated in the SumMe [40] and self-created Dataset.
- The experiments analysis demonstrates that the proposed method works better than cutting-edge techniques in the field of VS.

The remainder of the article is structured as follows: The literature review for departing strategies is described in Section II. The proposed VS method that describes the video summarizing technique is presented in Section III. Section IV discussed the experimental analysis and results. However, the comparative analysis is performed in Section V. Section VI provides an overview of the graphical user interface. Finally, Section VII addresses the conclusion and future work.

II. Literature Review

Several VS approaches have been put forth in the literature. A technique for summarizing the video developed by Ngo et al. [33] is based on content balance and perceptual quality. The task was completed by immediately identifying moving objects, which were then used to apply video optimization. An event-based video clarification approach has been presented by Damnjanovic et al. [19]. The method first totals the absolute difference in pixel values between the current frame and the reference frame before calculating each frame's energy. All of the current events in the frames are identified in this manner. The technique for video summarization is then used to extract keyframes. Using three processes to produce the video summary: extracting visual elements, summarizing the movie, and filtering it.

A two-stage technique is presented by Miniakhmetova and Zymbler [10] to produce a personalized summary of the video. The first step is video structure, which involves using different scene identification algorithms to produce a video summary. Using the detection bank, items are picked out of the frames of videos in the second step. The most influential sequences in which items are recognized, which later form a region of the user's interest, are included in the video summary that is produced. Three primary steps—shot boundary identification, redundant frame reduction, and stroboscopic imaging—can be used to summarize video according to a method suggested by Varghese and Nair [8].

The neighboring frame is compared to the current frame to determine the shot boundary. After that, the Structural Similarity Index (SSI) is adopted to eliminate repeating frames. The strobe is also used to display the activities that are already taking place in the film and to grasp the common backdrop. In comparison to the original video, the summarized video's overall volume has decreased by 55%. Lai et al. [15] developed a frame re-composition-based technique utilizing a clustering algorithm, optical flow, and background reduction with the goal of recognizing foreground elements. The foreground object has been identified thanks to the fusion of several pixels. Once the objects or actions have been seen, a sliding window has been utilized to integrate the recognized elements in succeeding frames to produce a spatiotemporal trajectory. The full spatiotemporal trajectory is combined to produce the video summary, and the algorithm has a 97% accuracy rate.

Three factors may be employed to determine a video's summary, according to Srinivas et al. [17]. First, each frame is given a score based on a variety of factors, such as color,

statistical attention, quality, demonstration, temporal segment, and uniformity. After that, it assigns a weight to each value based on the position of the attribute for producing key frames. The weighting is determined using the standard deviation. Lastly, the repeated frames are removed, with frames being gathered in ascending order by considering score. In comparison to previous strategies, this keyframe selection method yields outcomes that are 1.8% better. Frame selection in lecture clips has been studied by Davila and Zanibbi [32], who focused on segmentation by reducing content section conflicts, deleting objects, and re-building every frame to produce a summary of the video. The Kalman filter has been used to monitor human movements in Ajmal et al.'s [27] approach to determining the trajectory. The properties of color are useful for video since the color histogram may be utilized to identify shots and provide a synopsis of the video.

To identify an aberrant frame and eliminate noisy information from the video, Ma et al. [9] have developed a shared representation of neighboring frames. Keyframes are chosen using minimal sparse reconstruction to minimize noise and preserve critical information. A keyframe is the frame with a significant aberrant representation inaccuracy. The average percentage of reconstruction (APOR) and the sparse border are used to manage the keyframe count in a greedy iterative technique for model optimization. A cloud-based system called HOMER was introduced by Meyer et al. [41] for the creation of video highlights. With this technology, the user's emotions may be detected in order to provide a video summary. A dataset captured using a dual-camera system and a video of a home randomly chosen from Microsoft's Video Titles in the Wild (VTW) dataset are both used for experimental research. As a result, HOMER improved by 38% above baseline. Uncertainty detection and image processing technique in decision making has been presented [42-43]. ResNet 152 and a Gated Recurrent Unit (GRU) were used in tandem to summarize a movie, according to Afzal and Tahir [44]. The deep features that were present in the movie are extracted using ResNet 152 in this technique. Similarly, a GRU is utilized to increase the approach's performance and resilience. Utilizing the F-measure 43.7 and the SumMe dataset, an experimental study is conducted. A brief overview of current VS approaches is discussed in Table I.

The majority of current solutions remove unnecessary frames and a few key frames that can lose crucial information pertaining to a user's interest. In order to monitor suspicious actions like mobile theft, terrorism, robberies, etc., where the information contained in each single frame is crucial, numerous surveillance cameras have been erected in public spaces. These methods are unable to yield meaningful results because of the restriction (the disappearance of objects and events). Additionally, no method produces a summary of a video according to the user's specifications, such as one based on a single item (person, car, etc.). The proposed VS method is straightforward and incredibly reliable; it quickly and accurately generates a summary of a video depending on the user's requirements. The user chooses the UOoI as an input in the proposed method, and the algorithm generates the output in accordance with the user's requirements.

TABLE I. BRIEF OVERVIEW OF EXISTING VS METHODS

Sr. No.	Authors	Methodology	Remarks
1	Varghese and Nair [8]	For the purpose of inspecting the common backdrop frames, stroboscopic effect is used.	55% reduces of video duration.
2.	Lai *et al.* [15]	By using frame re-composition, it is deleting the irrelevant spatio-temporal segments. For the purpose of creating a video summary, the extracted items are reconnected in the spatiotemporal trajectory.	Only a stationary camera is capable of detecting objects.
3.	Ma *et al.* [9]	To optimize the model based on the adjacent frame, utilize the iteration method that use the average percentage of frame reconstruction.	Dedicated only to fixed-size frames.
4	Davila and Zanibbi [32]	Focusing the lecture video on the hand-written material that was present on the whiteboard and summarizing the film by removing any uncertainty between the topic sections.	Lower in term of accuracy.
5	Damnjanovic *et al.* [19]	Identifying and grouping the events shown in CCTV footage. Additionally, two summary types static and dynamic were added.	The main drawback is the possibility of falsely detecting events when the backdrop environment changes.
6	Ngo *et al.* [33]	This method of video summary caught both attention values and the structure of the video. Video can be organized in a hierarchical tree depending on scenes, groups, etc. to eliminate redundancy.	Low summarization rate about 10-15%
7	Miniakhmetova and Zymbler [10]	Make a description of the video based on user comments that include likes, dislikes, and neutral criticism in light of aspects influencing the scenario, including item appearance.	Prototype is missing.
8	Ajmal *et al.* [27]	The Support Vector Machine (SVM) classifier may be used to recognise the individual using a Histogram of Oriented Gradient (HOG), and the Kalman filter can be used to monitor mobility.	By making browsing quickly, the technology decreases video storage and saves time.

III. PROPOSED METHODOLOGY

Fig. 2 depicts the architecture of the proposed VS system based on the UOoI. The following key modules make up the suggested system:

- UOoI detection module: detect UOoI in videos using deep learning.
- Dictionary: The UOoI's data repository.
- The video summary is produced by the video summarization module using the frames having a UOoI in the video.

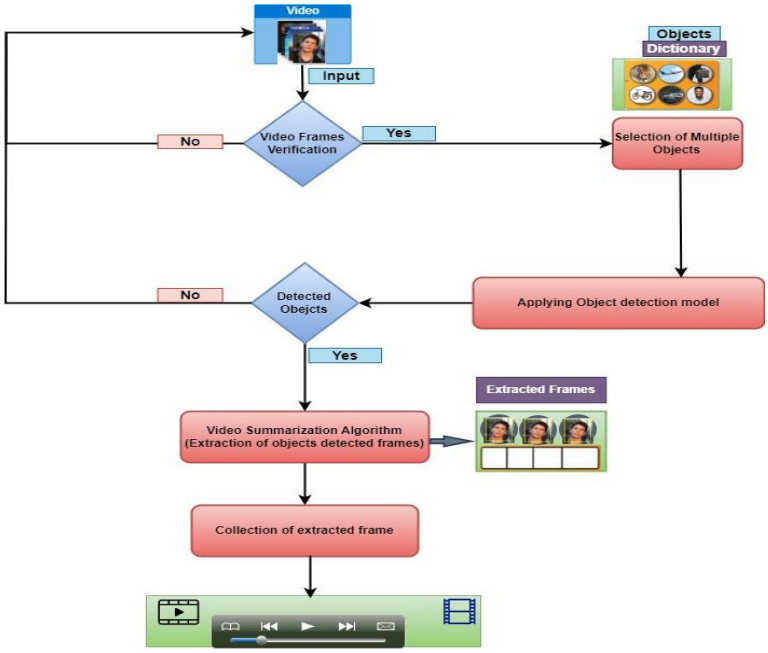


Fig. 2. Proposed method flow diagram.

A. User Object of Interest (Repository)

Defining the UOoI is the initial stage for VS. An 80-item UOoI dictionary is made specifically for this purpose. The COCO dataset, which comprises 330 thousand pictures and more than 200 thousand labeled images, is used to define the UOoI. It also offers 80 categories for things like cars, people, and handbags [45].

B. Object Type Detection

Yolo (You Look Only Once) v3 is employed in the intended attempt to identify the OoI. The position of the scene and picture where the UOoI is detected and categorized according to the category, such as a person, automobile, bicycle, etc., is determined by the object's detection. Yolo v3 employs a 53-layer modified darknet that is trained on Imagenet. In addition, 53 more layers have been added for job identification, giving Yolo v3's underlying architecture a total of 106 layers. To avoid losing low-level data, there is no pooling layer, and the feature maps are down-sampled using a convolutional layer with stride 2. YOLOv3 is substantially quicker at identifying objects than other object recognition methods [46]. The entire video is processed by Yolo v3 using just one neural network. The network divides the images into areas and generates bounding boxes and probabilities for each region. Logistic regression is used in YOLO v3 to forecast each class score, and a threshold may be used to predict an object's multiple labels. The courses that have scores over a certain level, however, are put in the box [47]. Fig. 3 describes the prediction of bounding box.

where, the bounding-box's x and y dimensions are (b_x,b_y). However, four coordinates predicted by YOLO v3 such t_x, t_y, t_w, t_h for each bounding box. The predictions are shown as follows if the cell is offset by (C_x,C_y) from the image's top-left corner and the bounding box prior has dimensions of p_w,p_h:

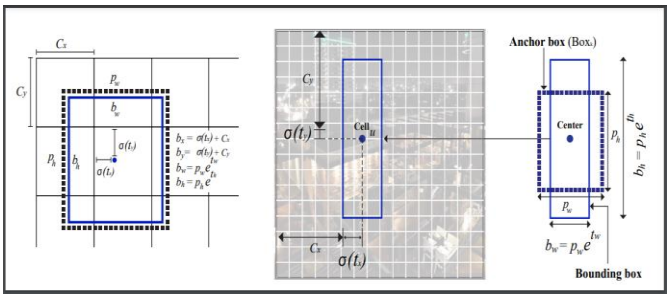


Fig. 3. Prediction of the bounding box.

$$b_x = \sigma(t_x) + c_x \quad (1)$$

$$b_y = \sigma(t_y) + c_y \quad (2)$$

$$b_w = p_w\, e^{t_w} \quad (3)$$

$$b_h = p_h\, e^{t_h} \quad (4)$$

$$Cofindence = \Pr(Obj).IOU^{truth}_{Prediction} \quad (5)$$

$$IOU = \frac{area(box_{prediction} \cap (box_{truth}))}{area(box_{prediction} \cup (box_{truth}))} \quad (6)$$

However, Pr(obj) is the value of the probability that an object existed in the grid. The value of Pr(obj) as well as confidence are dependent on object existence in grid. For example, the score of Pr(obj) is 1 if the object is in a grid. Similarly, the score is 0 if the confidence is 0. Eq. (6) describes the $IOU^{truth}_{prediction}$ which is the ratio between predicted objects and real objects. $(box_{prediction} \cap (box_{truth})$ describes the area of the intersection between predicted and real objects; whereas $area(box_{prediction} \cup (box_{truth})$ describe the area that is combined regarding predicted and real objects. Similarly, the object class is predicted $Pr(cls|Obj)$ and defined when it appears in the grid. In such a case, the confidence is measured by the multiplication of the predicted class by the probability of

an object with box convergence, as mentioned in the following equations:

$$Confidence(M) = \Pr(cls|Obj).\Pr(Obj).IOU^{truth}_{Prediction} \quad (7)$$

$$= \Pr(cls_M).IOU^{truth}_{Prediction} \quad (8)$$

C. Comparison in Terms of Speed and Accuracy

In terms of speed, when compared to other models, Yolo v3 is the best object detection model. YOLO v3 processes data at a rate of 45 fps, which is rather fast in contrast to single-shot detectors (SSD), Faster-RCNN, and R-FCN. YOLO v3's speed performance against other object identification models is shown in Fig. 4. Accuracy, taken into account for the comparison, is another crucial element, as mentioned in Fig. 5. The model Faster-RCNN executes with an accuracy rate greater than YOLO v3, but YOLO v3 has significantly better accuracy than most of the other models. It operates in a realistic situation considerably more slowly than Yolo v3 does. Many other algorithms, like the R-CNN family and SSDs, operate similarly but take longer to complete because of their numerous, intricate phases. On the other hand, YOLO v3 uses single-stage detection to complete the same task using a single neural network. When compared to other models, YOLO v3 operates precisely and executes more quickly, for example, detecting 45 frames per second as opposed to the Faster-RCNN family's detecting just five frames per second [48–49].

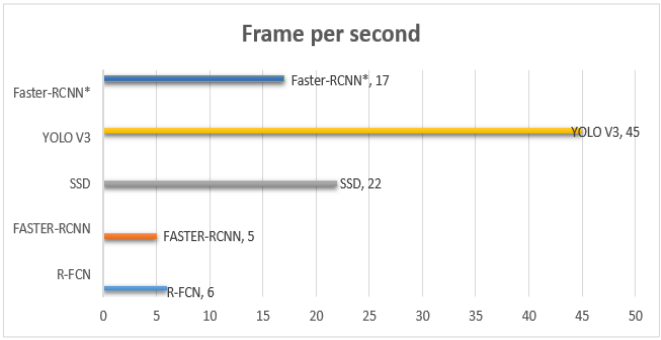


Fig. 4. Comparison of object detection models in term of speed.

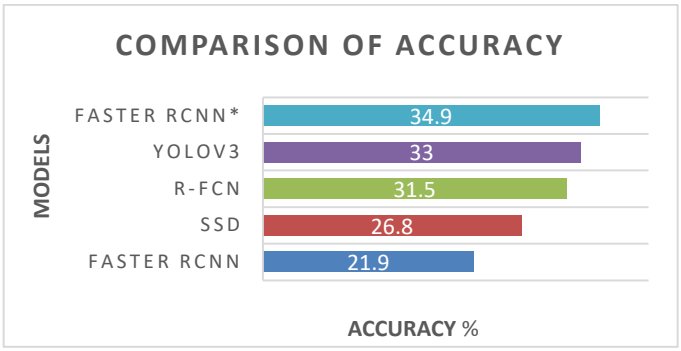


Fig. 5. Comparison of object detection models in term of accuracy coco dataset [48-49].

D. Video Summary Generation Algorithm

The key collection process based on objects of interest in videos using the YOLOv3 deep learning model can be described mathematically as follows:

Let:

Vs be the collection of videos, it is denoted by vs_i, i ∈ [1, n].

UO be the collection of desired object, it is denoted by uo_j, j ∈ [1, m].

Fm(vs_i) be the set of frames in video v_i, where each frame is denoted by fm_{ik}, k ∈ [1, p].

The YOLOv3 model calculates the bounding-boxes (b_{ik}) and associated class-probabilities (p_{ik}) for each frame fm_{ik} in the video v_i. The item's size, location, and bounding box coordinates (x, y, w, h) are displayed, and the likelihood that it belongs to a certain class is indicated by the class probabilities. Confidently selecting the frames that include interesting items is a necessary step in the key collection process. This may be achieved by setting a threshold for the class probability. The symbol α will be used to represent this threshold, where $0 \leq \alpha \leq 1$.

The mathematical method for key collection using YOLOv3 depending on desired object is defined below:

$$\text{Key} = \{(vs_{ik}, fm_{ik}) | vs_{ik} \in \text{Vs}, fm_{ik} \in \text{fm}(vs_i), \exists\, uo_j \in \text{O}, p_{ik}(uo_j) \geq \alpha\} \quad (9)$$

where:

Key is the set of key frames containing user objects of interest. $p_{ik}(uo_j)$ represents the class probability of user object uo_j in frame fm_{ik}.

In this equation, each video vs_i and its frames fm_{ik} are repeated a number of times. We include the video-frame pair (vs_i, fm_{ik}) in the set of key frames K if an object uoj appears in frame fm_{ik} with a class probability $p_{ik}(uo_j)$ greater than or equal to the threshold. Additionally, to summarize the video using all crucial frames, we may alter the equation mentioned earlier as follows:

$$\text{Smv} = \{(vs_{ik}, fm_{ik}) | vs_{ik} \in \text{Vs}, fm_{ik} \in \text{fm}(vs_i), \exists\, uo_j \in \text{uO}, p_{ik}(uo_i) \geq \beta\} \quad (10)$$

where:

Smv is the collection of frames containing interesting items whose class probabilities are larger than or equal to the summary threshold β, and where $0 \leq \beta \leq 1$.

Now that we have set a threshold on the class probabilities, we may collect n important frames and also summarize the video by looking at all frames that meet the threshold and include relevant elements. With these adjustments, the key frame collection procedure is more adaptable, and YOLOv3-based movie summaries are now possible. Utilizing the suggested architecture, the process for creating video summaries is depicted in Algorithm.

Algorithm : YOLOv3-Based Video Key Frame Selection and Summarization

Algorithm: YOLOv3-Based Video Key Frame Selection and Summarization

Input

Vs: Collection of videos (vs_i, $i \in [1, n]$)
UO: Collection of desired objects (uo_j, $j \in [1, m]$)
Fm(vs_i): Set of frames in video vs_i (fm_{ik}, $k \in [1, p]$)
α: Class probability threshold for key frame selection ($0 \leq \alpha \leq 1$)
β: Class probability threshold for video summarization ($0 \leq \beta \leq 1$)

Output

Key: Set of key frames containing desired objects
Smv: Summarized collection of frames containing desired objects

Algorithm:

1. ***for***
2. *Initialize an empty set Key to store key frames.*
3. *Initialize an empty set Smv to store summarized video frames.*
4. ***for***
5. *Iterate over each video vs_i in Vs:*
6. *Iterate over each frame fm_{ik} in Fm(vs_i):*
7. *Load object detection model YOLOv3 on frame (fm_{ik}).*
8. *For each detected object uo_j with its corresponding class*
9. *$pik(uo_j)$:*
10. ***If** ($pik(uo_j) \geq \alpha$):*
11. *Add the video-frame pair (vs_i, fm_{ik}) to set Key.*
12. ***If** ($pik(uo_j) \geq \beta$):*
13. *Add the video-frame pair (vs_i, fm_{ik}) to set Smv.*
13. ***else***
15. *Frame discarded*
16. ***end if***
17. ***end for***
18. ***end for***
20. ***Output [Sets Key and Smv]***

IV. Experimental Analysis and Result

Python is used as the programming language, and all experiments are done on a computer with specifications such as an Intel Core i5 6th generation with 8 GB of RAM.

In this study, the effectiveness of the proposed method is assessed using a subjective technique. Each test stream has a summarized movie that is produced both manually (using the video editing application Davinci) and automatically using the proposed scheme. In simple terms, this is considered a frame-level comparison. Precision, recall, F1-score, and accuracy are used to assess the performance of the proposed method. The following equations provide the mathematical expressions for various evaluation parameters:

$$Precision(P) = \frac{TP}{TP+FP} \quad (11)$$

$$Recall(R) = \frac{TP}{TP+FN} \quad (12)$$

$$F1 - Score(F1) = \frac{2\times P\times R}{P+R} \quad (13)$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (14)$$

Two distinct datasets, the SumMe dataset and the author's dataset, are utilized to evaluate and compare the effectiveness of our approach with the manual method. 25 videos that have each had at least 15 human video tags are included in a video summarizing dataset called SumMe. The videos gathered from various sources are included in our own dataset. The videos come in numerous sizes, including 320 x 240, 352 x 240, 640 x 360, 854 x 480, and 1920 x 1080, and are in the AVI and MP4 formats. The example test video sequences from the SumMe dataset and our self-created datasets are listed in Tables II and III, along with their parameters.

TABLE II. SumMe Dataset Statistics

Sr. No.	Sequences	Duration	No. of Frames
1	River Crossing	408 sec	10,200
2	Playing ball	104 sec	3,120
3	Kids playing	106sec	3,180
4	St Maarten Landing	70 sec	1750
5	Documentary1	74sec	2220

The efficiency of our approach is assessed by a number of tests on video of various lengths and resolutions. The following sections discuss the evaluation of both datasets.

A. Evaluation of SumMe dataset

For the evaluation of the SumME dataset, different scenarios have been taken from it, as mentioned in Table III. However, the first scenario belongs to a river crossing where several people are crossing the river. In which some of them have a handbag. So, in this scenario, collect all those scenes where a handbag (a user object of interest) appears.

TABLE III. Self Created Dataset Statistics

Sr. No.	Sequences	Duration	No. of Frames
1	Car mirror breaking	10 sec	300
2	Robbery	7 sec	210
3	Dog Activity	9 sec	900
4	Street video	10 sec	300
5	Person Activity	15 sec	450

Similarly, the second scenario is related to playing ball, in which a dog is playing with the ball, so in this video, keep tracking all the movements of the dog. In kid's scenarios, a bicycle appears for a limited duration, so it is taken as an object of interest. In the next video, St. Martin is taken as UOoI, and the final video is related to the documentary Under Water, where people are searching for different things, so here the person is taken as UOoI. Fig. 6 describes these scenarios, and Fig. 7 shows the efficiency of our approach by presenting UOoI-detection shots.

1) Results of SumMe dataset: As can be seen, the proposed method showed tremendous results on the SumMe dataset. However, some frames can be falsely predicted as well as missed, as mentioned in Scenario 2 of Fig. 7. This is because of distortion in the video, so such frames can only be viewed with the naked eye. In the best case, like Scenario 5, all the frames are properly detected by the proposed method and achieve the highest accuracy. Fig. 8 shows the confusion matrices. The SumMe dataset results are mentioned in Table IV.

TABLE IV. RESULT OF PROPOSEED METHOD ON SUMME DATASET

Sr. No.	Scenarios	UOoI	Duration of Video	Duration of summary	P (%)	R (%)	F1 score (%)	Accuracy (%)
1	River Crossing	Handbag	408 sec	7.8 sec.	100	79.5	88.64	98.82
2	Playing ball	Dog	104 sec	20.17 sec.	95.8	100	97.89	98.40
3	Kids playing	Bicycle	106sec	8.02 sec.	93.6	96.4	95.03	98.54
4	St Maarten Landing	Airplane	70 sec	13.21 sec.	100	100	100	100
5	Documentary 1	Person	74sec	4.13 sec.	100	83.2	90.84	97.74
			Total Time 762 seconds	**Total video summary = 49.20 seconds (93.5%)**	**97.8%**	**91.8%**	**94.48%**	**Overall accuracy 98.7%**

B. Evaluation of Self Created Dataset

For the evaluation of the self-created dataset, different scenarios have been taken from online repositories, as mentioned in Table III. However, the first scenario belongs to a car mirror breaking, in which a person broke the car mirror and took the car from it, so the handbag is considered an object of interest. The second scene belongs to a robbery, so the person is taken as UOoI. The third scene is related to monitoring dog activity, so the dog is UOoI. In the fourth and fifth scenarios, bicycles and people are taken as UOoI. Fig. 9 describes these scenarios along with UOoI detection shots in order to show the efficiency of our method. Fig. 10 shows UOoI-based shot detection in order to show the efficiency of the proposed method.

1) Result of self created dataset: On a self-created dataset, as can be seen, the proposed strategy yields superior results. However, some frames can be falsely predicted as well as missed because of low resolution or low light in the video, so such frames can only be viewed with the naked eye. In the best case, like Scenarios 1, 3, and 5, all the frames are accurately detected by our method with the highest accuracy. The confusion matrices are shown in Fig. 11. Table V describes the result of the self-created dataset.

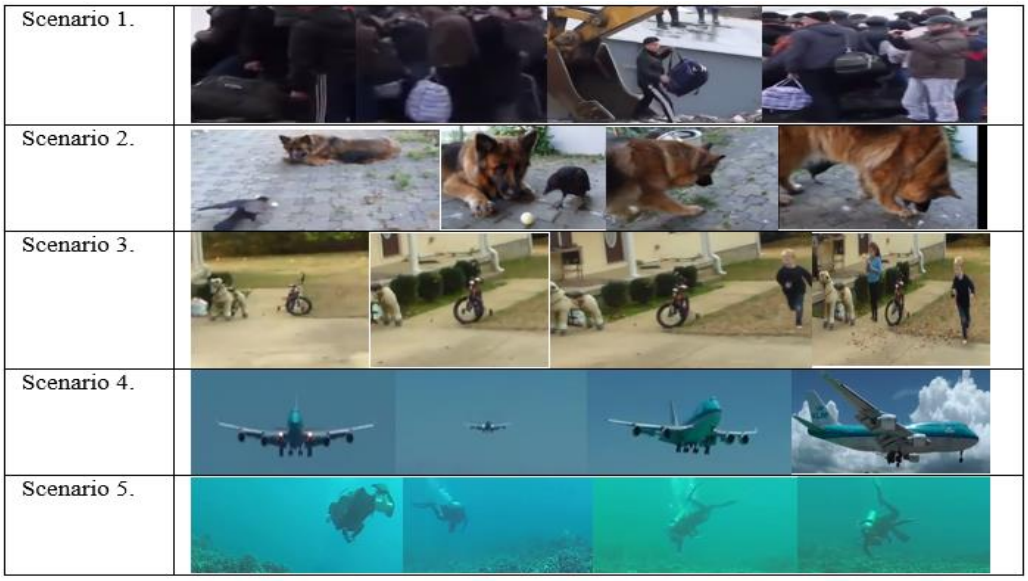


Fig. 6. Sample shot of SumMe dataset.

TABLE V. RESULT OF PROPOSEED METHOD ON SELF-CREATED DATASET

Sr. No.	Scenarios	UOoI	Duration of Video	Duration of summary	P (%)	R (%)	F1 score (%)	Accuracy (%)
1	Car mirror breaking	Car	10 sec	10 sec.	100	100	100	100
2	Robbery	Person	7 sec	1.17 sec.	100	100	100	100
3	Dog Activity	Dog	9sec	1.07 sec.	100	83.12	90.7	95.17
4	Street video	Bicycle	10 sec	3.20 sec.	100	82.7	90.5	92.48
5	Person Activity	Person	15 sec	1.23 sec.	100	100	100	100
			Total Time 51 seconds	**Total video summary = 16.67 seconds (67.3%)**	**100%**	**93.1%**	**96.2%**	**Overall accuracy 97.5%**

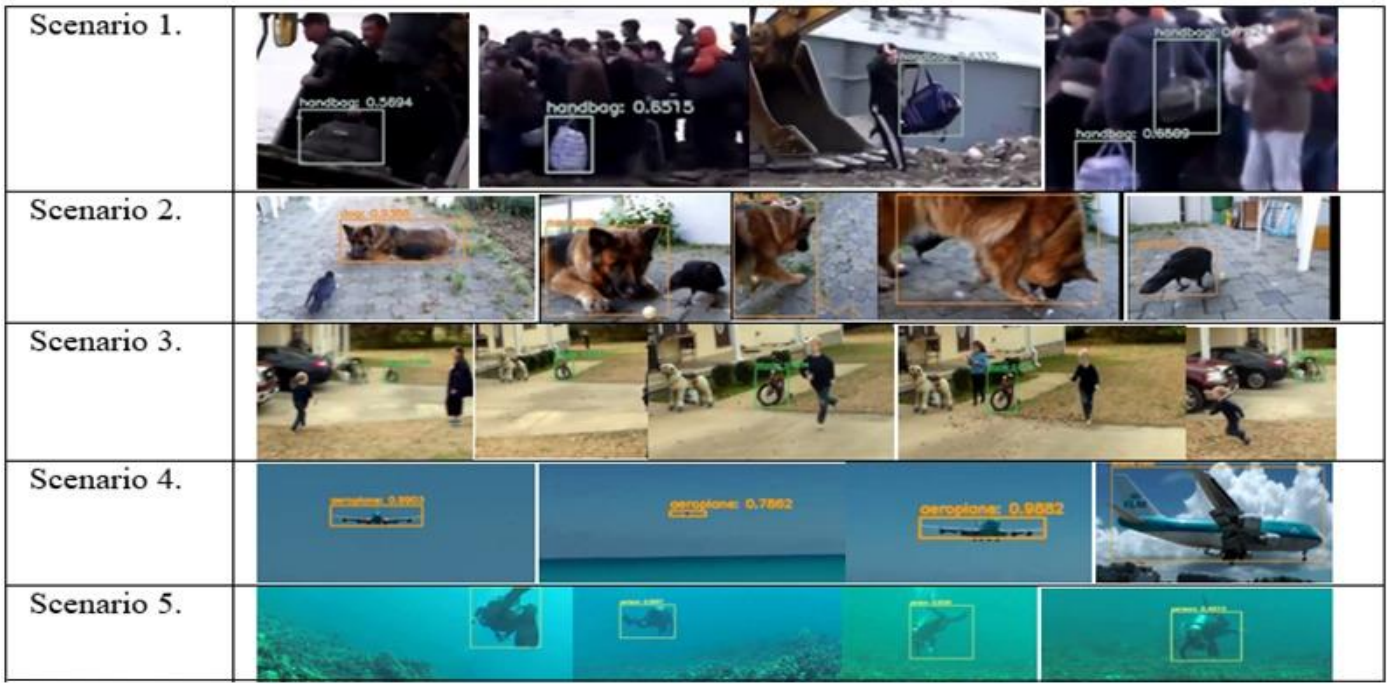


Fig. 7. Detection of UOoI on SumMe dataset.

Scenario 1 (Handbag)

TARGET / OUTPUT	UOoI	Outlier	SUM
UOoI	468 4.59%	0 0.00%	468 100.00% 0.00%
Outlier	120 1.18%	9612 94.24%	9732 98.77% 1.23%
SUM	588 79.59% 20.41%	9612 100.00% 0.00%	10080 / 10200 98.82% 1.18%

Scenario 2 (Dog Activity)

TARGET / OUTPUT	UOoI	Outlier	SUM
UOoI	1160 37.18%	50 1.60%	1210 95.87% 4.13%
Outlier	0 0.00%	1910 61.22%	1910 100.00% 0.00%
SUM	1160 100.00% 0.00%	1960 97.45% 2.55%	3070 / 3120 98.40% 1.60%

Scenario 3 (Bicycle)

TARGET / OUTPUT	UOoI	Outlier	SUM
UOoI	440 13.95%	30 0.95%	470 93.62% 6.38%
Outlier	16 0.51%	2669 84.60%	2685 99.40% 0.60%
SUM	456 96.49% 3.51%	2699 98.89% 1.11%	3109 / 3155 98.54% 1.46%

Scenario 4 (St Maarten Landing)

TARGET / OUTPUT	UOoI	Outlier	SUM
UOoI	780 44.57%	0 0.00%	780 100.00% 0.00%
Outlier	0 0.00%	970 55.43%	970 100.00% 0.00%
SUM	780 100.00% 0.00%	970 100.00% 0.00%	1750 / 1750 100.00% 0.00%

Scenario 5 (Documentary1)

TARGET / OUTPUT	UOoI	Outlier	SUM
UOoI	248 11.22%	0 0.00%	248 100.00% 0.00%
Outlier	50 2.26%	1912 86.52%	1962 97.45% 2.55%
SUM	298 83.22% 16.78%	1912 100.00% 0.00%	2160 / 2210 97.74% 2.26%

Fig. 8. Confusion matrices of SumMe dataset.

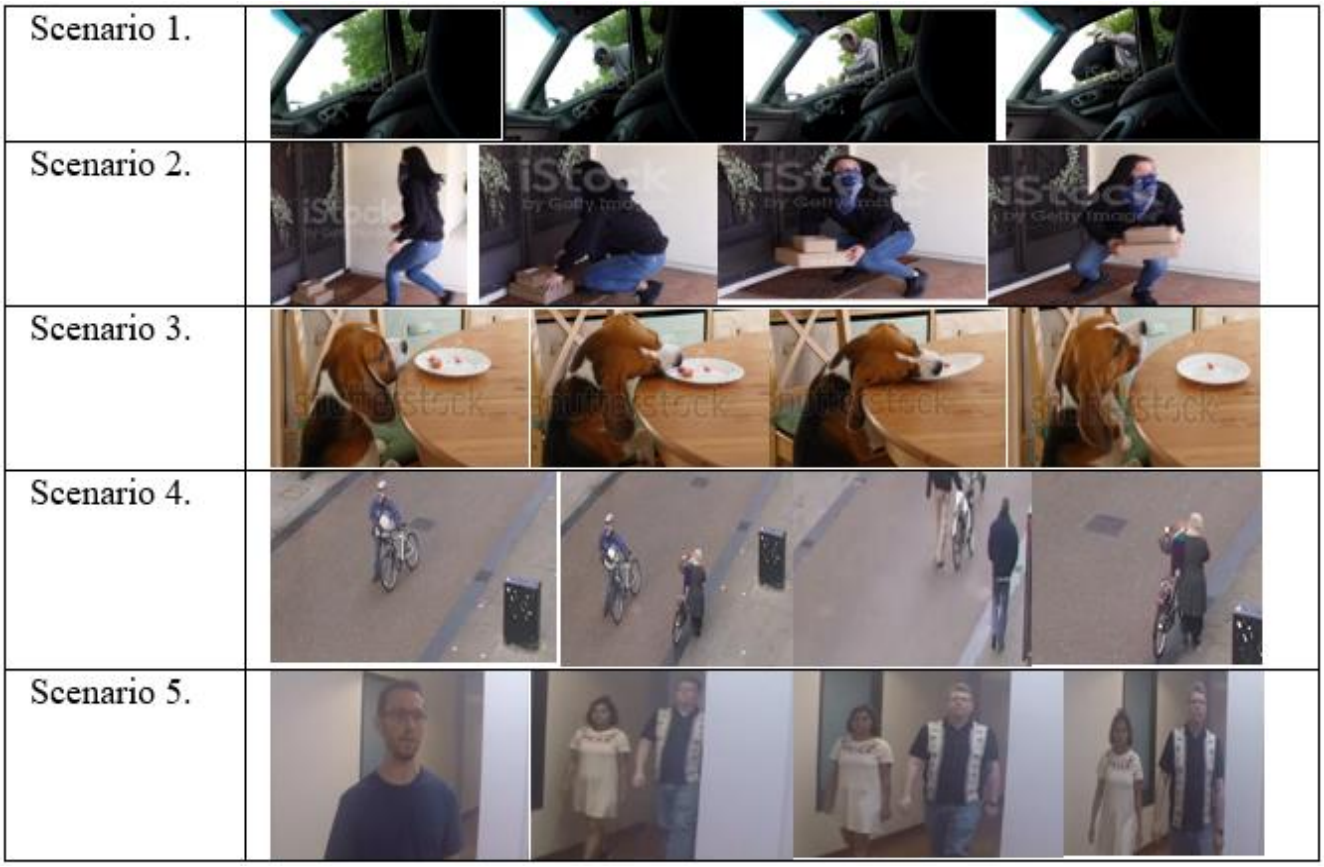


Fig. 9. Sample shot of own dataset.

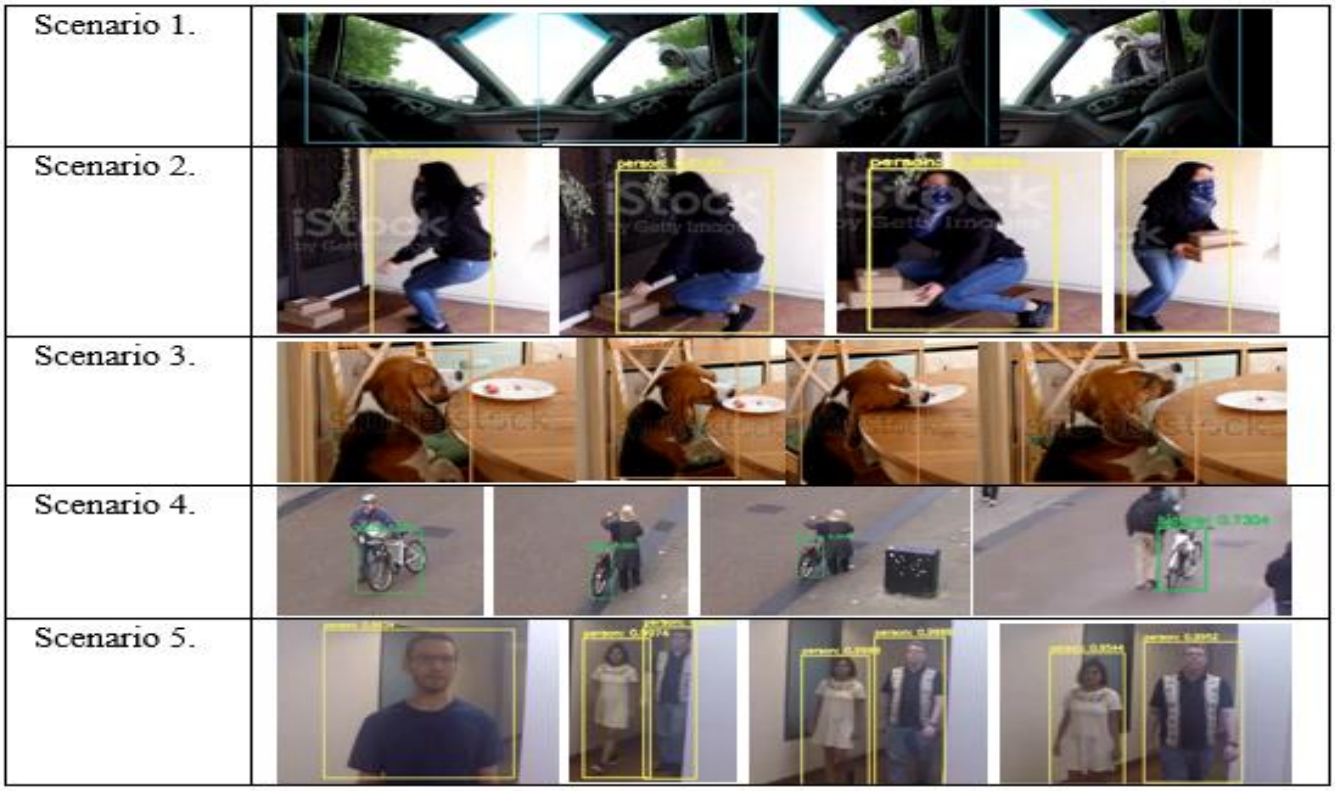
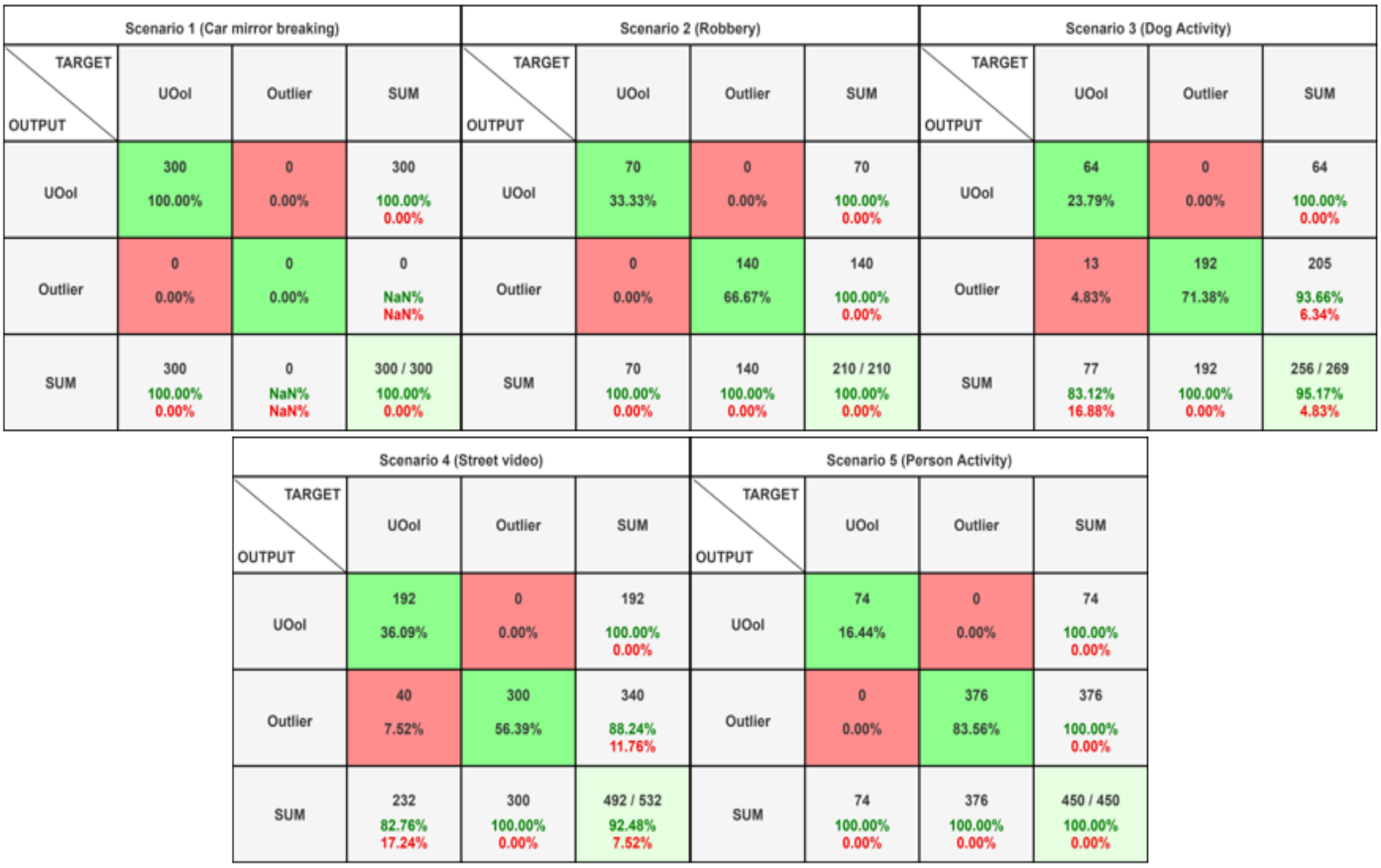


Fig. 10. Detection of UOoI on self-created dataset.



Scenario 1 (Car mirror breaking)			
OUTPUT \ TARGET	UOoI	Outlier	SUM
UOoI	300 100.00%	0 0.00%	300 100.00% 0.00%
Outlier	0 0.00%	0 0.00%	0 NaN% NaN%
SUM	300 100.00% 0.00%	0 NaN% NaN%	300 / 300 100.00% 0.00%

Scenario 2 (Robbery)			
OUTPUT \ TARGET	UOoI	Outlier	SUM
UOoI	70 33.33%	0 0.00%	70 100.00% 0.00%
Outlier	0 0.00%	140 66.67%	140 100.00% 0.00%
SUM	70 100.00% 0.00%	140 100.00% 0.00%	210 / 210 100.00% 0.00%

Scenario 3 (Dog Activity)			
OUTPUT \ TARGET	UOoI	Outlier	SUM
UOoI	64 23.79%	0 0.00%	64 100.00% 0.00%
Outlier	13 4.83%	192 71.38%	205 93.66% 6.34%
SUM	77 83.12% 16.88%	192 100.00% 0.00%	256 / 269 95.17% 4.83%

Scenario 4 (Street video)			
OUTPUT \ TARGET	UOoI	Outlier	SUM
UOoI	192 36.09%	0 0.00%	192 100.00% 0.00%
Outlier	40 7.52%	300 56.39%	340 88.24% 11.76%
SUM	232 82.76% 17.24%	300 100.00% 0.00%	492 / 532 92.48% 7.52%

Scenario 5 (Person Activity)			
OUTPUT \ TARGET	UOoI	Outlier	SUM
UOoI	74 16.44%	0 0.00%	74 100.00% 0.00%
Outlier	0 0.00%	376 83.56%	376 100.00% 0.00%
SUM	74 100.00% 0.00%	376 100.00% 0.00%	450 / 450 100.00% 0.00%

Fig. 11. Confusion matrices of self created dataset.

V. Comparative Analysis

The comparative analysis of the proposed method with the existing state-of-the-art method is done in this section. The following core characteristics serve as the foundation for the comparison analysis:

- C1.Customised User object type (UOoI),
- C2. Frame Extraction based on UOoI,
- C3. Accuracy.
- C4: Rate of Summarization

Table VI demonstrates that the majority of the strategies now in use focus on the general detection of objects rather than one particular, specific object (UOoI). Similar to this, numerous algorithms extracted the video summary by eliminating unnecessary frames and scenes rather than concentrating on the objects. This research demonstrates that our method is distinctive in that it includes the most important qualities for VS. Like the proposed method, it considers the user's input to summarize the video and produce the output according to the user. So the proposed method extracted those frames that were in the region of the user's interest. Furthermore, the proposed method is more accurate and achieved 98.7% accuracy with the highest summarization rate of 93.5% as compared to existing state-of-the-art methods.

Table VII provides another comparison of the proposed work with the existing method.

TABLE VI. Comparative Analysis with Existing Methods based on Factors

Authors	C1	C2	C3	C4
Srinivas et al. [17]	✗	✗	✗	1.8 % Improved block frame method
Ma at el. [9]	✗	✗	✗	35%-48.28%
Varghese and Nair [8]	✗	✗	✗	55%
Ngo et al. [33]	✗	✗	✗	25%
Davila and Zanibbi.[32]	✗	✓	✗	50%
Wang and Ngo [16]	✗	✓	94%	50%
Proposed Model	✓	✓	**98.7%**	**93.5%**

TABLE VII. ANALYSIS OF THE PROPOESED METHOD'S EFFICIENCY IN COMPARISON TO MODERN TECHNIQUES

Authors	Precision (%)	Recall (%)	F1-Score (%)
FSM[16]	40.73	54.43	46.59
SVM[16]	49.7	71.2	58.53
A-HHMM[16]	77.2	74.83	75.99
DT[9]	42.57	32.04	35.30
STIMO[9]	41.12	47.81	42.50
VSUMM[9]	50.43	45.34	46.51
MSRa[9]	40.03	52.05	43.56
SOMP[9]	41.83	55.02	45.33
AGDS[9]	41.35	58.40	46.27
CRavg[9]	44.94	56.44	48.28
DSNET on SumMe[50]	50.8	51.9	51.2
Proposed method on SumMe dataset	**97.8**	**91.8**	**94.48**
Proposed method on Self- created dataset	**100**	**93.1**	**96.2**

VI. GRAPHIC USER INTERFACE (GUI APPLICATION)

In the current study, a desktop application is also created utilizing PYQT5 and a Python-based GUI to give users an interactive interface for performing VS after finding objects. The interface of the application created for the selection of input video is shown in Fig. 12. Additionally, it does validation to verify the supplied input's format and contains information about the input. The system requires a video file in MP4 or AVI format as input. The explanation is that MP4 and AVI are both standardized file types. The application does not regard the input as a video if it has fewer than two frames. As a result, it issues a warning notice to the user. The next step is choosing the object type (UOoI) that will be recognized in the input video after the video has been chosen. There are several possibilities for choosing an object in this section. As a result, a user may choose UOoI with ease based on his or her preferences.

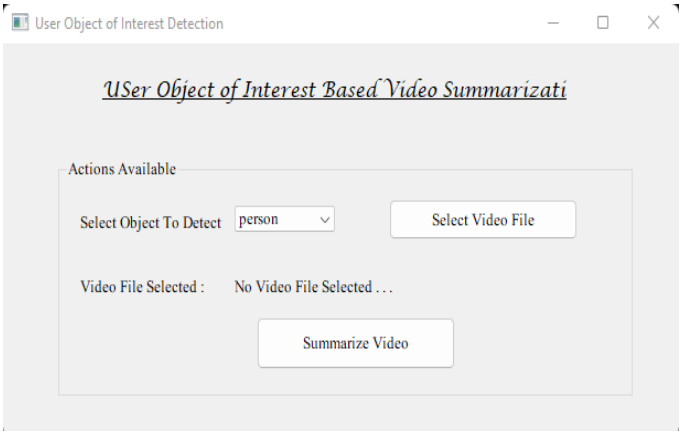


Fig. 12. Graphical user interface.

VII. CONCLUSION AND FUTURE DIRECTION

This article provides a useful VS technique for summarizing videos using the UOoI. The proposed approach is notably more efficient, optimal, and quick as compared to current state-of-the-art techniques for summarizing the video. The UOoI-based solution increases the user's ability to reliably and flexibly construct the pertinent video summary. The proposed method can detect diverse object types accurately and efficiently using YOLOv3. The proposed approach is extensively tested on two different datasets, including the SumMe dataset and my personal dataset. The proposed approach achieves an accuracy of 98.7% with a quick processing rate and a time savings of 93.5% when the complete video is viewed to detect the UOoI on the SumME dataset. Accuracy is 97.5% on the self-created dataset, and overall time reductions are 67.3%. Similarly, a comparative analysis has been performed that shows the proposed work contains novelty with the highest accuracy as well as the highest summarization rate. Furthermore, a GUI that provides ease and configurable object selection is also developed. Future work on this project will expand it to include multiple objects of interest and concentrate on improving its accuracy and summary rate.

ACKNOWLEDGMENT

"This work was supported by the Petchra Pra Jom Klao Ph.D research scholarship under Agreement No. "14/2565", by King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand".

REFERENCES

[1] E. Cosgrove, "One billion surveillance cameras will be watching around the world in 2021, a new study says. CNBC". https://www.cnbc.com/2019/12/06/one-billion-surveillance-cameras-will-be-watching-globally-in-2021.html (2019).

[2] Data generated by new surveillance cameras to increase exponentially in the coming years. SecurityInfoWatch.Com. https://www.securityinfowatch.com/video-surveillance/news/12160483/data-generated-by-new-surveillance-cameras-to-increase-exponentially-in-the-coming-years (2016).

[3] Duarte, F. T. Amount of Data Created Daily. https://explodingtopics.com/blog/data-generated-per-day (2023).

[4] L. Ceci, Statista, https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/.

[5] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, "STIMO: STIll and MOving video storyboard for the web scenario," Multimedia Tools Appl. vol. 46, no. 1, pp. 47–69 (2010).

[6] R. Kansagara, D. Thakore, M. Joshi, “A study on video summarization techniques”. International journal of innovative research in computer and communication engi-neering, (2014).

[7] A. Bora, S. Sharma, “A Review on Video Summarization Approcahes: Recent Advances and Directions”. In 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), pp. 601-606 (2018).

[8] J. Varghese, K. R. Nair, “An Algorithmic Approach for General Video Summarization”. In 2015 Fifth International Conference on Advances in Computing and Communications (ICACC), pp. 7-11 (2015).

[9] M. Ma, S. Mei, S. Wan, Z. Wang, D. D. “Feng. Robust video summarization using collaborative representation of adjacent frames”. Multimedia Tools and Applications, 78(20), pp. 28985-29005 (2019).

[10] M. Miniakhmetova & M. Zymbler, “An approach to personalized video summarization based on user preferences analysis”. In Application of Information and Communication Technologies (AICT), 2015 9th International Conference. pp. 153-155 (2015).

[11] Hafiz Burhan Ul Haq, M.Asif, Maaz Bin Ahmad, “Video Summarization Techniques: A Review”, International Journal of Scientific & Technology Research, Vol. 9(11),(2020).

[12] S. Uchihachi, J. Foote, L. Wilcox, “Automatic Video Summarization Using a Measure of Shot Importance and a Frame Packing Method”. United States Patent 6, pp.535-639 (2003).

[13] Z. Lu, K. Grauman, “Story-Driven Summarization for Egocentric Video”. 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2714–2721 (2013) [doi:org/10.1109/CVPR.2013.350].

[14] Y. Jiang, K. Cui, B. Peng, & C. Xu. Comprehensive Video Understanding: Video summarization with content-based video recommender design. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. (2019).

[15] P. K. Lai, M. Décombas, K. Moutet, R. Laganière, “Video summarization of surveillance cameras”. In 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS,) pp. 286-294 (2016).

[16] F. Wang, C.W. Ngo, “Summarizing rushes videos by motion, object and event understanding”. IEEE Transactions on Multimedia, (2012).

[17] M. Srinivas, M. M. Pai, & R. M. Pai, “An Improved Algorithm for Video Summarization–A Rank Based Approach”. Procedia Computer Science. 89, pp. 812-819 (2016).

[18] K. Kumar, D. D Shrimankar. & N. Singh. “Event BAGGING: A novelevent summarization approach in multiview surveillance videos”. In Innovations in Electronics, Signal Processing and Communication (IESC), 2017 International Conference, pp. 106-111, IEEE (2017).

[19] U. Damnjanovic, V. Fernandez, E. Izquierdo, “Event Detection and Clustering for Surveillance Video Summarization”, In: Proceedings of the Ninth International Workshop on Image Analysis for Multimedia Interactive Services. IEEE Computer Society, Washington, USA (2008).

[20] S. S. Thomas, S. Gupta, V. K. Subramanian, “Perceptual video summarization—A new framework for video summarization”. IEEE Transactions on Circuits and Systems for Video Technology. 27(8), pp. 1790-1802 (2016).

[21] F. Cricri, S. Mate, I. D. Curcio & M. Gabbouj. “Salient event detection in basketball mobile videos”. In Multimedia (ISM), 2014 IEEE International Symposium, pp. 63-70, IEEE (2014).

[22] S.J. Andaloussi, A. Mohamed, N. Madrane & A. Sekkaki. “Soccer video summarization using video content analysis and social media streams”. In Proceedings of the 2014 IEEE/ACM International Symposium on Big Data Computing, IEEE Computer Society,pp. 1-7 (2014).

[23] M. Cote, F. Jean, A. B. Albu & D. Capson. “Video summarization for remote invigilation of online exams”. In Applications of Computer Vision (WACV), 2016 IEEE Winter Conference ,pp. 1-9, (2016)

[24] R. Agyeman, R.Muhammad, & G.S. Choi. "Soccer video summarization using deep learning". In 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 270-273. IEEE. (2019).

[25] E. Bulut, T. Capin, “Key Frame Extraction from Motion Capture Data by Curve Saliency”. In: Proceedings of 20th Annual Conference on Computer Animation and Social Agents, Belgium (2007).

[26] Tonge, A and Sudeep D. Thepade, “Creating Video Visual Storyboard with Static Video Summarization using Fractional Energy of Orthogonal Transforms” International Journal of Advanced Computer Science and Applications(IJACSA), 13(9), (2022). http://dx.doi.org/10.14569/IJACSA.2022.0130931.

[27] M. Ajmal, M. Naseer, F. Ahmad, A. Saleem. “Human Motion Trajectory Analysis Based Video Summarization”. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 0-103 (2017).

[28] Yasmin S. K. and Soudamini P., “Video Summarization: Survey on Event Detection and Summarization in Soccer Videos” International Journal of Advanced Computer Science and Applications(IJACSA), 6(11), 2015. http://dx.doi.org/10.14569/IJACSA.2015.061133.

[29] E. Apostolidis, E. Adamantidou, A.I. Metsai, V. Mezaris, and I. Patras, I.. Video Summarization Using Deep Neural Networks: A Survey. arXiv preprint arXiv:2101.06072. (2021).

[30] Al-Musawi, N.J. and Hasson S.T., “Improving Video Streams Summarization Using Synthetic Noisy Video Data” International Journal of Advanced Computer Science and Applications(IJACSA), 6(12), 2015. http://dx.doi.org/10.14569/IJACSA.2015.061233.

[31] A. A. Rav, Y. Pritch, S. Peleg, Making a long video short: Dynamic video synopsis. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06).vol. 1, pp. 435-441 (2006).

[32] K. Davila , R. Zanibbi. “Whiteboard Video Summarization via SpatioTemporal Conflict Minimization”. In Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference. vol 1. pp. 355-362(2017).

[33] C. W. Ngo, Y. F. Ma, H. J. Zhang, “Video summarization and scene detection by graph modeling”. IEEE Transactions on circuits and systems for video technology, 15(2), pp. 296-305 (2005).

[34] M. Agrawal and D. S. Niranjan, "Video Summarization using Machine Learning Mechanism: A Comprehensive Review," 2021 International Conference on Advances in Technology, Management & Education (ICATME), Bhopal, India, 2021, pp. 31-36, doi: 10.1109/ICATME50232.2021.9732735.

[35] A. Tonge and S. D. Thepade, "A Novel Approach for Static Video Content Summarization using Shot Segmentation and k-means Clustering," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022, pp. 1-7, doi: 10.1109/MysuruCon55714.2022.9972379.

[36] K. Kumar, D.D. Shrimankar, N. Sing. "Equal partition based clustering approach for event summarization in videos". In: 2016 12th international conference on signal-image technology & internet-based systems (SITIS), IEEE, pp 119–126. (2016)

[37] K Kumar, D.D. Shrimankar, N. Singh, "Eratosthenes sieve based key-frame extraction technique for event summarization in videos". Multimed Tools Appl. 2018;77(6), pp. 7383–7404. doi: 10.1007/s11042-017-4642-9, (2018).

[38] [OFFICIAL] Wondershare Filmora - Easy, Trendy and Quality Video Editing Software. Filmora. Retrieved June 17, 2020, from https://filmora.wondershare.net/.

[39] OpenShot Studios, LLC. OpenShot Video Editor Free, Open, and Award-Winning Video Editor for Linux, Mac, and Windows! Openshot. Retrieved June 17, 2020, from https://www.openshot.org/.

[40] DaVinci Resolve logo June 22, 2023, from https://www.blackmagicdesign.com/products/davinciresolve/studio

[41] H. Meyer, P. Wei, & X. Jiang, “Intelligent Video Highlights Generation with Front-Camera Emotion Sensing”. Sensors, 21(4), pp. 1035 (2021).

[42] M. Saqlain. "Sustainable Hydrogen Production: A Decision-Making Approach Using VIKOR and Intuitionistic Hypersoft Sets". Jounal of intellegent management decision, 2(3), pp. 130-138, (2023). https://doi.org/10.56578/jimd020303

[43] H. B. U. Haq, and M. Saqlain. "Iris detection for attendance monitoring in educational institutes amidst a pandemic: A machine learning approach." Journal of Industrial Intelligence, 1, no. 3, pp. 136-147, 2023.

[44] M. S. Afzal, & M.A. Tahir, “Reinforcement Learning based Video Summarization with Combination of ResNet and Gated Recurrent Unit”. In VISIGRAPP (4: VISAPP), pp. 261-268 (2021).

[45] COCO - Common Objects in Context. COCO - Common Objects in Context. Retrieved July 1, 2020, from https://cocodataset.org/#home.

[46] A. Kathuria, What’s new in YOLO v3? Towards Data Science. https://towardsdatascience.com/yolo-v3-object-detection-53fb7d3bfe6b (2018).

[47] J. Redmon, & A. Farhadi, “Yolov3: An incremental improvement”. arXiv preprint arXiv:1804.02767, (2018).

[48] J. Hui, Object detection: speed and accuracy comparison (Faster R-CNN, R-FCN, SSD, FPN, RetinaNet and YOLOv3). Medium. https://medium.com/@jonathan_hui/object-detection-speed-and-accuracy-comparison-faster-r-cnn-r-fcn-ssd-and-yolo-5425656ae359 (2018).

[49] YOLO v3 theory explained, Medium. https://medium.com/analytics-vidhya/yolo-v3-theory-explained-33100f6d193 (2019).

[50] W. Zhu, J. Lu, J. Li, J. Zhou, "DSNet: A Flexible Detect-to-Summarize Network for Video Summarization." IEEE Transactions on Image Processing, pp. 948-962, 2020