2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) | 979-8-3503-3250-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/UPCON56432.2022.9986376

Object Detection based Approach for an Efficient Video Summarization with System Statistics over Cloud

1st Alok Negi
Department of Computer Science and Engineering
National Institute of Technology Uttarakhand
Srinagar, India
aloknegi.phd2020@nituk.ac.in

2nd Krishan Kumar (*SMIEEE MACM*)
Department of Computer Science and Engineering
National Institute of Technology Uttarakhand
Srinagar, India
k2b@ieee.org

3rd Parul Saini
Department of Computer Science and Engineering
National Institute of Technology Uttarakhand
Srinagar, India

4th Shamal Kashid
Department of Computer Science and Engineering
National Institute of Technology Uttarakhand
Srinagar, India

***Abstract*—The tremendous volume of video data generated by industrial surveillance networks presents a number of difficulties when examining such videos for a variety of purposes, including video summarization (VS), analysis, indexing and retrieval. The task of creating video summaries is extremely difficult because of the huge amount of data, redundancy, interleaved views and light variations. Multiple object detection and identification in video is difficult for machines to recognize and classify. To address all such issues, multiple low-feature and clustering-based machine learning strategies that fail to completely exploit VS are recommended. In this work, we achieved VS by embedding deep neural network-based soft computing methods. Firstly, the objects in extracted frames are detected using YOLOv5, and then the frames without objects (useless frames) are removed. Video summary generation occurs with the help of frames containing Objects. To check the quality of the proposed work Summary length, precision, recall, PR curve, and mean average precision (mAP) are used and system resource utilization during the model training are also tracked. As a result, the proposed work was able to identify the most effective video summarization framework with best summary length under varying conditions.**

***Index Terms*—Mean average precision (mAP), Object Detection, Precision, Recall, Video Summarization (VS), YOLOv5**

I. Introduction

With the fast development of interactive multimedia, vast database systems of images and videos necessitate algorithms that are relevant to browsing and accessibility of the data pursued. The majority of visual data offered by videos is certainly redundant, and we need to figure out how to preserve only the information strictly required for responsive browsing and querying.

Video summarization (VS) is one of the most widely used mechanisms for developing an effective video archiving system [1]. VS is indeed a mechanism that generates a short summary of video, which can be a pattern of stationary or moving images. It aims to reduce the quantity of data that has to be analyzed in order to retrieve a specific chunk of information in a video, and is thus an important task in video analysis and indexing applications [2]. The emerging field of automatic VS is broad and includes the use of these technologies by media organizations to generate trailers or teasers for films and shows; trying to present the highlights of an event (e.g., a sporting event, a music group performance, or an open debate); and generating a video short summary with primary activities that occurred, e.g., the last 24 hours of CCTV surveillance recordings for time-efficient monitoring or safety reasons [3], [4].

The generation of video summary from static frames is known as key-frames. Generally, a key frame extraction approach should be completely automated and use the video's information to construct a summary [5]. Theoretically, important frames should be retrieved using high-level properties like objects, activities, and events [6], [7]. Whereas video skimming delivers graphical, motion, and auditory information whereas still images capture video content more quickly and compactly. Key frames allow users to understand the whole content relatively faster than when viewing a series of video scenes [8], [9]. Video abstraction is viewed as the challenge of mapping a full segment (including static and moving content) to a limited number of relevant images in key frame-based visual representation. The extraction of key frames should be automated and content-based in order to preserve the video's salient content while eliminating all repetition. However, key frame retrieval based on high-level characteristics is mainly limited to certain applications, while low-level features are frequently used [10]. Color histograms, correlation, moments, edges, and motion features are a few examples of low-level features that are often employed. These low-level characteristics can subsequently be used to construct domain-specific applications by deriving high-level features [11]–[13].

The human visual system is capable of instantaneously and precisely distinguishing a given visual along with its content, position, and nearby visuals, but human-made computer vision-enabled machines are very slow in speed and accuracy. As a result, real-time object identification has become a critical topic in the ongoing automation or substitution of manual effort. The primary purpose of object detection is to identify instances in images and videos [14], [15]. Object detection in the context of CV refers to the process of finding objects of interest at certain locations in an image [16], [17]. Object detection has several applications in artificial intelligence and computer vision, involving human-robot interaction, safety, surveillance, and virtual reality. Object detection essentially includes another CV method, i.e., image classification. It means that object detection algorithms initially attempt to determine whether an object of interest exists in the image. If affirmative, the very next step is to determine the object's coordinates and build a bounding box around it. In the meantime, constructing a bounding box around an object is a regression problem in itself. Building the bounding box involves predicting the pixels that encompass the object of interest [18].

Object detection is important in a variety of applications, including surveillance, cancer identification, vehicle tracking, and underwater object detection. For various purposes, several strategies have been employed to identify objects precisely and effectively. However, many proposed solutions continue to have issues with accuracy and efficiency [19]. To address these object detection issues, machine learning and deep neural network approaches are more successful in object detection. In this regard, we proposed object of interest based technique in which we initially extract frames at 25 frames per second, then objects are detected from the extracted frames using YOLOv5 to reduce superfluous frames. Finally, a video summary is generated using frames containing objects. The tracking of the system resource utilization during model training over cloud is the major contribution for the proposed work.

II. Related Work

In this multimedia world, video summarization has emerged as an intriguing research issue for both the computer vision communities and the multimedia fields. It has not only provided new solutions for social media analytics and apps, but it has also motivated us to assess tag localization methods, which are in great demand [20]. For a comprehensive review of summarization, Jiang et al. [21] developed a high-level semantic video summarizing system based on a set of consumer-oriented recommendations. Hannon et al. [22] offered real-time online user-generated content that provided vital insights into real-world events using time-stamped comments from social networking sites for writing soccer match recaps.

There are several kinds of representation of video summary including keyframes [9], [23], video skimming [24], storyboard [25]. The author Krishan Kumar [9] highlighted a HCS-based key-frame extorted model for VS for creating simply and intelligently event summary (ES), in which a spatiotemporal similarity function was devised to generate a similarity matrix using visual attributes. ES improved users' access to vast amounts of video content in an efficient and timely manner, which was dependent on the extracted key-frames, thus the summary had a small number of frames.

In the continuation Hussain et al. [26] proposed a CNN and DB-LSTM-based MVS framework. The framework initially divided the multi-view videos into segments based on human and vehicle appearance. The segmented images captured in the online tier are saved in a lookup table with a date and then communicated to the cloud for transmission of the valuable and needed data to the cloud, which played an important role in reducing bandwidth and processing resources. Then, DB-LSTM, which was trained to learn informative and non-informative frame sequences, outputs class probabilities. Finally, the final summary produced by the sequences having the highest probabilities of being informative.

Video summarization (VS) is creating a summary of extensive video content by detecting and presenting relevant material to the potential users that are most informative and contain up-to-date information. The researchers have made several efforts in this regard such as Ul Haq et al. [27] offered an effective VS framework that is based on Object of Interest(OoI). OoI selection was conducted from the dictionary to exclude extraneous noisy objects that were required for object segmentation. Tahir et al. [19] proposed four pipelines of object detection in satellite imagery as faster RCNN (faster region-based convolutional neural network), YOLO (you only look once), SSD (single-shot detector) and SIMRDWN (satellite imagery multiscale rapid detection with windowed networks).

Zhu et al. [28] presented object recognition on drone-captured instances related to high-speed and low-altitude. The suggested Transformer Prediction Heads (TPH-YOLOv5) utilized cutting-edge approaches, such as a transformer encoder block and attention zone detection in circumstances with crowded objects. The authors also developed a convolutional block attention model (CBAM). In continuation, Jung et al. [29] proposed YOLOv5 that enhanced by replacing the focus layer with a 6*6 Conv2D layer and the SPP layer with an SPPF layer which detected objects in a variety of environmental and meteorological circumstances, including clear, cloudy, rainy, snowy days, evening, night, low altitude, and high altitude.

The rest of the paper is structured as follows: Section 3 introduces the proposed approach for video summarization based on object detection using YOLOv5, while Section 4 provides experiment analysis to demonstrate the accuracy and efficiency of the proposed algorithm followed by the conclusion in Section 5.

III. Proposed Work

The proposed framework works on a YOLOv5 based object of interest and collects all the frames with targeted objects such as person for the video summarization process. The Figure 1 shows the architecture for the proposed work. Firstly the frames are extracted from the Video and then object of interest

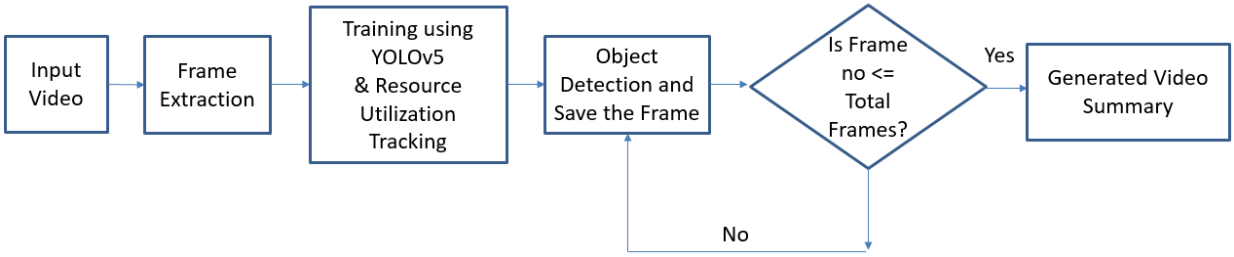


Fig. 1: Architecture of the Proposed work

detection is performed on the frames using Yolov5 as shown in Figure 2.

YOLO the one stage detector with advanced loss function is presented that performs dense detection (localization and classification) at the same time after feature extraction. YOLO returns coordinates of the bounding box and class of detected objects with the help of different parameters that are listed in Table I.

TABLE I: Different Notations

Notations	*Description*
x_{min},y_{min}	top left corner coordinate
x_{max},y_{max}	bottom right corner coordinate
W,H	width and height of an image
W_p,H_p	width and height of prediction box
x_c,y_c	normalized (center) coordinates
w_c,h_c	normalized width and height

Data deterioration prevention must be a fundamental responsibility in the object detection process, and this is accomplished by normalization. As a result, the YOLO parameters are normalized in relation to the center of the bounding box, as given in equation 1. While performing many tasks in computer vision, we use a "backbone" which is usually pre-trained on ImageNet. The backbone (Darknet 52) is used as a feature extractor, which gives us a feature map representation of the input. Neck is the subset of backbone which enhances the feature of discrimination. It also tells us about the ability and robustness of our set and head handles the prediction. The proposed work used PANet as a neck during object localization process and dense prediction for the single stage.

$$B_{YOLO} = \begin{cases} x_c \rightarrow \left(\frac{x_{max}+x_{min}}{2}\right) \\ y_c \rightarrow \left(\frac{y_{max}+y_{min}}{2}\right) \\ w_c \rightarrow \frac{W_p}{W} \\ h_c \rightarrow \frac{H_p}{H} \end{cases} \quad (1)$$

If an object of interest is found then the proposed work selects the frames with desired object (person) else discard it. After this object localization process, it generates the video summary for the obtained targeted frames. This approach is able to detect multiple objects from the extracted video frames.

IV. Result and Discussion

A. Dataset and Data Preprocessing

The proposed work used the Office and Lobby dataset in order to maintain the quality of summarization. The Office dataset has the four Non-Synchronized views with 3016 seconds duration while the Lobby dataset has three not fixed and Crowded views with a total 1482 seconds duration. The model training is performed on the both dataset while the summarization process is evaluated on the Lobby Dataset. Firstly the frame extraction is performed from both the dataset with 25 FPS. Then, some of the extracted frames are annotated using roboflow which have objects in it with their respective classes. The dataset is divided into training and validation sets with 281 and 65 images respectively. All the images are resized into 416×416 from 650×480 before the model training.

B. YOLOv5 based Model Training

The model training is performed on the divided dataset using the backbone Darknet 52 which uses the spatial pyramid pooling. The model is trained for 150 epochs using yolo 5s and batch size 8 over the google colab. The total duration for the training was recorded 0.399 hours over cloud and it used 157 layers, 7012822 parameters, 0 gradients and 15.8 GFLOPs as a model summary. The precision, recall, precision recall (PR) curve and mean average precision (mAP) are used for the model evaluation which are denoted by the Equation 2, 3, and 4. Precision is the ratio between correctly detected positive samples to the total number of positive samples while recall is the ratio between correctly detected positive samples to the actual positive samples. The PR curve shows the plotting between the precision and recall values and helps to select the best value that maximizes both precision and recall value. The mAP shows the area under the PR curve. The overall performance metrics are summarized in the Table II and plotted in figure 3.

$$Precision = TP/(TP + FP) \quad (2)$$

$$Recall = TP/(FN + TP) \quad (3)$$

$$mAP = (1/N) \times sum(AveragePrecision) \quad (4)$$

Fig. 2: Object Localization

where N denotes the number of classes.

TABLE II: Proposed Model Experimental Summary

S.No	*Parameter*	*Score*
1	best/epoch	118
2	best/mAP 0.5	0.98899
3	best/mAP 0.5:0.95	0.64733
4	best/precision	0.96926
5	best/recall	0.93643
6	metrics/mAP 0.5	0.98889
7	metrics/mAP 0.5:0.95	0.64712
8	metrics/precision	0.96926
9	metrics/recall	0.93663
10	train/box loss	0.01826
11	train/cls loss	0.0
12	train/obj loss	0.00919
13	val/box loss	0.03264
14	val/cls loss	0.0
15	val/obj loss	0.0097

C. System Statistics for Model Training

The proposed work also tracked the system resource utilization during the model training. The resource utilization is tracked using the GPU power usage, GPU memory access percent, GPU memory allocated percent, GPU temperature and network traffic as shown in Figure 4. GPU power usage denotes the GPU is busy less than 60 percent while GPU memory access is quite low which means it is spending the time for computation rather than fetching the data from memory. The GPU memory allocation has also recorded less then 20 percent over time during the training. GPU temperature is also recorded less than 75 °C which shows that it's not being overheated. The network traffic is traced in bytes which is also quite low due to single machine only.

The proposed model is validated on the Lobby dataset for three views. After Applying the trained model, it eliminated 2077 frames which have no object for the first view. It discarded 2185 and 1820 frames for the second and third views respectively. However, the proposed model captured the target object with satisfactory statistics and effective resource utilization but duplicate frames can be removed to generate better summary.

D. Comparison with Related Works

This section describes the comparative analysis of the proposed work with the existing approaches. Most of the approaches are based on the key frame selection and scene elimination rather than focusing on the object of interest. This shows the uniqueness of the proposed work for the video summarization. The proposed model found the 2870, 2759 and 3124 frames with the targeted object for the view 1, 2 and 3 of the lobby dataset. The generated summary duration for all the views are 114.8 110.36 124.96 seconds respectively which are best among all the existing approaches as shown in Table III.

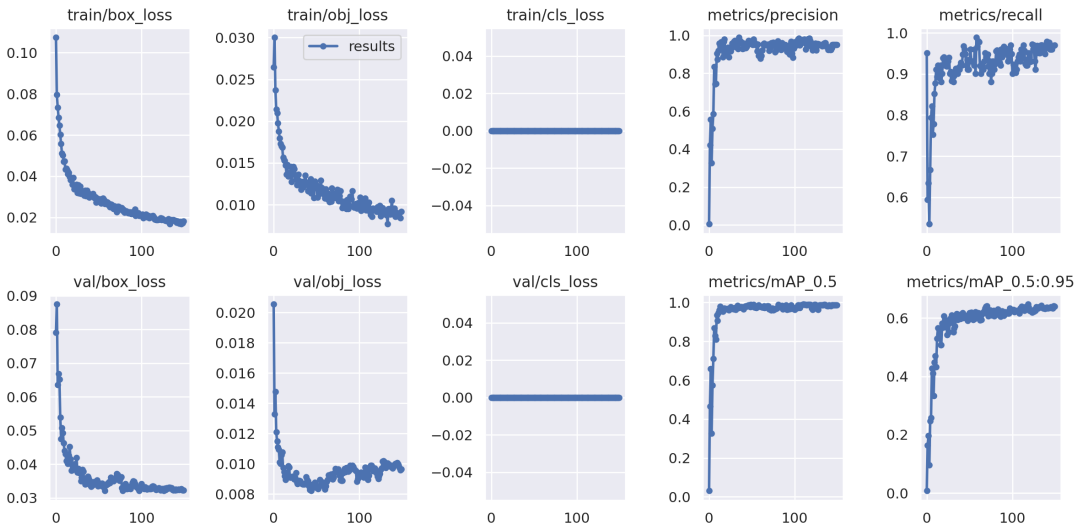
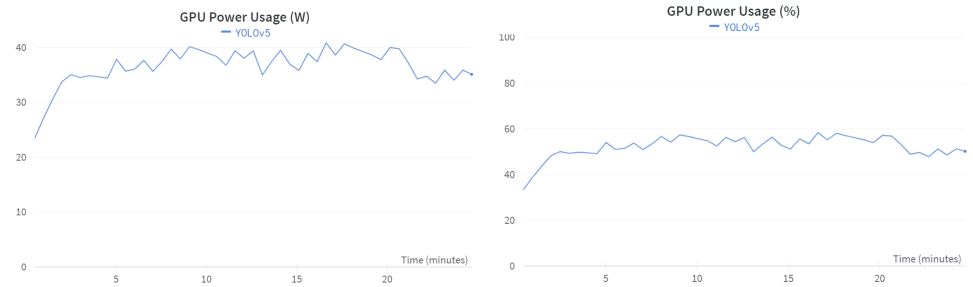
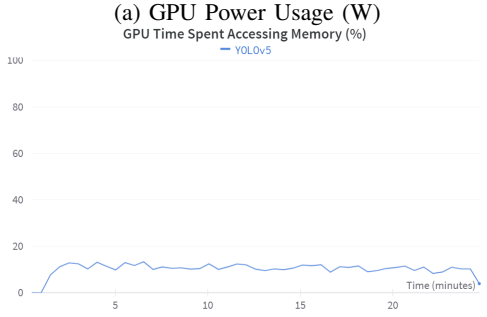


Fig. 3: Visulization of Performance Metrics

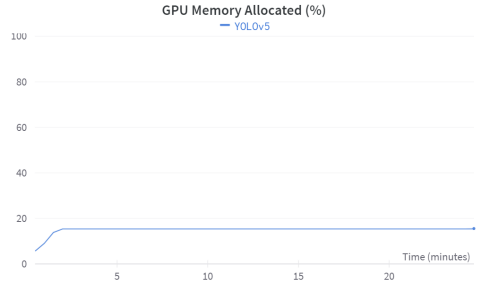


(a) GPU Power Usage (W)

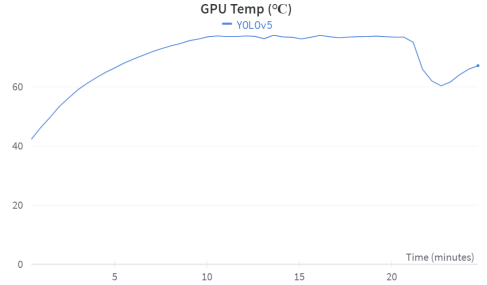
(b) GPU Power Usage (percent)



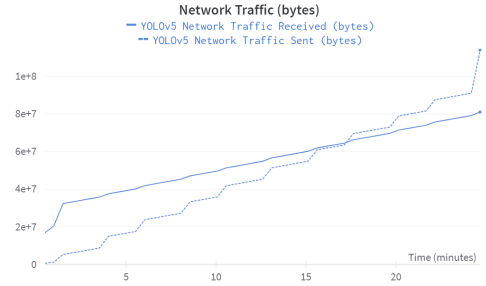
(c) GPU Time Spent Accessing Memory (percent)



(d) GPU Memory Allocated (percent)



(e) GPU Temp (°C)



(f) Network Traffic (bytes)

Fig. 4: System Resource Utilization during Training

TABLE III: Comparison with the Related Work

Views	Methods	Summary Length(s)
Lobby–0	[30]	225
Lobby–0	[23]	300
Lobby–0	[8]	113
Lobby–0	**Proposed**	**114.8**
Lobby–1	[30]	198
Lobby–1	[23]	303
Lobby–1	[8]	161
Lobby–1	**Proposed**	**110.36**
Lobby–2	[30]	206
Lobby–2	[23]	290
Lobby–2	[8]	150
Lobby–2	**Proposed**	**124.96**

V. Conclusion

Surveillance networks are pretty ubiquitous in today's modern world. These networks create 24-hour videos on a regular basis with substantial redundancy, wasting storage resources and making analysis harder. In this paper, we suggested an effective YOLOv5-based VS framework in response to these problems. Our framework initially extracts frames at 25 frames per second. YOLOv5 is used to identify the extracted frames that include objects. As a result, our system only communicates useful and needed data, which helps to save efficiency, usage, and processing resources. Furthermore, in the final summary, sequences with object-oriented frames of informativeness are included. Extensive studies, tracking of system resource utilization during training and comparisons shows that the propose work outperforms the other state–of–the–art models with best summary length. In future, the large duration multi-view videos can be processed and deployed in real time environment using distributed computing over cloud.

References

[1] S. Chakraborty, "A graph-based ranking approach to extract key-frames for static video summarization," *arXiv preprint arXiv:1911.13279*, 2019.

[2] D. M. Davids and C. S. Christopher, "An efficient video summarization for surveillance system using normalized k-means and quick sort method," *Microprocessors and Microsystems*, vol. 83, p. 103960, 2021.

[3] M. Birinci and S. Kiranyaz, "A perceptual scheme for fully automatic video shot boundary detection," *signal processing: image communication*, vol. 29, no. 3, pp. 410–423, 2014.

[4] A. Sahu and A. S. Chowdhury, "Summarizing egocentric videos using deep features and optimal clustering," *Neurocomputing*, vol. 398, pp. 209–221, 2020.

[5] E. Naveed and W. B. Sung, "Weighting low level frame difference features for key frame extraction using fuzzy comprehensive evaluation and indirect feedback relevance mechanism," *International Journal of Physical Sciences*, vol. 6, no. 14, pp. 3377–3388, 2011.

[6] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video summarization using deep neural networks: A survey," *Proceedings of the IEEE*, vol. 109, no. 11, pp. 1838–1863, 2021.

[7] S. Rani and M. Kumar, "Social media video summarization using multi-visual features and kohnen's self organizing map," *Information Processing & Management*, vol. 57, no. 3, p. 102190, 2020.

[8] K. Kumar and D. D. Shrimankar, "F-des: Fast and deep event summarization," *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 323–334, 2017.

[9] K. Kumar, "Evs-dk: Event video skimming using deep keyframe," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 345–352, 2019.

[10] J. Ren, J. Jiang, and Y. Feng, "Activity-driven content adaptation for effective video summarization," *Journal of Visual Communication and Image Representation*, vol. 21, no. 8, pp. 930–938, 2010.

[11] S. Antani, R. Kasturi, and R. Jain, "A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video," *Pattern recognition*, vol. 35, no. 4, pp. 945–965, 2002.

[12] C. Fredembach, M. Schroder, and S. Susstrunk, "Eigenregions for image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 12, pp. 1645–1649, 2004.

[13] R. Schettini, C. Brambilla, C. Cusano, and G. Ciocca, "Automatic classification of digital photographs based on decision forests," *International journal of pattern recognition and artificial intelligence*, vol. 18, no. 05, pp. 819–845, 2004.

[14] H. S. Munawar, "Reconfigurable origami antennas: A review of the existing technology and its future prospects," *Int. J. Wirel. Microw. Technol*, vol. 10, pp. 34–38, 2020.

[15] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf, "The isprs benchmark on urban object classification and 3d building reconstruction," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3 (2012), Nr. 1*, vol. 1, no. 1, pp. 293–298, 2012.

[16] L. Zhang, Z. Shi, and J. Wu, "A hierarchical oil tank detector with deep surrounding features for high-resolution optical satellite imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 10, pp. 4895–4909, 2015.

[17] M. Radovic, O. Adarkwa, and Q. Wang, "Object recognition in aerial images using convolutional neural networks," *Journal of Imaging*, vol. 3, no. 2, p. 21, 2017.

[18] T. Ahmad, Y. Ma, M. Yahya, B. Ahmad, S. Nazir *et al.*, "Object detection through modified yolo neural network," *Scientific Programming*, vol. 2020, 2020.

[19] A. Tahir, H. S. Munawar, J. Akram, M. Adil, S. Ali, A. Z. Kouzani, and M. P. Mahmud, "Automatic target detection from satellite imagery using machine learning," *Sensors*, vol. 22, no. 3, p. 1147, 2022.

[20] N. B. Aoun, M. Mejdoub, and C. B. Amar, "Graph-based approach for human action recognition using spatio-temporal features," *Journal of Visual Communication and Image Representation*, vol. 25, no. 2, pp. 329–338, 2014.

[21] W. Jiang, C. Cotton, and A. C. Loui, "Automatic consumer video summarization by audio and visual analysis," in *2011 IEEE international conference on multimedia and expo*. IEEE, 2011, pp. 1–6.

[22] J. Hannon, K. McCarthy, J. Lynch, and B. Smyth, "Personalized and automatic social summarization of events in video," in *Proceedings of the 16th international conference on Intelligent user interfaces*, 2011, pp. 335–338.

[23] S.-H. Ou, C.-H. Lee, V. S. Somayazulu, Y.-K. Chen, and S.-Y. Chien, "On-line multi-view video summarization for wireless video sensor network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 1, pp. 165–179, 2014.

[24] Y. Zhang, G. Wang, B. Seo, and R. Zimmermann, "Multi-video summary and skim generation of sensor-rich videos in geo-space," in *Proceedings of the 3rd Multimedia Systems Conference*, 2012, pp. 53–64.

[25] J. Kwon and K. M. Lee, "A unified framework for event summarization and rare event detection," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1266–1273.

[26] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. de Albuquerque, "Cloud-assisted multiview video summarization using cnn and bidirectional lstm," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 77–86, 2019.

[27] H. B. Ul Haq, M. Asif, M. B. Ahmad, R. Ashraf, and T. Mahmood, "An effective video summarization framework based on the object of interest using deep learning," *Mathematical Problems in Engineering*, vol. 2022, 2022.

[28] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2778–2788.

[29] H.-K. Jung and G.-S. Choi, "Improved yolov5: Efficient object detection using drone images under various conditions," *Applied Sciences*, vol. 12, no. 14, p. 7255, 2022.

[30] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *European conference on computer vision*. Springer, 2014, pp. 540–555.