INTERNATIONAL JOURNAL OF ADVANCED INNOVATIVE TECHNOLOGY IN ENGINEERING

Published by Global Advanced Research Publication House Journal

Home page: www.ijaite.co.in

Video Object Detection and Human Action Recognition: Techniques, Challenges, and Future Trends

[1]Kanchan S. Tidke, [2]Meenal G. Kachhavay, [3]Ashvini D. Nakhale

[1]Department of Electronics and Telecommunication Engineering, Dr. Rajendra Gode Institute of Technology and Research, Amravati, Maharashtra, India

[2]Department of Computer Science and Engineering, Dr. Rajendra Gode Institute of Technology and Research, Amravati, Maharashtra, India

[3]Department of Applied Science and Humanities, Dr. Rajendra Gode Institute of Technology and Research, Amravati, Maharashtra, India

[1]kanchantidke11@gmail.com, [2]minalashwinthakur15@gmail.com, [3]engineeringmathm@gmail.com

Article History

Received on: 25 Dec 2024

Revised on: 15 Jan 2025

Accepted on: 28 Jan 2025

Keywords: Object Detection, Human Recognition, Traffic Monitoring, Video Detection Systems

e-ISSN: 2455-6491

DOI: 10.5281/zenodo.14831793

Production and hosted by
www.garph.org

ABSTRACT

Video object detection and human action recognition have become essential components of modern computer vision systems. These techniques are widely utilized in scenarios such as traffic monitoring, industrial safety, and public surveillance. However, challenges like motion blur, occlusion, and video defocus hinder the accuracy and reliability of detection systems. This paper provides a comprehensive review of video object detection techniques, covering frame-based approaches, one-stage and two-stage detection algorithms, and mixed-stage methods. It also examines commonly used datasets, including ImageNet VID and YouTube-Bounding Boxes, to highlight their strengths and limitations. Furthermore, advancements in object detection, including deep learning methods like YOLO and Vision Transformers, are discussed. By identifying challenges such as real-time detection in resource-constrained environments and ethical concerns like data privacy, this study explores emerging trends and future directions in video detection systems. The findings aim to guide researchers and practitioners toward developing efficient, robust, and ethical video object detection solutions.

1. INTRODUCTION

Video object detection and human action recognition are used in many scenarios. These include recognizing vehicle plate numbers, detecting dangerous driving, spotting red-light violations, and identifying abnormal behaviors in industries, stations, and airports. Video detection faces challenges like defocus, motion blur, and occlusion. Video defocus occurs during focusing,

and object motion adds to this blur [1]. Occlusion happens when objects overlap. Additionally, object shapes may change with camera distance, making video detection harder than image detection.

Most video detection methods work by analyzing video frames. These methods break videos into frames and apply image detection techniques. The speed of video detection depends on the speed of frame detection. Some methods operate directly on videos but still rely on adjacent frames.

Earlier methods for image detection included HOG, SIFT, and Haar-like features. HOG extracts object outlines, calculates gradient histograms, and combines them into descriptors. SIFT identifies key points in images and is stable under lighting, noise, and transformation changes. Haar-like features use templates to slide over images and recognize features using classifiers like SVM and Random Forest. SVM uses hyperplanes for classification and offers high accuracy but requires more computation and storage. Random Forest uses multiple decision trees for classification and often achieves high accuracy [2].

Deep learning introduced loss functions like cross-entropy for classification. These functions detect and recognize classes by minimizing errors. Before deep learning, methods like SIFT and HOG lacked translation invariance. They extracted simpler features compared to deep learning models [3].

Detectors can be one-stage or two-stage. One-stage detectors are faster but less accurate. They combine feature extraction and classification. Two-stage detectors are slower but more accurate. They have separate feature extractors, often called backbones, and classifiers.

Video detection methods are categorized into three types. These include detecting frames using image detectors, focusing on key frames, or using temporal data between frames. Frame-by-frame methods rely on image detection and often skip key frame extraction.

2. OBJECT DETECTION SYSTEM

Recognizing objects plays a critical role in human daily life. Researchers have focused extensively on the challenging task of recognizing humans due to its wide range of applications across various fields. Different gestures convey specific messages; for instance, quick head movements often signify surprise or alarm, while visual attention within a group can indicate communication cues. Head pose estimation plays a significant role in speech recognition, helping identify messages from individuals with speaking or listening impairments.

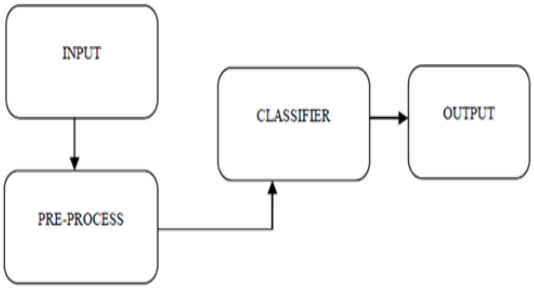

Figure 1: Basic step of object detection System

Visual Monitoring is a significant concern in computer vision research to identify, recognize, and track data across a series of images and understand and explain object detection by replacing the outdated method of human operators operating cameras [4]. A computer vision system can detect immediate unauthorized activity, and long-term suspicious activities are also possible, alerting a human operator to conduct a more thorough investigation. Video surveillance can be manual, semi-automatic, or fully automated. In a manual system, the control operator conducts all tasks while observing the visual information from the different cameras. In a semi-automatic system, the computer vision system assists the operator in certain tasks, and in a fully automated system, the computer vision system performs all tasks without human intervention.

A. Video-Based Datasets for Object Detection and Action Recognition

The commonly used video classification datasets are as follows: ImageNet VID dataset [5], which has 3862 snippets for training, 555 snippets for validation, and 937 snippets for testing. The dataset has 30 classes. These classes are carefully selected considering different factors, such as motion type, background interference, average number of objects, etc. Each frame of the video is annotated. Another video object detection dataset is the YouTube-Objects dataset [6], which was collected from YouTube and has 10 object classes. The videos in the dataset are formed as frames; these frames can be restored to videos if necessary. A video object dataset with artificial bounding boxes is the YouTube-Bounding Boxes dataset [7], which contains 380,000 19-s-long videos with 23 classes of objects. The quality of the video is similar to that of a mobile phone. Google Brain makes the project, and the dataset has 5.6 million human-annotated bounding boxes. A video object detection dataset used for urban geographic recognition is the Apolloscape dataset [8], which is provided by Baidu and includes RGB videos with high-resolution images and per-pixel annotations. The dataset defines 26 objects, such as cars, bicycles, pedestrians, buildings, street lights, etc.

CDnet2014 [9] is a video change detection dataset with 11 categories.

B. Video Frame-Based Object Detection Algorithms

Most video detection methods decompose the video into frames and use the image detection model to detect. Therefore, almost all image detectors can be applied for video detection. The other video detection methods utilize the correlation between frames and operate on adjacent frames. Some of the methods which operate on adjacent frames use LSTM-like models. The following discusses in detail.

One-Stage Video Object Detection: The current object detection methods are divided into two categories, one-stage object detection and two-stage object detection. In the two-stage object detection, feature extraction is the first stage, the classification is the second stage. One-stage object detection methods include YOLO [10], SSD [11] and RetinaNet [12]. Their common point is that the detection speed of a single frame is very fast, and real-time video detection can be implemented.

You Only Look Once (YOLO): YOLO [10] makes the object classification as "regression". In the training, YOLO will resize the images to a specific size, which can be set in the program. In the model, the nonlinear mapping between image features and neural network parameters is established. In the detection, the image or video is performed. The network structure of YOLO uses the structure of GoogLeNet [13] for classification, and replaces the inception modules of GoogLeNet with 1 × 1 and 3 × 3 convolutional layers, in order to simplify the structure and improve the detection speed. YOLO has 24 convolutional layers and 2 fully connected layers.

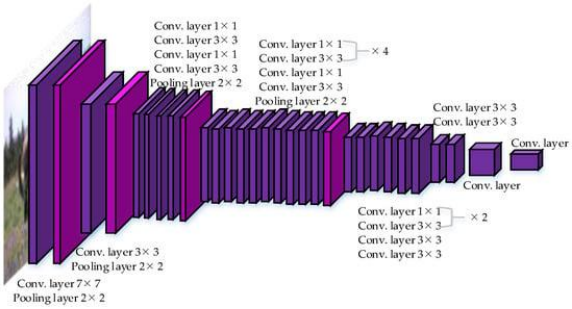

Figure 2: The network structure of YOLO

YOLO9000 (YOLOv2): YOLOv2 [14] uses a series of methods to improve detection accuracy and speed, and adopts strategies to enable YOLOv2 to detect more than 9000 objects. In addition, the basic framework of YOLOv2 is similar to YOLOv1. YOLOv2 uses the following methods to improve the detection speed: (A) YOLOv2 adopts Darknet19 as the detection neural network, which has 19 convolutional layers with 3 × 3 filter and 5 max pooling layers with doubling the number of channels compared with the previous layer. (B) YOLOv2 follows almost every 3 × 3 convolution layer with a 1 × 1 convolution layer, which may reduce the complexity of network computing and improve the detection speed. (C) YOLOv2 does not use the dropout layer, which may reduce the network computational complexity and help increase the network speed.

YOLOv3: YOLOv3 [15] still uses the framework of DarkNet, and the network uses the residual module and the multi-scale prediction. The multi-scale prediction is similar to Feature Pyramid Networks (FPN) [16]. Compared with YOLOv2, YOLOv3 uses more residual skip modules which reduces the loss of the information caused by convolution and pooling, making the network deeper, which can extract more advanced semantic features and improve the recognition accuracy. YOLOv3 uses multi-scale prediction to enhance the detection accuracy.

YOLOv4: The detection speed and detection accuracy of YOLOv4 [17] are improved, compared with YOLOv3. YOLOv4 has three parts: backbone, neck and head. The backbone is used for extracting features. The neck is used for transmitting the extracted features to the part of head. The head is used for object classification and bounding box regression.

YOLOv4 uses Cross Stage Partial Networks (CSP Darknet) [18] as the backbone. CSPNet solves the problem of gradient information duplication in other backbones, and integrates the gradient changes into the feature map, therefore, YOLOv4 reduces the parameter amount and FLOPS of the model, improves the detection speed and accuracy, and reduces the size of the model. CSPNet is based on the idea of DenseNet. CSPNet uses the shortcut connections for reducing the information loss in the transmission, effectively alleviates the gradient disappearance.

YOLOv4 uses PANet [19] as the neck. The neck can generate the feature pyramids. PANet is based on Mask R-CNN [20] and FPN [16]. The neck adopts a kind of FPN structure that enhances the bottom-up transmission, which improves the transmission of the bottom features. YOLOv4 uses the YOLOv3 detector as the head. The characteristic of the head is fast detection speed and high detection accuracy. In the head, each object class generates three kinds of anchor boxes, corresponding to the three different object scales and sizes. The structure of YOLOv4 is shown in Figure 3.

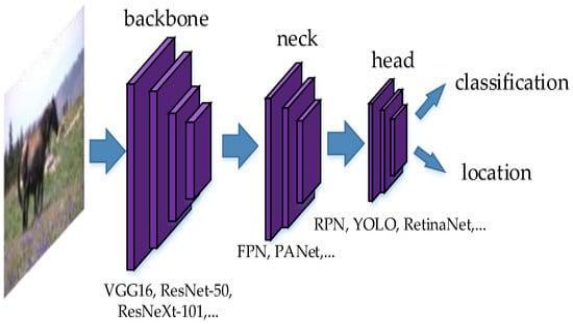

Figure 3; Structure of YOLOv4

C. Two-Stage Video Object Detection

Since video is composed of frames, theoretically, all two-stage image detection methods could be used for video detection by detecting the frames. In general, since the detection speed of the two-stage detector would be not very fast, this form of video detection cannot implement real-time detection.

Two-stage object detection has a separate module for extracting features and region proposals, which is called backbone. Therefore, the detection speed is slower than the one-stage detector, although the detection accuracy is always higher than the one-stage detector.

D. Mixed-Stage Video Object Detection

The mixed-stage object detection is a mixture of one-stage detection and two-stage detection, or other video detections which could not be classified as one-stage or two-stage detection. Minimum Delay video object detection [64] uses one-stage and two-stage image detector simultaneously, which can achieve real-time detection speed. The idea of Minimum Delay is the quickest detection theory. The "quickest detection" is to realize fast detection with a probability, by calculating the distribution variation of the video sequences. The "quickest detection" is implemented as the cumulative sum (CUSUM) algorithm. The algorithm of CUSUM integrates the feature map sampling values of the video sequence, and can aggregate the small deviations of the video sequence into a fluctuation. Therefore, CUSUM can detect the changes of the average value of the observed video sequences, and can overcome the signal-to-noise ratio threshold effect. The framework of Minimum Delay Video Object Detection is composed of CNN detector which is implemented frame by frame, an NMS module which is used to filter the inaccurate candidate boxes, the CUSUM module to implement the accurate and minimum delay detection. The CNN detector adopts ResNet [21], SSD, RetinaNet [12] in the experiments. The method improves the detection accuracy without reducing the detection speed. When using one-stage detector as the CNN detector, the framework can achieve real-time detection speed.

Usha Rani et. al. (2023) presented the study in-depth analysis of such video surveillance systems and presents a full assessment of methods and data sets utilized in human (object) detection. The most significant analyses of these systems are provided along with the employed architectures. To provide a clearer image and a comprehensive overview of the system, existing surveillance systems were compared in terms of their features, advantages, and challenges. These comparisons are summarized in this document. Future trends are also examined, laying the groundwork for new study avenues [22].

K. Visakha et. al. (2018) presented a video surveillance system is directed on automatic identification of events of interest, especially on tracking and classification of moving objects. A video surveillance system consists of three phases: moving object recognition, tracking, and decision making. This study focuses on detection of human beings in a scene, and tracking those people as long as they stay in the scene by identifying individual persons. Automating the video surveillance process will help in effortless monitoring of the sensitive areas with less human resource utilisation [23].

3. CURRENT TRENDS IN OBJECT DETECTION

Object detection has seen rapid advancements in recent years, particularly with deep learning. Some notable trends include:

Vision Transformers (ViT): Emerging models like DETR (DEtection TRansformer) are revolutionizing object detection by using transformers to directly predict bounding boxes without region proposal stages.

Advances in YOLO: The YOLO family (e.g., YOLOv7 and YOLOv8) has introduced features like improved detection speed, support for smaller objects, and multi-scale prediction enhancements.

Lightweight Models for Edge AI: Models like MobileNet and EfficientDet focus on balancing accuracy with computational efficiency for deployment on edge devices like drones and IoT cameras.

4. CHALLENGES AND RESEARCH GAPS

Object detection still faces several unsolved challenges, such as:

Real-Time Detection in Resource-Constrained Environments: Achieving real-time performance on devices with limited computational power.

Handling Occlusion and Motion Blur: Dealing with overlapping objects or motion artifacts in dynamic scenes.

Data Scarcity and Annotation Costs: The need for large, annotated datasets for training and the high cost of creating such datasets.

Domain Adaptation: Ensuring models trained on specific datasets (e.g., urban environments) perform well in other scenarios (e.g., rural or aerial views).

5. Application-Specific Customizations

Object detection models are often tailored to meet specific application requirements. Examples include:

Traffic Monitoring: Models like Faster R-CNN customized for detecting license plates or vehicles in crowded scenes.

Surveillance: Enhancements for low-light detection using thermal imaging combined with YOLO or SSD.

Healthcare: Modifications to detect tumors in medical imaging, such as combining ResNet with attention mechanisms for more accurate results.

6. Evaluation Metrics

Evaluation metrics are critical for comparing the performance of detection models. Include:

Mean Average Precision (mAP): Measures precision across different confidence thresholds.

$$mAP = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{AP_i}$$

$$Average\ Precision\ (AP) = \int_{0}^{1} \mathrm{P(R)dR}$$

Frames Per Second (FPS): Assesses the real-time capability of the model.

$$FPS = \frac{\mathrm{T}}{\mathrm{Time}}$$

Precision and Recall: Evaluate the model's ability to identify true positives and avoid false negatives.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

IoU (Intersection over Union): Measures the overlap between predicted and ground truth bounding boxes.

$$\mathrm{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

7. Emerging Technologies

Edge AI: Using lightweight models (e.g., MobileNet) for real-time detection on edge devices.

Federated Learning: Collaborative training of models across devices without sharing sensitive data.

Multimodal Detection: Combining video, audio, and other sensor data to improve detection accuracy.

8. Ethical and Privacy Concerns

Data Privacy: Protecting sensitive information in surveillance datasets.

Bias in Detection Models: Datasets often have biases (e.g., over-represented demographic groups), which can lead to unfair results.

Misuse of Technology: Risks of deploying detection systems in ways that infringe on individual freedoms or enable surveillance abuse.

9. Cross-Domain Applications

Explore how object detection techniques are applied in diverse fields, such as:

Healthcare: Tumor detection, analyzing X-rays, or tracking patient activity.

Sports Analytics: Tracking players or objects like balls during games.

Disaster Management: Using drones for detecting survivors or hazards in affected areas.

10. Future Trends and Directions

Identify potential areas for future research, such as:

3D Object Detection: Using LiDAR or stereo cameras to detect objects in 3D space for applications like autonomous vehicles.

Quantum Computing in Detection: Leveraging quantum algorithms to speed up detection tasks.

Zero-Shot and Few-Shot Learning: Enabling models to detect unseen or rarely-seen objects with minimal labeled data.

Sustainability in AI: Developing energy-efficient algorithms to reduce the environmental impact of large-scale detection models.

Conclusion

Video object detection continues to evolve, driven by advancements in deep learning and the growing demand for automated monitoring systems. This review highlights the significant progress achieved through algorithms such as YOLO, SSD, and RetinaNet, as well as the integration of temporal data for improved detection accuracy. However, challenges such as data scarcity, domain adaptation, and real-time processing on edge devices remain unsolved. Moreover, ethical considerations like bias in detection models and

privacy concerns in surveillance systems must be addressed. Future research should focus on developing lightweight and energy-efficient models, exploring multimodal detection approaches, and enhancing the generalization capability of detection systems. By addressing these challenges, the field can enable the creation of reliable, scalable, and socially responsible solutions that cater to diverse applications, including healthcare, urban safety, and disaster management.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

FUNDING SUPPORT

The author declare that they have no funding support for this study.

REFERENCES

[1] Bouafia, Y., Guezouli, L. & Lakhlef, H. Human Detection in Surveillance Videos Based on Fine-Tuned MobileNetV2 for Effective Human Classification. Iran J Sci Technol Trans Electr Eng 46, 971–988 (2022). https://doi.org/10.1007/s40998-022-00512-6

[2] Nouar AlDahoul, Aznul Qalid Md Sabri, Ali Mohammed Mansoor, "Real-Time Human Detection for Aerial Captured Video Sequences via Deep Models", Computational Intelligence and Neuroscience, vol. 2018, Article ID 1639561, 14 pages, 2018. https://doi.org/10.1155/2018/1639561

[3] Duan, Genquan & Ai, Haizhou & Lao, Shihong. (2010). Human Detection in Video over Large Viewpoint Changes. 6493. 683-696. 10.1007/978-3-642-19309-5_53.

[4] Brox, T.; Malik, J. Object Segmentation by Long Term Analysis of Point Trajectories. In Proceedings of the Computer Vision—ECCV 2010, Berlin/Heidelberg, Germany, 5–11 September 2010; pp. 282–295.

[5] Kristan, M.; Pflugfelder, R.; Leonardis, A.; Matas, J.; Cˇ ehovin, L.; Nebehay, G.; Vojírˇ, T.; Fernández, G.; Lukežicˇ, A.; Dimitriev, A.; et al. The Visual Object Tracking VOT2014 Challenge Results. In Proceedings of the Computer Vision—ECCV 2014 Workshops, Cham, Switzerland, 6–7 September 2014; pp. 191–217.

[6] Dendorfer, P.; Osep, A.; Milan, A.; Schindler, K.; Cremers, D.; Reid, I.; Roth, S.; Leal-Taixé, L. Motchallenge: A benchmark for single-camera multiple target tracking. Int. J. Comput. Vis. 2021, 129, 845–881

[7] Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.

[8] Kliper-Gross, O.; Hassner, T.;Wolf, L. The action similarity labeling challenge. IEEE Trans. Pattern Anal. Mach. Intell. 2011, 34, 615–621.

[9] Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-Scale Video Classification with Convolutional Neural Networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.

[10] Huang, X.; Cheng, X.; Geng, Q.; Cao, B.; Zhou, D.; Wang, P.; Lin, Y.; Yang, R. The ApolloScape Dataset for Autonomous Driving.In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1067–10676.

[11] Yavariabdi, A.; Kusetogullari, H.; Cicek, H. UAV detection in airborne optic videos using dilated convolutions. J. Opt.-India 2021, 50, 569–582

[12] Yavariabdi, A.; Kusetogullari, H.; Celik, T.; Cicek, H. FastUAV-NET: A Multi-UAV Detection Algorithm for Embedded Platforms. Electronics 2021, 10, 724.

[13] Wong, S.C.; Gatt, A.; Stamatescu, V.; McDonnell, M.D. Understanding Data Augmentation for Classification: When to Warp? In Proceedings of the 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Gold Coast, Australia, 30 November–2 December 2016; pp. 1–6.

[14] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.

[15] Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.

[16] Wang, Y.; Jodoin, P.; Porikli, F.; Konrad, J.; Benezeth, Y.; Ishwar, P. CDnet 2014: An Expanded Change Detection Benchmark Dataset. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Washington, DC, USA, 23–28 June 2014; pp. 393-400.

[17] He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2015, 37, 1904–1916.

[18] Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.

[19] Lao, D.; Sundaramoorthi, G. Minimum Delay Object Detection from Video. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 5096–5105.

[20] Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; Van Gool, L. The 2017 davis challenge on video object segmentation. arXiv 2017, arXiv:1704.00675.

[21] Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Gool, L.V.; Gross, M.; Sorkine-Hornung, A. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 724–732.

[22] Usha Rani, J., Raviraj, P. Real-Time Human Detection for Intelligent Video Surveillance: An Empirical Research and In-depth Review of its Applications. SN COMPUT. SCI. 4, 258 (2023). https://doi.org/10.1007/s42979-022-01654-4

[23] K. Visakha and S. S. Prakash, "Detection and Tracking of Human Beings in a Video Using Haar Classifier," 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2018, pp. 1-4, doi: 10.1109/ICIRCA.2018.8597322.