The Scientific Temper (2024) Vol. 15 (3): 2644-2650
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.32
https://scientifictemper.com/

RESEARCH ARTICLE

An optimized real-time human detected keyframe extraction algorithm (HDKFE) based on faster R-CNN

Rajeshwari D.*[1], C. Victoria Priscilla[2]

Abstract

The aim of this project is to support criminal investigators by utilizing surveillance camera footage in their investigations. To apprehend the culprit, it is necessary to examine the video footage and extract the relevant and crucial information. Analyzing lengthier videos might provide challenges due to the time needed to process the entire video while maintaining its semantic features. In this situation, a dataset is collected in real time to aid in the criminal investigation, which consequently requires the use of keyframes. The study illustrates that content-based video retrieval (CBVR) enables the video analysis technique. Keyframe extraction is a significant component of video analysis. The main objective of key frame extraction is to reduce the amount of repetitive frames in a video, thereby improving the clarity and efficiency of the scenario. Moreover, it optimizes video sequences to expedite processing. The study paper introduces the human detected keyframe extraction algorithm (HDKFE), which utilizes a dual-stage methodological approach. The faster region-convolutional neural network (Faster R-CNN) detects humans in surveillance by identifying frames that contain humans and reporting them using an optimized threshold value. The frames then identify a suitable keyframe by recognizing local maxima through the absolute difference between frames in the subsequent phase. This significantly decreases the complexity of long-term criminal investigations. The experimental report reveals that the HDKFE approach achieves a precision of 98.87% while minimizing both space and time complexity.

Keywords: Keyframe extraction, Faster R-CNN, Closed-circuit television, HDKFE, CBVR, Crime scene investigation.

Introduction

Closed-circuit television (CCTV) monitors many regions of a place and documents incidents occurring at homes and commercial security systems to precisely rebuild a crime scene during an investigation. This footage also serves as a valuable source for forensic Intelligence, such as (1). Visual depictions help reconstruct events. (2) It assists the crime scene investigation. (3) Clothing, bags, and weapons are shown. (4) Before and after the occurrence, the system can track the suspects. (5) Traceable goods, such as automobile registration plates, can be traced. CCTV surveillance not only reports the most catastrophic incidents, such as kidnapping, murder, and other specific crimes, but it also reports the person who committed the crime and returns to the site ordinarily. Such incidents are reported to occur daily with a large storage capacity, as major and minor crime scenes consume an enormous amount of time for an investigator to investigate. Investigators also struggle to 1) acquire massive footage in days. 2) The suspect may have socialized. 3) Bad analogy surveillance camera photos. 4) Centralized footage monitoring and higher frame-per-second (FPS) diminishes detail. This takes time and effort because surveillance must evaluate both significant and minor incidents to find the offender. These issues can be resolved through Content-based video retrieval (CBVR) (Patel, 2012). The automatic video content analysis through CBVR represents simulating, searching, retrieving, and studying surveillance data for keyframe extraction (Sima, 2021). The keyframe extraction method represents the primary content of the video sequence, as well as screening out keyframes based on the user's preferences supporting the crime scenes. Therefore,

[1]Research Department of Computer Science, Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women, Affiliated to University of Madras, Chennai, Tami Nadu,India.

[2]PG Department of Computer Science, Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women, Affiliated to University of Madras, Chennai, India

***Corresponding Author:** Rajeshwari D., Research Department of Computer Science, Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women, Affiliated to University of Madras, Chennai, Tami Nadu,India., E-Mail: rajeshwari.d@sdnbvc.edu.in

How to cite this article: Rajeshwari, D., Priscilla, C. V. (2024). An optimized real-time human detected keyframe extraction algorithm (HDKFE) based on faster R-CNN. The Scientific Temper, **15**(3):2644-2650.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.32

Source of support: Nil

Conflict of interest: None.

Received: 07/07/2024 **Accepted:** 08/08/2024 **Published** : 10/09/2024

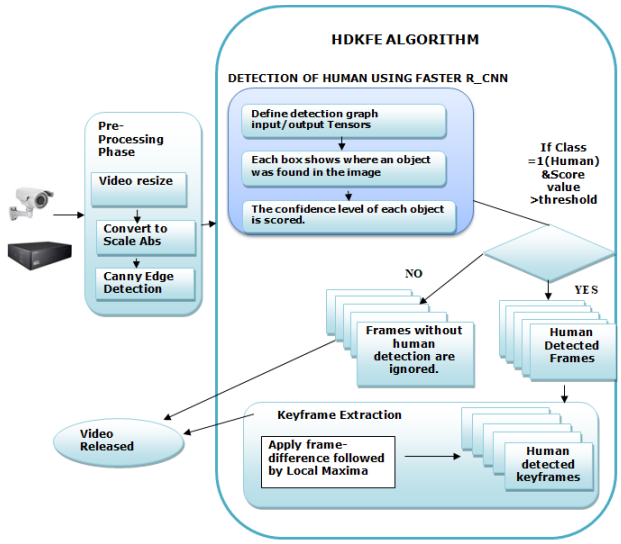

Figure 1: Overall framework of the enhanced methodology

the keyframe extraction technology can effectively eliminate redundant frames for subsequent information with sufficient storage space (Saoudi & Jai-Andaloussi, 2021).

The suggested HDKFE method analyses compressed real-time datasets of different frame sizes, extracting video frame identifiers and human motion data and delivering important keyframes. In this context, human motion detection is advanced using Faster RCNN, an essential element of video analysis that tracks and analyses motion by representing it as a rectangular box. The video, including human detection has been acquired, from which the human detection frames have been extracted. Thus, the keyframe extraction method is subsequently refined to eliminate redundant frames of real time dataset and reports on the essential human-detected keyframes, thereby streamlining the investigation process.

Methodology

The proposed methodology consists of two processes for extracting the keyframes that humans have detected. The following phases are outlined: The procedure consists of two phases: (1) the pre-processing phase and (2) the HDKFE phase, which involves the recognition of individuals, as depicted in Figure 1. Within this particular framework, pre-processing methods are utilized to enhance the quality of the video, then followed by the detection of moving objects utilizing Faster R-CNN. Video surveillance keyframes containing human subjects are recovered for criminal investigation purposes once the classification of humans is finished, distinguishing them from other moving objects using a different approach than the proposed one.

Pre-Processing Phase

The CCTV surveillance footage is now undergoing early pre-processing. During this stage, the video footage undergoes video scaling to a resolution of 640 × 480 in order to enhance the speed of detection (Joshi *et al*., 2023). The video undergoes additional processing to adjust contrast and correct brightness in the images. The alpha parameter governs the degree of contrast, whereas the beta parameter governs the level of brightness in the image. This approach employs the canny edge detection method, known for its high accuracy and robustness in recognizing edges while simultaneously reducing noise.

HDKFE Method

The human-detected keyframe extraction (HDKFE) method comprises two stages. (1) Optimized faster R-CNN and (2) Keyframe extraction method. The initial phase involves the utilization of the faster R-CNN with an optimized threshold value to identify frames, including humans. In the subsequent phase, a keyframe extraction approach is employed to calculate the necessary keyframe.

Faster R-CNN

The main goal of the faster R-CNN network is to create a comprehensive framework for accurately identifying and locating specific individuals in surveillance footage. The merging of convolutional neural networks (CNNs), region proposal networks (RPNs), and deep learning approaches into a single network improves both the speed and accuracy of the model (Tan *et al*., 2021). Figure 2 depicts the suggested framework. The approach consists of a total of three stages: the region-proposed network (RPN) is used for recommending novel regions, the region-based convolutional neural network (R-CNN) is employed for recognizing objects, and the process of training involves classifying individuals in video frames.

The initial stage involves extracting convolutional features from the input image through the faster R-CNN pipeline. This pipeline employs a pre-trained CNN backbone, such as ResNet or VGG. The RPN employs these characteristics to displace a condensed network, usually composed of a small number of convolutional layers, across the feature map in order to produce recommendations for specific regions. The RPN generates a probability of object scores and also creates a collection of anchor boxes. The ratings (Kim & Cho, 2021) represent the likelihood of an object, like a person, being found in each box. The RPN then sends the region concepts to the R-CNN for additional processing. The recommendation for each region is obtained by extracting it from the feature map and then resizing it to a certain dimension before being fed into the R-CNN. The R-CNN comprises an assortment of layers that are interconnected to extract features, terminating in a final softmax layer for categorizing objects (Akshatha *et al*., 2022). R-CNNs are tasked with discerning the presence of a human in each designated site for person detection.

The Faster R-CNN model is trained to improve the creation of region proposals and object classification simultaneously

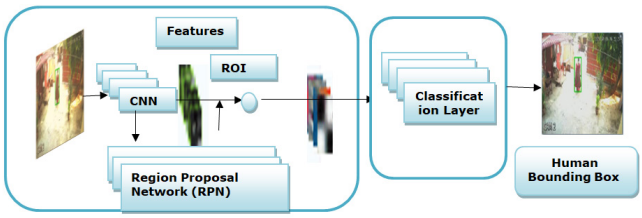

Figure 2: Faster R-CNN architecture for proposed methodology

(Angadi & Nandyal, 2021). A comprehensive collection of annotated college recordings, including human subjects, is currently being compiled for the intended research. The aim of this study is to discern the distinct attributes that distinguish individuals from the surrounding chaos in their surroundings. The Faster R-CNN model improves its capacity to accurately detect individuals in previously seen images by employing gradient descent and backpropagation with recurrent weight updates (Ushasukhanya & Karthikeyan, 2022).

Keyframe extraction

There are numerous benefits to using a faster R-CNN with keyframe extraction for object detection in videos (Sinulingga & Kong, 2023). Keyframe extraction reduces the computational load by analyzing a selection of frames, which increases the effectiveness of object detection, especially in lengthy video sequences (Man & Sun, 2022). Keyframes in videos are certain frames designed to highlight important visual information. They are very important in allowing faster R-CNN to precisely identify objects at pivotal periods in the video (Kumar *et al*., 2022).

The keyframe extraction process detects significant changes or deviations in visual content by means of the absolute difference approach, using local maxima. This work computes, in a video sequence (Bharathi *et al*., 2023) the absolute difference between consecutive frames. Local maxima are points when the absolute difference reaches its maximum point, therefore indicating important visual or content changes detectable via Local Maxima Detection. Using the local maxima in the distribution of absolute differences, one can identify frames with significant changes and thereby detect video transitions (Lv *et al*., 2021). Equation (1) generates the formula for computing the distribution of absolute differences in order to find local maxima between frames i and i-1.

$$AD(i) = F(i) - F(i-1) \quad (1)$$

Local maxima in the absolute difference distribution must be identified following the calculation of absolute differences for all subsequent frames. To identify local maxima, compare the absolute difference values at frame i to those at frames i-1 and i+1. A local maximum is present when the absolute difference at frame i is greater than that at frames i-1 and i+1. Figure 3 illustrates that if AD(i+1)is greater than AD(i-1) and AD(i)is greater than AD(i+1) then frame i may represent a keyframe and local maximum.

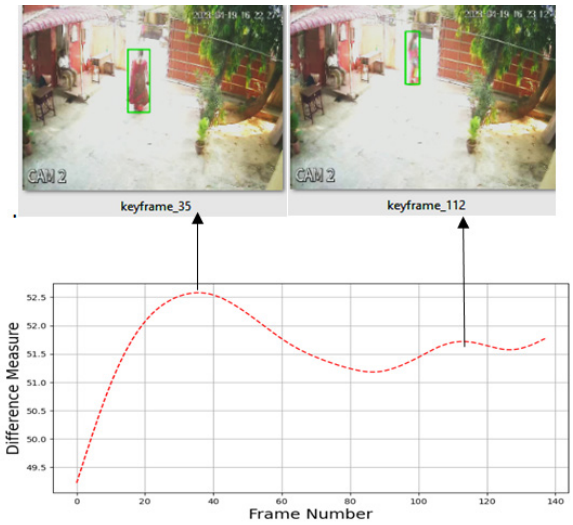

Figure 3: Keyframe selection using local maxima

Results and Discussion

Dataset

Our college site, namely Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women, is where the data was collected. Ten one-minute video recordings in total are used in the suggested method. Videos consist of a sequence of frames, with a total count of 7480. A 2.4MP infrared bullet camera with a 20 m range is used on campus by the CP-URC-TC24PL2-V3 CCTV system. It has a 1/2.7" 2.4MP CMOS image sensor that is 0.9407 cm in size. With an infrared range of 20 m, the device is fitted with a fixed 3.6 mm lens. The camera can capture images at a resolution of 2.4 MP at a rate of 25/30 frames per second.

Results and Evaluation

OpenCV also referred to as the Open Source Computer Vision Library, is a freely available software framework used in the fields of computer vision and machine learning. In this particular context, the software provides support for a programming language, especially Python, which is utilized to perform the required tasks.

In this study, the experimental analysis of real-time dataset is used, which is compared with three distinct approaches, including human detected (HD) - Histogram of oriented gradients-support vector machine (HOG-SVM) (Abed *et al*., 2021) with background subtraction, HD-Faster RCNN and HDKFE method. The suggested methodology consists of two components: optimized faster R-CNN

Table 1: Performance measures of human detected keyframes

Method	*Extracted frames*	*Keyframes reported*	*Human detected keyframes*
HD- HOG-SVM	7502	93	27
HD- Faster R-CNN	7500	88	48
HDKFE	3606	55	55

Table 2: Performance metrics of keyframes extracted

Surveillance footages	*Frames Obtained*	*HOG-SVM with background Subtraction*		*HD- Faster R-CNN*		*HDKFE*	
		Keyframe	*Human detected keyframe*	*Keyframe*	*Human detected keyframe*	*Keyframes Reported*	*Human detected keyframes*
Footage 1	489	6	3	4	2	7	7
Footage 2	643	6	2	6	3	8	8
Footage 3	598	15	3	13	6	5	5
Footage 4	639	7	6	7	7	6	6
Footage 5	709	9	5	10	9	10	10
Footage 6	956	10	2	10	7	8	8
Footage 7	742	9	1	8	5	3	3
Footage 8	886	11	2	9	2	2	2
Footage 9	730	11	1	10	3	3	3
Footage 10	747	9	2	11	3	3	3

and HDFKE algorithm. The findings presented in Table 1 demonstrate the setup with different outcomes.

This proposed methodology will be advantageous for criminal investigation, as evidenced by the real-time dataset, which exclusively emphasizes the human in keyframe reports, as the primary suspect to be identified by investigators is an individual. This also demonstrates the spatial complexity of the suggested methodology for identifying keyframes that contain human detection.

The performance metrics in Table 2, which together provide an extensive report, show significant differences between these three techniques. In this case, analyzing and presenting the most important keyframes takes a lot of time when there is more human presence in the video footage. This is achieved by utilizing the optimized threshold value, as demonstrated in the Faster R-CNN method. If the video has a minimal amount of human presence, fewer and more accurate keyframes are reported when compared to other methods. Thus, the optimized faster R-CNN effectively captures the significant frames without any unnecessary replication, identifying the keyframes where humans are spotted, as seen in Figure 3.

The comparison between the graphical representation of keyframe extraction and the human-detected keyframe in Figures 4 and 5 demonstrates that the suggested methodology yields higher accuracy in keyframe identification compared to the existing method.

Figure 6 represents one of the real-time footage, specifically video footage 8, which consists of a total of 886 video frames, with 131 frames being detected as containing humans using the HDKFE approach, among which only 2 are the most required human-detected keyframes, others are represented as redundant frames. Therefore, the suggested methodology efficiently aids in criminal investigation and adequately fulfills the necessary criteria.

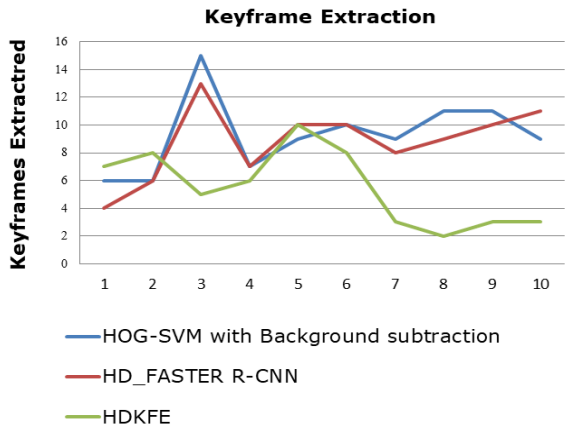

Figure 4: Metrics of keyframe extraction

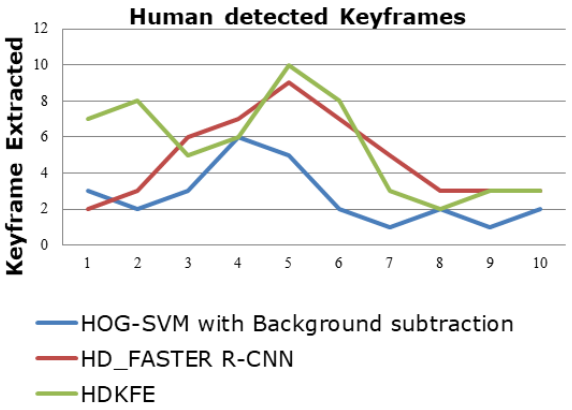

Figure 5: Metrics of human-detected keyframe extraction

This approach utilizes the VGG-16 pre-trained CNN model to train the frames that humans have spotted. The dimensions of all these frames are adjusted to 224×224×3 in order to match the input image, and then they are forwarded to the input layer. The hidden layer is subsequently convoluted

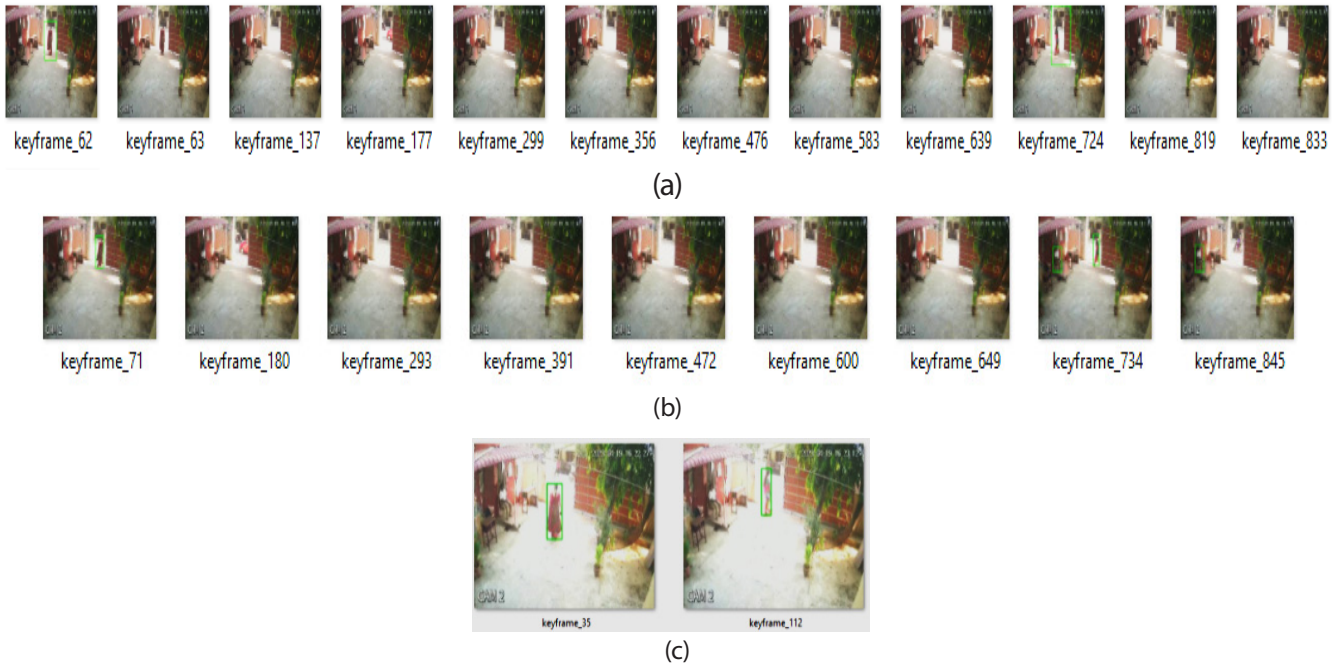

Figure 6: Resultant keyframes of video footage 8(a) HOG-SVM with background subtraction (b) Faster R-CNN (c) HDKFE method

three times with a dropout rate of 0.5, and the output layer is constructed using the Softmax function. The accuracy of the HDKFE model in Table 3 is much enhanced by compiling it using the Adam optimizer. This improvement from the obtained frames is reflected in the training frame as 70% and the testing frame as 30% in the stated precision, recall, and compression ratio.

Precision and recall

As determined by Eq. (2), precision denotes the percentage of positive instances that were accurately identified relative to the total number of instances predicted as positive.

$$Precision = \frac{TPH}{TPH+FPH} * 100\% \quad (2)$$

Here, TPH denotes true positive human and FPH denotes false positive human

Recall, as defined in Eq. (3), quantifies the ratio of correctly detected true positive occurrences to the total number of actual positive instances.

$$Recall = \frac{TPH}{TPH+ FNH} * 100\% \quad (3)$$

Here, FNH denotes False Negative Human.

Compression ratio

This metric quantifies the decrease in file size attained by predominantly representing the video content through keyframes as opposed to utilizing each frame. It showcases the efficacy of keyframe extraction in minimizing file size without compromising video quality, as represented by Eq. (4).

$$CR = 1 - \left\{\frac{HKF}{HF}\right\} * 100 \quad (4)$$

Here, HKF represents the number of human-detected keyframes and HF represents the total human-detected frames.

Consequently, the HDKFE method's accuracy metrics as in Table 3, produce an average compression ratio of 98.87%. As a result, the space complexity was demonstrated, as the resultant values of human-detected frames are lower than those of genuine frames.

Furthermore, upon analysis, the difference in processing time between the original video and the resulting video, as illustrated in Figure 7, reveals a significantly reduced degree of variation, indicating a lower level of time complexity. Thus, the experimental analysis of the HDKFE method demonstrates that it aids in criminal investigations with greater precision and reduced spatial and temporal complexity.

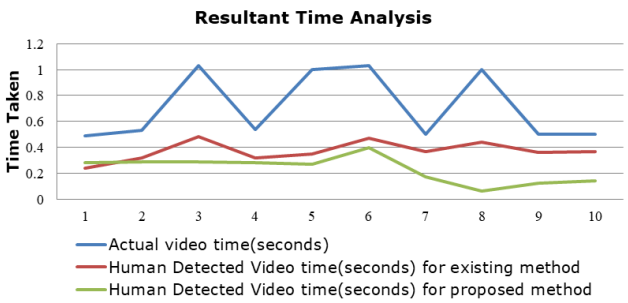

Figure 7: Resultant time analysis

Table 3: Performance metrics of HDKFE method

Surveillance Video	*Actual video time (seconds)*	*Resultant HD-Faster R-CNN video time (seconds)*	*HDKFE METHOD*					
			Resultant video time(seconds)	*Frames obtained*	*Human detected Frames*	*Precision (%)*	*Recall (%)*	*CR(%)*
Video 1	0.49	0.24	0.28	489	489	93.65	95.01	98.97
Video 2	0.53	0.32	0.29	643	292	93.84	91.04	98.86
Video 3	1.03	0.48	0.29	598	431	97.05	98.01	98.91
Video 4	0.54	0.32	0.28	639	424	99.02	99.02	98.78
Video 5	1.00	0.35	0.27	709	421	98.01	99.00	98.58
Video 6	1.03	0.47	0.40	956	625	92.05	99.29	99.01
Video 7	0.50	0.37	0.17	742	290	92.75	98.46	98.97
Video 8	1.00	0.44	0.06	886	131	90.62	96.67	98.96
Video 9	0.50	0.36	0.12	730	243	99.01	98.23	98.77
Video 10	0.50	0.37	0.14	747	260	95.16	98.33	98.90

Conclusion

This study presents a technique for extracting keyframes from real-time surveillance video datasets based on the presence of humans. The proposed HDKFE method employs two configurations, one utilizing faster R-CNN with an optimized threshold value and another one is the keyframe extraction using local maxima. This configuration selectively identifies only the frames containing humans and successfully extracts keyframes. This approach effectively solves the issue of excessive frame redundancy and provides a comprehensive evaluation of its space complexity. As a result, it reduces the duration of the resultant video and establishes the time complexity. Thus, the HDKFE method seeks to promptly identify humans and offers an effective technique for efficient video summarization and content representation in real-time datasets, as opposed to other methods. This HDKFE approach can be utilized for crime footage to facilitate the identification of suspects in criminal investigations by investigators. Consequently, this can improve the understanding and investigation of material in various video datasets.

Acknowledgment

I would like to express my gratitude to my supervisor, Dr. C. Victoria Priscilla, for her invaluable assistance and mentorship. The cooperation provided by Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women in collecting real-time datasets to support my research is very noteworthy.

References

Abed, R., Bahroun, S., & Zagrouba, E. (2021). KeyFrame extraction based on face quality measurement and convolutional neural network for efficient face recognition in videos. *Multimedia Tools and Applications*, *80*(15), 23157–23179. https://doi.org/10.1007/s11042-020-09385-5

Akshatha, K. R., Karunakar, A. K., Shenoy, S. B., Pai, A. K., Nagaraj, N. H., & Rohatgi, S. S. (2022). Human Detection in Aerial Thermal Images Using Faster R-CNN and SSD Algorithms. *Electronics (Switzerland)*, *11*(7), 1–15. https://doi.org/10.3390/electronics11071151

Angadi, S., & Nandyal, S. (2021). Human Identification Using Histogram of Oriented Gradients (HOG) and Non-Maximum Suppression (NMS) for Atm Video Surveillance. *International Journal of Innovative Research in Computer Science & Technology*, *9*(3), 1–10. https://doi.org/10.21276/ijircst.2021.9.3.1

Bharathi, S., Senthilarasi, M., & Hari, K. (2023). Key Frame Extraction Based on Real-Time Person Availability Using YOLO. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, *14*(2), 31–40. https://doi.org/10.58346/JOWUA.2023.I2.003

Joshi, S., Kadlag, G., & Deshmukh, S. (2023). Real-Time Detection of Anomalous Behavior in Cctv Videos Using Advanced Machine Learning Techniques. *International Research Journal of Modernization in Engineering Technology and Science*, *05*, 6447–6451. https://doi.org/10.56726/irjmets40069

Kim, J., & Cho, J. (2021). Rgdinet: Efficient onboard object detection with faster r-cnn for air-to-ground surveillance. *Sensors*, *21*(5), 1–16. https://doi.org/10.3390/s21051677

Kumar, S., Kumar, K., Menon, M. V., Reddy, K. A., Yadav, B. M., & Pavan, B. M. N. S. (2022). People Counting in Crowd: Faster R-CNN. *International Journal for Research in Applied Science and Engineering Technology*, *10*(6), 928–934. https://doi.org/10.22214/ijraset.2022.43989

Lv, C., Li, J., & Tian, J. (2021). Key Frame Extraction for Sports Training Based on Improved Deep Learning. *Scientific Programming*, *2021*. https://doi.org/10.1155/2021/1016574

Man, G., & Sun, X. (2022). Interested Keyframe Extraction of Commodity Video Based on Adaptive Clustering Annotation. *Applied Sciences (Switzerland)*, *12*(3). https://doi.org/10.3390/app12031502

Patel, B. V. (2012). Content Based Video Retrieval Systems. *International Journal of UbiComp*, *3*(2), 13–30. https://doi.

org/10.5121/iju.2012.3202

Saoudi, E. M., & Jai-Andaloussi, S. (2021). A distributed Content-Based Video Retrieval system for large datasets. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00479-x

Sima, M. (2021). Key frame extraction for human action videos in dynamic spatio-temporal slice clustering. *Journal of Physics: Conference Series*, *2010*(1). https://doi.org/10.1088/1742-6596/2010/1/012076

Sinulingga, H. R., & Kong, S. G. (2023). Keyframe Extraction for Reducing Human Effort in Object Detection Training for Video Surveillance. *Electronics (Switzerland)*, *12*(13). https://doi.org/10.3390/electronics12132956

Tan, L., Huangfu, T., & Wu, L. (2021). Comparison of YOLO v3, faster R-CNN, and SSD for real-time pill identification. *ArXiv*. https://doi.org/10.21203/rs.3.rs-668895/v1

Ushasukhanya, S., & Karthikeyan, M. (2022). Automatic human detection using reinforced faster-rcnn for electricity conservation system. *Intelligent Automation and Soft Computing*, *32*(2), 1261–1275. https://doi.org/10.32604/iasc.2022.022654