Key Frame Extraction Based on Real-Time Person Availability Using YOLO

S. Bharathi[1*], M. Senthilarasi[2] and K. Hari[3]

[1*]Assistant Professor, Department of Electronics and Communication Engineering,
Dr. Mahalingam College of Engineering and Technology, Pollachi, India.
bharathi_mani@yahoo.com, Orcid: https://orcid.org/0000-0001-9638-3779

[2]Assistant Professor, Department of Electronics and Communication Engineering,
Thiagarajar College of Engineering, Madurai, India. arasiece@gmail.com,
Orcid: https://orcid.org/0000-0001-9053-8485

[3]Junior Research Fellow, Department of Electronics and Communication Engineering,
Dr. Mahalingam College of Engineering and Technology, Pollachi, India.
harikrishnasamy@gmail.com, Orcid: https://orcid.org/0009-0003-7750-7707

Received: March 05, 2023; Accepted: May 12, 2023; Published: June 30, 2023

Abstract

Keyframe extraction plays a crucial role in summarizing lengthy videos, particularly in the context of surveillance footage with a fixed field of view that records events over extended periods. The process of manually reviewing such videos can be time-consuming and challenging to extract essential information effectively. To address this issue, a study was conducted to evaluate four distinct methods for keyframe extraction, with the aim of determining the most suitable approach for creating a people-based dataset. The four methods assessed in the study were absolute difference, entropy, optical flow, and object detection-based video summarization using the YOLO (You Only Look Once) algorithm. Each method offers a unique approach to identify keyframes that encapsulate critical instances within the video footage. Among all the four evaluated methods, the object detection-based video summarization approach stood out as particularly promising. This method employed the YOLO algorithm, which utilizes advanced object detection techniques to identify and track people within the video frames. With this approach only fewest numbers of frames are extracted but still capturing all the relevant instances featuring people. The results of this study suggest that object detection-based video summarization using the YOLO algorithm is a highly effective method for keyframe extraction in surveillance videos. By significantly reducing the number of frames while preserving all relevant instances, this approach offers a time-efficient solution for reviewing and analyzing extensive video footage, ultimately facilitating the creation of a people-based dataset for further research and applications in various domains.

Keywords: Key Frame, Surveillance Video, Absolute Difference, Entropy, Optical Flow, Yolo.

1 Introduction

The exponential growth of multimedia devices such as mobile phones, tablets, and personal computers, along with the increasing usage of social media, has resulted in a surge of videos available in the online

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 14, number: 2 (June), pp. 31-40. DOI: *10.58346/JOWUA.2023.I2.003*

*Corresponding author: Assistant Professor, Department of Electronics and Communication Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi, India.

environment (Survey on electronic gadgets and its effects, 2023). Video summarization, specifically key frame extraction, has become a crucial method in recent times. Key frame extraction is an essential technique that involves representing a video using as few frames as possible, thus reducing redundancy and computation. There are various methods for key frame extraction for different applications listed below.

Video compression: Here, key frame extraction helps in reducing the size of a video file by reducing the number of frames to be stored. This results in faster transmission and reduced storage space requirements. Video analysis: In the area of video analysis, key frames can be used to quickly and effectively analyze the content of a video. By selecting a representative frame for each significant change in the video, it becomes easier to understand the events that took place in a shorter amount of time (Lei, S., 2014). Video retrieval: In video retrieval the key frame extraction can be used to index and retrieve videos from large video databases. By extracting key frames and storing them as thumbnails, users can quickly preview the video content and decide if it is relevant to their needs. Surveillance: In the case of video surveillance, key frame extraction is useful for summarizing long hours of video recordings and reducing the time required to analyze important incidents.

The two common types of camera used for video recording are stable cameras and First-Person Vision (FPV) cameras. Stable cameras provide a static field of view with stable picture quality, while FPV cameras record rapidly changing scenes. Key frame extraction is commonly used in video surveillance cameras as it reduces the time required to analyze important incidents.

In many cases, video surveillance cameras are used for continuous recording and to provide critical information when needed. Analyzing each frame manually to decide the key frame is time-consuming and may result in missing important information. Key frames are typically extracted by identifying visual discontinuity in the video sequence using various methods. However, this method is not effective in capturing minimal changes in the frames.

This paper proposes key frame extraction from a surveillance video when people are present in a specific area. The extracted data will be used to summarize long surveillance videos and to create a people-based dataset for various surveillance applications. The remainder of the paper is structured as follows. Section II reviews the related research on key frame extraction. The proposed key frame extraction techniques are discussed in Section III and section IV presents the experimental results obtained with our model. Finally, Section V concludes the paper and highlights future work.

2 Related Work

The summarization of video is essential in surveillance video to reduce the time spent in analyzing the video to extract the required essential information. The general method of extracting key frames from a lengthy video is shown in the fig.1. The keyframe extraction methods are mainly used for various purposes but the main objective is to provide access to the video content in less time.

The key frame extraction different researchers used different methods: Deep prior and contextual saliency (Chen, J., 2021), canonical correlation and mutual entropy, difference of frames (Chen, L., 2017), block matching and spatial-temporal (Chen, C., 2017), background subtraction (Huang, C., 2019), low rank and sparse representation (Chen, Y., 2019) (Zhao, H., 2019), visual odometry (Lin, X., 2019) (Shih, H.C., 2013), LSTM and SLAM based, etc. but all these models follows the same pattern that's the video is converted into frames then additional information like motion sensor signal, etc., are added if need for extraction. There are various keyframe extraction models are used for extracting the significant frames.

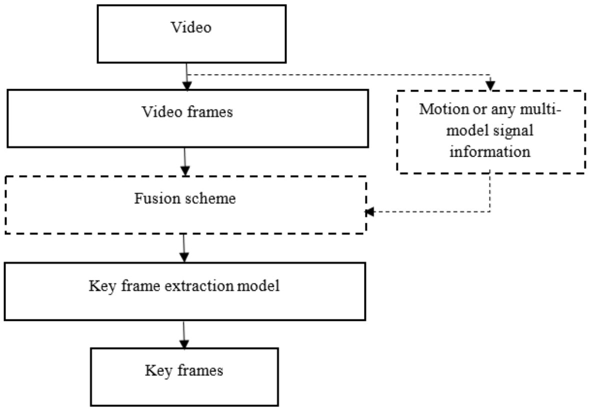

Video
Video frames
Motion or any multi-model signal information
Fusion scheme
Key frame extraction model
Key frames

Figure 1: General Structure of Keyframe Extraction

Out of various methods of key frame extraction few interesting methodologies are discussed below for creating surveillance video based keyframe extraction.

Yujie Li et al. (Li, Y., 2020) proposed novel key frame extraction for First Person Vision (FPV) videos by using multi-model sensor signals. Two effective models namely a sparse-based model and a graph-based model have been introduced and their proposed model uses the combination of visual scenes and motion information to provide better results for FPV videos. Here two datasets are used. Frist dataset is the Carnegie Mellon University Multimodal Activity (CMU-MMAC) database with multi-model information such as video, audio, motion capture and accelerations. There are five different recipes making is available in the dataset: brownie, salad, pizza, scrambled eggs, and sandwich; only brownie dataset used for evaluation. The second dataset is the Daily activity dataset which contains the activities of eight persons; all these test subjects wore wearable motion sensors that included three-axis accelerometers, gyroscopes, and a wearable camera. The results show that the model has produced enhanced performance over pure video data because of additional information like sensor signals and for integration of all these data, the model will take more time comparatively.

Qi Zhong et al. (Zhong, Q., 2020) proposed a Visual Geometry Group (VGG) based image saliency extraction model assisted with multi-feature and deep prior information. A new visual attention model is produced to capture moving objects clearly in sports videos and hence this is used for video post-processing. This algorithm is tested with the pedestrian video and easily grabbed the motion targets for post-processing. Their proposed algorithm performed well with the moving targets compared to the optical flow method and color histogram clustering.

Zhe Wang et al. (Wang, Z., 2020) introduced a novel motion vector based key frame extraction technique with rich motion information. Firstly, the entropy and two-dimensional entropy are calculated and a combination of these two is taken to know the difference between two frames and outliers are detected then using the Vibe algorithm, foreground objects in the frame are also detected. Finally, the motion vector is introduced for block matching and hence the active and inactive layers are detected in an adjacent frame then based on the similarities the keyframes are selected.

Xiaowei Gu et al. (Gu, X., 2020) proposed Sentiment key frame extraction using user generated macro-videos by using low rank and sparse representation. The sentiment features are extracted using Deep Senti bank fc7 layers and the global video information of distinct frames are represented using low rank and sparse representations and for C3D, LSTM or SVM are used for sentiment classification as happy, neutral, and sad. Yotube8 and MOUD datasets are used for evaluating the model. Due to the unsteady and continuous changing in the frames, the model produces below the expected results.

Kyung Min Han et al. (Han, K.M., 2020) presented a camera tracking system called "KeySLAM" that uses RGB-D cameras (cameras that capture both RGB image data and depth data) to robustly track the camera's 6-DoF pose in real-time. It combines an adaptive visual odometry (VO) approach with an optimal key-frame selection algorithm to improve tracking performance. The system adapts to changing environments and can handle occlusions, making it suitable for use in real-world applications.

Mohammad Shokri et al. (Shokri, M., 2020) focused on the task of salient object detection in video, which involves detecting the most visually prominent or "salient" objects in each frame of a video. The authors have proposed a deep non-local neural network (NLNN) architecture to perform this task, which leverages non-local operations to capture long-range dependencies in the video. The proposed method outperforms existing state-of-the-art methods in terms of accuracy and computational efficiency, making it suitable for real-time video processing applications.

Zheng Wanga et al. (Wang, Z., 2020) proposed a new method for improving the accuracy of salient object detection in video. The method uses a combination of global and local sensitivity to guide the re-augmentation of key salient objects in each frame of the video. The re-augmentation process involves adding additional information to the salient objects to improve their visibility, making them easier to detect. The proposed method is shown to outperform existing state-of-the-art methods in terms of accuracy and robustness, particularly in challenging video scenarios such as scenes with significant background clutter or occlusions.

G. Liu et al. (Liu, G., 2021) presented a new approach for video-based person re-identification, a task that involves identifying a target person across different camera views in a video. The method uses a graph neural network (GNN) architecture to model both intra-frame and inter-frame relationships between person instances. The intra-frame relationships capture the similarity between different body parts within a single frame, while the inter-frame relationships capture the temporal dynamics between frames. The authors show that the proposed GNN approach outperforms existing state-of-the-art methods in terms of accuracy and robustness, particularly in challenging video scenarios with variations in camera view, illumination, and appearance.

Seema Rani et al. (Rani, S., 2020) proposed a new method for summarizing videos shared on social media platforms. The method uses a combination of multi-visual features (e.g., color, texture, motion) and Kohnen's Self-Organizing Map (SOM) to generate a compact and informative summary of the video. The SOM is used to cluster similar frames in the video and select representative frames for the summary, while the multi-visual features are used to evaluate the visual quality of the frames. The authors show that the proposed method outperforms existing video summarization methods in terms of summary quality and computational efficiency, making it suitable for large-scale video analysis tasks on social media platforms.

The Salah Al-Obaidi et al. (Al-Obaidi, S., 2020) concentrated on the task of human action recognition in videos, with a focus on preserving visual anonymity. The authors propose a new method for modeling the temporal visual salience of human actions in video, which represents the most visually prominent parts of the actions. The method uses a deep neural network to predict the temporal visual salience of actions, while taking into account the need to preserve visual anonymity by blurring sensitive parts of the video such as faces. The authors show that the proposed method outperforms existing state-of-the-art methods in terms of action recognition accuracy while maintaining a high level of visual anonymity preservation.

3 Methodology

There are different methods used for keyframe extraction for various applications. In this work, the keyframe extraction is done for summarizing surveillance video when people are present in a particular location. In order to do this YOLO-based object detection (Shadadi, E., 2022) is used for extracting important frames. In order to have exact keyframe three different methodologies are used for key frame extraction and their results are analysed to identify the better method.

The Absolute Difference for Keyframe Extraction

The process of extracting key frames using absolute difference involves calculating the difference between consecutive frames. This difference is determined by counting the number of non-zero pixels in the frame and comparing it to a threshold value. The threshold is chosen based on the specific requirements and application, with a high threshold resulting in the recording of complete scene changes and a low threshold capturing even minor changes in the frames

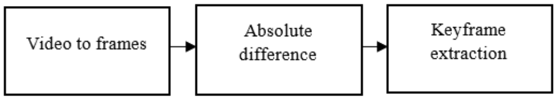


Figure 2: Absolute Difference for Keyframe Extraction

Entropy for Keyframe Extraction

In the entropy value based key frame extraction, entropy values of two adjacent frames are calculated. In order to do this first the histogram values and total no of pixels are calculated for each image and the probability was calculated by normalizing the histogram values of each pixels with the total no of pixels in the frame. Then the entropy was calculated using the formula depicted in the Equation 1.

$$Entropy\ (S) = -P \log P \ / \log 2 \qquad (1)$$

where, P is the normalized histogram value of each image. The difference between two entropy values of adjacent frames is calculated and lower entropy difference will result in the highest scene change.

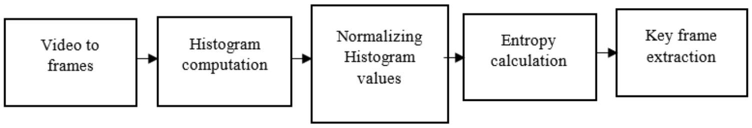


Figure 3: Entropy Based Keyframe Extraction

Optical-Flow Dense Keyframe Extraction

The optical flow in a video sequence was calculated based on comparing two frames and object movement at a particular time with the velocity and direction. Based on the optical flow calculations the keyframes are extracted. Here the real-time video is converted into frames then consecutive frames are converted to gray images for motion calculations. If the flow was high then the image was selected as keyframe.

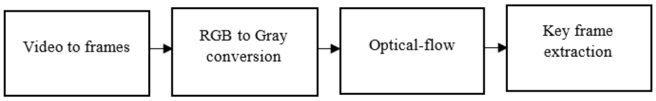


Figure 4: Optical-Flow Dense Keyframe Extraction

YOLO Algorithm for Keyframe Extraction

The YOLO algorithm is used for selecting the keyframes from a real-time surveillance video. Initially, the input video is converted into frames then preprocessing of frames are done before giving into YOLO algorithm to detect people. The frames are uniformly converted into the size of 416x416, confidence threshold set to 0.25 and IoU threshold with 0.45. In order to make algorithm unified for detecting only persons, the class value is set to be zero as default.

The Region of interest is selected in the frame from the Field of View (FoV) to select the appropriate frames with the exact object size in all the frames and this RoI can be selected in a different way for different applications. The model was proposed to select the keyframes when person is detected in the surveillance video and from the field of view best region is selected when person was clear and nearer to the camera. In proposed model setup Hikvision camera was used for surveillance, the camera can record video in the rate of 25 frames per second. If we process all the frames for detecting people then the computation complexity will get increased and redundant frames may be selected as keyframe. To eliminate the redundant frames only one frame is selected for every second for evaluation and the redundant frames are almost eliminated. The process of keyframe extraction using YOLO is shown in fig.5

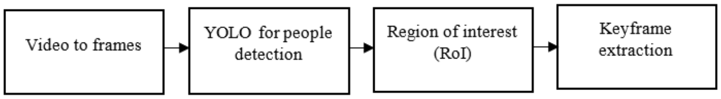


Figure 5: People Detection based Keyframe Extraction

Dataset Created

For evaluating the proposed method, from the surveillance videos collected from various places the person-based video dataset has been created with different scenarios like male, female, different outfits, carrying objects, persons very close together, multiple entries, many people, running persons, and persons with different age groups. This data set contains short videos of duration from 5 sec to 20 sec maximum, all videos contains people from 1 to 8 members in it and they will cover a distance of 60 feet, people are coming towards the camera in very few videos people will move right to left. The sample frames from the created video dataset is shown in fig 6. This dataset is really useful for validating the proposed keyframe extraction model.

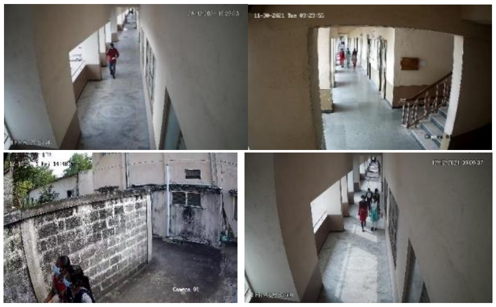


Figure 6: Sample Frames from Video Dataset Created

4 Experimental Results

The keyframe extraction for creating people-based datasets from surveillance video was done with four different methods and they are discussed below. For evaluating different models, the dataset has been created from the surveillance video based on person availability and considering different criteria. The experimental results of all evaluated models with different sample videos are briefly discussed below.

Here one sample video is taken to evaluate the different models and this sample video is created from the surveillance video and length of the video is 18 seconds and frames per second is 20 and this video contains 6 persons coming towards the camera. Fig .7 illustrates the keyframes extracted using the above four methods and their comparison is also depicted in the table.1

Optical Flow Based

The optical flow-based dense keyframe extraction was based on motion vector detection. Based on the movement of an object from frame to frame flow is measured and from that maximum, minimum and mean values of flows are calculated, and based on the mean value the keyframes are extracted. In this method, if an object is far from the camera then the mean flow will be very small so the keyframes will not be extracted and at the same time if an object is very near to the camera then the mean flow value will be high. Hence, the keyframes are extracted continuously due to high flow value and the neighbor frames also get saved. This will record important instances but more consecutive frames are saved, the number of frames extracted is more in this method. Fig 7(a) shows the key frames extracted from a particular video using an optical flow-based dense method is in total 14 frames for 6 persons.

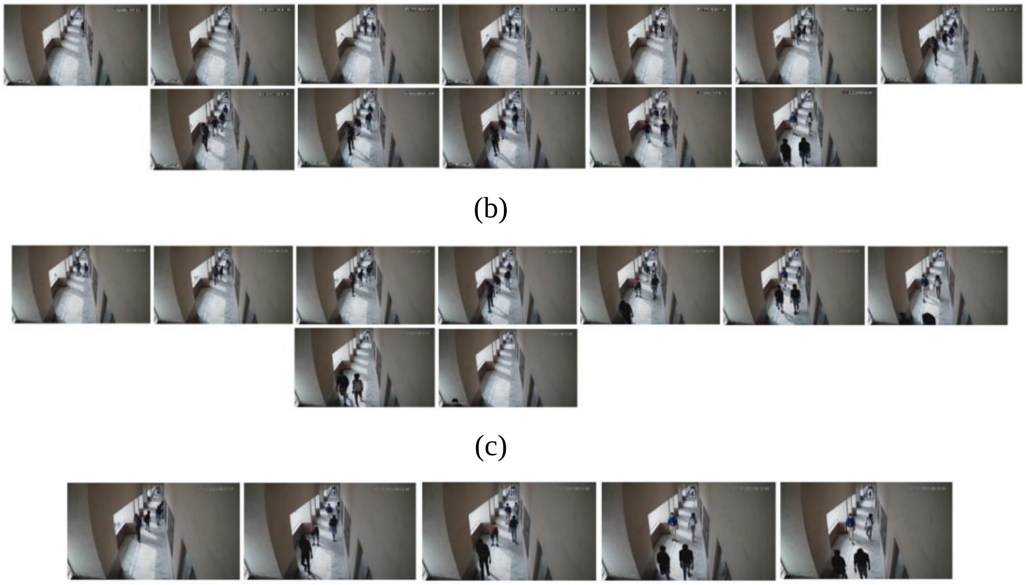
Entropy-Based

The entropy of each and every frame is calculated and based on the difference between two adjacent frame’s entropy and threshold value the keyframes are selected. The threshold value can be selected based on the scene difference that is to be recorded as a keyframe. In this paper the surveillance videos are used for keyframe extraction when person is detected, so the threshold has been set by calculating various entropy differences between an empty frame and frame with person. Here the keyframes are extracted based on entropy differences between two adjacent frames based on person availability. The extracted frames are 12, due to the entropy difference the frames are extracted in a uniform interval, from the person appearing in the FoV to leaving out of FoV. Due to this the frames extracted is more and shown in the fig.7(b).

Absolute Difference Based

The absolute difference-based method is using the same concept as the entropy-based method, here also the difference between two adjacent frames is calculated but in terms of the pixel value. In total 9 keyframes are extracted based on calculating zero pixels in the difference of frame. For extracting keyframes from surveillance video to make a people-based dataset, the person should be clearly visible and nearer to the camera. In this method based on the difference, the important frames are selected from surveillance video when a person appears at a very long distance, so the person is not clearly visible to make a dataset. Total frames extracted using absolute difference method is 9, fig. 7(c) illustrates the keyframes extracted.

(a)



(b)

(c)

(d)

Figure 7: Keyframes Extracted Using (a) Optical Flow (Motion Vector) Method (b) Entropy (c) Absolute Difference (d) YOLO

Table1: Comparison of Number of Key Frames Extracted Using Various Methods

S. No	Method	No. of keyframes extracted
1.	Optical flow	14
2.	Entropy	12
3.	Absolute difference	9
4.	Object detection	5

Object Detection-Based Model

The object detection-based model for keyframe extraction is a more suitable method for, creating a people-based dataset for various applications like people detection, gender recognition, identity recognition, etc. In this paper YOLO people detection based key frame extraction model is proposed, the algorithm is unified to detect persons only and the area of detection is also reduced to a required area, where the persons are clearly visible to make the dataset. If persons are coming near to the camera or moving away from the camera, person crossing the area marked the keyframes are extracted. Therefore, only important frames in the required area get collected due to which the redundant frames or continuous recording of frames will be eliminated and recording keyframes if a person is far away from the camera also eliminated. This people detection based keyframe extraction will give us exact keyframes with good person visibility for creating the person-based dataset. In this particular sample video, the total number of keyframes extracted is 5, when compared to all other methods without any redundancy the important frames alone captured using this YOLO-based model and it is shown in fig.7(d).

5 Conclusion

In the context of summarizing lengthy surveillance videos to generate a people-based dataset, several techniques were compared. The aim was to extract keyframes from specific areas with optimal visibility of individuals. Among the evaluated methods, the YOLO object detection-based approach showcased exceptional accuracy compared to other techniques. Its ability to precisely identify keyframes with individuals stood out. Consequently, this method holds promise for further implementation in gender recognition and women's safety applications. By leveraging the YOLO algorithm, such applications could benefit from efficient keyframe extraction, enabling enhanced monitoring and response mechanisms. With its high accuracy and potential for real-world implementation, the YOLO object detection-based keyframe extraction method could contribute significantly for improving women's safety and overall security. This advancement in video summarization holds great potential for creating more effective and targeted surveillance systems.

Acknowledgement

We thank Department of Science and Technology (DST), India- Science for Equity Empowerment Development (SEED) division for funding this research work (DST/SEED/WS/2019/203).

References

[1] Al-Obaidi, S., Al-Khafaji, H., & Abhayaratne, C. (2020). Modeling temporal visual salience for human action recognition enabled visual anonymity preservation. *IEEE Access*, *8*, 213806-213824.

[2] Arfizurrahmanl, M., Ahmad, M.S.H., Hossain, M.S., Haque, M.A., & Andersson, K. (2021). Real-Time Non-Intrusive Driver Fatigue Detection System using Belief Rule-Based Expert System. *Journal of Internet Services and Information Security (JISIS), 11*(4), 44-60.

[3] Chen, C., Li, S., Wang, Y., Qin, H., & Hao, A. (2017). Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. *IEEE transactions on image processing*, *26*(7), 3156-3170.

[4] Chen, J., Song, H., Zhang, K., Liu, B., & Liu, Q. (2021). Video saliency prediction using enhanced spatiotemporal alignment network. *Pattern Recognition*, *109*, 1-26.

[5] Chen, L., & Wang, Y. (2017). Automatic key frame extraction in continuous videos from construction monitoring by using color, texture, and gradient features. *Automation in Construction*, *81*, 355-368.

[6] Chen, Y., Huang, T., Niu, Y., Ke, X., & Lin, Y. (2019). Pose-guided spatial alignment and key frame selection for one-shot video-based person re-identification. *IEEE Access*, *7*, 78991-79004.

[7] Gu, X., Lu, L., Qiu, S., Zou, Q., & Yang, Z. (2020). Sentiment key frame extraction in user-generated micro-videos via low-rank and sparse representation. *Neurocomputing*, *410*, 441-453.

[8] Han, K.M., & Kim, Y.J. (2020). Keyslam: robust RGB-D camera tracking using adaptive VO and optimal key-frame selection. *IEEE Robotics and Automation Letters*, *5*(4), 6940-6947.

[9] Huang, C., & Wang, H. (2019). A novel key-frames selection framework for comprehensive video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(2), 577-589.

[10] Lei, S., Xie, G., & Yan, G. (2014). A novel key-frame extraction approach for both video summary and video index. *The Scientific World Journal*, *2014*, 1-9.

[11] Li, Y., Kanemura, A., Asoh, H., Miyanishi, T., & Kawanabe, M. (2020). Multi-Sensor integration for key-frame extraction from first-person videos. *IEEE Access*, *8*, 122281-122291.

[12] Lin, X., Wang, F., Guo, L., & Zhang, W. (2019). An automatic key-frame selection method for monocular visual odometry of ground vehicle. *IEEE Access*, *7*, 70742-70754.

[13] Liu, G., & Wu, J. (2021). Video-based person re-identification by intra-frame and inter-frame graph neural network. *Image and Vision Computing*, *106*.

[14] Rani, S., & Kumar, M. (2020). Social media video summarization using multi-Visual features and Kohnen's Self Organizing Map. *Information Processing & Management*, *57*(3).

[15] Shadadi, E., Ahamed, S., Alamer, L., & Khubrani, M. (2022). Deep Anomaly Net: Detecting Moving Object Abnormal Activity Using Tensor Flow. *Journal of Internet Services and Information Security*, *12*(4), 116-125.

[16] Shih, H.C. (2013). A novel attention-based key-frame determination method. *IEEE Transactions on Broadcasting*, *59*(3), 556-562.

[17] Shokri, M., Harati, A., & Taba, K. (2020). Salient object detection in video using deep non-local neural networks. *Journal of Visual Communication and Image Representation*, *68*, 1-24.

[18] Survey on electronic gadgets and its effects. https://realresearcher.com/media/survey-on-the-use-of-electronic-gadgets-and-its-effects/

[19] Wang, Z., & Zhu, Y. (2020). Video key frame monitoring algorithm and virtual reality display based on motion vector. *IEEE Access*, *8*, 159027-159038.

[20] Wang, Z., Zhou, Z., Lu, H., & Jiang, J. (2020). Global and local sensitivity guided key salient object re-augmentation for video saliency detection. *Pattern Recognition*, *103*.

[21] Zhao, H., Wang, W.J., Wang, T., Chang, Z.B., & Zeng, X.Y. (2019). Key-frame extraction based on HSV histogram and adaptive clustering. *Mathematical Problems in Engineering*, *2019*, 1-10.

[22] Zhong, Q., Zhang, Y., Zhang, J., Shi, K., Yu, Y., & Liu, C. (2020). Key frame extraction algorithm of motion video based on priori. *IEEE Access*, *8*, 174424-174436.

Authors Biography

S. Bharathi is currently working as an Assistant Professor in the Department of Electronics and Communication Engineering at Dr. Mahalingam College of Engineering & Technology, Pollachi, India. She holds a PhD degree in Information and Communication Engineering from Anna University, Chennai. She has published research papers in international journals and conference proceedings in the area of Image processing and biometric system. Also, she is an active reviewer in international journals and conferences. Her area of research includes image processing, Biometrics and Pattern recognition.

M. Senthilarasi is currently working as an Assistant Professor in the Department of Electronics and Communication Engineering at Thiagarajar College of Engineering, Madurai, India. She holds a PhD degree in Information and Communication Engineering from Anna University, Chennai. She has published research papers in international journals and conference proceedings in the area of Image processing and she is an active reviewer in international journals and conferences. Her area of research includes image processing and Computer vision.

K. Hari is working as Junior research fellow, in the Department of Electronics and Communication Engineering at Dr. Mahalingam College of Engineering & Technology, Pollachi, India. He holds his Master's degree in Communication Engineering from Coimbatore Institute of Technology, Coimbatore. He has published papers in international journals and conference proceedings. His area of research includes image and video processing.