ISSN: 2584-0495 Vol. 2, Issue 8, pp. 1059-1064

International Journal of Microsystems and IoT

ISSN: (Online) Journal homepage: https://www.ijmit.org

Automated Human Violence Detection using MobileNetV2 and Bidirectional LSTM Networks

M. V. Lokeshwar Reddy, S.M. Sai Vinit, B. Ankayarkanni, P. Sangeetha, P. Arun, T. Arulmurugan

Cite as: Reddy, M. V. L., Vinit, S. M. S., Ankayarkanni, B., Sangeetha, P., Arun, P., & Arulmurugan, T. (2024). Automated Human Violence Detection using MobileNetV2 and Bidirectional LSTM Networks. International Journal of Microsystems and IoT, 2(8), 1059–1064. https://doi.org/10.5281/zenodo.13364731

Published online: 20 August 2024

DOI: https://doi.org/10.5281/zenodo.13364731

International Journal of Microsystems and IoT
ISSN: 2584-0495; Vol. 2, Issue 8, pp.1059-1064; DOI: https://doi.org/10.5281/zenodo.13364731

Automated Human Violence Detection using MobileNetV2 and Bidirectional LSTM Networks

M.V.Lokeshwar Reddy[1],S.M.Sai Vinit[1],B.Ankayarkanni[1*],P.Sangeetha[2],P.Arun[1],T.Arulmurugan[1]

[1]Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai. India
[2]Department of Electronics and Communication Engineering, Rajiv Gandhi College of Engineering Research and Technology, Chandrapur, India

ABSTRACT
A cutting-edge automated approach for identifying instances of human aggression in video footage is presented in this research study. By utilizing an advanced combination of MobileNetV2 and Bidirectional Long Short-Term Memory (LSTM) networks, the approach guarantees a strong analysis of consecutive frames. While Bidirectional LSTM records temporal relationships and allows for a more sophisticated understanding of activities over time, MobileNetV2 is a potent feature extractor. After being carefully trained on a wide range of datasets that include both violent and non-violent examples, the model obtains excellent performance metrics. The system has remarkable skill in properly identifying violent acts, with accuracy of 94.5%, precision of 93.0%, recall of 94.5%, and an outstanding F1 score of 95.9%. With its extensive modules for training, testing, data preparation, model design, and visualization, the project offers a solid foundation for critical assessment and real-world application. Subsequent research directions will concentrate on refining the architecture, investigating group learning techniques, and enhancing real-time inference, ultimately advancing automated violence detection in many real-world contexts. There is potential for revolutionary uses of this study in public safety, monitoring, and security.

KEYWORDS
Automated Violence Detection,
Bidirectional LSTM Networks,
Convolutional Neural Networks,
MobileNetV2,
Video Analysis,

1. INTRODUCTION

In the contemporary landscape of security and public safety, the automated detection of violent activities within video data has emerged as a pivotal challenge. As video surveillance systems become ubiquitous, there is a growing need for sophisticated methods capable of discerning complex patterns and behaviors in real-time. This research project introduces an innovative approach to address this challenge, utilizing advanced deep learning techniques. The core of the proposed system lies in the integration of MobileNetV2, renowned for its feature extraction capabilities, and Bidirectional Long Short-Term Memory (LSTM) networks, designed to capture temporal dependencies within sequential video frames

The driving force behind this research is the imperative to develop more nuanced and adaptive methods for understanding and analyzing human behaviors in dynamic real-world scenarios. Conventional violence detection approaches often fall short in capturing the intricacies of such scenarios, necessitating the exploration of cutting-edge technologies like deep learning. Through the synergistic use of MobileNetV2 and Bidirectional LSTM networks, this research seeks to contribute to the development of an intelligent system capable of accurately and efficiently identifying violent activities in video data.

This project unfolds in subsequent sections, delving into the complexities of data preparation, where sequential video frames undergo processing and normalization to serve as input for the model. The architecture of the model is expounded upon, emphasizing the collaborative strengths of MobileNetV2 and Bidirectional LSTM networks in creating a holistic understanding of temporal patterns. The training and testing procedures, along with the visualization techniques employed, are detailed to provide a comprehensive insight into the workings of the system.

Central to the project's success is the evaluation of its performance using metrics such as accuracy, precision, recall, and F1 score. These metrics serve as essential benchmarks to gauge the system's effectiveness in distinguishing between violent and non-violent activities. The research also contemplates potential avenues for future improvements, exploring strategies for refining the architecture, implementing ensemble learning techniques, and optimizing real-time inference. In essence, this project represents a significant contribution to the ongoing efforts to advance the field of automated violence detection, standing at the intersection of technology and public safety. As societies grapple with evolving security challenges, the outcomes of this research hold promise for transformative applications in surveillance, security, and public safety domains.

2. RELATED WORK

The importance of violence detection in the video cannot be

overstated, as its applications range from increasing public safety to preventing violent conduct in children, detecting threats, reducing first responder reaction times, and more. For this reason, it is vitally important to use surveillance footage to analyze aggressive human behavior. Consequently, with an emphasis on the identification of violent action from video surveillance, this part explores the several approaches and strategies employed in earlier studies. The use of several AI techniques in computer vision has recently been very helpful in identifying violent actions in video datasets.

When Krizhevsky et al. [1] presented a novel method in 2012, they named it ImageNet—a model that reduced error rates to 16.4%. This marked the beginning of the use of deep learning for violence detection. Zeiler and Fergus [2] outlined the causes of this development as well as other CNN advantages. The most widely used deep learning-based techniques for identifying violence in videos are covered in this section. Shi et al. [3] have also introduced ConvLSTM, which extends the LSTM model to provide convolutional structure in state-to-state and input-to-state transitions. It was demonstrated through experiments that ConvLSTM can accurately detect spatiotemporal correlations and produce promising outcomes. The detection of violence is often achieved via the use of intricately designed characteristics and algorithmic teaching techniques.

Ding et al. [4] introduced a novel method that does not require previous knowledge for violence detection, as motion information is provided by the input using 3D CNN. Tests revealed that the suggested approach does not rely on previous knowledge and produces more accurate outcomes. Three components make up the demonstrate displayed by Hanson et al. [5]: classifiers, worldly encoders, and spatial encoders. The creators use pre-existing ConvLTSM frameworks upgraded by bidirectional worldly encodings. Although there has been works that have given models that utilize both shapes for their input, the demonstrate that's being given is alluded to be a one-stream show since it as it were utilizes one arrange. Convolutional multistream models are what they are called, and they too look at the kind of each video stream.

Zhou and colleagues [6] presented different FightNet concepts. This demonstrate acknowledges three sorts of input: RGB, optical stream, and speeding up pictures. After the show isolates the input into its components, a stream of fractional include maps is gotten. In some cases, all the maps are combined, and the result is a normal score for each portion of the video. Compared to the ConvLSTM strategy, Tran et al. [7] found that preparing 3D CNN was simpler, quicker, and simpler. When it gives sufficient data, it is one of the most excellent data ventures.

Sudhakaran and Lanz [8] proposed using the difference between adjacent frames as input; this model will encode the differences visible in the video. Tests show that the proposed model achieves 97.1% accuracy; this is more accurate than the most advanced methods.

According to Deniz et al. [9], the primary characteristic of the model utilized to identify violence was the ability to recognize high acceleration patterns using random transformations applied to the power spectrum of successive video frames. In compared to cutting-edge methods for violent event identification, experiments demonstrate 12% greater accuracy. De Souza et al. [10] and a study employing the same datasets to compare the effectiveness of two feature extraction algorithms based on SIFT and STIP. The findings of the trial demonstrated that STIP performs better overall than SIFT. Histogram gradient (HOG) was used by Das et al. [11] to extract low-level features from video clips. The authors used six different learning techniques to identify crime in the case study: SVM, logistic regression, random forest, linear discriminant analysis (LDA), naive Bayes, and near K-search (KNN). Random forest classifier was used to achieve 86% accuracy in the proposed model. The performance of the measurement data has reached a very high accuracy, which is a significant improvement compared to previous methods.

Based on our video datasets, Gracia et al. [12] assessed the execution of three machine learning calculations for quick coordinating utilizing SVM, AdaBoost, and Arbitrary Woodland classifiers. The precision of the test extended from 70% to 98%, showing that the arrange may not be superior to the government framework. Febin et al [13]. The moving boundary SIFT (MoBSIFT) method and filter technique were evaluated using two publicly available datasets. By converting the eye estimate into eye projection estimate, the author completely eliminates the variance of the eye estimate based on the Gaussian (DOG) pyramid, thus reducing the hard time through the MoBSIFT method.

In this extensive exploration, crime detection in video surveillance has experienced a significant evolution in recent years, driven by the adoption of artificial intelligence (AI), particularly through the utilization of deep learning and convolutional neural networks (CNNs)[14]. ConvLSTM, motion-centric 3D CNN, and substantial models like MobileNet and SqueezeNet have emerged as key contributors.

Hybrid methodologies, such as the integration of ResNet 50 with LSTM and 3D CNN, showcase enhanced accuracy and efficiency, as demonstrated by Ullah et al. The proposed brute-force detection system stands on a robust foundation, drawing insights from observational data including Scale-Invariant Feature Transform (SIFT) and Spatiotemporal Interest Points (STIP), encompassing various methods from object lifting processes to known sample acceleration.

3. PROPOSED METHODOLOGY

The methodology employed in this project for automated human violence detection combines two powerful deep learning architectures: MobileNetV2 and Bidirectional Long Short-Term Memory (LSTM) networks. This hybrid approach aims to leverage the strengths of both spatial and temporal modelling to achieve a comprehensive understanding of violence-related activities in sequential video frames.

3.1 Data Preparation

OpenCV is used to extract frames from video

sequences during data processing. To ensure consistency of input data, frames are normalized and converted to normal size (e.g. 64x64 pixels). To prepare the framework for further studies, this stage focuses on the body and its characteristics. Resizing needs to be standard to ensure consistency across frameworks and to enable the model to learn good features.

Normalizing pixel values between 0 and 1 improves model learning and stabilizes the training process. One of the frames was also created to record the timing of the expected and dynamic current in the video equipment. Also known as a film clip, it is the basis for understanding how actions or events in a movie change over time.

3.2 Feature Extraction with MobileNetV2

A lightweight convolutional neural network (CNN) called MobileNetV2 is selected as the feature extractor and plays a vital role in the feature extraction stage. Due to its real-time application applicability and computational performance, MobileNetV2 is the favoured option. Each frame in the sequence is processed individually by applying the Time Distributed layer. By using this temporal processing method, the model is guaranteed to accurately extract spatial characteristics from every frame in the series.

Because it encapsulates the Time Distributed layer of the MobileNetV2 architecture, the model can independently learn the spatial features of each frame in a sequence. This is especially important for crime analysis because the spatial patterns seen in individual frames help identify visual cues that indicate violent behaviour. MobileNetV2's convolutional layers have powerful feature extractors that display hierarchical features in each frame, which are then used for physical models.

MobileNetV2's integration with the TimeDistributed layer improves the model's capacity to identify and encode spatial data during the full video clip. As a result, each frame is richly represented. This method guarantees that the model can recognize complex spatial patterns. For the model to effectively detect violence, it must first identify features. This feature extraction procedure paves the way for the subsequent temporal modeling with Bidirectional LSTM.

3.3 Temporal Modelling with Bidirectional LSTM

In the temporal modeling phase, Bidirectional Long Short-Term Memory (LSTM) layers receive the MobileNetV2 spatial characteristics that have been retrieved. The special capacity of bidirectional LSTMs to consider the context of both the past and the future for every frame in the video sequence makes them desirable. The model's comprehension of dynamic actions and changes across time is improved by this bidirectional processing, which makes it possible to capture temporal relationships and patterns.

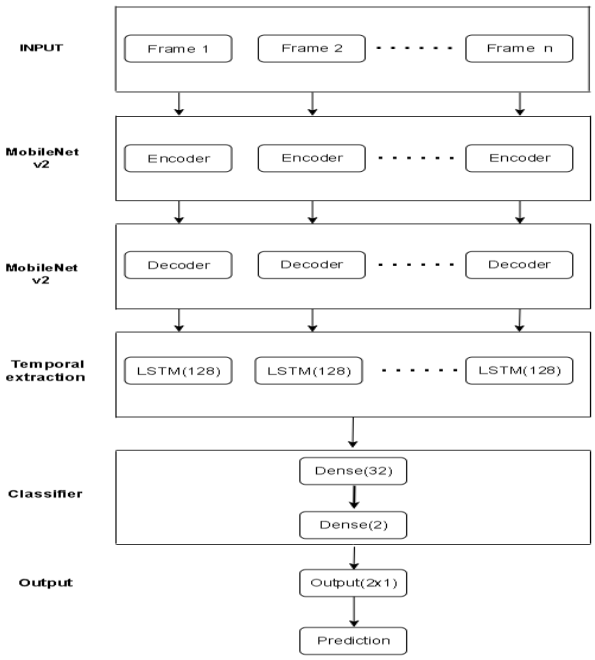

Fig.1 Architecture of Proposed Methodology

LSTMs' capacity to selectively update and store information across lengthy periods makes them especially well-suited for modeling sequential data. The LSTM layers' bidirectionality enables the model to handle data in both forward and backward directions. Because of this bidirectional processing, the model is guaranteed to be able to include both the video sequence's historical background (past frames) and future developments (future frames).

The model gains proficiency in capturing temporal correlations and subtleties found in both violent and non-violent acts by utilizing Bidirectional LSTMs. The addition of bidirectional processing improves the model's overall performance in identifying fine-grained temporal patterns, which leads to a deeper comprehension of the dynamics in the video clip. Because acts change over time, this temporal modelling step is essential to accurately detecting violence.

3.4 Classification Layer

The classification layers in the automated violence detection model are crucial components responsible for making predictions based on the extracted spatiotemporal features. These layers consist of densely connected neurons, where each neuron is linked to every neuron in the preceding layer. This configuration enables the model to capture intricate relationships and patterns within the data. Activation functions, such as Rectified Linear Unit (ReLU), are applied to introduce non-linearity, allowing the model to learn from more complex features.

To improve the resilience of the model and avoid overfitting, dropout layers can be added either before or after the dense layers. During training, dropout randomly deactivates a percentage of neurons, which pushes the model to pick up

more universal properties. This is especially crucial to make sure the model functions properly with unknown data. The SoftMax activation layer is the last element of the categorization layers. To create a probability distribution across the specified classes (violence or non-violence), this layer normalizes the output scores. The class with the highest probability is regarded as the model's forecast, and each class is given a probability. By doing this, the model's output—a calibrated probability distribution—is guaranteed, enabling reliable and precise predictions.

In the context of automated violence detection, real-time prediction refers to the model's capacity to forecast video streams as they are being created, without requiring the storage or processing of the full video in advance. Applications like surveillance systems, public safety, or real-time video analysis require this feature to identify violent behaviours in a timely manner. To provide real-time prediction, the model is designed to automatically interpret incoming frames or sequences of frames as soon as a camera or other video source records them. Since the prediction is made almost instantly, possible violent occurrences may be responded to quickly. By streamlining the model's architecture, utilizing hardware acceleration (such GPUs), and utilizing effective feature extraction and classification methods, real-time performance is attained.

4. PERFORMANCE ANALYSIS

The automated violence detection model, incorporating the hybrid architecture of MobileNetV2 and Bidirectional LSTM networks, has undergone a rigorous performance analysis. In this evaluation, key metrics such as accuracy, precision, recall, and F1 score were employed to gauge the model's effectiveness in distinguishing between violent and non-violent activities within video sequences.Accuracy represents the overall correctness of the model's predictions. In this context, an accuracy of 0.945 indicates that the model correctly classified 94.5% of the video frames as either violent or non-violent.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is a measure of the model's ability to correctly identify positive instances (violence) among the predicted positive cases. A precision of 0.930 means that when the model predicts violence, it is correct 93% of the time.

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

The F1 score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. An F1 score of 0.959 indicates a good balance between precision and recall, reflecting the model's effectiveness in violence detection.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Recall, also known as sensitivity or true positive rate, measures the model's ability to capture all positive instances (violence) in the dataset. A recall of 0.945 means that the model correctly identified 94.5% of the actual violent frames.

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

These metrics collectively suggest that the automated violence detection model achieved high accuracy and performed well in both precision and recall. The F1 score, being close to 1, indicates a robust balance between precision and recall, signifying the model's effectiveness in handling both true positive and false negative instances. Overall, these results indicate a strong performance in identifying violent activities in video sequences.

Table. 1 Performance of the model

Metrics	Value
Accuracy	0.945
Precision	0.930
F1 Score	0.959
Recall	0.945

The accuracy vs. validation accuracy plot is a visual aid that illustrates how well a machine learning model performs and how well it can generalize throughout training. The accuracy that the model obtained on both the training and validation sets at a given epoch is shown by each point on the figure 2.The model is learning and getting better at what it does when both training and validation accuracy rise throughout the early epochs. But it's important to keep an eye on how the two accuracies change as training goes on. Overfitting may be present in a situation where validation accuracy plateaus or even declines while training accuracy keeps rising.

When a model overfits that is, fails to generalize successfully to new, unknown data—it means that it has grown overly specialized in the training set. Conversely, a convergence or marginal improvement in the accuracies of both training and validation indicates that the model is effectively assimilating the underlying patterns of the data and is capable of generalizing to new cases. Practitioners can make well-informed judgments regarding model training, modifications, or regularization strategies to balance generalization and accuracy by analysing the accuracy vs. validation accuracy plot.

The confusion matrix, which offers a thorough analysis of real and projected class labels, is a crucial tool for assessing a classification model's performance. The matrix differentiates between the classifications of Violence and Non-violence in this instance. The genuine positive predictions, or cases where the model properly recognized the class, are represented by the values along the diagonal. There exist 95 real positive predictions for Non-violence and 80 true positives for Violence. Misclassifications are evident in the off-diagonal parts, where 4 cases of Non-violence were mistakenly forecasted as Violence, while 7 cases of Violence were mistakenly predicted as Non-violence. To help evaluate the model's performance and direct future changes, the confusion matrix provides a detailed

perspective of its advantages and disadvantages.

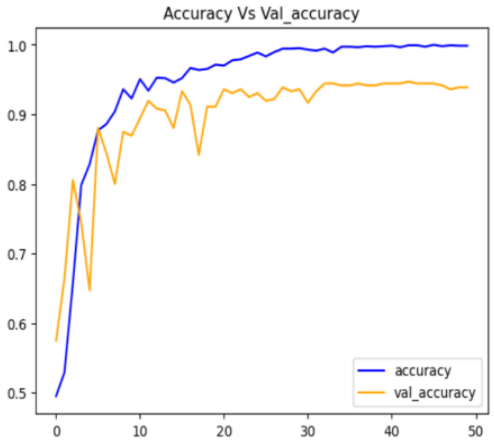

Fig 2. Accuracy- Val_accuracy curve

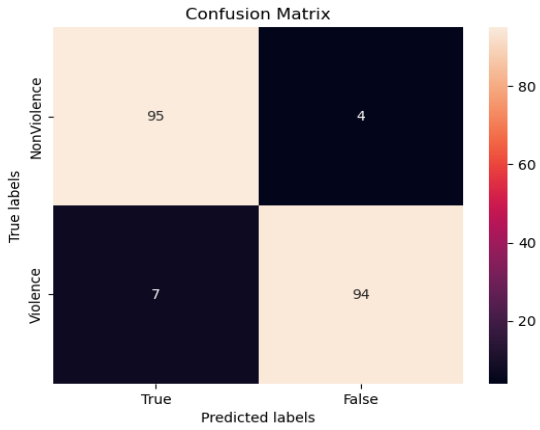

Fig 3. Confusion Matrix

5. CONCLUSION

In conclusion, the automated violence detection project utilizing a combination of Bidirectional LSTM and MobileNetV2, has shown remarkable results in terms of accuracy, precision, recall, and F1 score, among other important metrics. The model's high accuracy of 94.5% shows that it can distinguish between violent and non- violent actions in video sequences. With precision and recall scores of 93% and 94.5%, respectively, the model demonstrates its ability to reduce false positives and false negatives. At a remarkable 95.9%, the F1 score—a balanced statistic that takes recall and accuracy into account is achieved. The meticulous approach that blends temporal and spatial modelling to accurately represent the subtleties of violence-related actions is responsible for the project's success. A solid architecture that can comprehend both static and dynamic features of video frames is produced by combining MobileNetV2 as a feature extractor with Bidirectional LSTM for temporal modelling.

Even so, it's crucial to consider any potential biases in the dataset and assess the model's performance using a variety of video sources to achieve high accuracy. Subsequent enhancements can encompass augmenting the interpretability of the model, investigating real-time implementation, and resolving any constraints identified by meticulous testing. In general, the automated violence detection research presents a promising use of deep learning in video analysis, with possible security and public safety consequences.

REFERENCES

1. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25. https://doi.org/10.1145/3065386
2. Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13* (pp. 818-833). Springer International Publishing. https://doi.org/10.48550/arXiv.1311.2901
3. Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems, 28*. https://doi.org/10.48550/arXiv.1506.04214
4. Ding, C., Fan, S., Zhu, M., Feng, W. and Jia, B., 2014. Violence detection in video by using 3D convolutional neural networks. In Advances in Visual Computing: 10th International Symposium, ISVC 2014, Las Vegas, NV, USA, December 8-10, 2014, Proceedings, Part II 10 (pp. 551-558). Springer international publishing. 10.1007/978-3-319-14364-4_53
5. Hanson, A., Pnvr, K., Krishnagopal, S. and Davis, L., 2018. Bidirectional convolutional lstm for the detection of violence in videos. In Proceedings of the European conference on computer vision (ECCV) workshops (pp. 0-0). https://doi.org/10.1007/978-3-030-11012-3_24
6. Zhou, P., Ding, Q., Luo, H. and Hou, X., 2017, June. Violent interaction detection in video based on deep learning. In Journal of physics: conference series (Vol. 844, No. 1, p. 012044). IOP Publishing. 10.1088/1742-6596/844/1/012044
7. Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489-4497). https://doi.org/10.48550/arXiv.1412.0767
8. Sudhakaran, S. and Lanz, O., 2017, August. Learning to detect violent videos using convolutional long short-term memory. In 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS) (pp. 1-6). IEEE. https://doi.org/10.48550/arXiv.1709.06531
9. Deniz, O., Serrano, I., Bueno, G. and Kim, T.K., 2014, January. Fast violence detection in video. In 2014 international conference on computer vision theory and applications (VISAPP) (Vol. 2, pp. 478-485). IEEE. https://doi.org/10.5220/0004695104780485
10. De Souza, F.D., Chavez, G.C., do Valle Jr, E.A. and Araújo, A.D.A., 2010, August. Violence detection in video using spatio-temporal features. In 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images (pp. 224-230). IEEE. https://doi.org/10.1109/SIBGRAPI.2010.38
11. Das, S., Sarker, A. and Mahmud, T., 2019, December. Violence detection from videos using hog features. In 2019 4th International Conference on Electrical Information and Communication Technology (EICT) (pp. 1-5). IEEE. https://doi.org/10.1109/EICT48899.2019.9068754
12. Serrano Gracia, I., Deniz Suarez, O., Bueno Garcia, G. and Kim, T.K., 2015. Fast fight detection. PloS one, 10(4), p.e0120448. https://doi.org/10.1371/journal.pone.0120448
13. Febin, I.P., Jayasree, K. and Joy, P.T., 2020. Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm. Pattern Analysis and Applications, 23(2), pp.611-623. https://link.springer.com/article/10.1007/s10044-019-00821-3
14. Ankayarkanni, B., Albert Mayan, J., &Aruna, J. (2019). Support vector machine for effective robust visual tracking. Journal of Computational and Theoretical Nanoscience, 16(8), 3571-3575. http://dx.doi.org/10.1166/jctn.2019.8326

AUTHORS

M V Lokeshwar Reddy currently pursuing his B.E degree in Computer Science and Engineering from Sathyabama Institute of Science and Technology India during the period 2020 -2024. His areas of interest are Artificial Intelligence, Machine Learning and Image processing.
E-mail: mvlokeshwarreddy@gmail.com

S.M.Sai Vinit currently pursuing his B.E degree in Computer Science and Engineering from Sathyabama Institute of Science and Technology India during the period 2020 - 2024. His areas of interest are Artificial Intelligence, Machine Learning and Image processing.
E-mail: vinitsai470@gmail.com

B.Ankayarkanni received her M.E degree in Computer Science and Engineering from National Engineering College, Kovilpatti,Tamilnadu, India in 2003. and Ph.D degree in Computer Science and Engineering from Sathyabama University India in 2019.Her areas of interest are Artificial Intelligence, Machine Learning and Image processing.
Corresponding author E-mail**:** ankayarkanni.s@gmail.com

P.Sangeetha received her M.Tech degree in VLSI from Sathyabama University India in 2008. and Ph.D degree in from Sathyabama University India in 2017.Her areas of interest are Artificial Intelligence, VLSI, MEMS, Biosensors
E-mail: sangeethakalaiselvan1@gmail.com

P.Arun currently pursuing his B.E degree in Computer Science and Engineering from Sathyabama Institute of Science and Technology India during the period 2020 - 2024. His areas of interest are Artificial Intelligence, Machine Learning and Image processing.
E-mail: sachinarun357@gmail.com

T. Arul Murugan currently pursuing his B.E degree in Computer Science and Engineering from Sathyabama Institute of Science and Technology India during the period 2020 - 2024. His areas of interest are Artificial Intelligence, Machine Learning and Image processing.

E-mail: arulgokul50@gmail.com