ANOMALOUS ACTIVITY RECOGNITION IN VIDEO SCENES USING ARTIFICIAL INTELLIGENCE

Kanagamalliga S [1]*, Rajalingam S [2], Renugadevi R [3], Gunapati Teja [4], Pothala Srikanth [5], Bynagari Abraham Ron Clayton [6]

[1,3,4,5,6] Department of ECE, Saveetha Engineering College, Chennai - 602105, India.

[2] Department of EEE, Saveetha Engineering College, Chennai - 602105, India.

[1]*kanagamalligas@saveetha.ac.in*, [2]*rajalingams@saveetha.ac.in*, [3]*renugadevir@saveetha.ac.in*, [4]*tejagunapati@gmail.com*, [5]*pothalasrikanth45@gmail.com*, [6]*ronclayton538@gmail.com*

doi: 10.48047/ecb/2023.12.si4.857

Abstract: Even though numerous target tracking algorithms have been developed and proven successful, performance is still negatively impacted by occlusion and lighting changes. In this research, a successful tracking technique for non-rigid objects in video situations is described. The use of computer vision to address real-time violent activity is offered as an intelligent system algorithm. Different acts of violence occasionally occur in our daily lives when we are absent. Real-time violent activity detection is essential to a smart surveillance system. Since a video consists of several frames of pixels, classifying and analysing the video is a difficult investigation problem in the field of artificial intelligence. Convolutional Long Short Term Memory is utilized in this research to take into account every scenario that could arise in order to recognise real-life violence more correctly. Datasets gathered from diverse sources are compared to get a result with a sufficient level of accuracy. With multiple experiments employing various deep video analysis algorithms, the research effort came to a conclusion. The several deep learning models are compared, and the top candidate with a 96% accuracy rate is chosen. Real-time video has now been set up to categorise anomalous activity using the proposed model.

Key Words: Artificial intelligence, Video surveillance, Convolutional Neural Network, Person detection, Anomalous Activity Detection.

Introduction

Computer vision plays a key role in addressing the video's real-time violent activity (Aktı et. al., 2019). The primary area of action recognition is robotics and video surveillance. Because it provides pictorial connection that enables the processer to view without manual interference, it is fascinating and helpful for action recognition (Abhishek et. al., 2021). This might be a crucial component of the automated video surveillance system as the globe moves toward automation. Finding an association or pattern from numerous images while taking spatial features into account is the process of action recognition. The average number of 2D frames in a single period of the video is 30.

*Corresponding Author

Finding linkage and detecting action from numerous frames of film might be difficult when there is violence present or not (Jain et. al., 2020). Prior detection systems were primarily centred on the detection of various objects and activities, such as walking, jogging, cooking, and gestures (Amit et. al., 2020) rather than specifically on the detection of violence or crime.

In this investigation, the criminal scenes that actually take place in our daily lives are the main emphasis. In order to identify violent activities accurately, deep learning models are analysed (Nasaruddin et. al., 2020). Because there are many different situations in real life, such as people fighting, hijacking, using knives or guns, and munching. Violent behaviour could be detected in those videos by several CNN algorithms (Ke et. al., 2020) and certain violent behaviour required determining the frame movements feature over time. People carrying weapons like a rifle or a knife, for instance, could use 2D CNN (Md et. al., 2020) to quickly identify potential areas for aggressive action in a violent environment. And secondly, there are instances where people use violence by fighting without a knife or a gun (Nguyen et. al., 2019). The model prediction must establish the correlated motion with period in a solution.

According to the literature review, the difficult issues include changing object and scene appearance patterns, non-rigid human structures (Kanagamalliga et. al., 2018), camera motion, a dynamic environment, multiple objects, changing scene illumination, occlusion, and cluttered background (Kanagamalliga et. al., 2019). The suggested approach aims to solve the target recognition issues brought on by occlusion and non-rigid target structures. The pre-processed background image in this paper is obtained using the median filter. To determine the precise foreground area, segmentation is used. To estimate motion, the optical flow is specified. For applications involving video surveillance, deep learning classifiers using features are recommended because of their accuracy. We have known from studying prior studies that anomalous activity identification is a challenging work in artificial intelligence (Zufan et. al., 2020).

The majority of earlier study mainly examined human behaviour in daily living in the home (Kanagamalliga et. al., 2016). Pre-processing is carried out for additional analysis as this study focused on violent action and datasets. It is simpler to analyse and extract features from still photographs than it is to identify action in video. When a video has multiple frames and is linked to a spatial characteristic over time, the artificial intelligence model might appropriately take the factors into account. The artificial intelligence model required some period to train, and things become more difficult and time-consuming when the topic is video- based. On the other hand, some actions appear to be violent but are not, while some appear to be normal but may be violent.

Proposed Methodology

Our solution uses object tracking to follow the target in an ongoing surveillance video. Figure 1 depicts the proposed method's workflow. The backdrop modelling is a crucial step in object detection estimations. This study's major goal was to identify violent behavior in live video. The video will be examined using the Deep Learning algorithm to identify action violence. Once the model recognizes violence and sends the data to a mobile application, our algorithm extracts abnormal actions from surveillance videos. The geographical and temporal components of data are handled by CLSTM. There are many video frames in the video databases.

In order to effectively identify violence, numerous video frames must be analysed in order to determine how the motion changes over time.

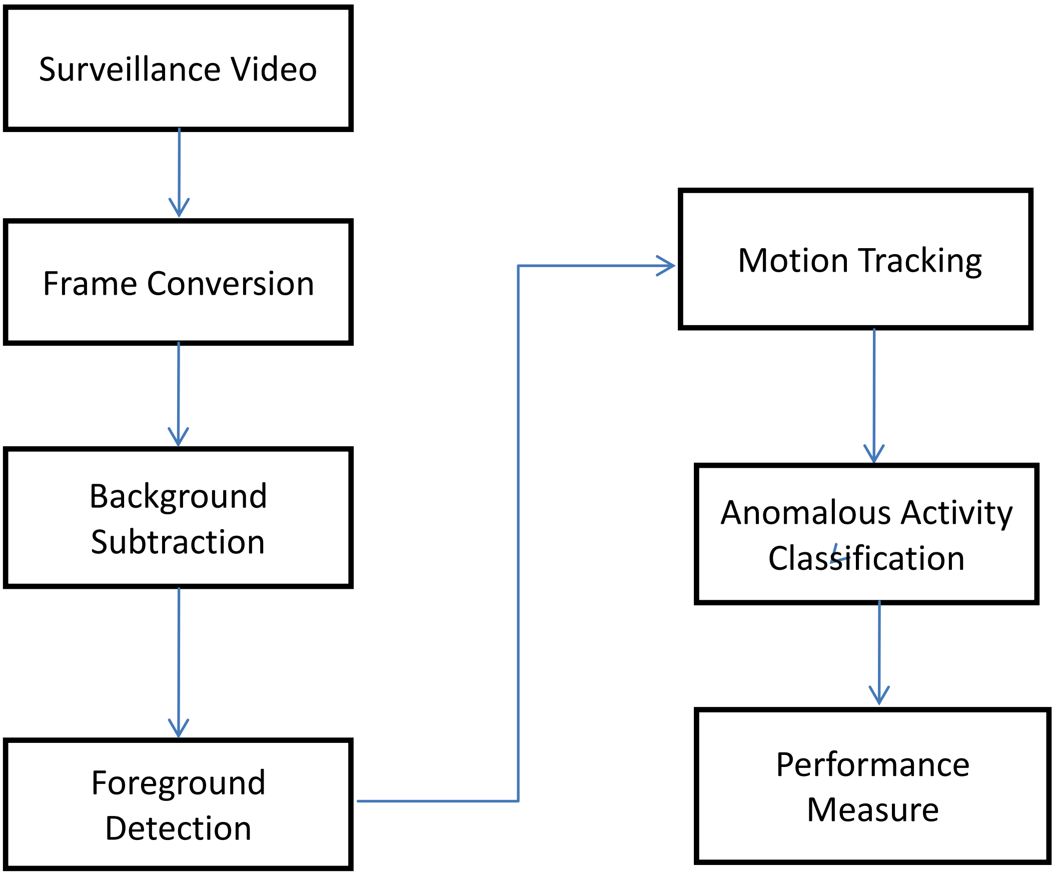


Figure 1: Basic workflow of proposed anomalous activity detection method

When evaluating a video, CLSTM encodes the spatial information with the time sequence while focusing on the spatial and temporal dimensions. Whole datasets are trained and encoded using a localised spatial-temporal feature after the procedure. Because it takes features from a fully linked layer, a standard LSTM was unable to localise the changes in spatial-temporal properties between frames. CLSTM is hence appropriate for our methodology. It employs the Convolutional LSTM, a structure created especially for spatiotemporal sequences. Convolution is used by CLSTM in its internal calculation. LSTM typically operates in the following 4 phases. Forget, save, and update the result. Before passing through a sigmoid gate, one must first forget the preceding stage's irrelevant history. Secondly, what portions of the old data and the portions of the new data are still relevant are stored in the cell state. Upgrading the condition of their interior cells. The output gate further regulates the final state. In contrast to LSTM, CLSTM used the convolution technique between those states.

Experimental Results and Analysis

For the datasets, the suggested method yields the most accurate model. Our train model's performance is examined using classification metrics, tables, and categorization reports. The model's most likely outcome for a study of violent acts is either true or unfavorable as shown in Figure 2. Table 1 provides a more detailed description.

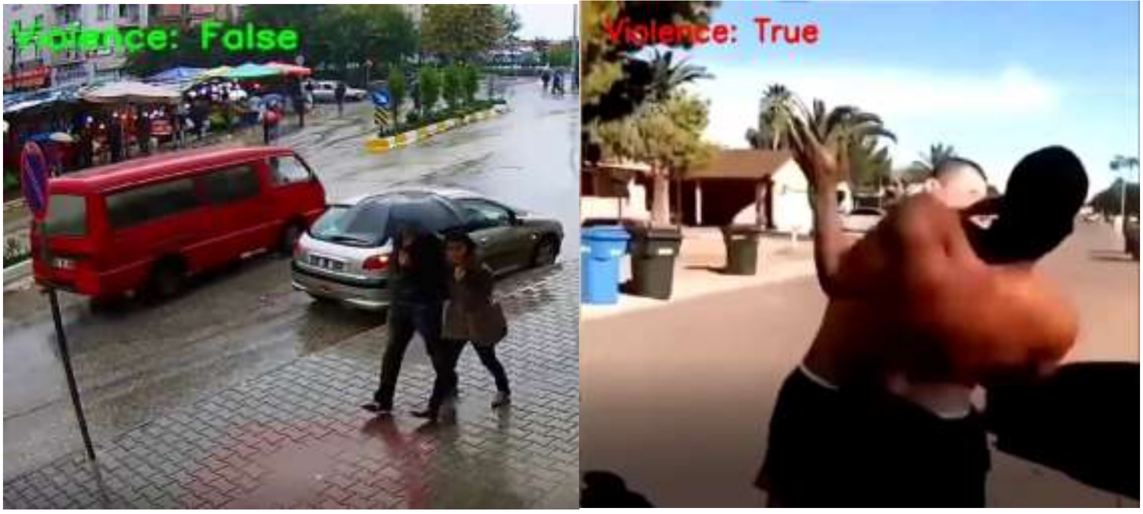
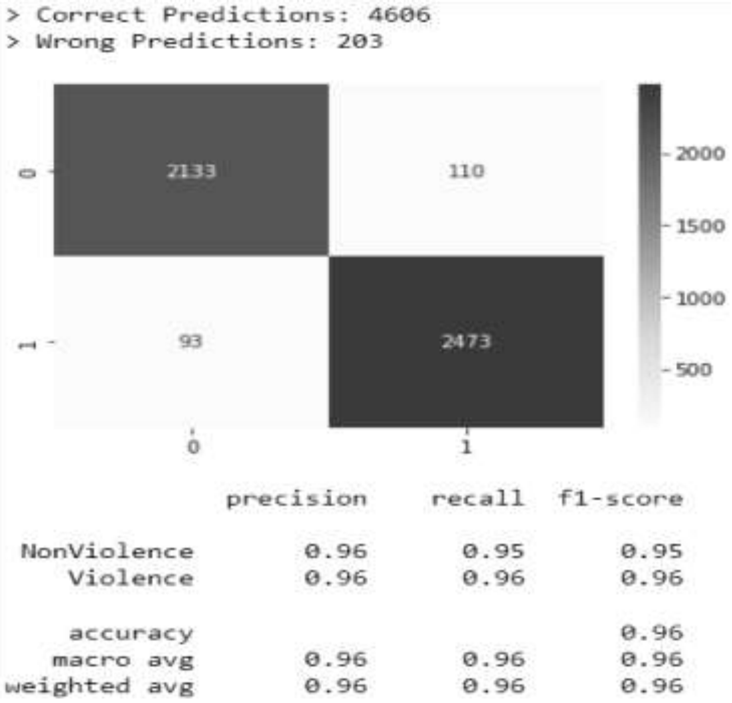


Figure 2: Anomalous activity detection results of proposed method, from left to right: normal activities video, violence detected scene.



```
> Correct Predictions: 4606
> Wrong Predictions: 203
```

```
                precision    recall  f1-score

 NonViolence       0.96      0.95      0.95
    Violence       0.96      0.96      0.96

    accuracy                           0.96
   macro avg       0.96      0.96      0.96
weighted avg       0.96      0.96      0.96
```

Figure 3: Accuracy measurement with the parameters precision, recall, and f1 score.

Table 1: Performance measure of existing and proposed anomalous activity detection methods for different surveillance videos.

Algorithms	*Pr (%)*		
	Dataset 1	*Dataset 2*	*Dataset 3*
3D CNN (Spiess et.al 2021)	86.79	87.88	87.41

CLSTM (Peker et.al 2022)	90.94	90.11	90.27
Proposed method	95.47	96.12	96.53

When the projected value coincides with the actual value, this is known as a true positive, which indicates that the model correctly predicts the value. When the projected value from the classification algorithm is negative and the actual value is negative, this is referred to as True Negative. There are two different sorts of errors: False Positive and False Negative. False Positive errors happen when the model predicts a positive estimation, but the real value is negative, and False Negative errors happen when the real data is positive, but the estimation produces a negative result.

The accompanying parameters Precision (Pr), Recall, and F1 are used to gauge how well the tracking method is being carried out. The condition Equation is used to acquire this parameter (1-3).

$$\mathrm{Recall} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{1}$$

$$\mathrm{Pr} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \tag{2}$$

$$\mathrm{F1}=2*precision*recall/precision+recall \tag{3}$$

Where,
TN - true negatives,
TP - true positives,
FN - false negatives, and
FP - false positives

The accurateness is determined by taking three ideas into account: measurement f1 score, recall, and precisionas shown in Figure3. precision as long as the overall percentage of cases that were accurately anticipated were positive. The accuracy value is used to gauge the model's dependability. Recall indicates a model's accuracy; the total positive examples indicate how well the model was able to accurately predict. F1-score is employed to obtain a combined understanding of the two variables. The harmonic mean is calculated using precision and recall values. My research has produced a useful model for identifying violent action. The model performs appropriately, beyond our expectations. CLSTM is used with a layer of 3D CNN to increase accuracy.

Conclusion and Future Work

The research on identifying violent behaviour is completed successfully. The trained model can analyse real-time video footage to identify violent or non-violent activities and send an alert to the closest authorised person for further action. In this study, many human action detection models and various computer vision analysis fields are examined. Since research is an ongoing process, newly developed methods may lead to future advancements. The CLSTM approach was used in this research to recognize action based on the spatial correlation between videos. However, as research is an ongoing process, different algorithms and methods could be used to improve the recognition's accuracy. Following the investigation, activity is discernible by estimating human stance. Limited datasets were used in this study, which were gathered from several online sources. With more particular category datasets, future study might be more effective. The two categories are used, one is for violence and another for no violence. The classification for violence includes footage from surveillance cameras showing fights, hijackings, and other violent acts. Newly gathered datasets could be used in future work to update it and produce a more precise and useful result.

Acknowledgements

Authors wish to thank the Department of Electronics and Communication Engineering and Department of Electrical and Electronics Engineering, Saveetha Engineering College for providing technical assistance throughout the research.

References

[1] Abhishek G, Alagan A, Ling G. and Ahmed S K (2021). Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues, Array. Volume 10, 100057.

[2] Aktı, G. A., Tataroğlu and Ekenel, H. K (2019). Vision-based Fight Detection from Surveillance Cameras. International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1-6.

[3] Amit S, Albert H, Alexandre A, Serena Y, Michelle G, Jill R G, William B, Terry P, Li F and Arnold M (2020). Automatic detection of hand hygiene using computer vision technology. Journal of the American Medical Informatics Association, 27 (8), pp. 1316–1320.

[4] Jain, A., and Vishwakarma, D. K(2020). State-of-the-arts Violence Detection using ConvNets. International Conference on Communication and Signal Processing, pp. 0813-0817.

[5] Kanagamalliga, S., & Vasuki, S. (2019). An efficient algorithm for tracking and counting pedestrians based on feature points in video surveillance applications. Journal of Intelligent & Fuzzy Systems, 36(1), pp. 67-78.

[6] Kanagamalliga, S., & Vasuki, S. (2018). Contour-based object tracking in video scenes through optical flow and gabor features. Optik, 157, pp. 787-797.

[7] Kanagamalliga, S., Vasuki, S., & Shanmugapriya, M. (2016). Foreground object detection using expectation maximization based effective Gaussian mixture model. Middle-East J Sci Res, 24(Special Issue on Innovations in Information, Embedded and Communication Systems), pp. 51-57.

[8] Ke R, Li W, Cui Z. and Wang Y (2020). Two-Stream Multi-Channel Convolutional Neural Network for Multi-Lane Traffic Speed Prediction Considering Traffic Volume Impact. Transportation Research Record, 2674(4), pp. 459-470.

[9] Md. Zabirul Islam, Md. Milon Islam and Amanullah, A (2020). A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. Informatics in Medicine Unlocked, Volume 20, 100412.

[10] Nasaruddin, N., Muchtar, K and Afdhal, A (2020). Deep anomaly detection through visual attention in surveillance videos. J Big Data 7, 87.

[11] Nguyen, T. N and Meunier J (2019), Anomaly Detection in Video Sequence with Appearance – Motion Correspondence. IEEE/CVF International Conference on Computer Vision, pp. 1273-1283.

[12] Zufan Z, Zongming L, Chenquan G and Qingyi Z (2020). Human action recognition using convolutional LSTM and fully connected LSTM with different attentions. Neurocomputing, Volume 410, pp. 304-316.