Undergraduate Conference on Intelligent Computing and Systems (UCICS 2025)
26-27 February, 2025; Varendra University, Rajshahi, Bangladesh

A Hybrid Deep Learning Framework For Suspicious Activity Detection in Video Surveillance

Romana Nourin Nipa
Dept. of Information and Communication Technology
Islamic University, Kushtia-7003, Bangladesh
romananourin@gmail.com

MD Jahid Hassan
Dept. of Information and Communication Technology
Islamic University, Kushtia-7003, Bangladesh
jahid.ict.iu@gmail.com

Md Faisal Ahammed
Dept. of Information and Communication Technology
Islamic University, Kushtia-7003, Bangladesh
faisal.ahammed.ict.iu@gmail.com

Ismatul Ferdush Liva
Dept. of Biomedical Engineering
Islamic University, Kushtia-7003, Bangladesh
livaismatul11@gmail.com

Md Abdullah Al Masud
Dept. of Computer and Communication Engineering
International Islamic University Chittagong
mdabdullahalmasud.ai@gmail.com

MD. Alamgir Hossain
Dept. of Information and Communication Technology
Islamic University, Kushtia-7003, Bangladesh
hossain@iu.ac.bd

***Abstract*—Over the last few years, video surveillance has become an essential means of ensuring security and safety in numerous domains. However, the detection of unusual and suspicious behaviors or movements in surveillance video still poses a significant problem due to the dynamically complex nature of human activity and context. While traditional methods are better at extracting spatial features, they don't include temporal attributes that are necessary to model how to find suspicious activity. To overcome these limitations, this paper proposes a novel deep-learning architecture called ConvGRU, combining Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU). This approach builds on the fact that CNNs are inherently capable of extracting spatial features, while GRUs seek to capture temporal dependencies; hence, the proposed method is more general in a framework than using CNNs only for crime and suspicious activity detection. In experiments on multiple video surveillance datasets, the model achieves an increased accuracy of 99.52%, in particular, it is effective in distinguishing among suspicious behaviors.**

***Index Terms*—ConvGRU, Suspicious Activity, Video Surveillance, Hybrid Deep Learning, CNN, UCF-Crime**

I. Introduction

Due to human behavior being so unpredictable, it can be difficult to determine whether anything is suspicious or normal [1]. Video surveillance is important and fascinating to industry and academia. Scientific interest in video surveillance has grown due to safety concerns in train stations, airports, military installations, and retail malls. Because building effective detection technologies requires knowledge of the target activity and environment, accurately detecting suspicious events is vital. A rare or irregular occurrence is referred to as a suspicious activity [2].

Unusual behavior is hard to spot. Video crime detection is popular due to the reliability of deep learning algorithms, especially CNNs. CNNs report video crimes without context or timing. CNN excels at geographic data extraction but not contextual and temporal correlations for suspicious behavior detection. Regular gun club shooting is dangerous. CNNs misread video surveillance abnormalities due to context ignorance. Abnormal behavior can occur without inherent causes [3]. Security cameras for theft or access are unsuitable for complicated and unexpected behavior.

From previous studies, it is found that CNNs, LSTMs, and GRUs misbalance spatial and temporal connections. CNNs extract features well but cannot identify sequential patterns, while GRU and LSTM models capture temporal connections but struggle with spatial information. Many solutions require labeled data or known traits, limiting their utility. Very few studies go beyond theft/intrusion. On surveillance tape, unexpected situations are hard to spot.

To overcome the shortcomings of individual GRU and CNN models, we introduced ConvGRU, which blends GRUs' temporal model with CNNs' spatial feature extractor. This improves the capacity to distinguish suspicious occurrences in surveillance footage by addressing spatial and temporal interactions.

Our key contributions are:

1) **Hybrid ConvGRU Model:** Integrates CNN's spatial feature extraction with GRU's temporal modeling for improved spatiotemporal learning.
2) **Improved Generalization:** Adapts to diverse suspicious activities beyond specific anomaly types.
3) **Enhanced Accuracy in Multiclass Suspicious Activity Detection:** Achieves higher accuracy in multiclass suspicious activity detection.

II. Literature Review

Video surveillance anomaly detection and suspicious human activity detection are popular research topics. From simple machine learning models to the most complicated deep learning architectures, each has pros and cons. To meet the changing nature of complex surveillance situations, this paper discusses recent improvements in techniques, their successes, limitations, and potential.

Kolaib et al. [4] proposed a deep learning based forensic criminal detection system with the help of UCF-Crime and DCSASS datasets. MobileNet-V2 and EfficientNet-B7 were outperformed by their proposed architecture with a DCSASS accuracy of 89% (88% F1 score) and UCF-Crime accuracy of 99.48% (99.44% F1 score). Yet, several issues are outstanding when deployed in real-time or when working with different datasets and for large-scale applications. We compared it with a recent work by Ahmed et al. [5] who employed a CNN-based autoencoder and GAN to detect suspicious actions in Sabbath videos with 97.5% on the UT Interaction dataset, 89.6% on HCA, and 47.34% on UCF-Crime. While it performed good on controlled data it fared very poorly on diverse data.

In 2024, Shrivas et al. [6] developed an enhanced ConvLSTM with spatial-temporal attention for the surveillance of educational facilities to identify and detect activities that are considered unusual to maintain order and security in the premises at 99% accuracy on UCF-Crime and 82.66% accuracies on the You Tube sourced educational anomaly videos. Nevertheless, they have the shortcomings of generalizing results to real-world situations. Later on, Pallewar et al. [7] worked on the concept of CNN-LSTM for unusual activities such as abuse, theft, and explosions and estimated 98.5% accuracy of the work in the real-world proof scenario for surveillance from the UCF-Crime dataset. Qasim et al. [8] developed an anomaly detection system using a CNN-SRU framework, where the model achieved a specificity of 88.92% with ResNet18+SRU, 89.34% with ResNet34+SRU, and 91.24% with ResNet50+SRU. Patwal et al. [9] proposed a DenseNet121-based CNN model for anomaly detection in CCTV footage, optimizing computational efficiency for real-time use. However, the lack of temporal modeling limits its ability to capture sequential dependencies, achieving 86.63% AUC on the UCF-Crime dataset.

III. Proposed Methodology

This section presents our proposed methodology, and Figure 1 illustrates the overall workflow of our study.

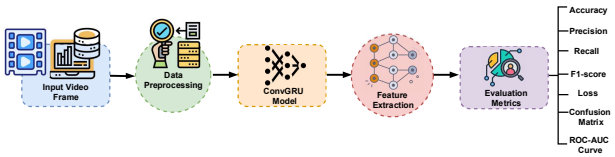


Fig. 1: Overall workflow of our study

A. Data Collection

This study uses the UCF-Crime dataset [10], which includes 128 hours of footage from 1,900 online RGB camera recordings in various circumstances. The database includes abuse, arrest, arson, assault, road accident, burglary, explosion, fighting, robbery, shooting, stealing, shoplifting, and vandalism. These movies may identify activity kinds and anomalies in each group in 13 real-world contexts. They then only annotated videos chronologically. We found irregularities in 338,784 frames from 12 suspicious activity types. The training and testing samples are as, Abuse: 15,558 train, 3,815 test; Arrest: 23,666 train, 6,096 test; Arson: 21,863 train, 5,351 test; Assault: 10,714 train, 2,600 test; Burglary: 37,766 train, 9,395 test; Explosion: 20,249 train, 5,014 test; Fighting: 20,623 train, 5,292 test; Vandalism: 11,798 train, 2,939 test; Stealing: 37,391 train, 9,395 test; Robbery: 33,850 train, 8,478 test; Shooting: 11,859 train, 2,911 test; Shoplifting: 25,991 train, 6,467 test, where the total dataset has been split into an 80:20 ratio.

B. Data Preprocessing

First, we turned videos into frames, regularly extracting images to produce a structured dataset for model development. The images were then grayscale, eliminating color information to simplify computation and highlight intensity-based patterns. Each image was then rescaled to consistent 50x50 pixels to guarantee standardizing feature dimensions for efficient learning and consistency over the dataset. To fit the input requirements of deep learning models, the processed images were last transformed into a structured four-dimensional format with a single grayscale channel.

C. CNN Block

The purpose of the CNN block is to identify suspicious activity patterns by extracting spatial information from the input images. The layers of the architecture are as follows:

1) **Convolutional Layers:**
   - **First Conv2D Layer:** Utilized 64 3x3 filters with the same amount of padding to maintain spatial dimensions. The following is a representation of the operation:
   $$Z = W * X + b \tag{1}$$
   where, W is the filter, X is the input, b is the bias, and $*$ denotes the convolution operation.
   - **Second Conv2D Layer:** Uses 128 filters with similar operations.
   - **Third Conv2D Layer:** Allows the model to capture more intricate patterns by increasing the number of filters to 256.
2) **Activation Function:** After every convolutional layer with α=0.1, LeakyReLU is added. This avoids the "dying neuron" issue by permitting slight gradients for negative values:
$$f(x) = \begin{cases} x & \text{if } x > 0, \\ \alpha x & \text{if } x \leq 0. \end{cases} \tag{2}$$

3) **Pooling Layers:** By choosing the maximum value in each 2x2 window, MaxPooling2D minimizes spatial dimensions while maintaining important features and lowering computing cost.
4) **Dropout:** used at rates of 0.25 and 0.4 to randomly deactivate a portion of neurons during training in order to avoid overfitting.
5) **Flatten and Dense Layer:**
   - **Flatten:** The 2D feature maps are flattened into a 1D vector.
   - **Dense Layer:** Leaky ReLU activation and 256 units make up this fully connected layer, which integrates spatial features into a compact representation.

D. GRU Block

The GRU block is crucial for identifying dynamic activity in videos because it records temporal connections in sequential data. The following are included in the architecture:

1) **GRU Layers:** $tanh$ activation produces a non-linear mapping by the sequential operation of two GRU layers with eight units each.

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tanh(W_h * [r_t \odot h_{t-1}, x_t]) \quad (3)$$

where, z_t is the update gate, r_t is the reset gate, and h_t is the hidden state at time t.
2) **Dense Layer:** Temporal information is compressed into a compact feature representation via a fully linked layer with four units and $tanh$ activation.
3) **Dropout:** To avoid overfitting and regularize the model, a rate of 0.2 is used.
4) **Flatten:** This actually transforms the GRU's output into a 1D vector so that it may be concatenated with the CNN block output.

E. ConvGRU Architecture

A fully connected layer with 12 units (for the 12 activity classes) and a softmax activation function receives the concatenated outputs of the CNN and GRU blocks:

$$y = \text{softmax}(W \cdot [f_{\text{CNN}}, f_{\text{GRU}}] + b) \quad (4)$$

The probability distribution across the activity classifications is output by this last layer, allowing for the accurate classification of questionable activity. Shown in Figure 2 the number of CNN layers for hierarchical spatial feature extraction and GRU units for temporal dependencies was tuned to balance model complexity, computational efficiency, and classification accuracy.

IV. Result and Discussions

A. Performance Evaluation and Comparative Analysis

The learning rate was varied throughout training, and the model was assessed at the rates in the table below. Training and assessment accuracy, precision, recall, F1-score, and loss for each learning rate. The accuracy and losses of training and testing are shown in Figure 3.

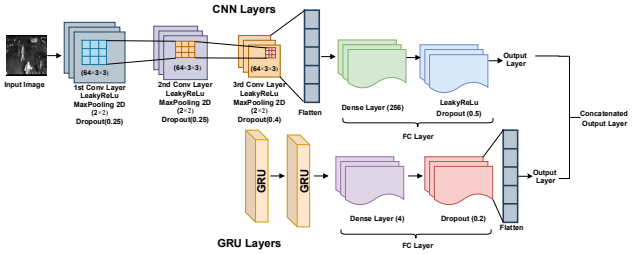
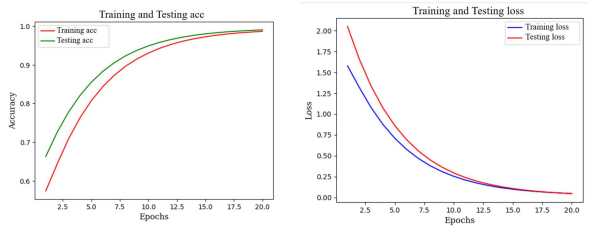


Fig. 2: Architecture of the proposed ConvGRU model

From Table I, for all learning rates the training and test accuracy is very high for the model. At learning rate of 0.001, we have the highest model performance where the test accuracy equaled to 99.52%, precision, recall, and f1=score equally equal to 99.52%, low loss of 0.0145.



(a) Accuracy (b) Loss

Fig. 3: Accuracy and Loss for learning rate 0.001

At a learning rate of 0.01, there is slight decrease in performance; test accuracy 98.26%; precision is 98.3%; recall is 98.26%; loss value is 0.0723. This implies that when using learning rate of 0.01 may overfit or may not converge optimally as intended during training.

Regardless of the inital learning rate as 0.0001, the model still performs well as it achieved a test set accuracy, precision, the recall, and F1 score of 99.47% and a low loss of 0.015. This shows that the lower learning rate offers a firm training straightforward and is again not that distinctive from the 0.001 learning rate when the test accuracy is considered.

To conclude, 0.001 is the best learning rate because it produces good results for performance metrics like loss value at points and total and ultimate loss for the data.

TABLE I: Performance metrics for different learning rates

Learning Rate	Accuracy (Train)	Accuracy (Test)	Precision	Recall	F1 Score	Loss
0.01	0.9859	0.9827	0.9831	0.9826	0.9827	0.0723
0.001	0.9968	0.9952	0.9952	0.9952	0.9952	0.0145
0.0001	0.9963	0.9947	0.9947	0.9947	0.9947	0.0150

From Table II our ConvGRU model outperforms other suspicious activity detection approaches. Singh et al. [11] and Nafim et al. [12] used DenseNet121 and ConvLSTM, respectively, to achieve 82.91% and 77%. Kumar et al. [13] combined CNN-LSTM but only got 49.04% due to temporal dependencies. In comparison, our ConvGRU model

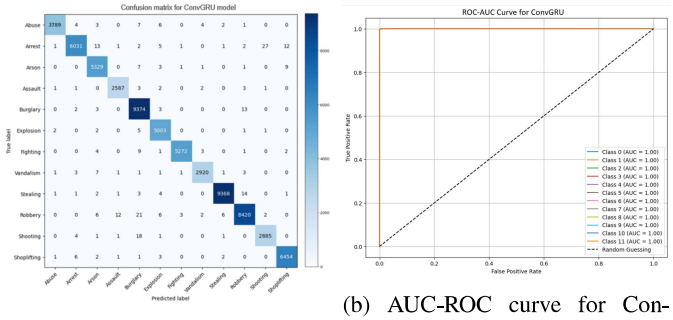
TABLE II: Comparison of different methods for suspicious activity detection

Reference	Method	Result
[11]	DenseNet121	82.91%
[12]	ConvLSTM	77%
[13]	CNN-LSTM	49.04%
Ours	ConvGRU	99.52%

captures geographical and temporal data with state-of-the-art 99.52% accuracy, proving its robustness in detecting suspicious surveillance video activity.

B. Error Analysis Using Confusion Matrix

The obtained confusion matrix as shown in Figure 4a indicates high degrees of classification accuracy and limited overlap of misclassified instances, which mainly happen between semantically related classes. For example, “Abuse” may sometimes belong to the category of “Arrest” or “Assault,” whereas “Arrest” may bear some similarity with “Arson,” and “Shooting.” That is why “Assault” can be easily mistaken for “Fighting,” and “Arson” has minor fuzz, being classified as “Explosion” or “Burglary.” Likewise, Shooting includes “Explosion” and, at times, Vandalism might fall under Abuse or Fighting. This is especially so because Misclassifications for the accusations “Stealing” and “Robbery” might include actions that are visually similar to ‘Shoplifting’ or ‘Shooting’ respectively. These errors are as a result of the similarities in patterns of these activity classes.



(a) Confusion Matrix

(b) AUC-ROC curve for ConvGRU model

Fig. 4: Confusion Matrix and AUC-ROC curve for ConvGRU model

C. AUC-ROC Curve

The ROC-AUC curve from Figure 4b illustrates perfect results for all the classes, with the AUC parameter having reached 1.000. This suggests that the class activity classification is perfect where the model no longer has any uncertainty when making the two classifications. The curve also demonstrates that the model is accurate and precise for all of the 12 classes, and even when making predictions for categories that may be difficult to predict with other AI algorithms or for which they may not obtain sufficient sample data to make accurate predictions, the model can still predict with acceptable precision.

V. Conclusion and Future Direction

The ConvGRU model has been implemented to give high understanding about video anomaly detection and the feature of integration between spatial and temporal related information have been also discussed. This work shows that architectures that incorporate CNNs and GRUs are better suited to identify sequences of events than the ones that do not combine these techniques. More precisely, our proposed hybrid method obtained performances concerning both false positive and false negative rates. Further, we noticed that the flexibility of the model enables it to solve different video scenes as the extraction and tracking behaviour are consistent with the model’s performance in this study. Epilogue While it is possible to argue that the presented model provides great results for lung cancer detection, there are some ideas for improving the future work such as the real-time capability of the system is an important aspect of using the deep learning model in surveillance settings, and further research should be aimed at increasing the processing speed of the system.

References

[1] K. Barsagade, S. Tabhane, V. Satpute, V. Kamble, Suspicious activity detection using deep learning approach, in: 2023 1st International Conference on Innovations in High Speed Communication and Signal Processing (IHCSP), IEEE, 2023, pp. 1–6.

[2] T. Sahar, M. Rauf, A. Murtaza, L. A. Khan, H. Ayub, S. M. Jameel, I. U. Ahad, Anomaly detection in laser powder bed fusion using machine learning: A review, Results in Engineering 17 (2023) 100803.

[3] M. Q. Gandapur, E2e-vsdl: End-to-end video surveillance-based deep learning model to detect and prevent criminal activities, Image and Vision Computing 123 (2022) 104467.

[4] R. J. Kolaib, J. Waleed, Crime activity detection in surveillance videos based on developed deep learning approach, Diyala Journal of Engineering Sciences (2024) 98–114.

[5] W. Ahmed, M. H. Yousaf, A deep autoencoder-based approach for suspicious action recognition in surveillance videos, Arabian Journal for Science and Engineering 49 (3) (2024) 3517–3532.

[6] A. M. Shrivas, D. K. Singh, R. Jain, S. S. Chandel, A. Sahu, Enhanced convlstm with hierarchical attention scheme for unusual activity detection in classrooms or laboratories, in: 2024 IEEE Region 10 Symposium (TENSYMP), IEEE, 2024, pp. 1–6.

[7] M. G. Pallewar, V. R. Pawar, A. N. Gaikwad, Unusual human behavior analysis using the deep learning, in: 2024 International Conference on Emerging Smart Computing and Informatics (ESCI), IEEE, 2024, pp. 1–5.

[8] M. Qasim, E. Verdu, Video anomaly detection system using deep convolutional and recurrent models, Results in Engineering 18 (2023) 101026.

[9] A. Patwal, M. Diwakar, V. Tripathi, P. Singh, An investigation of videos for abnormal behavior detection, Procedia Computer Science 218 (2023) 2264–2272.

[10] U. C. f. R. In Computer Vision, CRCV | Center for Research in Computer Vision at the University of Central Florida, [Online; accessed 31. Dec. 2024] (Dec. 2024).
URL https://www.crcv.ucf.edu/projects/real-world

[11] A. Singh, A. Bajaj, A. Singh, S. D. Saha, A. Sharma, Exploring anomaly detection techniques for crime detection, in: The International Conference on Recent Trends in Communication & Intelligent Systems, Springer, 2024, pp. 183–201.

[12] I. H. Nafim, S. A. Tonni, H. A. Runa, Criminal activity detection from videos under low light condition using deep neural network, Ph.D. thesis, Brac University (2024).

[13] M. Kumar, M. Biswas, Abnormal human activity detection by convolutional recurrent neural network using fuzzy logic, Multimedia Tools and Applications 83 (22) (2024) 61843–61859.