Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Suspicious Human Activity Recognition for Mobile Robot

Souhila Kahlouche[1], El-Fani Roufaida[2], Tazekritt Malika[2]
[1] *Centre de développement des Technologies Avancées (CDTA) Cité 20 août 1956 Baba Hassen, Alger, Algérie*
[2] *Université des Sciences et Technologies Houari Boumediene (USTHB)*
BP 32 Bab Ezzouar, 16111 - ALGER
skahlouche@cdtad.dz roufaida.elfani@gmail.com tazekrittmalika08@gmail.com

ARTICLEINFO

Received: 28 Dec 2024
Revised: 18 Feb 2025
Accepted: 26 Feb 2025

ABSTRACT

This work aims to develop a real-time application for surveillance robot in indoor environment, to recognize human suspicious activity using deep learning architecture applied on visual data. To learn different classes of activity, a combination of three deep-learning algorithms have been used, based on the idea that a set of classifiers improves machine learning accuracy. The ensemble of classifier has been trained on a collected public dataset containing videos of human activities with normal and suspicious behavior. The model is then, evaluated in real time scenarios, and the experimental results show that the proposed system has the potential to benefit applications in surveillance robots

Keywords: Surveillance robots, suspicious activity recognition, convolutional network, ensemble classifier, Video Surveillance.

1. INTRODUCTION

Nowadays, mobile robot are moving outside their classic passive role, where they can basically detect events and raise alarms[1], to active surveillance robots, that can interact with their environments[2] and with other robots for cooperative or collaborative task [3].

Figure 1 shows the main components included in surveillance mobile robot systems; The first one is a reactive navigation system in which the mobile robot moves, while avoiding obstacles in its environment, using the collected information provided by sensors. The second one uses camera for a real-time scene analysis and human suspicious activity recognition, for alerting the security expert, when abnormal behavior occurred.

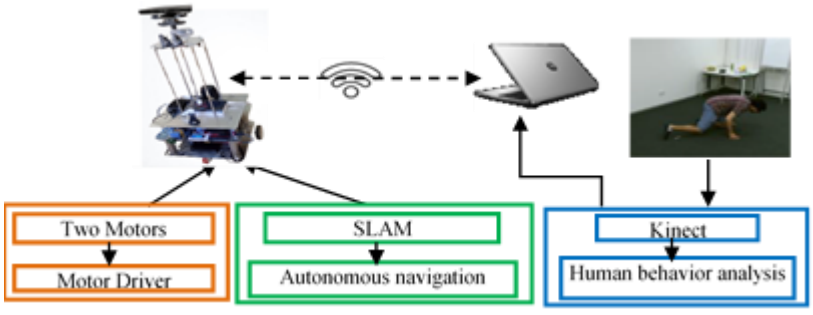


Figure1. System architecture [4]

Actually, it is difficult for a surveillance robot, to constantly monitoring public spaces to detect and identify suspicious or abnormal human behavior. Therefore, action recognition is the main processes in building intelligent machines that deal with real-world scenarios; this process is able to infer labels to actions from a series of observations. In this paper, we will focus on the development of an intelligent real-time framework that can spot suspicious activities.

The early works for suspicious action recognition was based on conventional hand-crafted approaches, such as K-Nearest Neighbour (KNN), Support Vector Machine (SVM), decision tree, K-means, or Hidden Markov Models

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

(HMMs)[5] .... However, these techniques require an expert to identify and define features and descriptors, and mostly, they are not suitable for real time applications.

Recently, deep learning and specially Convolutional Neural Network (CNN) [6], have been actively used and achieved surprising performance for various computer vision problems. Nevertheless, Convolutional Neural Network architectures are designed to learn spatial features, while Human Activity Recognition is time-series data including spatial and temporal information, which requires more robust model with the potential to extract both information at the same time. On the other hand, many studies demonstrated that even though some methods based on machine learning or deep learning have been proposed. The need for highly accurate, highly precise, low-false-positive and low-false-negative prediction system can be enhanced by the use of, hybrid ensemble learning algorithm in suspicious action recognition, which is a technique that combines the predictions of multiple classifiers to form a single classifier, which results in a higher accuracy than any of individual classifier [7] [8].

2. OBJECTIVES

In this study, we investigate different deep learning architectures that are diverse and yet accurate to extract both spatial and temporal features from visual data to recognize human suspicious activities such as kicking, pushing, staggering, vomiting, falling, and touch pocket. Then, we employ an ensemble model to combine the investigated models in order to achieve an improved performance of the system.

In summary, this study has the following main contributions:

(1) Preparing a dataset of six suspicious activities from three public datasets: NTU RGB dataset, SBUKINECTINTERACTIONS dataset and UR FALL datasets.

(2)The use of 3D convolution neural network architecture for extending the time dimension to capture motion information from video frames.

(3) Fine tuning 3D CNN and gated recurrent units (GRUs).

(4)Combining two CNNs for spatial and temporal data respectively. Therefore, in addition to the spatial RGB network, a temporal network for optical flow data is trained to learn the displacement field.

(5) The use of ensemble classifier to allow better predictive performance than single classifier method.

3. METHODS

The general suspicious human action recognition mechanism is illustrated in figure2. It has two components: The offline process where videos from the dataset are pretreated and then three deep learning architectures are trained. The online process where the real time data stream of the activity being performed is inferred using voting process of the above classifiers.

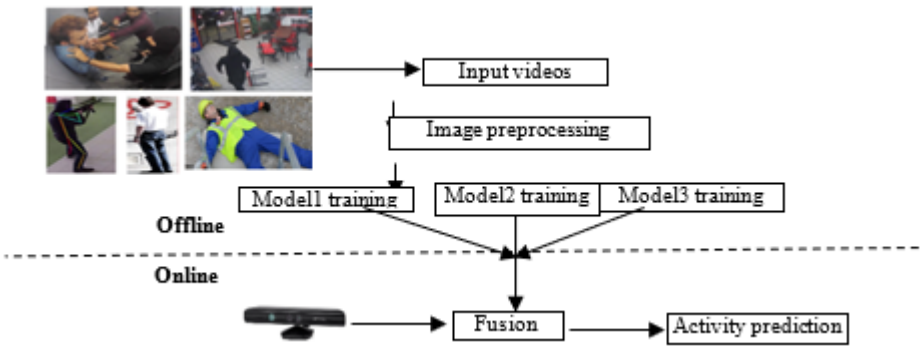


Figure 2. The block Diagram of our solution.

3.1 Dataset gathering: To train our models we have prepared a dataset of six suspicious activities from tree different datasets which are:

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

(A) NTU RGB dataset [9]:This dataset from Rapid-Rich Object Search Lab (ROSE) for action recognition, it contains 60 action classes within three major categories, which are: daily actions, mutual actions and medical conditions (see figure 3). We have chosen six (6) action classes of RGB video samples that we qualified as suspicious activity, these classes are: kicking, pushing, staggering, and vomiting, falling and touch pocket.

(B) SBU-Kinect-Interaction dataset [10]: The SBU Kinect interaction dataset consists of RGB, depth images, and tracked skeleton data acquired by an RGB-D sensor. It includes eight activities: *approaching, departing, pushing, kicking, punching, exchanging objects, hugging,* and *shaking hands* (see figure 4). From this dataset, we took videos of kicking and pushing activities.

(C) The UR Fall Dataset URFD [11]: This dataset contains 70 (30 falls + 40 activities of daily living) sequences. Fall events are recorded with two Microsoft Kinect (RGB + Depth) cameras and corresponding accelerometric data. Figure5 shows some example video frames.

FIG.3: Sample frames of "NTU RGB" dataset, FIG.4: Sample frames the UR Fall Dataset URFD, FIG. 5: Sample frames SBU-Kinect dataset

3.2. Data preprocessing

Converting videos into image frames: The video size samples are from 30 to 90 frames, which leads to high computational cost. Hence, to speed up the training process, we have used a fixed jump step. Based on our experiments, we took one in every seven (7) frames, this allows to remove the redundant images without significantly reduce the quality of the video. This fixed jump step provides good compromise between the minimum required frame to have enough information about the action a recognition accuracy.

Image resizing: Rescaling the frames size by 10% against their original shape, it is done to reduce the allocated space memory and to reduce the neural network size, thus, increasing both the space and time complexity in the training phase.

Normalization: The main advantage of normalization is to remove the intra-class variation between data of different persons (figure 6). It is computed as follow:

$$Nor(I_j)=(I_j-\min(I))/(\max(I)-\min(I))$$ Where: I_j is the pixel j of the image matrix I.

Figure 6: Example before and after normalization

3.3. Models building: Three deep architecture have been proposed:

(A) 3D Convolutional Neural Network (Conv3D) model: In order to add the temporal dimension to the feature map, a 3D convolution trained [12], it is realized by convolving a 3D kernel to a cube of stacked adjacent frames (see figure 7). Therefore, video frames were convolved by moving in three directions (x, y, t) to extract features from both spatial and temporal dimensions. Then multi-channel action features were convolved and pooled continuously (i.e.

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

sub sampled). Finally, each channel feature was further mapped and fused to obtain action feature representation, and then the features were classified.

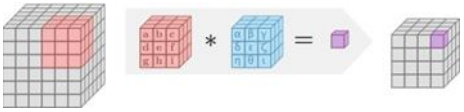


Figure 7: Conv3D illustration.

The convolution operation at position (x,y,z) in the j[th]feature map in the i[th] layer at time *t* is given below:

$$v_{ij}^{xyz} = f(b_{ij} + \sum_{m}\sum_{0}^{p}\sum_{0}^{q}\sum_{0}^{t} w_{i,j,m}^{p,q,t} * v_{(i-1)m}^{(x+p,y+q,z+t)})$$

f: non-linear activation function (Tanh, Sigmoid, Relu,...)

$w_{i,j,m}^{p,q,t}$: The kernel linked to the convolutional feature map in the previous layer and*t* is the 3D kernel size along the temporal axis.

*bij :*bias term.

(p,q,t): height, width and depth of the kernel.

Figure 15 shows the fundamental architecture of our model, it outlines:

- Four Conv3D layers to produce batches of 3D feature maps containing spatiotemporal features.

-Two (2) Max pooling layers of kernel size (1,2,2), to scale down the extracted feature maps.

- Flattened layer, which is a concatenation of the features into vector of one dimension.

-Two fully connected layers where RELU and Softmax scores are computed in the last layers for final classification. A detailed description of the layers proposed 3D-CNN architecture is shown in fig.8.

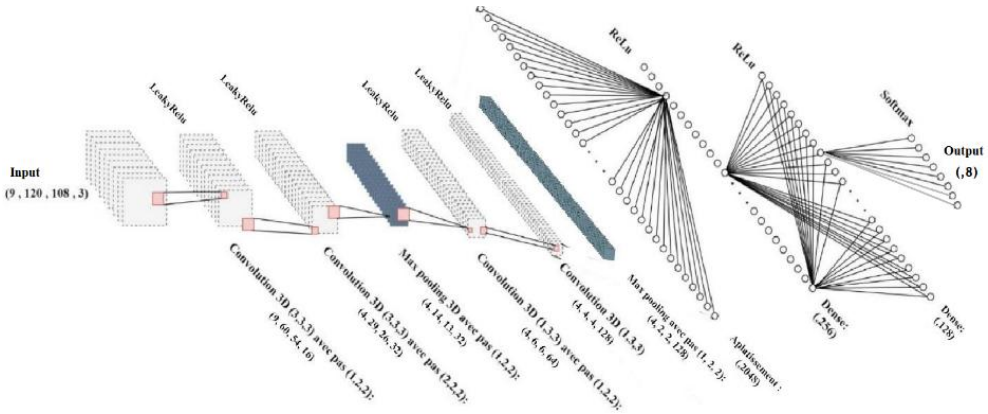


Figure 8: Architecture of 3D-CNN of RGB data.

(B)3D Convolutional Neural Network with Gated Recurrent Units (GRUs):GRU [13]is designed to model sequential data by allowing information to be selectively remembered or forgotten over time. I is done by processing sequential data one element at a time, updating its hidden state based on the current input and the previous hidden state. At each time step, the GRU computes a “candidate activation vector” that combines information from the input and the previous hidden state. This candidate vector is then used to update the hidden state for the next time step.

Our proposed architecture is similar to the previous 3D CNN architecture, to which we added Time distribution layer where features are separated into vectors to take advantages of time sequencing, followed by GRU layer of 256 units just before the fully connected layer (see figure 9), here RELU is used as activation function.

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

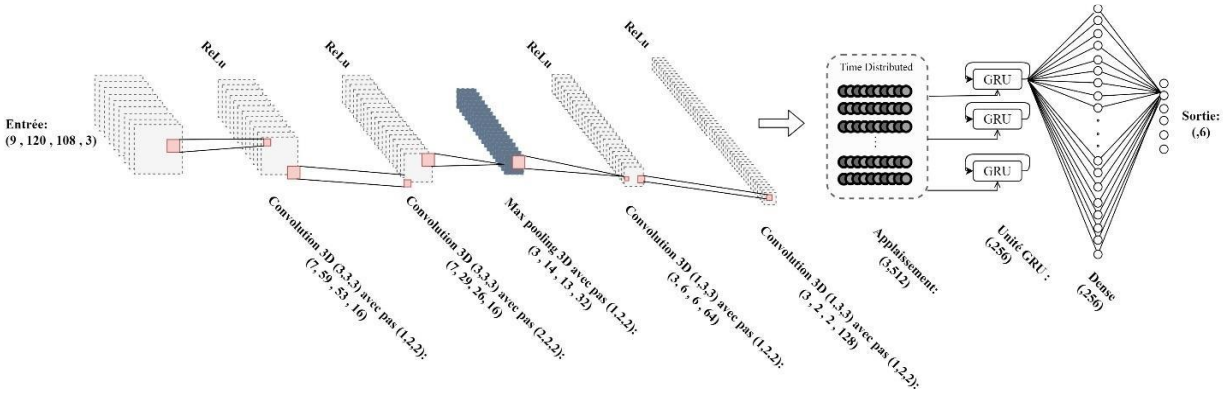


Figure 9: Architecture of 3D-CNN of RGB data.

(C) CNN-OF fusion model building**:** In order to extract the temporal and spatial features of action videos respectively, dual-stream network architecture is proposed, it uses two independent convolutional neural networks (CNN) for each stream (time stream and space stream).

For the time stream, dense optical flow [14] is calculated every two frames in the video and a convolutional neural network is trained on the optical flow data as an image. In parallel, a CNN model is trained on the RGB stream. Finally, the extracted features from both streams are fused and classified.

Figure 10 shows the fundamental architecture of the proposed model, it outlines the fusion of two deep ConvNet over Ten (10) layers. For each stream, we implemented:

- Four (4) 3D CNN layers with stride (1,2,2) for all convolution layer, the goal of these layers is to produce a batch of 2D feature maps.

-Two (2) Max pooling layers of kernel size (1,2,2), to scale down the extracted feature maps.

- Flattened layer, which is a concatenation of the features into vector of one dimension.

- Fully connected layer where Softmax score is computed in the output layer for final classification.

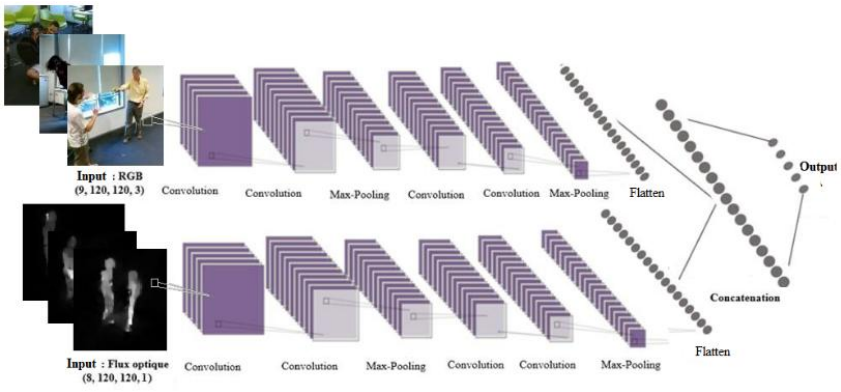


Figure 10: Architecture of 3D_CNN of RGB and Optical Flow data.

3.4. Models training and evaluating

To train our deep learning models we splitted 75% of the data for training and 25% for testing.We have used, for training, the Keras deep learning framework with a Tensor Flow backend. The network has been trained using Adam optimizer, and Categorical Cross Entropy as loss function. After several attempts of parameter tuning, the best results are obtained using a learning rate of 10-5 with an initial training rate set to 0.001, a batch size of 8.

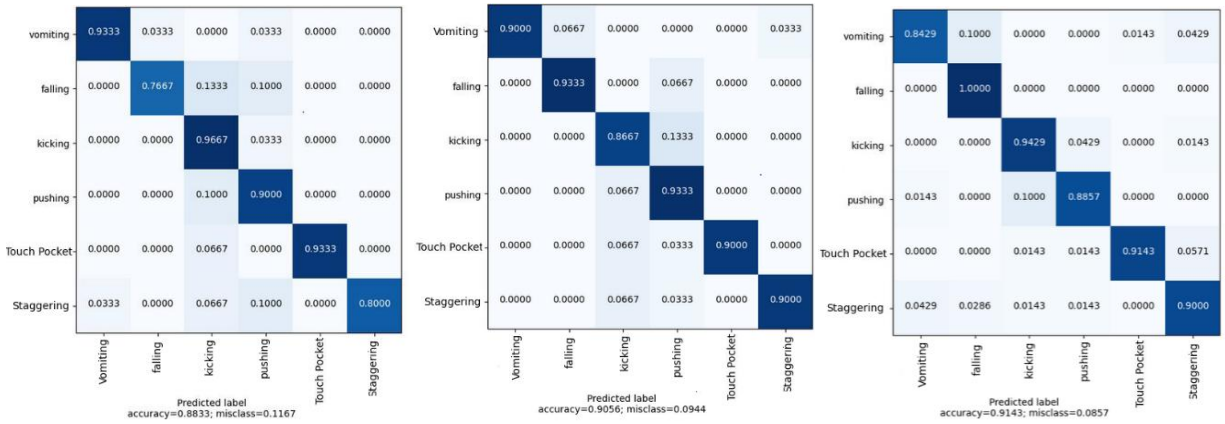
The performance of our modelshas been evaluated based on the accuracy percentage of activities that are correctly recognized. This metric reflect the model efficiency. It is computed as follow:

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Accuracy = (True Positive +True Negative)/ (True Positive +True Negative +False Positive + False Negative);

Figure 11 presents the confusions matrix of the three models. We noticed that 3D CNN and 3D CNN+ GRU models achieved accuracy of (88.33 %) and (90.56 %) respectively, while the combined 3D CNN and optical flow model accomplished better result with average accuracy of (91.66 %), where most actions have been correctly classified with height accuracy (>90%).



(a) (b) (c)

Figure 11: Confusions matrix of (a) 3D CNN, (b) 3D CNN+GRU, (c) 3D CNN+OF.

3.5. Ensemble learning: The use of an ensemble of classifiers model generally allows better predictive performance than the performance achievable with a single model [15]. Therefore, the previous three models are combined using the majority of outputs as ensemble methods. The majority of outputs decide the most voted class among outputs.

We have used majority-voting process as follows:

Let $M_i = \{m_1, m_2,\ldots,m_n\}$, be the ensemble of model trained and $C_j = \{c_1, c_2,\ldots, c_l\}$, be the classes.

The vote for each class j and each model is computed by:

$$v_{i,j}=\begin{cases} 1 & if\ mi\ votes\ for\ cj \\ 0 & else \end{cases}$$

The total vote for each class j is then calculated by $V_j=\sum_{i=0}^{n} v_{i,j}$ j=1,....,l

The final decision V_{final} is then the majority vote. $Vfinal=argmax(Vj)$.

Figure 12 shows the confusion matrix of suspicious activity recognition using voting process, which achieves accuracy of 94.66%.

The performance of the model has been evaluated using other metrics like precision, recall, and F-score.

Precision: this measure refers to the number of actual positive classes out of all the positive classes that was predicted correctly. It is calculated using the following equation:

Precision =True Positive / (True Positive +False Positive)

Recall: this measure refers to how much correct predictions are made out of all the positive classes.

Recall = True Positive/ (True Positive+False Negative)

F-measure is a measure of the accuracy of the test. It is defined as a weighted mean of precision and recall. Its best value is 1 and its worst value is 0.

*F-measure = 2∗(Precision * Recall) / (Precision+ Recall)*

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 1 shows a summary of the above performance metrics (%) for all activities; Accuracy, Precision, Recall and F-measure.

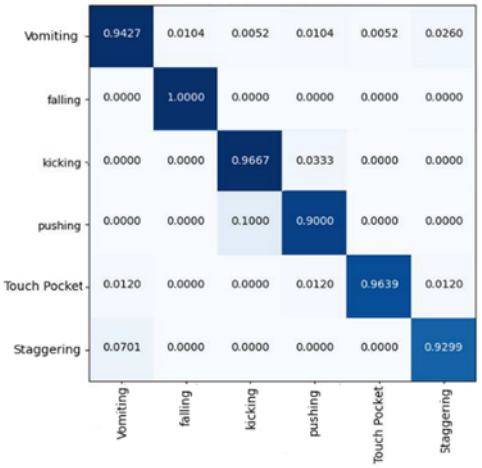


Figure 12 : Confusions matrix of ensemble classifier model.

Table 1: Performance by activity

Activity	Accuracy	Precision	Recall	F-measure
Vomiting	94	96	94	95
Falling	100	100	100	100
Kicking	96	95	96	95
Pushing	90	90	90	90
Touch pocket	96	95	96	96
Staggering	92	93	92	92

4. RESULTS

In order to demonstrate the performance of our model in real time, it has been evaluated in real scenarios on unseen data, Kinect camera has been used to capture video stream at refresh rate of 30 frames per second. For real time prediction, we have used the last 30 frames recorded during one second. In all experiments, we used a laptop with an i5-2320 (3.00GHz) CPU.

Figure 14 shows the obtained results with activities: kicking, Vomit, Falling, Staggering and touch pocket respectively.

As we can see, the proposed framework has been able to recognize different suspicious activities in real time, which confirm the usefulness of our proposed approach to recognize accurately the performed suspicious human activity in real time.

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

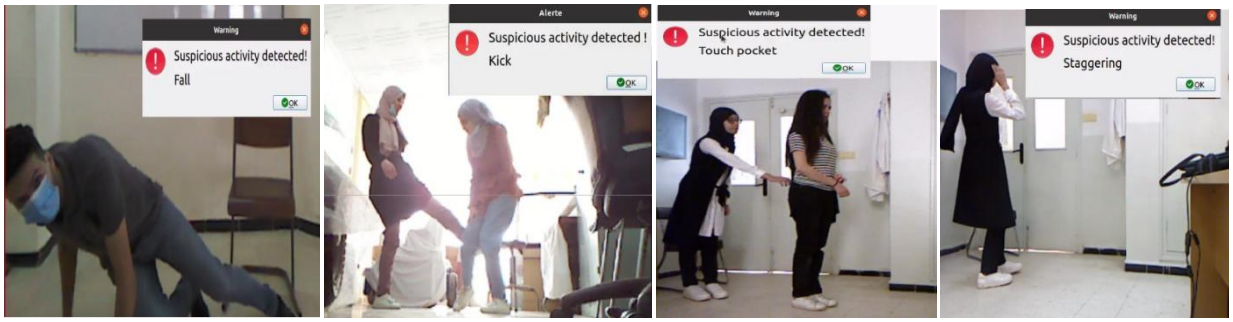


Figure 14: Examples of suspicious activity recognition in real time

5. DISCUSSION

To ensure better security in indoor environment, and to help the surveillance team in detecting abnormal human behaviors, the main challenge was the integration of high-level decision-making for different operative scenarios. Therefore, a module using video camera for suspicious activity recognition has been developed, for this purpose a large dataset of abnormal behavior in different scenarios and under varying lightening conditions has been collected and used for training and evaluating the model. Initially, three deep learning architectures have been proposed using 3D CNNs, GRU, and Optical flow respectively, and then a model combining them to take advantages of their diverse strengths and to compensate their weaknesses is proposed using the majority voting mechanism. This model improved the capacity to detect anomalies more accurately across varied video surveillance scenarios. Our future work aims to create our own dataset for suspicious activities including criminal actions like shoot with gun, wield knife, etc.,

REFRENCES

[1] Bai, Y. W., Xie, Z. L., & Li, Z. H. (2011). Design and implementation of a home embedded surveillance system with ultra-low alert power. IEEE Transactions on Consumer Electronics, 57(1), 153-159.

[2] Zhang, P., Zhang, Y., Thomas, T., & Emmanuel, S. (2014). Moving people tracking with detection by latent semantic analysis for visual surveillance applications. Multimedia tools and applications, 68, 991-1021.

[3] Chen, X., Zhang, P., Du, G., & Li, F. (2019). A distributed method for dynamic multi-robot task allocation problems with critical time constraints. Robotics and Autonomous Systems, 118, 31-46.

[4] S. Kahlouche, D. Dellaa and N. Hamdaoui, "ROS-Based Indoor Surveillance Mobile Robot", 2024 2nd International Conference on Electrical Engineering and Automatic Control (ICEEAC), pp. 1-6.

[5] M. Keyvanpour and F. Serpush, “ESLMT: A new clustering method for biomedical document retrieval,” Biomedical Engineering, vol. 64, no. 6, pp. 729–741, 2019.

[6] Hascoet, T., Zhuang, W., Febvre, Q., Ariki, Y. and Takiguchi, T. (2019) Reducing the Memory Cost of Training Convolutional Neural Networks by CPU Offloading. Journal of Software Engineering and Applications, 12, 307-320. doi: 10.4236/jsea.2019.128019

[7] Kahlouche, S., &Belhocine, M. (2021). Human Activity Recognition Based on Ensemble Classifier Model. In Proceedings of the 4th International Conference on Electrical Engineering and Control Applications: ICEECA 2019, 17–19 December 2019, Algeria (pp. 1121-1132). Springer Singapore.

[8] Zahid, Y., Tahir, M. A., Durrani, N. M., &Bouridane, A. (2020). Ibaggedfcnet: An ensemble framework for anomaly detection in surveillance videos. IEEE Access, 8, 220620-220630.

[9] A. Shahroudy, J. Liu, T. Ng, and G. Wang, “NTU RGB+D: A large scale dataset for 3d human activity analysis,” CoRR, vol. abs/1604.02808, 2016. [Online]. Available: http: //arxiv.org/abs/1604.02808

Research Article

[10] Kiwon Yun, Jean Honorio, DebaleenaChattopadhyay, Tamara L. Berg, and Dimitris Samaras, The 2nd International Workshop on Human Activity Understanding from 3D Data at Conference on Computer Vision and Pattern Recognition (HAU3D-CVPRW), CVPR 2012.

[11] BogdanKwolek, Michal Kepski, Human fall detection on embedded platform using depth maps and wireless accelerometer, Computer Methods and Programs in Biomedicine, Volume 117, Issue 3, December 2014, Pages 489-501, ISSN 0169-2607

[12] Ji, S., Xu, W., Yang, M., &Yu, K. (2012). 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1), 221-231.

[13] Dua, N., Singh, S. N., &Semwal, V. B. (2021). Multi-input CNN-GRU based human activity recognition using wearable sensors. Computing, 103(7), 1461-1478.

[14] Zach, C., Pock, T., &Bischof, H. (2007). A duality based approach for realtimetv-l 1 optical flow. In Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29 (pp. 214-223). Springer Berlin Heidelberg.

[15] Diao, R., Chao, F., Peng, T., Snooke, N., & Shen, Q. (2013). Feature selection inspired classifier ensemble reduction. IEEE transactions on cybernetics, 44(8), 1259-1268.