Deep-3DConvNet: A Network to Detect Abnormal Activities at Megastores

Mohd. Aquib Ansari
CSED, MNNIT Allahabad,
Prayagraj, India
mansari@mnnit.ac.in

Dushyant Kumar Singh
CSED, MNNIT Allahabad,
Prayagraj, India
dushyant@mnnit.ac.in

***Abstract*— These days, there has been a rapid increase in cases of abnormal human behavior at megastores/shops, where people commit theft by stealing, consuming, or unwrapping packets when no one is seeing and then leaving the place without paying. Such unusual actions cause huge losses in business. Therefore, there is an urgent need to attract the research community's attention to detect abnormal events at megastores. To address this issue, we have designed an advanced three-dimensional convolutional neural architecture to identify abnormal activities at megastores. The proposed network is 15 layers deep, takes a video stream of resolution 120×120 as input, and produces classification results as output. It extracts fine-tuned as well as general details from the video feed using small and large-sized 3D convolutional filters and categorizes them into respective classes. The proposed architecture is trained and tested on a synthesized action dataset that consists of human actions distributed into five classes: normal, stealing, eating, drinking, and damaging acts. Experimental results show that our model outperforms other state-of-the-art approaches with an accuracy of 88.88%.**

Keywords*— *Intelligent video surveillance, Human activity recognition, Video sequences, spatiotemporal features, Convolutional neural network (3DCNN).

I. Introduction

Nowadays, video surveillance has been widely adopted by various shopkeepers across the world. The shop owners cover almost every corner of the shop through CCTV cameras so that everyone in the shop can be kept under watch. Here, the control room personnel closely examine the video stream received from the CCTV cameras and alert the store owner if there is anything related to abnormal behavior occurs. The efficacy and trustworthiness of such systems are determined primarily by the alertness and ability of the person examining the array of video frames. Looking at video footage extensively for hours leads to weakness and loss of focus, which can lead to neglecting some critical frames and being left unsupervised by security personnel. Hence, the monitoring process requires to be automated [1, 2].

In shopping or any commercial environment, expected normal behavior can be recorded as taking the item off the shelf, inspecting the items, returning the item to the shelf, wandering on the floor, talking with others, etc. However, any deviation from the expected normal behavior is called an abnormality, including hiding items inside clothing or bags, consuming edible items, and opening the packet. Such kinds of abnormal activities cause huge losses to the retailers in the business. Therefore, an automated and proactive video surveillance system [3, 4, 5] is in serious demand to seamlessly examine video feeds and produce an alert when anyone is found engaged in such kind of abnormal or unusual behavior at megastores.

This article mainly focuses on accessing the above-mentioned unusual scenarios on megastores, aiming to improve store item's security so that no one can steal, consume, or damage them. In order to get the solution, we propose an advanced three-dimensional convolutional network to analyze the video feeds and predict human behavior seamlessly. The network is 15-layers deep that utilizes 3x3, 5x5, and 7x7 filters to extract low-level and generic details from the video feeds. In addition, a video dataset presented here is synthesized at the laboratory, including five classes: normal, stealing, eating, drinking, and damaging. The proposed model is evaluated on this synthesized dataset, showing effective outcomes with an accuracy of 88.88%.

This article is divided into five parts. After summarizing the problem in Section I, we discuss the related work on human behavior analysis in Section II. Section III briefly overviews the advanced 3D ConvNet-15 architecture used for identifying abnormal events at megastores. The next part of the article consists of the experimental results assessed on synthesized training and testing inputs. Finally, the last section concludes the work.

II. Related Work

This section critically reviews various state-of-the-art techniques related to human behavior analysis. These techniques can be feature learning based and deep learning based. Feature learning based approaches use manually engineered features to extract particular information from video sequences and then build any classifier on such features to perform action classification. In contrast, Deep learning-based approaches automatically extract the relevant features from video feeds and then classify them into one of the respective classes. Some of the HAR based approaches are as follows:

978-1-6654-9291-1/22/$31.00 ©2022 IEEE

Dwivedi et al. [6] propose a feature learning based activity classification method. This method first uses silhouette information to capture skeleton data and then uses morphological operations to extract relevant features. Here, the random forest classifier utilizes such feature sets for action classification. Aly and Sayed [7] propose a HAR method for identifying human actions using motion features. The method first creates the motion energy images using the frame differencing algorithm. It then uses the Zernike moment descriptor to extract local and motion patterns from these energy images. Finally, such features are used to build SVM classifiers for action classification. Jayaswal and Dixit [8] propose a deep learning based method to classify abnormal human actions in public places. This method first uses the exception model to extract relevant features from the video sequences and then LSTM networks to build spatial relationships between those features. This method provides well-suited results on the synthesized dataset but cannot handle occluded action accurately. Yue-Hei Ng et al. [9] analyze two streams (i.e., Raw and optical flow) to classify human behavior in video feeds. This method uses convolutional neural networks to extract contextual features from both streams, which are combined to get single-length vectors. Recurrent neural layers are then set on top of the concurrent features to classify the video sequences.

Apart from these, a three-dimensional convolutional network (3DCNN) has become a prevalent approach in the field of human activity recognition. Unlike 2DCNN, it can extract spatiotemporal features from video sequences that are used to encode structural dynamics for any given action. Some of the research based on the 3DCNN architecture to trace human activities are as follows:

Arunnehru et al. [10] use 3DCNN with 3D motion cuboids for activity representation. This method first utilizes a frame differencing algorithm to generate 3D cuboids from video sequences and then uses 3DCNN architecture for activity classification. Two sets of convolution with max-pooling operations are part of this 3DCNN architecture. Here, Weizmann and KTH dataset is used to measure the efficacy of the proposed method. Vrskova et al. [11] use two sets of convolutions with max-pooling and two sets of two convolutions with max-pooling operations for representing 3DCNN architecture. Going a little deeper with 3D convolutions, this method turned out to be good for recognizing human behavior in real time. Kanagaraj and Priya [12] use the 3DCNN architecture to achieve encouraging outcomes for classifying multimedia events. They use three sets of two 3DConvolutions with max-pooling operations to build the 3D CNN architecture. Donahue et al. [13], Roy and Mishra [14], and Anjali and Beena [15] also utilize 3DCNN architecture for recognizing human activities in real time.

In existing research, no work has been found that can classify human theft-related actions such as stealing, eating, drinking and damaging store items at megastores. Therefore, we focus on dealing with such kinds of cases using advanced 3DCNN architecture. Other side, the reason behind choosing 3DCNN network for performing activity classification is that 3DCNN can represent spatiotemporal features to encode the action dynamics accurately, but 2DCNN cannot. Therefore, we built an advanced deep architecture of 3DCNN to correctly represent the action dynamics presented in video sequences.

III. Proposed Method

This work is intended to identify various abnormal scenarios at megastores by analyzing human behavior. Figure 1 presents the overall workflow of our proposed method with the proposed 3DConvNet architecture. Here, the camera module captures the video feeds from the input scenes. Further, N frames are inputted to the proposed three-dimensional convolutional network (3DConvNet), where N is set to 145. This network then classifies the video feeds into one of these five classes: normal, shoplifting, eating, drinking and damaging, as shown in Fig. 1. Fig. 2 shows the deep architecture of the proposed 3DConvNet, which is 15 layers deep. It comprises nineteen 3DConvolution operations to extract spatiotemporal information from the set of N video sequences and five 3DMaxPooling operations to reduce the structural representation. This network also comprises two concatenation layers to concat the branches of convolutions, three batch normalization layers to standardize the inputs to the subsequent layer for each mini-batch and two fully connected layers with one output layer to learn the pattern and classify the instances accordingly.

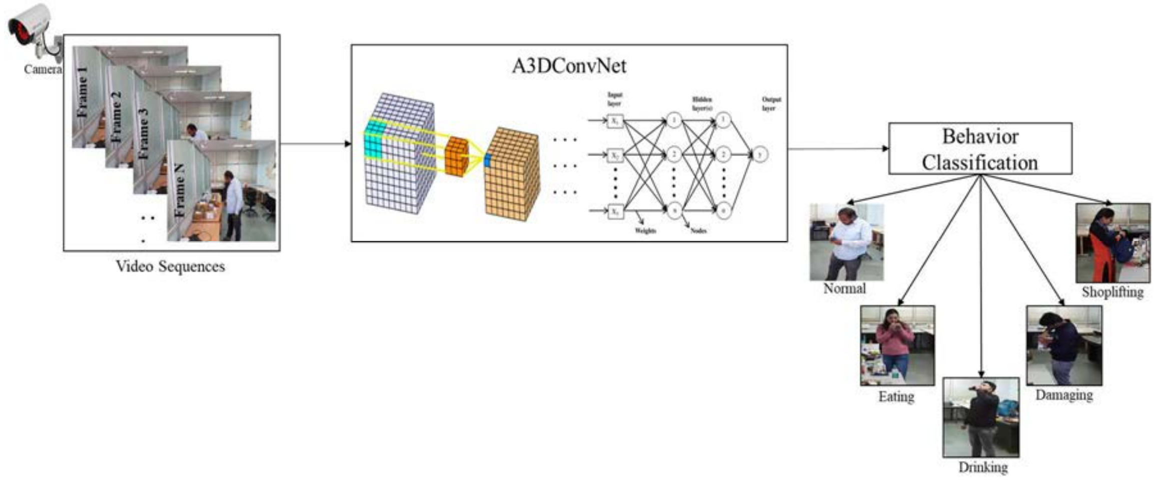

Fig. 1. Workflow of the proposed methodology

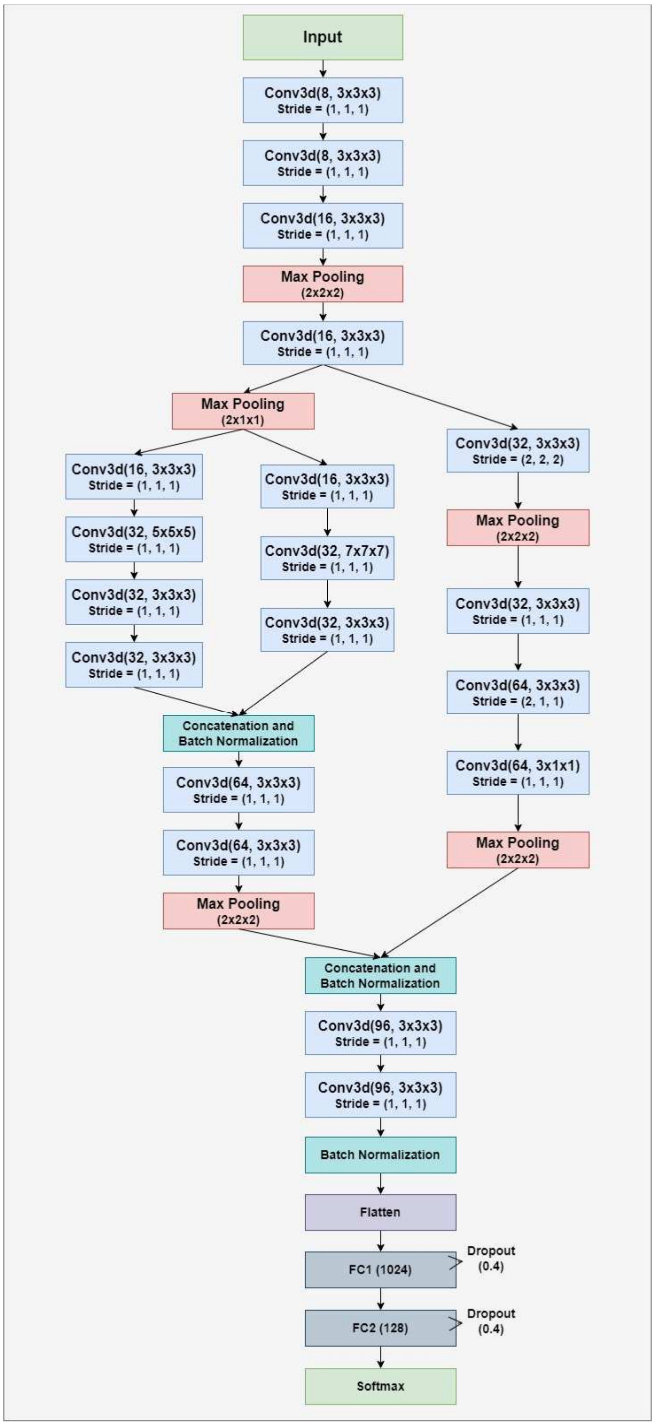

Fig. 2. Proposed 3DConvNet architecture

This network utilizes different sizes of 3Dconvolution operations such as (3×3×3), (5×5×5) and (7×7×7) to extract the pertinent features from the feeds. In this network, we have taken advantage of both small and large-sized convolutions. Here, convolutional operations with small filter sizes such as (3×3×3) evaluate the low-level or fine-tuned details. Other side, convolutional operations with larger filter sizes, such as (5×5×5) and (7×7×7) evaluate more generic features from the video sequences. In this network, low-level features are evaluated on the one hand and generic features are evaluated on the other, then they are merged accordingly, as shown in Fig. 2. Finally, the features extracted from the proposed architecture are used to represent action dynamics for any action.

IV. Results and Discussion

Experiments have been done on a machine running a Windows environment with Core i7, 16GB RAM, 512GB hard disk, and 4GB GTX NVIDIA 1050 Ti GPU capacity. Python 3.6 with tensor flow 1.14 is used as a programming environment.

Here, all the experiments have been performed on the self-synthesized dataset, which was created in the computer vision laboratory of MNNIT Allahabad. The clips present in this dataset comprise clearly performed and distinguishable human actions. This dataset consists of 400 video clips, where each clip is 10 seconds long. This dataset consists of five classes: normal, shoplifting, eating, drinking and damaging, where each class consists of 88, 85, 75, 72, and 80 clips, respectively. Out of 400 clips, 301 clips are used in training, and the rest of the clips are used for testing purposes, as presented in Table I.

TABLE I. Dataset's train-test distribution

Categories	Distribution	Clips	Total Clips
Train	Normal	66	301
	Shoplifting	64	
	Eating	57	
	Drinking	54	
	Damaging	60	
Test	Normal	22	99
	Shoplifting	21	
	Eating	18	
	Drinking	18	
	Damaging	20	

We have conducted experiments over different sizes of input data, i.e., 80×80, 120×80, 120×120, 160×120, and 160×160. The result infers that data samples with a resolution of 120×120 have acquired up to 88.88% detection accuracy, which is the optimum. Decrease or increase the resolution from 120×120, the detection accuracy starts to drop, as presented in Table II.

A confusion matrix is generally used to describe the performance of a classification model on a set of test data for which the true values are known. Table III shows the confusion matrix obtained over a synthesized dataset of 120×120 input size using the proposed method. Here, the method discriminates eating and drinking classes more accurately with very few false positive instances than other classes. Other side, normal, shoplifting and damaging classes incur a few more false positive instances during the validation process.

TABLE II. ACCURACY OBTAINED OVER DIFFERENT RESOLUTIONS OF INPUT DATA

Metrics	**Resolution (Pixels)**				
	80 × 80	**120×80**	**120 × 120**	**160 × 120**	**160 × 160**
Training Accuracy	94.83%	92.23%	96.20%	96.85%	95.28%
Validation Accuracy	84.84%	87.87%	88.88%	85..85%	82.82%

TABLE III. CONFUSION MATRIX

Class		**Predicted Values**				
		Normal	**Shoplifting**	**Eating**	**Drinking**	**Damaging**
Actual Values	**Normal**	19	2	0	0	1
	Shoplifting	1	18	0	0	2
	Eating	0	0	17	1	0
	Drinking	0	0	1	17	0
	Damaging	2	0	1	0	17

Precision is the measure of quality that represents correctly classified examples out of a total of predicted positive examples.

$$Precision = \frac{TP}{TP + FP} = 89.19\%$$

Recall is the measure of the quantity that represents the correctly classified examples out of the total actual positive examples.

$$Recall = \frac{TP}{TP + FN} = 89.05\%$$

Other hand, accuracy provides the ratio of the total number of correct predictions:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = 88.88\%$$

Finally, as exposed in the results, the proposed model achieves a good precision rate, recall rate, and classification accuracy on the synthesized training input, which are up to 89.19%, 89.05%, and 88.88%, respectively.

Fig. 3 shows a comparative analysis of the proposed method with existing state-of-the-art approaches. Here, the existing state-of-the-art approaches utilize sequential 3DCNN architectures, comprising 3D convolution and max-pooling operations of kernel size 3x3x3 for evaluating fine-tuned features in the local neighborhood. In contrast, our model consists of 15 layers of deep architecture of 3DConvNet that can extract fine-tuned and generic features using lower and larger kernel sizes (i.e., 3x3x3, 5x5x5, and 7x7x7). The performance of state-of-the-art approaches has also been evaluated on the same synthesized training inputs. The evaluated results show that our method outperforms other state-of-the-art approaches.

V. CONCLUSIONS

This paper provides a comprehensive overview of abnormal activity detection at megastores that the computer vision community has recently addressed. It can be used to prevent various intentional theft-related acts such as stealing, eating, drinking, and damaging store items at megastores. We have utilized an extended 3DCNN architecture that is 15 layers deep. The proposed method can extract spatiotemporal features from video sequences and classify them into one of the related classes, namely normal, stealing, eating, drinking and damaging. Evaluated results from extensive experiments show accurate results of up to 88.88% accuracy, 89.19% precision, and 89.05% recall on the synthesized dataset, indicating that our method is capable of modeling an individual's abnormal behavior in a theft-related context at megastores. The method may be expanded in the future to support the fast detection of actions in a real-time environment. In addition, more classes could be added to the dataset to detect a wider range of theft incidents at Megastores. In addition, more advanced deep convolutional architectures, such as quantum convolution, can be deployed to accurately understand action dynamics in the videos.

REFERENCES

[1] Manaf, Abdul, and Sukhwinder Singh. "Computer Vision-based Survey on Human Activity Recognition System, Challenges and Applications." 2021 3rd International Conference on Signal Processing and Communication (ICPSC). IEEE, 2021.

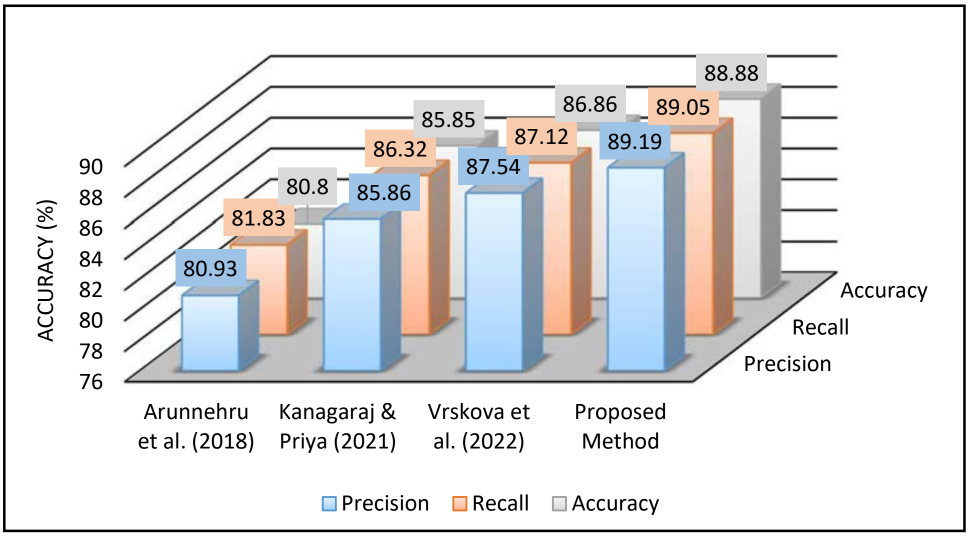

Fig. 3. Comparison with existing methods

[2] Singh, Dushyant Kumar, et al. "Human crowd detection for city wide surveillance." Procedia Computer Science 171 (2020): 350-359.

[3] Chen, Kaixuan, et al. "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities." ACM Computing Surveys (CSUR) 54.4 (2021): 1-40.

[4] Singh, Dushyant Kumar, and Dharmender Singh Kushwaha. "Tracking movements of humans in a real-time surveillance scene." Proceedings of fifth international conference on soft computing for problem solving. Springer, Singapore, 2016.

[5] Ansari, Mohd, and Dushyant Kumar Singh. "An Expert Eye for Identifying Shoplifters in Mega Stores." International Conference on Innovative Computing and Communications. Springer, Singapore, 2022.

[6] Dwivedi, Neelam, Dushyant Kumar Singh, and Dharmender Singh Kushwaha. "Orientation invariant skeleton feature (oisf): a new feature for human activity recognition." Multimedia Tools and Applications 79.29 (2020): 21037-21072.

[7] Aly, Saleh, and Asmaa Sayed. "Human action recognition using bag of global and local Zernike moment features." Multimedia Tools and Applications 78.17 (2019): 24923-24953.

[8] Jayaswal, R. and M. Dixit, "A framework for anomaly classification using deep transfer learning approach," Revue d'Intelligence Artificielle 35.3 (2021): 255-263.

[9] Yue-Hei Ng, Joe, et al. "Beyond short snippets: Deep networks for video classification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[10] Arunnehru, J., G. Chamundeeswari, and S. Prasanna Bharathi. "Human action recognition using 3D convolutional neural networks with 3D motion cuboids in surveillance videos." Procedia computer science 133 (2018): 471-477.

[11] Vrskova, Roberta, et al. "Human Activity Classification Using the 3DCNN Architecture." Applied Sciences 12.2 (2022): 931.

[12] Kanagaraj, Kaavya, and G. G. Priya. "A new 3D convolutional neural network (3D-CNN) framework for multimedia event detection." Signal, Image and Video Processing 15.4 (2021): 779-787.

[13] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko and T. Darell, "Long-term recurrent convolutional networks for visual recognition and description," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625-2634, 2015.

[14] Roy, Aprameyo, and Deepak Mishra. "ECNN: Activity Recognition Using Ensembled Convolutional Neural Networks." TENCON 2019-2019 IEEE Region 10 Conference (TENCON). IEEE, 2019.

[15] Anjali, C., and M. V. Beena. "Human Activity Recognition using Convolutional 3D Network." Int. J. Res. Eng. Sci. Manag 2 (2019): 832-836.