Journal of Electrical Engineering, Vol. 74, No. 3, 2023, pp. 140-153

sciendo

PAPER

Detecting abnormal behavior in megastore for intelligent surveillance through 3D deep convolutional model

Mohd. Aquib Ansari[1], Dushyant Kumar Singh[2] and Vibhav Prakash Singh[3]

The use of neural networks in a range of academic and scientific pursuits has introduced a great interest in modeling human behavior and activity patterns to recognize particular events. Various methods have so far been proposed for building expert vision systems to understand the scene and draw true semantic inferences from the observed dynamics. However, classifying abnormal or unusual activities in real-time video sequences is still challenging, as the details in video sequences have a time continuity constraint. A cost-effective approach is still demanding and so this work presents an advanced three-dimensional convolutional network (A3DConvNet) for detecting abnormal behavior of persons by analyzing their actions. The network proposed is 15 layers deep that uses 18 convolutional operations to effectively analyze the video contents and produces spatiotemporal features. The integrated dense layer uses these features for the efficient learning process and the softmax layer is used as the output layer for labeling the sequences. Additionally, we have created a dataset that carries video clips to represent abnormal behaviors of humans in megastores/shops, which is a consequent contribution of this paper. The dataset includes five complicated activities in the shops/megastores: normal, shoplifting, drinking, eating, and damaging. By analyzing human actions, the proposed algorithm produces an alert if anything like abnormalities is found. The extensive experiments performed on the synthesized dataset demonstrate the effectiveness of our method, with achieved accuracy of up to 90.90%.

Keywords: Video Surveillance, Human Detection, Human activity recognition (HAR), Abnormal behavior, 3D Convolutional Neural Network (CNN), Deep Neural Architecture

1. Introduction

In today's world, surveillance has evolved into an essential component in monitoring day-to-day activities to ensure human safety and asset protection. It is available in almost all the places like parking, hotels, stores, airports, railways, business enterprises, etc. These places are being monitored by security personnel through CCTV-based surveillance. However, active surveillance with almost 100% attentiveness is next to impossible because of human body constraints. Other side, monitoring reports [6] also confirm that the misclassification rate may rise up to 95% after 22 minutes in recognizing activity happening in video footage obtained through CCTV. Hence, the process can be automated, which could help in an active security exercise and allow improvement in the surveillance system.

Video sequences acquired from the camera are the main constituent of the surveillance system. These systems use computer vision algorithms to identify activities in the video sequences and raise alarms when some unsafe events are recorded. Computer vision algorithms mostly use activity detection methods [1-3] to identify person's usual or unusual behavior in the video sequences. Even though a lot of work has been done to recognize activities over the years, it is still an open and demanding topic. Earlier, the researches were focused on modest datasets with controlled settings to recognize human activities. However, with the rapid evolvement of video content, more realistic HAR datasets were developed to understand the content of realistic videos. The video-based HAR remains unsatisfactory despite of tremendous progress over the past decade. The difficulties associated with pattern analyses of spatiotemporal data in real-time videos are believed to be the major problem, including viewpoint variations, camera jitters, occlusion, etc. Training deep models over a huge amount of videos too is a daunting task. In order to address these challenges, researchers have proposed various handcrafted and deep learning based models. Handcrafted methods represent human action using expertly designed features. Histogram of Oriented Gradient [4], Motion Boundary Histogram Features [5] and Histogram of Optical flow [6] are the popular handcrafted features. R. Arroyyo et al. [5], Hiroaki and This [13], N Dwivedi et al. [14], and Tam and Ngoc [15] used handcrafted features for human action representation. Other side, deep neural based models like Convolutional Neural Networks (CNNs) [2, 7, 21] perform end-to-end learning where feature extraction and classification tasks are done internally. Their associated complexities allow models to derive more benefit from training inputs. Fatemah and Mahdi [8], A.B. Sargano et al. [10],

[1] School of Computer Science Engineering and Technology, Bennett University, Greater Noida, India
[1,2,3] Department of Computer Science & Engineering, MNNIT Allahabad, Prayagraj, India
mansari.aquib@gmail.com[1], dushyant@mnnit.ac.in[2], vibhav@mnnit.ac.in[3]

https://doi.org/10.2478/jee-2023-0020, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X

Donahue et al. [7], Kanagaraj and Priya [11], Guillermo A. et al. [12], H Riaz et al. [16], and Ruchi and Manish [17] have taken advantage of various deep neural architectures to encode action dynamics in accordance with human behavior.

The research presented in this paper focuses on detecting the possibly abnormal behavior of humans in megastores/malls through video surveillance. This is because so far many such incidents have come to the fore in stores/malls, due to which the business has suffered a lot. The National Retail Federation (NRF) [18] also sees theft as a severe problem in retail shrinkage. Therefore, this research focuses on detecting the abnormal acts that can harm the retailers in the business. These abnormal behaviors include shoplifting, eating, drinking and damaging. Here, shoplifting is a type of theft in which a person steals the shop's goods and leaves the store without any payment. The shoppers are as well found picking and eating the eatable products from the store like chips, fruits, cokes, etc., and leaves without paying for it. The another case could be damaging items, where the shoppers damage the products by tearing the product packaging. All cases likely happen in megastores when no one is watching. Such human conduct is a crime and could lead to a loss in business enterprise. Therefore, there is a need for an automated mechanism to identify these people by analyzing their abnormal actions and raising the alarm on the occurrence of such events. In order to get to this objective, we propose a deep neural architecture named advanced three-dimensional convolutional network (A3DConvNet) to facilitate abnormal human behavior in megastores. This network uses 3D convolutional operations of different sizes to extract meaningful information from the video sequences, which is then used by feed-forward neural layers to perform further classification tasks.

The prime contributions of this paper are listed as follows:

1. This paper presents a 15 layers deep architecture of three-dimensional convolutional network to accurately detect abnormal human acts in megastores.
2. This paper introduces a store realistic videos (SRV) dataset created in-house, which carries video clips consisting of five actions like normal, shoplifting, eating, drinking and damaging, and performs a wide range of experiments.
3. This paper also presents a wide range of experiments and the related analysis to see the effectiveness of the proposed model.

Finally, the rest of the paper is framed as follows. Section 2 presents the literature survey on existing HAR methods. Our methodology, along with modified three-dimensional convolutional neural architecture, is presented in Section 3. Section 4 details about the experimentation, results analysis and related discussions. Finally, the last chapter concludes the paper.

2. Related works

We review various HAR methods to track person's usual or unusual behavior for surveillance videos in the literature. The literature to explore usual/common scenarios like daily life activities, sports activities and others is discussed as follows:

In [8], a hierarchical method to recognize complex human actions has been proposed using Background Subtraction (BG), Histogram of Oriented Gradient (HOG), Deep Neural Networks (DNN) and Skeletal Modelling. Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) networks are used to retrieve the reduced feature sets rather than the complete feature set. Here, the KNN-Softmax classifier is used to label the actions into respective classes. Da Silva and Marana [9] proposed a HAR system to recognize the daily life activities of people's using postural data analysis. The system encodes the postural information into parameter space to retrieve pertinent features. The extracted features are further used by fitting GMM and SVM classifiers to categorize their actions into their respective classes. A. B. Sargano et al. [10] have categorized sports activities using a HAR framework. This framework uses a deep CNN model (AlexNet) and a deep rule-based classifier to classify 50 sports activities, making it more transparent and interpretable than others. Donahue et al. [7] have encoded real-time visuals using CNN and LSTM networks. This network maps variable-length videos to text as output by modeling complex temporal dynamics. Kanagraj and Priya [11] detect multimedia events using a modified 3D CNN structure and achieve encouraging performance in event classification.

Added to it, the literature related to recognition methodologies for crime intentional unusual behavior in surveillance videos is presented here. Guillermo A. [12] uses a basic structure of the 3D-CNN network for criminal intention detection, like shoplifting. The model comprises four 3D convolution layers (to capture long-term dependencies), two 3D Maxpooling layers and two fully connected layers. This model is trained and tested for the UCF crime dataset and found to be efficient for classifying suspicious behavior in shops/stores. The trials performed reveal that this model performs better for both balanced and unbalanced datasets. Hiroaki and Thi [13] proposed a framework to detect chain snatching through a surveillance camera. The framework first uses background detection and pedestrian tracking algorithms to track each individual. A feature extraction algorithm is then used to extract meaningful information from the tracked pedestrians by the weighted decision fusion method performed on different parameters like appearance, motion and area features. Finally, classifier algorithms are used to detect the chain-snatching events in video footage. Unfortunately, this is a passive approach in which the method only gives an alert only after losing the person's belongings. R. Arroyo et al. [5] have designed a surveillance system to categorize various unsafe threats in stores/shops by analyzing customer's behavior. The system first detects humans using the proposed human detection scheme, where HOG, LBP and GCH are used to extract features and an SVM classifier is used to categorize humans in video sequences. Thereafter, the video processing system processes the resulting trajectories of persons to analyze their behavior and identify likely alarming circumstances in stores. These potential circumstances include entry or exit events, unattended cash desks and loitering behavior of people.

N. Dwivedi et al. [14] suggested an efficient method to detect unattended object through CCTV surveillance. The proposed method utilizes the background subtraction algorithm to extract foremost objects, which are further differentiated into static and dynamic objects. The discovered object is labeled as abandoned when it does not move over a predetermined time interval and its dimensions are between predefined values. In [15], a spatiotemporal feature based model has been proposed to detect abnormal movements in video sequences. The model uses HOG, MBH and HOF descriptors to extract features and an improved support vector descriptor (DVDD) to categorize the activities. The model proposed can identify anomalies in an occluded environment at a low computational cost. H. Riaz et al. [16] detect unusual behavior of candidates in examination halls by cascading deep CNN models. The methodology uses the OpenPose algorithm to extract key body points, which are further used to retrieve patches. Thereafter, these patches are used by the dense CNN model to discriminate anomalous actions. In [17], an anomaly classification model has been developed to classify various violent activities in public places using deep learning approaches. The model utilizes a deep Xception model to mine relevant features and a deep LSTM network to detect violence in video sequences. It has achieved a remarkable accuracy of up to 97.25% over the self-synthesized HBD21 dataset.

The literature, as presented above, shows different machine learning-based and deep learning-based approaches for action representation. Since machine learning-based approaches rely heavily on handcrafted features, scaling up activity recognition to a complex high-level understanding is difficult and time-consuming. Other side, researchers have proposed deep learning-based approaches for getting better HAR models. Generally, deep HAR models use two-dimensional CNNs to extract spatial features from input scenes. However, 3D-CNN is an advanced CNN architecture that can represent an action using spatiotemporal features. Therefore, we present here an advanced 15 layers-deep 3DCNN architecture to get better spatiotemporal relation among sequences than existing 3DCNNs. Added to it, our supportive approach helps to prevent unusual cases like shoplifting, eating, drinking and damaging acts in stores/shops. The proposed method detects such abnormal human behaviors using deep neural architecture and generates an alert if anything related to the above-listed crimes happens. This is the first work that analyzes human behavior to predict different kinds of potential theft in stores/shops.

3. Proposed methodology

The proposed methodology for monitoring and identifying actions leading to criminal activities in shops or megastores is illustrated in Fig. 1. The approach conferred uses N frames/video sequences taken from the CCTV cameras as input for the later part of the algorithm. The value of N is set here to 145, which means that portraying a complete action is done using 145 frames. A deep multistage network is proposed to exploit multi-dimension features relevant to activity identification. Single Shot MultiBox Detector (SSD) is used in the first stage to spot the presence of the human subject and a 15 layers deep advanced 3D ConvNet architecture is used in the later stage to represent and categorize the scene semantics with action dynamics. Generating alerts on the screen could be a visual output in the end when anything unusual event is recorded.

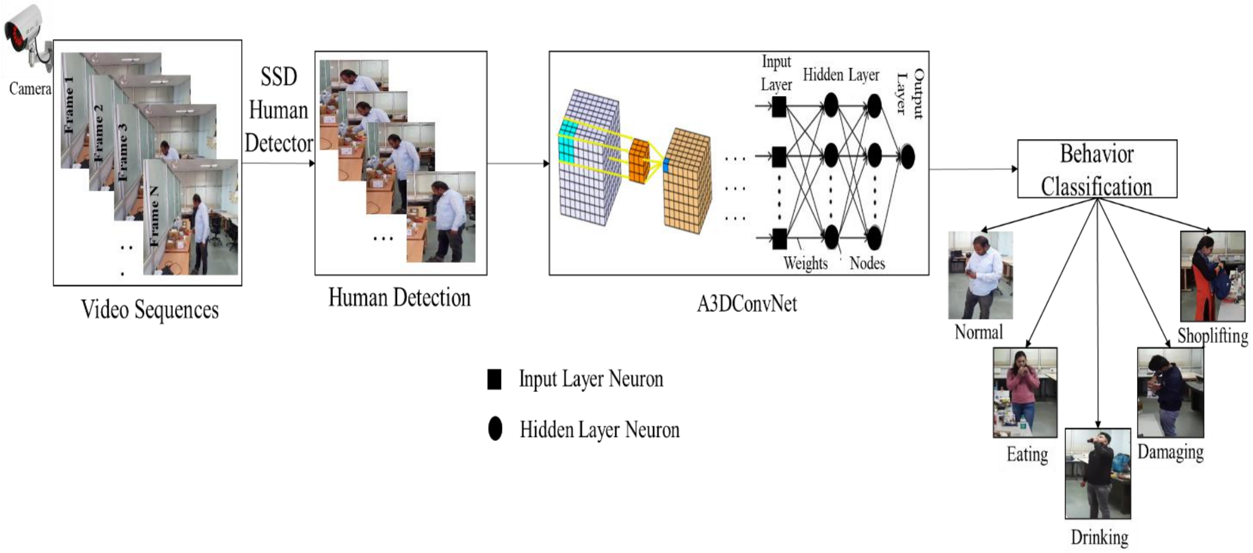


Fig. 1. Workflow of the proposed methodology

Earlier HAR-based researches [6, 11, 12, 17, 20] processed entire video sequences to analyze the individual's behavioral patterns. However, too much background information presented in video sequences can lead to inappropriate results. To deal with this issue, we first segment the object of interest (i.e., person) from each sequence and then pass them to the proposed 3DCNN network. A pre-train Single Shot MultiBox Detector (SSD) network [19] is used to segment the foreground and detect human subjects in real time. SSD network uses feed-forward convolutional network to localize and classify objects in one pass. It gives accurate outcomes close to two-stage detectors. SSD network discretizes the bounding box's space into defaults boxes for different scales and aspect ratios. It then generates a score for each default box, which is further adjusted to match the object shape. At last, SSD checks human presence for the detected object and then we perform segmentation to get the object of interest (i.e., person) from each sequence. After getting the segmented sequences, the next stage uses the A3DConvNet model to identify unusual events. The discussion related to the advanced A3DConvNet architecture is presented as follows.

3.1. Advanced three-dimensional convolutional network (A3DConvNet)

After the great sensation of 2DCNN, 3DCNN [7, 18] has also excelled in video processing tasks. 3DCNN works in single-stage pipelining, where feature extraction and classification tasks are performed in a single module. It can build a spatiotemporal relationship for the sequence predicting problems more efficiently, whereas 2DCNN can't. The convolutional operation performed over three dimensions space (x, y, z) is the strength of the 3DCNN network, where the kernel revolves in 3 dimensions and captures better dependencies among sequences. Equation 1 shows the value at each location of the feature map in the respective layers.

$$Out_{pq}^{x,y,z} = \tanh\left(b_{pq} + \sum_{r} \sum_{s=0}^{S_i - 1} \sum_{t=0}^{T_j - 1} \sum_{u=0}^{U_i - 1} W_{ijr}^{stu} v_{(p-1)m}^{(x+s)(y+t)(z+u)} \right) \tag{1}$$

Here, U_i is the 3D kernel's size and W_{ijr}^{stu} is the kernel's value associated with the feature map in the preceding layer.

The proposed advanced three dimensional convolutional architecture that is 15 layers deep, called A3DConvNet-15, is presented in Fig. 2. It contains a series of 3D convolution layers integrated with 3D max-pooling, concatenation, batch normalization, fully connected and output layers, which are discussed as follows:

- **3DConvolution layer:** It is used to extract pertinent information/features from the inputted representation in three-dimensional space. A three-dimensional kernel filter is used to perform a 3D convolution operation to evaluate the low-level representation.
- **MaxPooling layer:** It reduces the size of the image representation. A 3-D max-pooling layer performs downsampling by splitting the three-dimensional input into cuboidal pooling regions and computes the maximum of each region.
- **Batch Normalization layers:** It is used to regularize the preceding layer for each batch. It transforms mean activation to zero and standard deviation to one.
- **Concatenation layer:** It is used to concatenate two layers.
- **Flatten layer:** It converts or flats the matrix to the vector.
- **Fully connected layers:** It is the regular and deeply connected neural network used to categorize the data.
- **Output layer:** It uses softmax function to produce output.

The existing 3DConvNet architectures [7, 11] follow sequential networks. However, the proposed A3DConvNet evaluates fine-grained features more deeply and comprehensively than others do. The network is 15 layers deep, comprising nineteen 3D convolution operations and five 3D max-pooling operations. This network includes two concatenation layers, three batch normalization layers, a flatten layer, two fully connected layers and an output layer. The number of filters and their size varies for each layer. The architecture uses different sizes of kernels, such as $(3 \times 3 \times 3)$, $(5 \times 5 \times 5)$ and $(7 \times 7 \times 7)$, to perform convolution operations for getting fine-tuned details. The filter of size $(3 \times 3 \times 3)$ is used in most convolution operations because it can capture fine and gradual details at a low level. In addition, it diminishes computation costs and weight-sharing, which ultimately leads to lesser weights for back propagation. However, kernels of size $(5 \times 5 \times 5)$ and $(7 \times 7 \times 7)$ can capture generic details from the data representation. The features obtained from convolutional operation through kernels of small and large sizes make this method more intuitive than others.

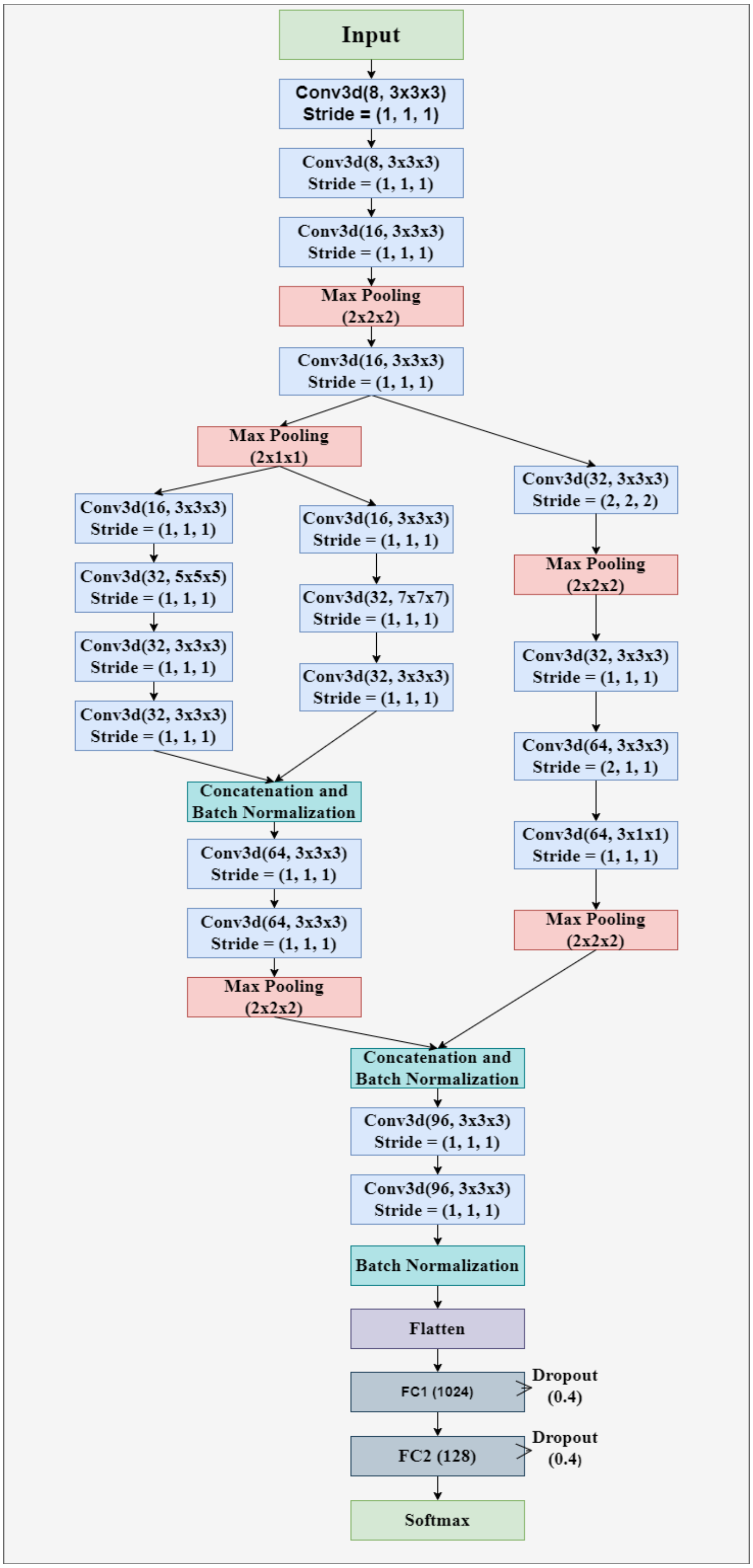

Input
Conv3d(8, 3x3x3) Stride = (1, 1, 1)
Conv3d(8, 3x3x3) Stride = (1, 1, 1)
Conv3d(16, 3x3x3) Stride = (1, 1, 1)
Max Pooling (2x2x2)
Conv3d(16, 3x3x3) Stride = (1, 1, 1)
Max Pooling (2x1x1)
Conv3d(16, 3x3x3) Stride = (1, 1, 1)
Conv3d(32, 5x5x5) Stride = (1, 1, 1)
Conv3d(32, 3x3x3) Stride = (1, 1, 1)
Conv3d(32, 3x3x3) Stride = (1, 1, 1)
Conv3d(16, 3x3x3) Stride = (1, 1, 1)
Conv3d(32, 7x7x7) Stride = (1, 1, 1)
Conv3d(32, 3x3x3) Stride = (1, 1, 1)
Conv3d(32, 3x3x3) Stride = (2, 2, 2)
Max Pooling (2x2x2)
Conv3d(32, 3x3x3) Stride = (1, 1, 1)
Conv3d(64, 3x3x3) Stride = (2, 1, 1)
Conv3d(64, 3x1x1) Stride = (1, 1, 1)
Max Pooling (2x2x2)
Concatenation and Batch Normalization
Conv3d(64, 3x3x3) Stride = (1, 1, 1)
Conv3d(64, 3x3x3) Stride = (1, 1, 1)
Max Pooling (2x2x2)
Concatenation and Batch Normalization
Conv3d(96, 3x3x3) Stride = (1, 1, 1)
Conv3d(96, 3x3x3) Stride = (1, 1, 1)
Batch Normalization
Flatten
FC1 (1024)
Dropout (0.4)
FC2 (128)
Dropout (0.4)
Softmax

Fig. 2. Proposed A3DConvNet-15 architecture

The input of the proposed architecture is the images of size $(120width \times 120height \times 3channels)$. The first four 3D convolutions have 16 filters with a kernel size of $(3 \times 3 \times 3)$. After that, feature evaluation for different filters of sizes like $(3 \times 3 \times 3)$, $(5 \times 5 \times 5)$ and $(7 \times 7 \times 7)$ is done in different branches and then they have combined accordingly, as shown in Figure 2. The final features are acquired by flatting the 3D convolved representation into the vectors, which are further passed to two fully connected layers of size 1024 and 128, respectively, followed by an output layer. A 40% dropout rate is used in both fully connected layers as a regularizer to prevent the network from overfitting. The output layer uses the softmax function to label the input video in its respective class, as presented in equation 2.

$$P\left(\frac{C}{S}\right) = \frac{P\left(\frac{\mathrm{S}}{\mathrm{c}}\right) \times \mathrm{P(c)}}{\sum_{k=1}^{c} P\left(\frac{S}{k}\right) \times P(k)} \tag{2}$$

Here, $P\left(\frac{C}{S}\right)$ denotes conditional probability. $\mathrm{P(c)}$ refers to class prior probability and the value of C indicates the total number of classes.

4. Experimental setup, results and discussion

This section explains how our model can be used to depict human behaviors in surveillance environments of megastores. We apply our proposed method to the synthesized SRV dataset, which is discussed in the later subsection. Quantitative comparisons are also part of this research, where the performance of our proposed method is explored and compared with other cutting-edge methods.

4.1. SRV dataset

We synthesize SRV (Store Realistic Videos) dataset in the CV laboratory of MNNIT Allahabad. Here, videos are recorded at a resolution of 640×480 with a frame rate of 30FPS. The SRV dataset consists of 5 classes, including normal, shoplifting, eating, drinking, and damaging. The person's usual behaviors/actions, like viewing and examining store's items, walking, talking, etc., are expressed in the normal class. Stealing actions like placing store items in bags or pockets are presented in the shoplifting class. Eating actions while taking eatable items (like chips, fruits, chocolates, etc.) and drinking actions while consuming drinkable items (like cold drinks, water, etc.) are included in eating and drinking classes. On the other hand, Destructive acts like tearing packages, trying to open caps of water or cold drink bottles, etc., are part of the damaging class. Table 1 presents the overall distribution of SRV dataset. It comprises 400 videos, where normal, shoplifting, eating, drinking and damaging classes include 88, 85, 75, 72 and 80 video clips, respectively. Each clip in the SRV dataset is 10 seconds long, including clear and distinguishable human acts. Out of 400 clips, the training process uses 301 clips and the testing process uses the remaining clips.

Tab. 1. SRV dataset distribution

Categories	Distribution	Clips	Total Clips
Normal	Train	66	88
	Test	22	
Shoplifting	Train	64	85
	Test	21	
Eating	Train	57	75
	Test	18	
Drinking	Train	54	72
	Test	18	
Damaging	Train	60	80
	Test	20	

Figure 3 presents the instances of the SRV dataset's video clips. The normal, shoplifting, eating, drinking and damaging actions are present in the first, second, third, fourth and fifth rows in the same figure. These instances show human actions with naked faces, partially covered faces and fully covered faces.

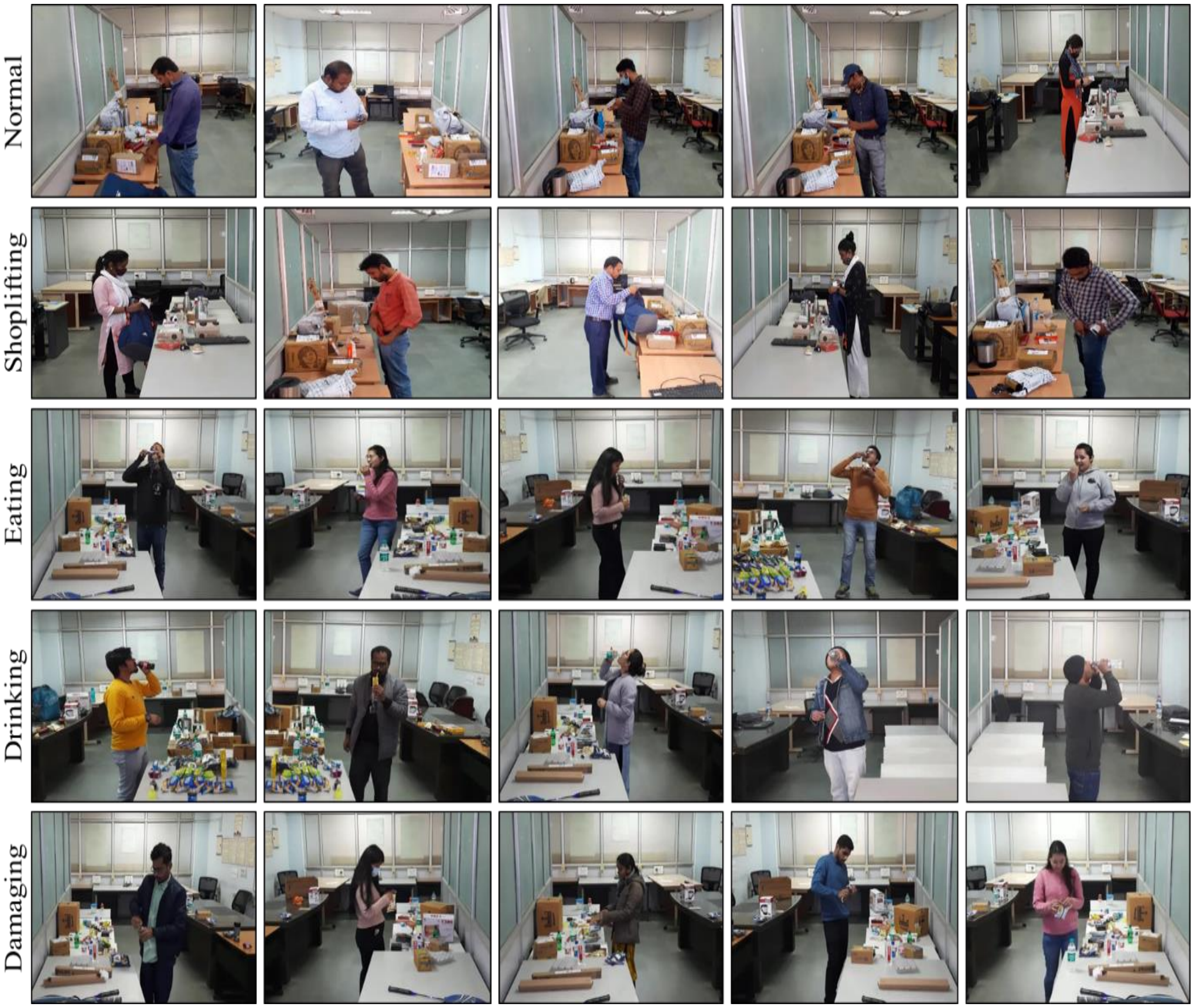


Fig. 3. Instances of SRV dataset

4.2. Evaluation metrics

We considered accuracy, f1 score, precision and recall as decisive metrics [6, 12, 17, 22, 25, 26] to evaluate the performance of the proposed method. Accuracy shows the network's overall performance representing correct hits, which is the proportion of all properly categorized instances to all instances. Precision measures the proportion of properly categorized positive instances among all the instances that were classified correctly. Recall measures the proportion of properly categorized positive instances out of all actual positive instances. Other side, F1 score calculates the harmonic means of precision and recall, which provides a more accurate assessment of misclassified cases than the Accuracy metric does.

These metrics can be represented as follows:

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{True\ Positive\ (TP) + True\ Negative\ (TN) + False\ Positive\ (FP) + False\ Negative\ (FN)} \tag{3}$$

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)} \tag{4}$$

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \tag{5}$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \tag{6}$$

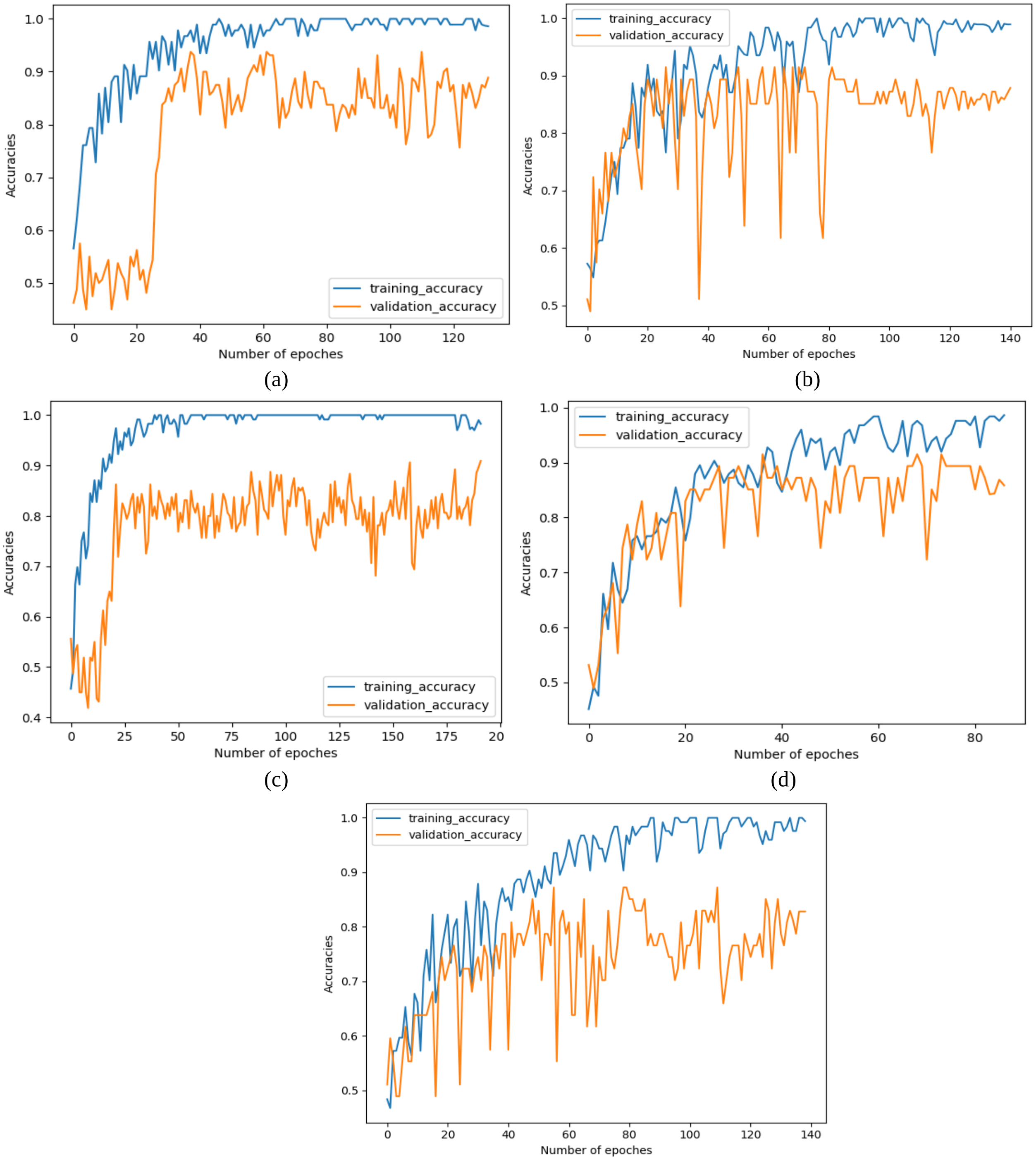
4.3. Results and analysis

The proposed method is assessed by performing experimentations on the machine running a Window 11 environment containing Core i5 Processor, 8 GB RAM, 256 GB SSD, 1 TB Hard Disk and 4GB NVIDIA 1650 GPU. The implementation was done on Python 3.6 with Tensorflow-GPU 1.14 learning framework. However, this machine configuration doesn't have much to do with the proposed methodology over modeling and model accuracies. High-end machine with dedicated graphic processor (with multicore support) only facilitates faster processing, i.e., training of the model. We evaluate our model on the SRV dataset, in which 75% of clips are used for the training process and the rest are used for the validation process. Table 2 presents the proposed models' performance over different resolutions of input data, which includes $160 \times 160 \times 3$, $160 \times 120 \times 3$, $120 \times 120 \times 3$, $120 \times 80 \times 3$ and $80 \times 80 \times 3$. After the investigation of results, it is found that the input data samples trained on higher resolution, such as $160 \times 160 \times 3$, introduce larger parameters that lead to a longer time in the training process. Here, the training time can be cut down by reducing the resolution of the input samples. This is because the model generates fewer training parameters for low-resolution images, which takes less time to train. Added to it, the model with input samples of $120 \times 120 \times 3$ resolution achieves the highest detection accuracy of up to 90.90% and training accuracy of up to 99.10%. Upon analyzing the frames at resolutions lower than $120 \times 120 \times 3$, a significant amount of fluctuation in the validation accuracy can be observed here due to loss of information, i.e., the lower resolution provides less information. Therefore, the model gets the lowest detection accuracy of 82.82% with a training accuracy of 99.38% for 80 × 80 × 3 resolution of input samples.

Tab. 2. Accuracy across various resolutions of input data

Metrics	Resolution (In Pixels)				
	$160 \times 160 \times 3$	$160 \times 120 \times 3$	$120 \times 120 \times 3$	$120 \times 80 \times 3$	$80 \times 80 \times 3$
Training Parameters	39,614,361	25,851,805	17,004,445	8,157,085	4,225,501
Training Time	20 hours:39 minutes	18 hours:55 minutes	17 hours:01 minutes	14 hours:12 minutes	12 hours:51 minutes
Training Accuracy	98.94%	98.59%	99.10%	98.68%	99.38%
Validation Accuracy	87.87%	88.88%	90.90%	85.85%	82.82%

Figure 4 shows the trade-off curves between training and validation accuracies obtained from the proposed model for different resolutions (a) 160 × 60 × 3, (b) 120 × 160 × 3, (c) 120 × 120 × 3, (d) 80 × 120 × 3, (e) 80 × 80 × 3 of inputs. In experiments, the value of the patience variable is set to 50, which controls the network from being overtrained. It automatically terminates the training process if validation loss does not decrease until 50 iterations. The network trained over the proposed model runs for 191 iterations for 160 × 160 × 3, 121 iterations for 160 × 120 × 3, 131 iterations for 120 × 120 × 3, 86 iterations for 120 × 80 × 3 and 138 iterations for 80 × 80 × 3 resolutions of inputs. Upon analyzing each of the trade-off curves, we find smooth training and validation patterns for resolutions of $160 \times 160 \times 3$, $120 \times 160 \times 3$, $120 \times 120 \times 3$ and $80 \times 120 \times 3$. However, in case of $80 \times 80 \times 3$ resolution, training and validation patterns are not smooth because a large gap is present between training and validation curves, leading to slight overfitting.



(a)

(b)

(c)

(d)

(e)

Fig. 4. Accuracy trade-off for (a) 160×160×3, (b) 120×160×3, (c) 120×120×3, (d) 80×120×3, (e) 80×80×3

Table 3 shows the performance metrics evaluated over the SRV dataset using the proposed A3DConvNet model over 120 × 120 × 3 of resolution. We find that the model achieves promising detection accuracy up to 90.90% with 91.17% precision, 91.21% recall, and 91.19% f1 score. Here, instances of eating and drinking classes are classified correctly, with few false positive instances. However, damaging class incurs somewhat large false positive instances, where they are misclassified in normal, shoplifting and eating classes.

Tab. 3. Performance metrics for SRV dataset using proposed 3D-CNN model

Categories		**Predicted**					**Metrics**
		Normal	**Shoplifting**	**Eating**	**Drinking**	**Damaging**	
Actual	**Normal**	20	1	0	0	1	Accuracy : 90.90%, Precision : 91.17%, Recall : 91.21%, F1 score: 91.19%,
	Shoplifting	1	18	0	0	2	
	Eating	0	0	17	1	0	
	Drinking	0	0	0	18	0	
	Damaging	2	0	1	0	17	

The investigational outcomes of various proposed models and our proposed model for SRV input samples of size 120×120×3 are presented in Tab. 4. In this regard, Arunnehru et al. [24] use the simplest form of 3DCNN architecture that takes two sets of 3DConvolutional layers followed by a 3DPooling layer to represent human act. Guillermo et al. [12] also use a slightly modified 3D-CNN architecture to encode an action in the spatiotemporal domain. Kanagaraj and Priya [11] adjust the existing 3DCNN architecture [12] by adding extra convolution layers to build a proficient HAR system. R Vrskova et al. [23] explore the spatiotemporal features using eight-layer deep 3DCNN architecture (comprising three sets of three convolutional and a pooling layer) for performing the same task. Arroyo R et al. [5] utilizes bag of words and C3D method to explore the features that are further used by deep MIL ranking model for performing classification task. Donahue et al. [7] take advantage of CNN and LSTM networks for modeling the visuals. Ruchi and Manish [17] use Xception and LSTM networks for temporal sequencing analysis. Ansari and Singh [18] also take advantage of InceptionV3 and LSTM modules to do the same. Our method uses a deep structure of 3DCNN to build a spatiotemporal relationship on the segmented forepart of the video sequences and identify abnormal person movements. Our method has reported an accuracy of up to 90.90% on the SRV dataset, outperforming other existing methods.

Tab. 4. Comparison of existing methods on SRV dataset

Models	**Model Description**	**Precision**	**Recall**	**Accuracy**	**F1 Score**
Arunnehru J. [24] (2018)	3DCNN (3DConv + 3DMaxPool + 3DConv + 3DMaxPool + flatten + Dense + Softmax)	76.64%	76.21%	75.75%	76.42%
Arroyo R et al. [5] (2018)	Bag of Words, C3D Features,Deep MIL Ranking Model	79.54%	79.07%	78.78%	79.30%
Guillermo et al. [12] (2021)	3DCNN (3DConv + 3DConv + 3DMaxPool + 3DConv + 3DConv + 3DMaxPool + flatten + Dense + Softmax)	82.35%	82.45%	81.81%	82.40%
Kanagaraj & Priya [11] (2021)	3DCNN (3DConv + 3DConv + 3DMaxPool + 3DConv + 3DConv + 3DMaxPool + 3DConv + 3DConv + 3DMaxPool + flatten + Dense + Softmax)	86.08%	86.12%	85.85%	86.10%
R Vrskova et al. [23] (2022)	3DCNN (3DConv + 3DMaxPool + 3DConv + 3DMaxPool + 3DConv + 3DConv + 3DMaxPool + 3DConv + 3DConv + 3DMaxPool + flatten + Dense + Softmax)	88.22%	88.03%	87.87%	88.12%
Jayaswal & Dixit [17] (2021)	Xception + LSTM	87.54%	87.12%	86.86%	87.32%
Donahue et al. [7] (2015)	CNN + LSTM	88.29%	88.24%	87.87%	88.26%

Ansari & Singh [18] (2022)	InceptionV3 + LSTM	88.31%	88.19%	87.87	88.25%
Proposed Model	A3DCNN-15	91.17%	91.19%	90.90%	91.18%

The resultant examples of video sequences assessed on the proposed methodology are shown in ***Figure 5***. In each video sequence, a person can perform certain types of actions, which are further identified by the proposed system. The proposed system generates alerts with a red rectangle and text on the screen when anything is found related to unusual activity. The visual consequences prove that our method performs well and provides more accurate results for these sets of video sequences.

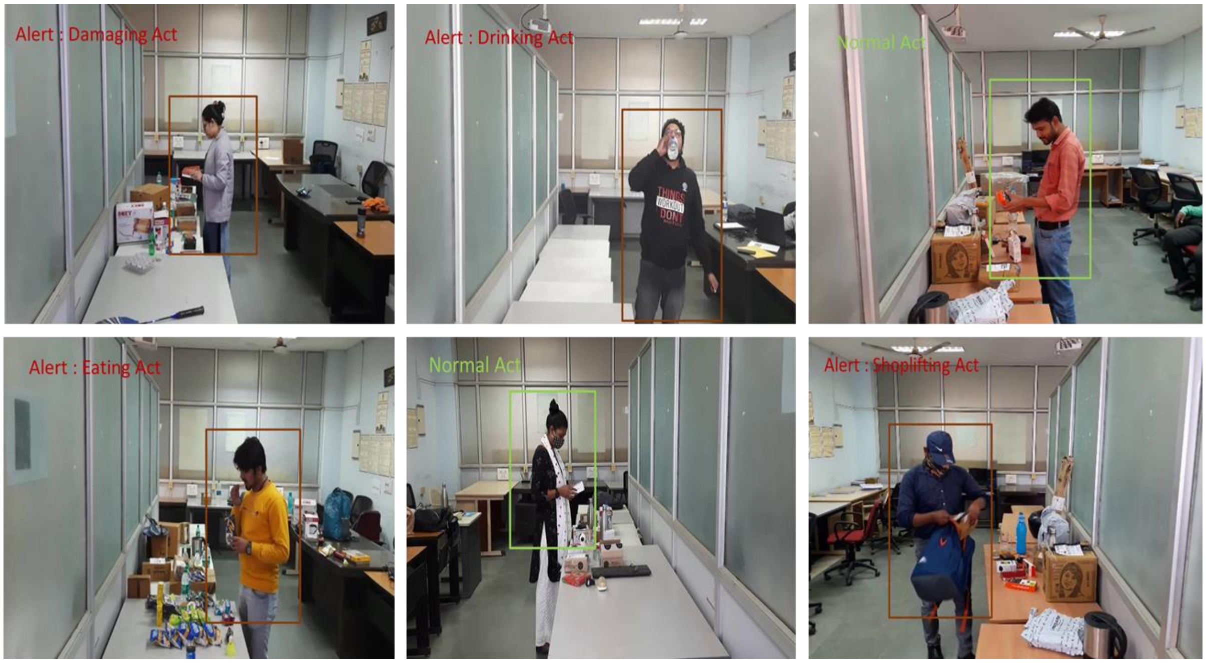


Fig. 5. Resulting sequences of our proposed model

4.4. Testing on real-store video clips

After examining how people behave in artificial situations, we evaluated how well the suggested technique performed on a variety of real shop camera footage collected from the UCF crime dataset [6] and YouTube videos. We tested our suggested approach on a total of 50 videos. They are categorized into five classes, i.e., normal, shoplifting, eating, drinking and damaging, each comprising ten videos. The performance consequences of the suggested strategy for classifying real-store recorded videos are shown in ***Table 5***. Upon analysis, we infer that videos containing typical human and drinking actions are accurately classified. On the other hand, the shoplifting and eating class examples yield few false positive examples during the test phase, while the damaging action produces a larger number of false positive examples than others. The proposed network tested over real-store video clips provides encouraging results, up to 86% accuracy, which means the proposed network can be practiced to test real-time videos.

Tab. 5. Performance Metrics for real store videos using proposed 3D-CNN model

Categories		Predicted					Metrics
		Normal	Shoplifting	Eating	Drinking	Damaging	
Actual	Normal	9	1	0	0	0	Accuracy : 86%, Precision : 86.73%, Recall : 86%, F1 score: 86.36%,
	Shoplifting	0	9	0	1	0	
	Eating	0	0	8	0	2	
	Drinking	0	0	0	10	0	
	Damaging	1	2	0	0	7	

Figure 6 depicts activity detection over real-world store-recorded clips. These captured movies show the typical behavior of customers ranging from normal to abnormal in megastore and supermarkets. Also, there are somewhat variations in the actual in-store clip that was captured, such as changes in lighting, poses, occlusion, etc. After investigation, we found that our suggested strategy successfully classifies occurring human actions with an accuracy rate of 86% and gives promising test results on real-store recorded videos.

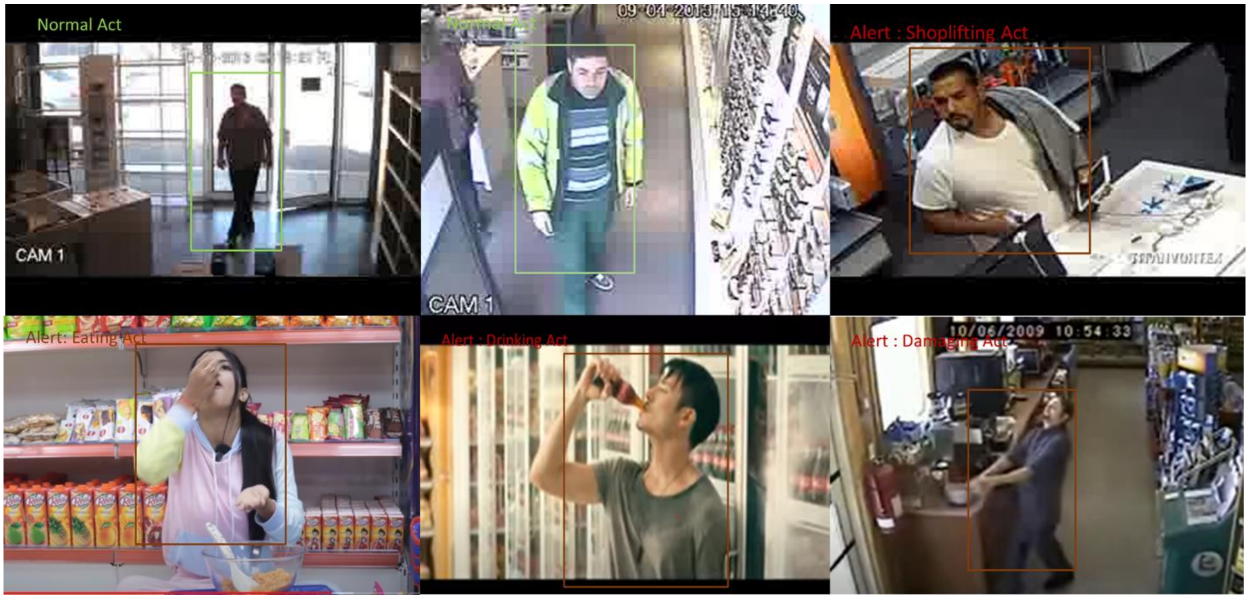


Fig. 6. Resulting sequences of our proposed model tested over real-store video clips

The test results of different state-of-the-art methods and our proposed model for an input sample size of 120 × 120 × 3 are presented in Fig. 7, where the accuracy comparison curve shows the behavior of each method for the real-world store instances. It is found that our model with 15 layers of deep architecture secures the highest accuracy of up to 86% compared to others. The performance of the proposed method is relatively better for synthesized training inputs than the real-world store videos. This is because the recorded clips of real-world store scenarios comprise too much variation in illumination, cluttered background and occluded actions. Despite these inconsistencies, the accuracy achieved on store clips recorded in the real world is excellent, which ensures that the model can be used in real-time to detect different abnormal happenings in megastores.

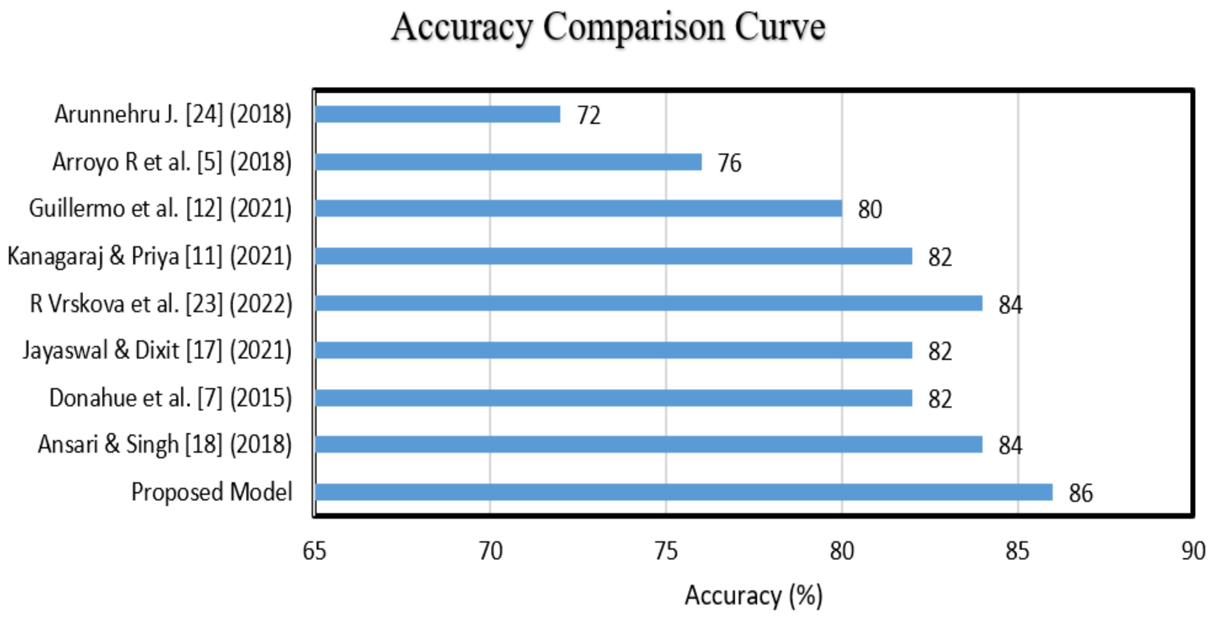


Fig. 7. Comparison with current cutting-edge methods evaluated in real-store videos

5. Conclusion

Motivated by recent advances in neural networks, we successfully applied SSD to segment the person and deep 3D-CNN architectures to detect abnormal human behavior in megastores/shops. Generating an alert on the screen whenever an event connected to shoplifting, eating, drinking and damaging happens is a consequent result of this research. In particular, we introduce 15 layers of deep 3DCNN architecture suitable for extracting more abstract features from video sequences. This representation can encode the dynamics of time-critical actions more effectively. Experiments performed on the SRV dataset show that our method is capable of detecting abnormal activities in indoor surveillance with an accuracy of up to 90.90%, which comparatively outperforms other state-of-the-art methods. Additionally, the performance of the proposed method is also evaluated for real-time recorded store videos, where we find another remarkable achievement in terms of accuracy of up to 86%. In the future, we can develop Inception, ResNet, and Xception like architecture to build an advanced 3DCNN structure for getting impressive performance in the HAR system. On the other hand, challenges like detecting new classes, occlusion and enclosing facts in existing models are more attractive avenues that require extensive exploration. Therefore, new methods can be developed in the future to address the HAR problem to meet these goals.

Acknowledgments

The authors are grateful for the funding support from the Department of Science & Technology - Science and Engineering Research Board (SERB) India released in the name of Dr. Dushyant Kumar Singh (PI), under the project titled “Development of an Integrated Intelligent Surveillance System for Suspicious Behavior Analysis” (Project file number CRG/2020/001982).

References

[1] Ray, Abhisek, et al. "Transfer Learning Enhanced Vision-based Human Activity Recognition: A Decade-long Analysis", International Journal of Information Management Data Insights, vol. 3, no.1, p. 100142, 2023.

[2] Beddiar, Djamila Romaissa, et al. "Vision-based human activity recognition: a survey", Multimedia Tools and Applications, vol. 79, pp. 30509-30555, 2020.

[3] Varshney, Neeraj, "Combining electrocardiogram signal with Accelerometer signals for Human Activity Recognition using Convolution neural network", Journal of Physics: Conference Series, vol. 1947, no. 1, IOP Publishing, 2021.

[4] Tripathi, Rajesh Kumar, Anand Singh Jalal, and Subhash Chand Agrawal, "Suspicious human activity recognition: a review", Artificial Intelligence Review, vol. 50, pp. 283-339, 2018.

[5] Arroyo, Roberto, et al. “Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls”, Expert systems with Applications, vol. 42, no. 21, pp. 7991-8005, 2015.

[6] Ansari, Mohd Aquib, and Dushyant Kumar Singh, "An expert video surveillance system to identify and mitigate shoplifting in megastores", Multimedia Tools and Applications, pp. 1-29, 2022.

[7] Donahue, Jeffrey, et al. "Long-term recurrent convolutional networks for visual recognition and description", Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.

[8] Serpush, Fatemeh, and Mahdi Rezaei, "Complex human action recognition using a hierarchical feature reduction and deep learning-based method", SN Computer Science, vol. 2, pp. 1-15, 2021.

[9] da Silva, Murilo Varges, and Aparecido Nilceu Marana, "Human action recognition in videos based on spatiotemporal features and bag-of-poses", Applied Soft Computing, vol. 95, p 106513, 2020.

[10] Sargano, Allah Bux, et al. "Human action recognition using deep rule-based classifier", Multimedia Tools and Applications, vol. 79, pp. 30653-30667, 2020.

[11] Kanagaraj, Kaavya, and GG Lakshmi Priya, "A new 3D convolutional neural network (3D-CNN) framework for multimedia event detection", Signal, Image and Video Processing, vol. 15, pp. 779-787, 2021.

[12] Martínez-Mascorro, Guillermo A., et al. "Criminal intention detection at early stages of shoplifting cases by using 3D convolutional neural networks", Computation, vol. 9, no. 2, 2021.

[13] Tsushita, Hiroaki, and Thi Thi Zin, "A study on detection of abnormal behavior by a surveillance camera image", Big Data Analysis and Deep Learning Applications: Proceedings of the First International Conference on Big Data Analysis and Deep Learning, Springer Singapore, pp. 284-291, 2019.

[14] Dwivedi, Neelam, Dushyant Kumar Singh, and Dharmender Singh Kushwaha, "Weapon classification using deep convolutional neural network", 2019 IEEE Conference on Information and Communication Technology, IEEE, 2019.

[15] Nguyen, Tam N., and Ngoc Q. Ly, "Abnormal activity detection based on dense spatial-temporal features and improved one-class learning", Proceedings of the 8th International Symposium on Information and Communication Technology, 2017.

[16] Riaz, Hamza, et al. "Anomalous human action detection using a cascade of deep learning models", 2021 9th European Workshop on Visual Information Processing (EUVIP), IEEE, 2021.

[17] Jayaswal R, Dixit M, “A framework for anomaly classification using deep transfer learning approach”, Revue d'Intelligence Artificielle, vol. 35, no. 1, pp. 255-263, 2021.

[18] Singh, Dushyant Kumar, Anshu Kumar, and Mohd Aquib Ansari, "Robust Modelling of Static Hand Gestures using Deep Convolutional Network for Sign Language Translation", 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), pp. 487-492, IEEE, 2021.

[19] Liu, Wei, et al. "Ssd: Single shot multibox detector", Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer International Publishing, 2016.

[20] Singh DK, “Human action recognition in video”, International Conference on Advanced Informatics for Computing Research, Springer, pp. 54-66, 2018.

[21] Varshney N., “Deep learning in human activity recognition from videos: a survey”, In Advances in Computational Intelligence and Communication Technology: Proceedings of CICT 2021, pp. 335-346, 2022.

[22] Singh, Dushyant Kumar, et al. "Human crowd detection for city wide surveillance", Procedia Computer Science, vol. 171, pp. 350-359, 2020.

[23] Wang, Tian, et al. "Internal transfer learning for improving performance in human action recognition for small datasets", IEEE Access, vol. 5, pp. 17627-17633, 2017.

[24] 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, vol. 133, pp. 471-477, 2018.

[25] Gupta, Somya, et al. "Artificial intelligence adoption in the insurance industry: Evidence using the technology–organization–environment framework", Research in International Business and Finance, vol. 63, p. 101757, 2022.

[26] Goel, Dhruv, and Rahul Pradhan, "A comparative study of various human activity recognition approaches", IOP Conference Series: Materials Science and Engineering, vol. 1131, no. 1, IOP Publishing, 2021.

Received 2 February 2023