BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **23**, No 1
Sofia • 2023 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2023-0008

Optimized Parameter Tuning in a Recurrent Learning Process for Shoplifting Activity Classification

Mohd Aquib Ansari, Dushyant Kumar Singh

CSED, MNNIT Allahabad, Prayagraj, India

E-mails: mansari.aquib@gmail.com dushyant@mnnit.ac.in

Abstract: *From recent past, shoplifting has become a serious concern for business in both small/big shops and stores. It customarily involves the buyer concealing store items inside clothes/bags and then leaving the store without payment. Unfortunately, no cost-effective solution is available to overcome this problem. We, therefore intend to build an expert monitoring system to automatically recognize shoplifting events in megastores/shops by recognizing object-stealing actions of humans. The method proposed utilizes a deep convolutional-based InceptionV3 architecture to mine the prominent features from video clips. These features are used to custom Long Short Term Memory (LSTM) network to discriminate human stealing actions in video sequences. Optimizing recurrent learning classifier using different modeling parameters such as sequence length and batch size is a genuine contribution of this work. The experiments demonstrate that the system proposed has achieved an accuracy of 89.36% on the synthesized dataset, which comparatively outperforms other existing methods.*

Keywords: *Computer vision, Deep neural network, Video analysis, Activity classification, Recurrent neural network.*

1. Introduction

Nowadays, theft of retail products from megastores/shops by the customers is increasing massively. Some buyers commit theft by concealing the items in bag or under clothes, when nobody is watching. They then leave the store/shop without paying for it. The individual involved in these thefts is called a shoplifter and the act is entitled shoplifting [1-3]. In simple terms, shoplifting is an act of taking items from the open retail establishment with no intention of paying, in such a way this is not noticed. Shoplifters are the foremost menace to retailers, which causes a massive loss to the business. Since no research has been done so far to recognize shoplifters, we aim to develop an automated indoor surveillance system for shops/stores to identify shoplifting incidents in real-time scenarios.

Shoplifting is a kind of suspicious activity [4, 6, 7] in which the person does not behave normally. For example, a person's expected normal behavior in a megastore is recorded like walking, examining items, billing, talking, etc. On the contrary, the actions involved in shoplifting are different from the normal ones, which are recorded like putting items under clothing or in the bags in a way when security personals are not seeing them. According to the National Association for Shoplifting Prevention (NASP) [3], American retailers lose almost $50 billion worth of merchandise annually, which is approximately $45 million each day. The NASP has also found that 1 in 11 people commit shoplifting crimes at some point in their life. On the other hand, the National Retail Federation (NRF-2020) report [5] still sees shoplifting as a foremost cause for influencing inventory shrinkage, where the inventory shrinkage is a merchandise loss due to shoplifting, error, fraud, employee theft, etc. Shoplifting causes tremendous losses and disruptions for retailers. Therefore, an early earnestness is needed on this issue to minimize the losses caused to retailers.

Nowadays, the retailers are using CCTV surveillance infrastructure for keeping their eyes on each buyer. Here, the control room personals attentively examine the video stream received from CCTV infrastructure. Due to the human's limited strength, one cannot keep their eye to see an array of video frames very long and lapses are usually possible. This process reduces losses to some extent but is still not very effective in dealing with shoplifting as it requires manual intervention. Therefore, an automated CCTV-based surveillance system [8-10] could be the right solution to automatically monitor each person's activity in the stores/shops.

So far, a cost-effective solution is still missing to identify the shoplifters. To fulfil the requirement, this article proposes an expert Human Activity Recognition (HAR) system, which automatically examines the real-time video stream to identify human stealing actions for surveillance. The system is able to generate a warning in the form of a message on the screen when a shoplifter commits an act of theft. The proposed scheme takes the benefit of deep convolutional neural network and recurrent learning network to make an intelligent HAR system. Here, the proposed system follows a three-step framework. The first step extracts the frames/sequences from the video stream. The second step extracts relevant features from each input sequence using the deep convolutional network. Finally, in the third step, the extracted features build a recurrent classifier for decision-making purpose, i.e., classifying normal and shoplifting events. The proposed system can reduce future crimes related to shoplifting in megastores/shops by identifying human stealing actions in real-time.

The key contributions of the presented work are presented as follows:

- We propose an expert HAR system to access shoplifting scenarios using deep neural strategies.
- We introduce a self-synthesized shoplifting dataset that involves person's normal and shoplifting activities. The clearly visualized human acts in recorded video clips make this dataset more affluent and informative than the prior available shoplifting dataset, thereby improving the efficiency of the proposed method.
- Subsequently, the impact of modeling parameters such as Sequence Length (S.L.) and batch size has been extensively analyzed, both for existing and synthesized training datasets.

The experimental outcomes reveal that our network proposed is well suited for predicting human behavior in indoor surveillance. The rest of the paper is structured in the following manner. Section 2 comprises the literature review with a brief knowledge of existing HAR systems. Section 3 is styled to discuss our proposed methodology along with some technicalities in brief. Experimentation for the proposed approach exercised with two earlier mentioned datasets is detailed in Section 4. The last section concludes the paper.

2. Literature review

The research trend in HAR from video sequences has encountered progressing interest over the past decade, while performance is satisfactory despite the enlightenment of more challenging datasets. This section considers the related works in the field of HAR related to learning based paradigms.

The studies in the literature [12] show that the action and activity are interrelated with some overlaps. Action is a single-person activity in which a person performs simplistic moves like waving, sitting, bending, etc., at intervals of a few seconds. On the other hand, a sequence of certain actions done by a person or group of persons represents an activity. This includes, for instance, shaking hands, assaulting individuals, leaving an unattended object in public places, etc. According to research in [13], the vision-based HAR system is fashioned into three-phase process: detection, tracking and understanding the action or activity. Recent research has focused on using marker-less computer vision and non-intrusive approaches to detect and understand ongoing human activities in the presence of natural and realistic scenes.

Nguyen and Ngoc [14] proposed handcrafted features based activity recognition system. Here, proposed system uses a frame differencing algorithm for moving object detection. In the next stage, the relevant features are calculated for each ROI using Motion Boundary Histogram (MBH), Histogram of Optical Flow (HOF) and Histogram of Oriented Gradient (HOG) descriptors. These features are used to build a Support Vector Data Descriptor (SVDD) classifier for distinguishing human activities. K. Arati, A. Khare and M. Khare [15] have proposed fusion based approach to encode human activity using optical flow motion vectors and HOG gradient vectors. The fused features are passed to the Support Vector Machine (SVM) classifier for classification purpose. In study [16], it has been found that optimization algorithms show a significant role in the feature selection, by reducing the number of input vectors when developing a predictive model. Sanal and Bhavani [18] have optimized the handcrafted features using Genetic optimization algorithm. They have found that optimized features could suitably encode activity characteristics and improve the predictive model performance.

Recently, Convolutional Neural Networks (CNNs) [6, 9] are widely getting used to solve various vision-based tasks with great success. However, only two-dimensional CNNs are unable to resolve video-related tasks. Guillermo et al. [20] have mitigated such deficiency by proposing an advanced HAR framework using a three-dimension convolutional neural network. This system derives pertinent features from spatiotemporal dimensions using 3D convolutions, encoding motion evidence in several contiguous frames.

Yue-Hei Ng et al. [21] propose one more attention-grabbing methodology for the video classification problem. They have employed a recurrent layer on the top of the convolutional network to classify video scenes efficiently. The suggested approach appears to be more effective, but it still requires some improvement. Donahue et al. [22] have proposed a long-term recurrent convolutional network for visual recognition that can efficiently learn compositional representation in space and time. The model proposed is directly associated with modern visual convolutional networks and can be trained to gain experience from convolutional perceptual representations and temporal dynamics. Farzan Majeed et al. [23] have taken advantage of pose assessment techniques with motion features to build an efficient activity recognition system. The motion features extract the relevant features to track each of the keypoints by considering their movements in successive frames. The work explores deep recurrent neural networks for HAR by considering motion features and presents competitive outcomes. Waqas, Chen and Mubarak [2] suggest a framework to distinguish real-world incongruities through multiple instance-ranking models by leveraging weakly labeled training samples. They have found that the deep anomaly-ranking model could efficiently foresee larger anomaly scores to segment the anomalous videos.

In [8], research investigations have reported the appropriateness of using in-depth learning methods to classify human visual actions and emphasize solving the difficulties involved in temporal dimensions of video's visuals. Asadi-Aghbolaghi et al. [24] have presented a comprehensive study by surveying various deep learning approaches for classifying human activities. The survey classifies deep learning approaches based on 3D models, motion-based descriptors and temporal based methods. They have found that 3D networks can learn temporal patterns more effectively over a long sequence, and LSTM can build a long-range temporal relationship than RNN. Jayaswal and Dixit [25] propose a framework to distinguish real-time anomalies using deep neural network based approaches. The framework utilizes fine-tuned deep Xception network for relevant features evaluation and a recurrent learning network for anomaly classification. The network proposed can deal with indoor and outdoor scenarios with different lighting conditions. However, this method cannot distinguish the anomalies, which involve much articulated and non-rigid nature of actions.

The literature related to the present work, as mentioned above, includes handcrafted features based HAR systems and deep model based HAR systems. Table 1 includes the comparative analysis among various activity recognition based researches. The handcrafted features based HAR systems rely heavily on handcrafted descriptors and sometimes cannot mine the pertinent information from an image due to structure losses in global data. In contrast, deep learning based HAR systems use deep Convolutional Neural Networks (CNNs) to learn spatial hierarchies of features based on the salient information available in video sequences. The CNN retains the relevant information in the form of the features from the video sequences automatically. Most researchers use CNNs with machine learning classifiers to build

HAR systems, but they could utilize very small scale temporal relationships. Later on, researchers use CNNs with recurrent learning processes like Recurrent Neural Networks (RNNs) to build expert HAR systems, which can remember every information along time domain with the help of memory components. However, RNN cannot persist long-term dependencies.

Table 1. Comparative analysis among various activity recognition based research

References	Techniques involved	Database used	Characteristics	Accuracy
Nguyen and Ngoc [14]	Frame Differencing, HOG, HOF, MBHH, SVDD	UCD, UMN	The generative model based anomaly detector can deal with various situations like overlapping between objects, human crowd and low resolutions	93.70% and 97.66%
K. Arati, A. Khare and M. Khare [15]	Optical Flow, HOG, SVM	UT Interaction, CASIA, HMDB51	This method is robust to illumination variation and view invariant	99.31%, 97.95% and 97.18%
Sanal and Bhavani [18]	Median Filter, HOG, Color, GiST, Genetic Algorithm, RF	Multimodal Egocentric Dataset	It can efficiently deal with missing data and maintains accuracy. It can balance the error in the unbalanced dataset	87.51%
Guillermo et al. [20]	3DCNN	UCF Crime	It can discover spatiotemporal features using three dimensions	84.04%
Yue-Hei Ng et al. [21]	Convolutional temporal feature pooling, LSTM	UCF101	It can detect human actions even in poor illumination	88.6%
Donahue et al. [22]	Long term recurrent deep CNN	UCF101	It learns convolutional perceptual over temporal dynamics to represent human actions accurately	82.66%
Farzan Majeed et al. [23]	2D Pose, temporal features, LSTM	MHAD	The proposed traceability factor can handle partially occluded human actions more efficiently	92.4%
Waqas, Chen and Mubarak [2]	C3D features, Bag of Words	UCF Crime	It represents anomalies more accurately using C3D representation	75.41%
Jayaswal and Dixit [25]	Xception, LSTM	HBD21	This method can detect human acts carrying indoor and outdoor scenarios with different lighting conditions	97.25
Asadi-Aghbolaghi et al. [24]	Surveyed various deep learning based approaches		3D networks can learn effective temporal patterns over a long sequence	-

To cover the still prevailing problems of gradient exploding and vanishing gradients during extensive training on the huge dataset, the LSTM is used here to achieve more efficient time-domain learning. The pipeline configuration as proposed in the manuscript is an exclusive contribution to the work, to model spatiotemporal characteristics of the activities. We extract spatial features using Deep Inception V3 architecture in the feature extraction pipeline for existing and synthesized training datasets. In addition to the above proposal, optimizing the impact of various modeling parameters towards gaining higher performance is another rational contribution of this proposed work. Sequence length and batch size are used to create different modeling cases, comprehensively investigating the proposed model behavior. The experimental outcomes ascertain that the proposed approach based on Inception V3 and LSTM networks can identify human stealing actions more accurately than others can.

3. Proposed methodology

This article suggests a HAR system built on deep neural networks to recognize shoplifting events in stores. The method proposed learns deep convolutional perceptual over long-term temporal dynamics to represent any activity. The block diagram of the proposed methodology is presented in Fig. 1.

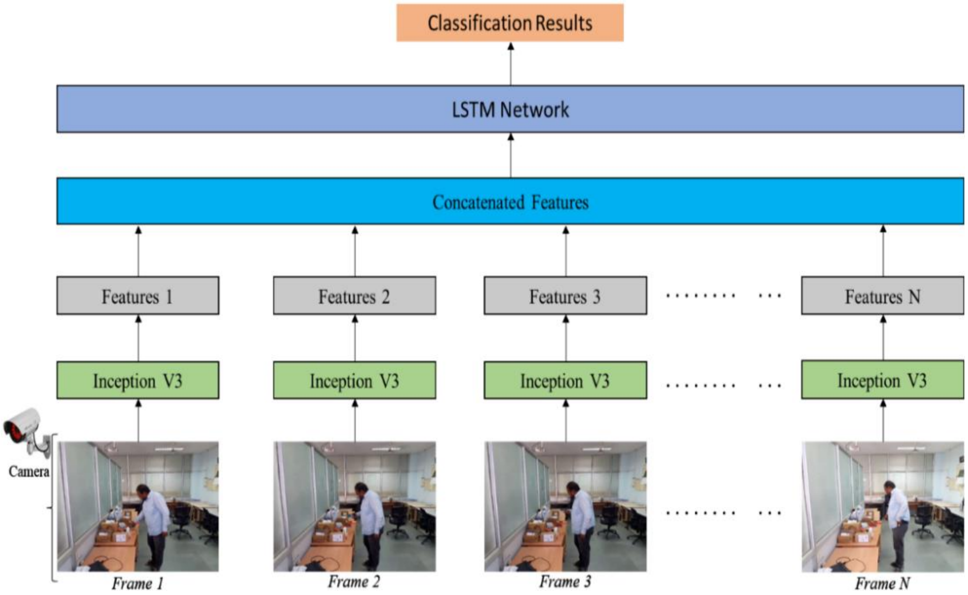


Fig. 1. Workflow of proposed methodology

As mentioned earlier, the proposed system follows a three-step framework: image capturing, feature extraction and classification. In the first step, the system captures the real-time sequences/frames from input scenes using a CCTV camera. Here, captured frames reflect appearance or spatial information in RGB format. In the next step, these frames are passed to the feature extraction module to extract relevant features. This module uses a deep multilevel CNN model to mine spatial information as features from each frame. Inception V3 architecture is part of this CNN model. Next, the extracted features are concatenated into a vector for N frames, where the value of N equals to sequence length (number of frames used to represent an action). The last step is to build a classifier using extracted features from previous step. Long Short Term Memory (LSTM) is used as the last stage classifier to categorize human activities like normal and shoplifting in real-time scenarios.

Algorithm 1. Shoplifting Detection Mechanism

```
Step 1. Input: Video frames, Sequence Length (S.L.)
Step 2. Output: Classification Outcome (Normal/Shoplifting)
Step 3. Procedure: Detection_Mechanism ()
Step 4. Initialize previous observation ($h_{t-1}$) to NULL
Step 5. While A is not completely explored do
Step 6.     Set tracker with $A[0]$
Step 7.     For each substream of length S.L. do
Step 8.         Attain tracking consequences R in frame
Step 9.         For each frame in sub-stream do
Step 10.            $x_t$ = InceptionV3(frame)
Step 11.            $h_t$ = LSTM_Network($x_t, h_{t-1}$)
Step 12.            $h_{t-1} = h_t$
Step 13.            Add $h_{t-1}$ to R
Step 14.        End For
Step 15.    Add R to Final Outcome
Step 16.    End For
Step 17. End While
Step 18. End Procedure
```

Let us system takes a video stream (A^i) as input. Substreams of size equal to Sequence Length (S.L.) are fragmented from the inputted video stream and then passed one by one to the classification module. Here, the action recognition module first extracts features from each sequence of sub-stream and uses the LSTM network to discriminate the person's acts. This process is repeated for each substream until the entire video stream is processed. The algorithm presented next shows the overall working process of the proposed model.

As depicted in the proposed algorithm, feature extraction and classification are two key components to build a realistic HAR system. Accordingly, deep learning techniques as utilized for structuring an intelligent surveillance system are discussed in the following subsections.

3.1. Inception V3

The Inception model has been first summarized by GoogLeNet/Inception V1 [11] and it achieved a great milestone in the arena of image analysis and object detection. The Inception model is heavily engineered and goes deeper for finding relevant information from an image. Inception V1 takes an image as input and undergoes 1×1, 3×3 and 5×5 convolution and maximum-pooling, which provides the convected vectors as output. Inception V2 [26] introduces batch normalization that makes the output more regular and balanced. It also replaces a 5×5 convolution with two 3×3 convolutions, which diminishes the parameter's size and maintains the receptive field's size. Inception V3 [17] is a popularly adopted image recognition model that incorporates a more simplified architecture than Inception V2. Inception

V3 model is 42 layers deep and it achieves an accuracy 78% approx. on the ImageNet dataset. It comprises factorization and grid size-reduction blocks, including convolutions, max pooling, average pooling, concatenation, dropouts and fully connected layers. It shows significant advancement in parameter reduction using the convolutional factorization concept. The factorization process replaces the bigger $n \times n$ convolution with two $1 \times n$ and $n \times 1$ convolution. For example, one 7×7 convolution that comprises 49 parameters is replaced by two 1×7 and 7×1 convolutions consisting of $\{(1 \times 7) + (7 \times 1)\}$, i.e., 14 parameters. This process reduces the parameter count to a large extent and reduces the probability of overfitting during training. The grid size-reduction module in Inception V3 makes the network less expensive and more effective than Inception V1 and V2. Furthermore, auxiliary classifiers are used as regularizers to deepen the convergence of the network and label smoothing is used as a regularization technique to regularize the classification layer for getting proficient outcomes.

3.2. Long Short Term Memory (LSTM)

Recurrent Neural Network (RNN) [2, 8] has proven to be very effective for modeling temporal sequences-based problems. These problems include generating image descriptions, video tagging, machine translation, speech recognition and many more. However, RNN is unable to deal with long-term dependencies. For that reason, Hoeckwriter and Schmiduber [19] have proposed an improved variant of RNN to tackle the long-term dependencies, called Long Short Term Memory (LSTM) [6, 23, 27]. A single unit of LSTM consists of three gates and a cell state, which can learn as well as retain information from each of the units selectively. The gates, namely Input gate (I), Forget gate (F) and Output gate (O) control network's flow, while Memory Cells (C) retain the information over an arbitrary time interval.

At timestamp T, X_T, H_{T-1}, H_T and C_T denote input data points, previous observation, current observation and cell state, respectively. LSTM updates network parameters in the following phases:

First, LSTM decides the amount of information regulated through the cell state and determines what new information is going to place in the cell state, as shown in the next three equations,

(1) $$F_T = \sigma[W_F \,.\, (H_{T-1}, X_T) + B_F],$$

(2) $$I_T = \sigma[W_I \,.\, (H_{T-1}, X_T) + B_I],$$

(3) $$\tilde{C}_T = \tanh[W_C \,.\, (H_{T-1}, X_T) + B_C].$$

In second phase, the old cell state is updated to the new cell state, as shown in the next equation,

(4) $$C_T = F_T * C_{T-1} + I_T * \tilde{C}_T.$$

The third phase provides a cell stage based filtered version of the output, as shown in the next equations:

(5) $$O_T = \sigma[W_O \,.\, (H_{T-1}, X_T) + B_O),$$

(6) $$H_T = \tanh(C_T) * O_T,$$

where, I_T, O_T and F_T represent the input layer, an output layer and forget layer, respectively. W_I, W_O, W_F and W_C denote to weight matrices and B_I, B_O, B_F and B_C denote biases in respective layers. C_T, C_{T-1} and $\tilde{C}_T$ represent current cell state, old cell state and fresh candidate value, respectively. The symbol σ represents the sigmoid layer and tanh represents tanh layer in deep recurrent networks. As recurrent in nature, LSTM passes the previous state's output to the current stage's input, as depicted in the next equation:

$$H_T = F_W(X_T, H_{T-1}). \tag{7}$$

At the final phase, the last or output layer uses the Softmax function to estimate the probability score for the respective classes.

The work described here combines Inception V3 and the LSTM network to create an intelligent HAR system for detecting human stealing acts in real-time scenarios. The next section presents the investigated results of method being proposed, evaluated over two shoplifting datasets. In addition, comparative analysis has also been made for different cases of modeling parameters to see how well the proposed solution can detect acts of human theft.

4. Experiments & Outcomes

This section presents the wide range of experimentations done in support of the proposed approach and their analysis in detail. The experimentations have been done on the machine configured with Core i7-4005 CPU, 500GB Disk, 16 GB RAM and 4 GB NVIDIA 1050 Ti GPU. For faster sample processing, we have configured the GPU device with Cuda version 10.0.130 and CUDNN version 7.6.3. In context to the programming environment, Python 3.6 with Keras 2.2 and Tensor Flow-GPU 1.14 packages have been used to perform training and testing on the input samples.

The parameter tuning in this case has been done during the experimentations to analyze the predictive model behavior in different states. The proposed model is tuned with two parameters: sequence length and batch size. The Sequence Length (S.L.) parameter represents the number of frames/sequences that delineate a complete action. Here, extensive experiments are done for different values of S.L. The last parameter, namely "batch size", controls the network's stability. It involves the number of samples to be processed before updating the model's internal parameter. The training of the network is done here by considering different batch sizes (i.e., 4, 8, 16 and 32). By considering these batches, the corresponding outcomes have been evaluated. Additionally, another parameter called "patience" is used to represent the stopping criteria of the running program. We set the patience parameter fixed to 50 for all the experiments, which implies that the model terminates its training program automatically when it sees no gain in validation losses for 50 iterations. This protects the model from overtraining.

Taking these considerations, the proposed model explores various cases of S.L. Further, each case is trained and validated on both shoplifting datasets for different batch sizes. The investigated results of the proposed model for different cases are deliberated in Subsection 4.2, while shoplifting datasets are mainly deliberated in Subsection 4.1.

4.1. Datasets

The evaluation of proposed system is done over two datasets, one of which is available on the UCF repository and the other one is synthesized by ourselves. The related description of these two datasets are as follows:

4.1.1. Existing shoplifting dataset

The existing shoplifting dataset is derived from the UCF crime dataset [2], which is openly accessible on the UCF repository **(https://webpages.uncc.edu/cchen62/dataset.html)**. UCF crime dataset consists of 128 hours of untrimmed crime videos, with 13 realistic anomalies, which are 1900 min duration. These anomalies include untrimmed real-world surveillance videos like Vandalism, stealing, Abuse, Arrest, Burglary, Stealing, Shooting, Robbery, Shoplifting, Fighting, Burglary, Arson and Explosion.

Table 2. Details of existing shoplifting dataset

Distribution	Category	Clips	Number of frames
Test	Normal	54	15,660
	Shoplifting	24	6960
Train	Normal	110	31,900
	Shoplifting	64	18,560
Total instances		**252**	**73,080**

To fulfil our objective, the normal and shoplifting video clips have been separated from the arbitrarily arranged CCTV video's in UCF crime dataset. Accordingly, these videos have ben manually trimmed to short videos of 5 s that represent an entire human action. The manually trimmed shoplifting dataset is here termed as the existing shoplifting dataset in this manuscript. It contains two different classes representing stealing and normal actions. Originally, this dataset has comprised 126 clips, with 82 clips representing normal/normal events and 44 clips representing shoplifting incidents. Each clip in the dataset is captured at 30 frames per second with a resolution of 320×240. As dataset size is relatively small and deep neural networks require large labeled inputs for efficient learning therefore, flip augmentation is used here to double the video clips presented in the dataset, which helps to train the network more deeply and make dataset rotation (left & right) invariant. After the augmentation, the dataset comprises 252 clips in which normal and shoplifting clips in the dataset are 164 and 88, respectively. The network is trained using 174 video clips out of 252, while the rest (78 video clips) are utilized for testing. The clips distribution in the existing shoplifting dataset is presented in Table 2.

Fig. 2 depicts image instances of shoplifting and normal classes for the same dataset. The first row in the figure illustrates some sample frames of shoplifting events in which a person is committing theft by concealing store items. On the other hand, the next row depicts the sample frames of normal events like walking, seeing, or examining products in the enterprise.

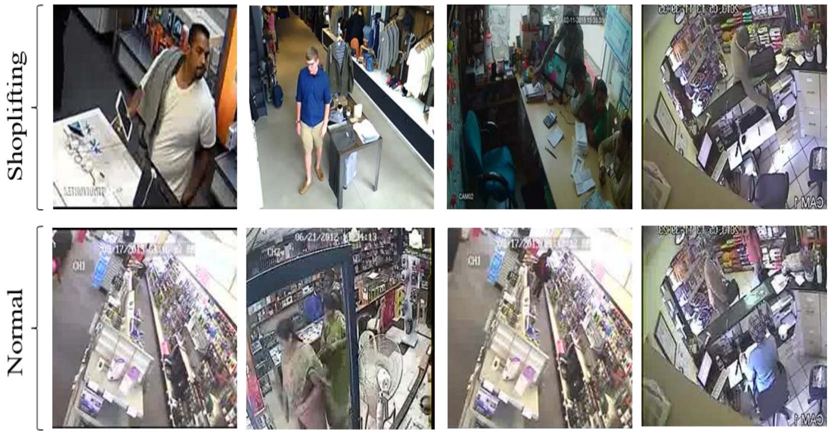


Fig. 2. Image instances of existing shoplifting dataset

4.1.2. Synthesized shoplifting dataset

On scrupulously analyzing each clip of the existing shoplifting dataset, it is found that various discrepancies exist in the video clips, which are in brief as follows.

- **Inter-object occlusion and self-occlusion.** Inter-object occlusion takes place when an object occludes another object, i.e., human stealing action is not clearly visible due to the presence of other stationary object or moving object (like person) or baggage or luggage, etc. Self-occlusion occurs when an object occludes itself, i.e., stealing action is not clear due to human stealing action is obstructed by clothing or other body parts.
- **Viewpoint variation.** The variation in views has been noticed in video clips as well. In the dataset, some clips are captured from too nearer to the camera and some clips are captured from far away of the camera. Therefore, the clips that are captured near to the camera may or may not represent an action more precisely due to occluded action or occluded body part. Similarly, the clips that are captured from far away from the camera contain actions with large background information, which may distort the results. In this case, an action may or may not be recognized.
- **Lightening variation.** Lightening is an essential concept in visual arts. The lightning problem decides the object’s visibility or change in the object’s appearance with varying lighting conditions. The clips presented in the existing shoplifting dataset consist of lightning variations such as low or high lightning in the frames that may create a hindrance for human action recognition.

The aforementioned inconsistencies present in the existing shoplifting dataset may create difficulty in recognizing human actions accurately. Keeping these issues in mind, we have created a video dataset synthesized in the laboratory. The best possible practice has been made to give a real touch to the shoplifting activities. Here, the objective of stealing and hiding is transparently modeled with the synthesized dataset. We have used 32 megapixels mobile camera to record those scenarios like normal and shoplifting kinds of activities. The video clips have been recorded at low resolution 640×480 of size 10 seconds each.

Like the existing dataset, our synthesized dataset also involves two categories: normal class and shoplifting class. The videos involved in normal class contain usual human activities like walking, interacting with others, checking items, etc. On the other side, the shoplifting class contains videos that express human stealing actions like hiding store items under clothes or baggage. The dataset comprises a total of 175 clips, with 88 clips representing normal actions and 87 clips representing shoplifting actions. After flip augmentation as performed in the existing shoplifting dataset, the dataset comprises a total of 350 video clips in which shoplifting and normal classes involve 176 and 174 clips. Here, 256 clips are used in the training process, while the rest of the clips are used for testing purposes. Table 3 exhibits the overall arrangement of the synthesized shoplifting dataset.

Table 3. Details of synthesized shoplifting dataset

Distribution	Category	Clips	Number of frames
Test	Normal	48	13,920
	Shoplifting	46	13,340
Train	Normal	128	37,120
	Shoplifting	128	37,120
Total instances		**350**	**101,500**

Fig. 3 presents the sample frames of the self-synthesized shoplifting dataset. The sample frames in the first row show human stealing actions/shoplifting events performed by the buyers. In contrast, the next row depicts the human's normal behavior, usually seen in stores/shops.

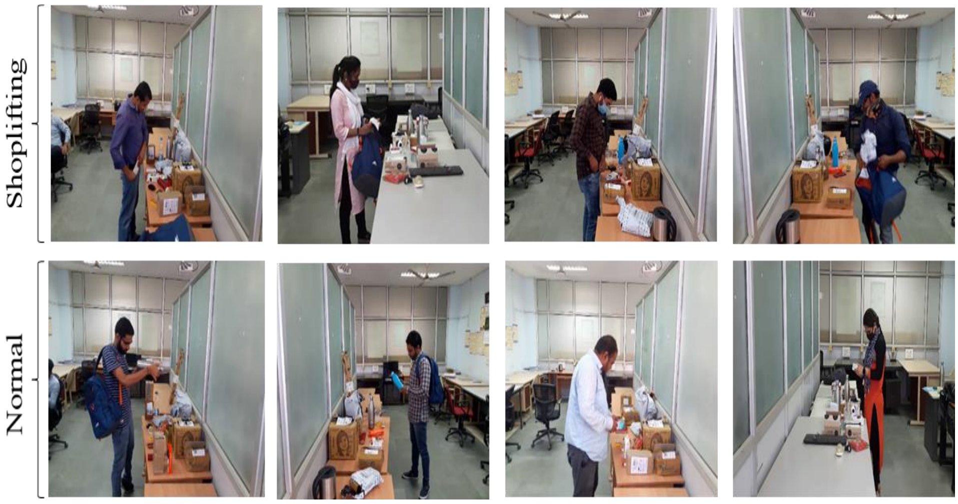


Fig. 3. Sample frames of self-synthesized shoplifting dataset

4.2. Experimental outcomes

As mentioned earlier, S.L. and batch size are the main part of parameter modeling. Considering these parameters, wide ranges of experimentations have been done to analyze the proposed model's performance for different modeling cases. Here, actual S.L. is set to 125 for existing shoplifting clips and 290 for synthesized shoplifting clips. The actual S.L. considers every frame of each clip in experiments. Other side S.L. is set to S.L./2 (consider every second frame of each clip), S.L./3 (consider every third frame of each clip) and S.L./4 (consider every fourth frame of each clip) in the experiments. The investigated results of the proposed model for different cases are deliberated as follows:

Feature extraction is a primary job to structure a smart human action recognition system. Thereby, the proposed network uses the deep Inception V3 architecture to mine pertinent evidence from the sequences, which takes a considerable amount of time. Fig. 4 compares the timing involved in the feature extraction phase for the proposed model trained on the existing and synthesized shoplifting datasets. On exploration, it is observed that the proposed model for the existing shoplifting dataset with sequence length of 125 takes a long time in feature extraction. In turn, the proposed model for synthesized shoplifting dataset with sequence length of 290 takes a much longer time during the feature extraction phase.

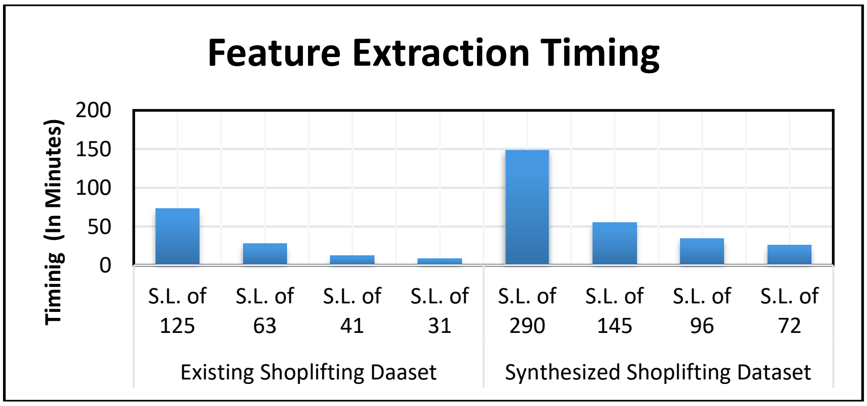


Fig. 4. Feature extraction timing for both datasets

After extracting the features, the classification module is required to categorize the video sequences. The proposed model uses the LSTM network for classifying normal and shoplifting anomalies. Table 4 presents the results obtained from the proposed model trained on the existing shoplifting dataset for different modeling cases. In terms of getting high accuracy for each case, we have discovered that the proposed model, consisting of S.L. of 125, provides up to 100% training accuracy and 73.07% validation accuracy for batch size of 32. Whereas, the model for S.L. of 63 offers up to 100% training accuracy and 80.76% validation accuracy for batch sizes of 16. On the other side, the model with S.L. of 41 attains up to 99.42% training accuracy and 80.76% validation accuracy for batch size of 4. In addition, the proposed model, consisting of S.L. of 31, reaches up to 99.42% training accuracy and 78.21% validation accuracy for batch size of 32.

Table 4. Evaluated results for existing dataset

S.L.	Batch size	Epochs	Training accuracy	Validation accuracy	Training loss	Validation loss
125	**4**	72	100	68.42	0.0005	2.642
	8	76	100	72.36	0.0001	1.386
	16	77	100	68.42	0.0003	1.596
	32	82	100	**73.07**	0.0016	1.212
63	**4**	106	100	73.07	0.0003	1.727
	8	86	99.42	76.92	0.0303	0.622
	16	78	100	**80.76**	0.0013	0.673
	32	81	100	78.21	0.0001	1.047
41	**4**	62	99.27	**80.76**	0.0231	1.562
	8	71	98.26	73.07	0.0548	1.723
	16	72	98.84	75.64	0.0603	1.045
	32	84	98.84	79.49	0.0303	0.988
31	**4**	65	97.69	75.64	0.0047	1.485
	8	67	98.84	70.51	0.0279	1.113
	16	71	97.67	73.08	0.0535	1.223
	32	75	99.42	**78.21**	0.0380	1.105

Table 5. Evaluated results for synthesized dataset

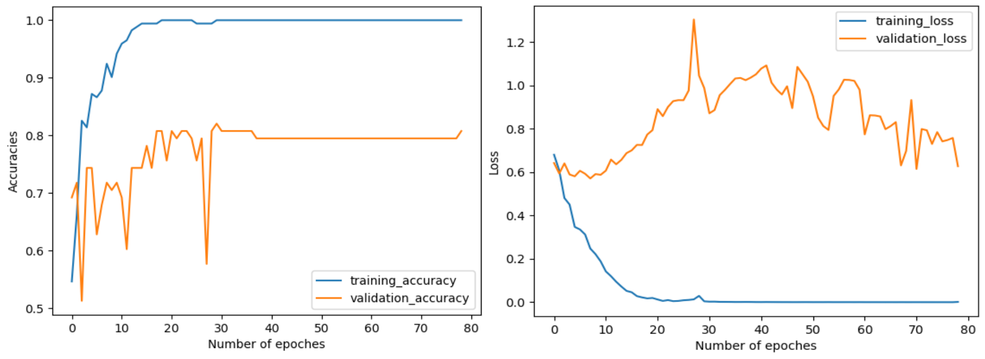
S.L.	Batch size	Epochs	Training accuracy	Validation accuracy	Training loss	Validation loss
290	**4**	117	85.58	79.78	0.194	0.557
	8	147	88.30	84.04	0.282	0.515
	16	292	92.33	**87.23**	0.218	0.484
	32	138	85.48	82.97	0.320	0.518
145	**4**	122	88.30	82.97	0.276	0.467
	8	201	93.14	**89.36**	0.227	0.411
	16	149	87.90	86.10	0.336	0.474
	32	156	87.05	87.23	0.320	0.401
96	**4**	143	100	**88.30**	0.002	0.885
	8	107	97.60	86.10	0.058	0.640
	16	82	98.00	87.23	0.041	0.578
	32	117	98.40	85.11	0.043	0.669
72	**4**	84	99.60	**88.30**	0.027	0.644
	8	85	98.40	88.30	0.055	0.600
	16	87	98.40	87.23	0.039	0.648
	32	92	95.60	84.04	0.124	0.572

Table 5 presents the evaluated results of the proposed model assessed over a self-synthesized shoplifting dataset for various modeling cases. In terms of getting high accuracy for each case, it has been seen that the proposed method, consisting of S.L. of 290, attains up to 92.33% training accuracy and 87.23% validation accuracy for batch size of 16. The model with S.L. of 145 gives training and testing accuracy up to 93.14% and 89.36%, respectively, for batch size of eight. On the other side, the model with S.L. of 96 produces up to 100% and 88.30% training and validation accuracy, respectively, for batch size of 4. In addition, the model with S.L. of 72 scores up to 99.60% training accuracy and 88.30% validation accuracy for batch size of 4.

After getting the results from the existing and synthesized shoplifting dataset, we have evaluated different performance measures like confusion matrix, precision, recall, f1-measure and specificity for the most accurate variants of each sequence length, as presented in Table 6. On analyzing, we found that the proposed model generates the highest accuracy of up to 80.76% and moderate F1-Measure of up to 71.72% for the existing shoplifting dataset (for S.L. of 63), where the class imbalance factor is 1:2. Due to an unbalanced imbalance factor, the model produces moderate precision, recall and specificity values too. On the other side, synthesized shoplifting improves the class imbalance factor to 1:1. Based on that, our model scores the highest accuracy and f1-measure values for the synthesized shoplifting dataset (for S.L. of 145), which is up to 89.36% and 89.13%, respectively.

Table 6. Performance of proposed model for existing and synthesized inputs

Dataset	S.L.	TP	TN	FP	FN	Precision	Recall	F1-Measure	Specificity	Accuracy
Existing shoplifting dataset	**125**	18	39	15	6	54.54	75.00	63.15	72.22	73.07
	63	20	43	11	4	64.51	83.33	71.72	79.62	80.76
	41	18	45	9	6	66.66	75.00	70.58	83.33	80.76
	31	17	44	10	7	62.96	70.83	63.17	81.48	78.21
Synthesized shoplifting dataset	**290**	41	41	7	5	85.41	89.13	87.23	85.41	87.23
	145	41	43	5	5	89.13	89.13	89.13	89.58	89.36
	96	40	43	5	6	88.88	86.95	87.90	89.58	88.30
	72	39	44	4	7	90.69	84.78	87.63	91.66	88.30

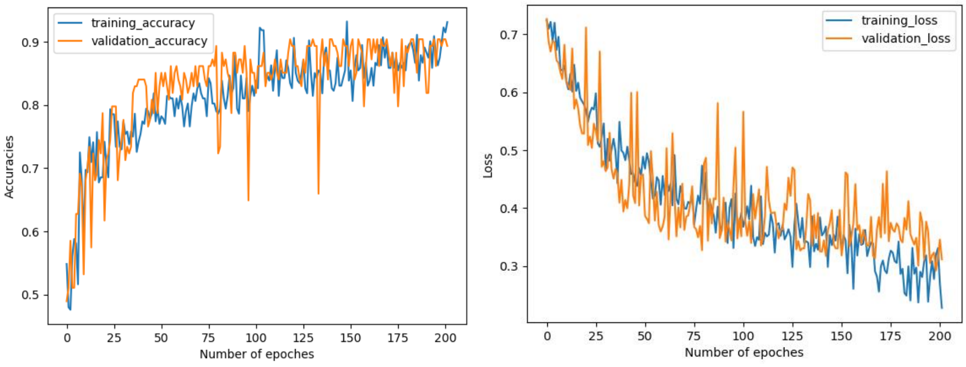


(a) (b)

Fig. 5. Comparison of (a) accuracies and (b) losses for proposed model (S.L. = 63, batch size = 16) over existing shoplifting dataset

Upon analyzing the results of the proposed model for the existing shoplifting dataset, it is observed that the model with S.L. of 63 produces higher training and validation accuracy for batch size of 16 compared to others. Therefore, a comparison of accuracies and losses for the aforementioned cases of the proposed model over the existing shoplifting dataset is presented in Fig. 5. Fig. 5a illustrates the trade-off curve between training and validation accuracy, which increases training accuracy and moderate validation accuracy. Fig. 5b depicts the trade-off curve between training loss and validation loss, which shows a comparatively lower training loss and a moderate validation loss.

On analyzing the proposed model results for the synthesized dataset, we found that the model with S.L. of 145 for batch size of 8 achieves higher validation accuracy than other cases. Therefore, a comparison of training and validation trade-offs for the same variant of the proposed model over the synthesized dataset is depicted in Fig. 6, which shows the decent learning and good performance ability of the proposed model. Fig. 6a shows the accuracy trade-off representing training and validation accuracy for 201 epochs with validation accuracy of up to 89.36%. Fig. 6b illustrates the loss trade-off for training and validation for the aforementioned cases of the proposed model, which incurs lower losses during the training and validation process.



(a) (b)

Fig. 6. Comparison of (a) accuracies and (b) losses for proposed model (S.L. of 145, batch size of 8) over synthesized shoplifting dataset

4.3. Experimental analysis

An epoch is a basic building block to update the internal parameter of the network that defines how the learning algorithm works through the entire training dataset. We have evaluated the average epoch timing for the proposed model training over both datasets. Fig. 7 illustrates the average epoch timing comparison curve for existing and synthesized shoplifting datasets, in which the impact of hyper-parameters to process an epoch is clearly presented.

Analyzing the above graph reveals that the proposed model with actual S.L. (125 for existing dataset and 290 for synthesized dataset) takes a longer time to complete each epoch that enlarges the training time. On the other side, the proposed model with S.L./4 (72 for synthesized dataset and 30 for existing dataset) takes less time to complete each epoch and spends comparatively much lesser time in training than others. For example, in the existing shoplifting dataset, the proposed model with actual S.L. and batch size of 4 spends on an average 32.39 s to evaluate an epoch and takes up to 40 min and 7 s in complete training. On the other hand, the model with S.L./4 and batch size of 4 takes on an average of 2.7 s to evaluate an epoch and spends almost 3 min and 25 s in training. In the case of synthesized shoplifting dataset, the model with actual S.L. and batch size of four spends on an average 150 s in evaluating an epoch, which takes a longer time up to 292 min and 51 s in complete training. The same model with S.L./4, and batch size of 32 spends on an average 4.21 s to evaluate an epoch and takes 6 min and 47 s to train the network. In addition, we also found that smaller batches take longer to train than larger batches. It signifies that the batch size can also affect the training time of the learning algorithm.

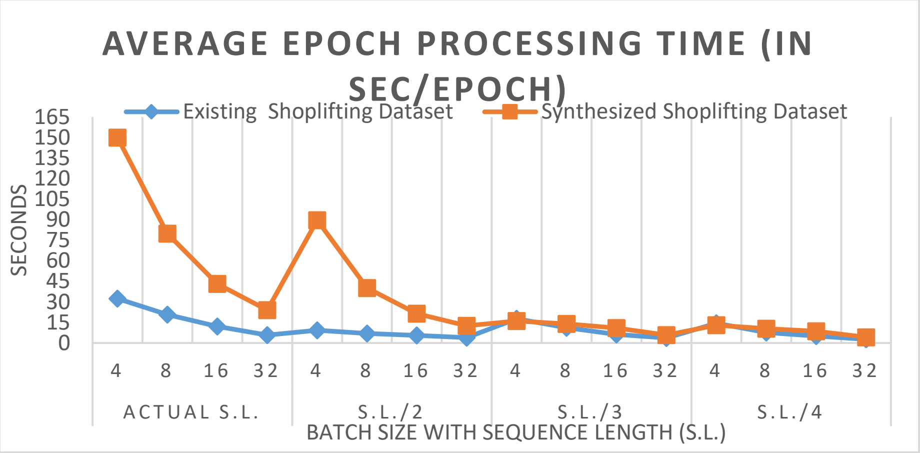


Fig. 7. Average epoch processing time comparison for both datasets

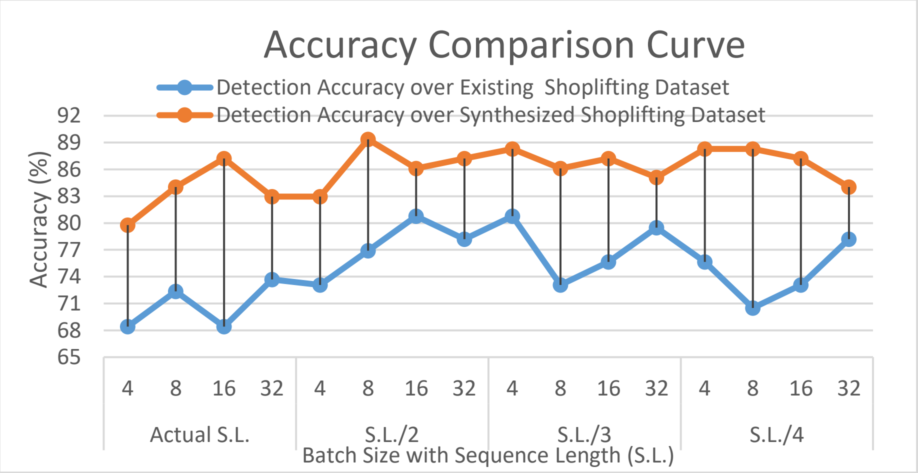


Fig. 8. Comparison of accuracy for different model cases

Finally, the proposed system assessed on two datasets provides decent results for different modeling cases. Fig. 8 displays accuracy comparison curve for both datasets. Upon analysis of the curve, it is seen that the model trained with S.L./2 attains utmost outcomes for both datasets than other cases of sequence length. In the case of S.L./3, the model offers good results for the existing dataset, while the model for S.L./4 shows promising results for both datasets. Finally, the obtained outcomes assessed on the validation process reveal higher detection accuracy for the synthesized dataset (up to 89.36%) than the existing shoplifting dataset (up to 80.76%).

4.4. Resulting samples

In the testing phase, we found that the model trained on the existing shoplifting dataset produces some false positives in the resulting sequences due to discrepancies presented in the dataset. The resulting instances of the video clips evaluated over the testing phase are depicted in the first row of Fig. 9. On the other side, the trained model over synthesized shoplifting dataset scores much more proficient outcomes and provides accurate predictions. The resultant instances of the video clips experimented on the synthesized shoplifting dataset are depicted in the second row of the same figure.

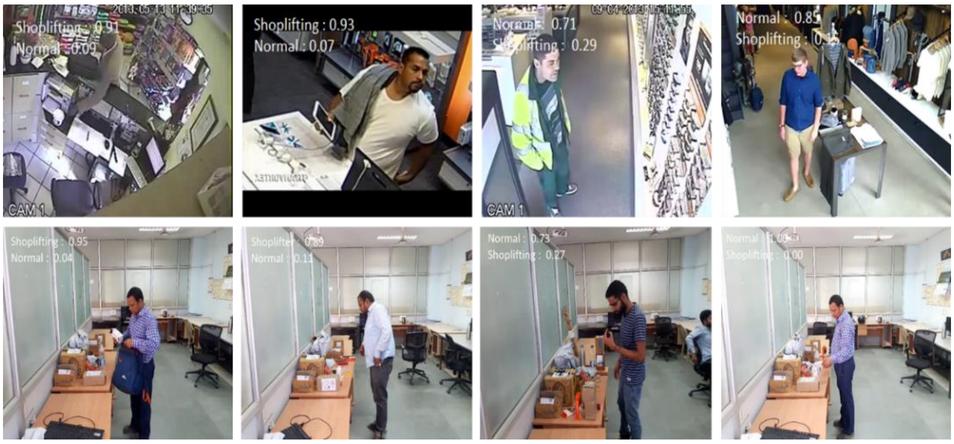


Fig. 9. Resultant sequences evaluated over both of Datasets

4.5. Comparison with existing approaches

In the end, we compare our approach to existing approaches, as shown in Table 7. Here, Guillermo et al. [20] take advantage of three Dimension Convolutional Neural Network (3D-CNN) for recognizing human acts. Jayaswal and Dixit [25] use Xception model with a recurrent learning classifier to identify involved activities in video sequences. Donahue et al. [22] encode human acts using Long Term Recurrent Convolutional Neural Network (LRCN). In addition to these, the proposed model binds the spatial evidence with the temporal domains to encode person actions using deep Inception V3 and LSTM network. Finally, we conclude that the proposed approach trained on both datasets encompasses more precise outcomes than existing approaches.

Table 7. Comparison with existing approach

Approaches	Accuracy (%)	
	Existing shoplifting dataset	Synthesized shoplifting dataset
Guillermo et al. [20]	76.92	84.04
Jayaswal and Dixit [25]	78.20	87.23
Donahue et al. [22]	79.48	88.59
Proposed model	**80.76**	**89.36**

5. Conclusion

Inspired by recent advancements in Deep Neural Networks, we use deep Inception V3 and LSTM design with different modeling parameters to identify shoplifting events in megastores/shops. Moreover, this article presents a synthesized shoplifting dataset consisting of normal and human stealing actions. The research is mainly intended to diminish the merchandise loss due to shoplifting. The model proposed here is capable of encoding human posture dynamics into respective actions, which further discriminates the different actions with the help of the deep LSTM network. In experiments, it is found that the proposed model performs remarkably well for the case of S.L./2. The examined results reveal that the proposed method has achieved a detection accuracy of up to 80.76% and 89.36% on the existing and synthesized shoplifting datasets, which outperforms other HAR-based existing systems. In future, more classes may be introduced in the synthesized shoplifting dataset containing clearly performed human stealing actions. As several HAR-based methods are available to interpret the scene's semantics, we may introduce other deep learning-based HAR techniques to better characterize human pose semantics for discriminating different human acts.

References

1. Arroyo, R., et al. Expert Video-Surveillance System for Real-Time Detection of Suspicious Behaviors in Shopping Malls. – Expert Systems with Applications, Vol. **42**, 2015, No 21, pp. 7991-8005.
2. Waqas, S., C. Chen, S. Mubarak. Real-World Anomaly Detection in Surveillance Videos. – In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2018.
3. National Association for Shoplifting Prevention (NASP). Shoplifting Statistics, 2020. **http://www.shopliftingprevention.org/what-we-do/learning-resource-center/statistics**
4. Singh, D. K., D. S. Kushwaha. Tracking Movements of Humans in a Real-Time Surveillance Scene. – In: Proc. of 5th International Conference on Soft Computing for Problem Solving, Singapore, Springer, 2016.
5. National Retail Security Survey 2020, National Retail Federation. Appriss Retail, 13 July 2020. **https://nrf.com/research/national-retail-security-survey-2020**
6. Xia, K., J. Huang, H. Wang. LSTM-CNN Architecture for Human Activity Recognition. – IEEE Access, Vol. **8**, 2020, pp. 56855-56866.
7. Ansari, M. A., D. K. Singh. ESAR, an Expert Shoplifting Activity Recognition System. – Cybernetics and Information Technologies, Vol. **22**, 2022, No 1, pp. 190-200.
8. Koohzadi, M., N. M. Charkari. Survey on Deep Learning Methods in Human Action Recognition. – IET Computer Vision, Vol. **11**, 2017, No 8, pp. 623-632.

9. Ansari, M. A., D. K. Singh. Human Detection Techniques for Real Time Surveillance: A Comprehensive Survey. – Multimedia Tools and Applications, 2020, pp. 1-50.
10. Singh, D. K., et al. Human Crowd Detection for City Wide Surveillance. – Procedia Computer Science, Vol. **171**, 2020, pp. 350-359.
11. Szegedy, C., et al. Going Deeper with Convolutions. – In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2015.
12. Poppe, R. A Survey on Vision-Based Human Action Recognition. – Image and Vision Computing, Vol. **28**, 2010, No 6, pp. 976-990.
13. Ladjailia, A., et al. Human Activity Recognition via Optical Flow: Decomposing Activities into Basic Actions. – Neural Computing and Applications, Vol. **32**, 2020, No 21, pp. 16387-16400.
14. Nguyen, T. N., Q. L. Ngoc. Abnormal Activity Detection Based on Dense Spatial-Temporal Features and Improved One-Class Learning. – In: Proc. of 8th International Symposium on Information and Communication Technology, 2017.
15. Arati, K., A., A. Khare, M. Khare. Human Activity Recognition Algorithm in Video Sequences Based on Integration of Magnitude and Orientation Information of Optical Flow. – International Journal of Image and Graphics, 2021, 2250009.
16. Abualigah, L., et al. The Arithmetic Optimization Algorithm. – Computer Methods in Applied Mechanics and Engineering, Vol. **376**, 2021, 113609.
17. Szegedy, C., et al. Rethinking the Inception Architecture for Computer Vision. – In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2016.
18. Sanal, K. K. P. S., R. Bhavani. Human Activity Recognition in Egocentric Video Using HOG, GiST and Color Features. – Multimedia Tools and Applications, Vol. **79**, 2020, No 5, pp. 3543-3559.
19. Hochreiter, S., J. Schmidhuber. LSTM Can Solve Hard Long Time Lag Problems. – Advances in Neural Information Processing Systems, 1997, pp. 473-479.
20. Guillermo, M.-M. G. A., et al. Criminal Intention Detection at Early Stages of Shoplifting Cases by Using 3D Convolutional Neural Networks. – Computation, Vol. **9**, 2021, No 2, 24.
21. Yue-Hei Ng, J., et al. Beyond Short Snippets: Deep Networks for Video Classification. – In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2015.
22. Donahue, J., et al. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. – In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2015.
23. Farzan Majeed, N., et al. A Robust Human Activity Recognition Approach Using Openpose, Motion Features, and Deep Recurrent Neural Network. – In: Proc. of Scandinavian Conference on Image Analysis. Cham, Springer, 2019.
24. Asadi-Aghbolaghi, M., et al. A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences. – In: Proc. of 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG'17), IEEE, 2017.
25. Jayaswal, R., M. Dixit. A Framework for Anomaly Classification Using Deep Transfer Learning Approach. – Revue d'Intelligence Artificielle, 2021. **https://doi.org/10.18280/ria.350309**
26. Ioffe, S., C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. – In: Proc. of International Conference on Machine Learning, PMLR, 2015.
27. Barot, V., V. Kapadia. Long Short Term Memory Neural Network-Based Model Construction and Fne-Tuning for Air Quality Parameters Prediction. – Cybernetics and Information Technologies, Vol. **22**, 2022, No 1, pp. 171-189.

Received: 16.09.2022; Second Version: 22.11.2022; Accepted: 06.01.2023