BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **22**, No 1
Sofia • 2022 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2022-0012

ESAR, An Expert Shoplifting Activity Recognition System

Mohd. Aquib Ansari, Dushyant Kumar Singh

CSED, MNNIT Allahabad, Prayagraj, India
E-mails: mansari@mnnit.ac.in dushyant@mnnit.ac.in

***Abstract*:** *Shoplifting is a troubling and pervasive aspect of consumers, causing great losses to retailers. It is the theft of goods from the stores/shops, usually by hiding the store item either in the pocket or in carrier bag and leaving without any payment. Revenue loss is the most direct financial effect of shoplifting. Therefore, this article introduces an Expert Shoplifting Activity Recognition* (*ESAR*) *system to reduce shoplifting incidents in stores/shops. The system being proposed seamlessly examines each frame in video footage and alerts security personnel when shoplifting occurs. It uses dual-stream convolutional neural network to extract appearance and salient motion features in the video sequences. Here, optical flow and gradient components are used to extract salient motion features related to shoplifting movement in the video sequence. Long Short Term Memory* (*LSTM*) *based deep learner is modeled to learn the extracted features in the time domain for distinguishing person actions (i.e., normal and shoplifting). Analyzing the model behavior for diverse modeling environments is an added contribution of this paper. A synthesized shoplifting dataset is used here for experimentations. The experimental outcomes show that the proposed approach attains better consequences up to 90.26% detection accuracy compared to the other prevalent approaches.*

***Keywords*:** *Automated surveillance system, Human Activity Recognition* (*HAR*), *Histogram of Oriented Gradient* (*HOG*), *Optical Flow, Convolutional Neural Network* (*CNN*), *Long Short Term Memory* (*LSTM*).

1. Introduction

Every year, people shoplift commodities worth billions of dollars in the world. Shoplifting [1, 2] is a crime that decreases the profitability of retailers. It is an act of stealing goods from the established retailer when no one is observing and leave the store without paying for them. The National Retail Security Survey [3] found shoplifting to be the main cause of enlarging the inventory shrinkage, introducing a massive loss of about $50 billion to the USA retail industry in 2016. As per the Global Retail Theft Barometer [4], global shrinkage during 2014-2015 was about $123.4 billion. The National Association for Shoplifting Prevention revealed that shoppers typically steal items from all kinds of retail markets, such as supermarkets, departmental stores, small shops and drug stores. The same report also indicates that

nearly one in eleven shoppers in USA commits shoplifting. Consequently, every affected trader is interested in curbing shoplifting.

Some countries have enacted specific statutes to deal with shoplifting and punish shoplifters according to their laws. India treats shoplifting as an offense against property and deals with shoplifters under the Indian Penal Code (IPC) Chapter XVII [5]. Section 379 of IPC quotes punishment for theft as imprisonment for up to three years, or fine, or both. In recent times, Closed-Circuit TeleVision (CCTV) surveillance infrastructure [6, 11] has been adopted by retailers to monitor each buyer in stores/shops. However, real-time analysis of video footage produced by cameras is a tedious task because the human gaze cannot proactively examine each frame continuously, as omissions are usually possible. Therefore, an advanced monitoring system could be the veracious solution for identifying shoplifting by analysing human stealing actions in stores and shops.

As of now, there has not been any cost-effective and accurate performing approach for detecting shoplifting in real-time. Therefore, this article proposes an advanced and automated Human Action Recognition (HAR) system [10, 18] to identify shoplifting in megastores/shops by analysing human stealing actions. Generating a warning message on the screen when shoplifting occurs could be a consequent outcome of the system. The proposed system leverages the deep Convolutional Neural Network (CNN) to mine pertinent features and the deep LSTM network to classify the performed human acts. Fusing dual-stream network features is a novelty of this work in which CNN extracts spatial features from RGB streams and motion features from another stream. Optical Flow vectors (i.e., angle and magnitude) and HOG vectors here represent this encoded salient information. The fusion of this dual-stream of features is done and used to build the LSTM network for classifying normal and shoplifting events in real-time video sequences. The suggested method has the potential to reduce shoplifting-related offenses in the future and can reduce business losses in merchandise.

The following are the contributions of this paper:

- We propose a dual-stream fusion based activity recognition scheme to recognize shoplifting events in real-time scenarios;
- We perform a wide range of experiments to analyse the behaviour of the proposed model on synthesized data.

The rest of the paper is structured as follows. Section 2 involves the correlated works of existing activity recognition systems. The procedure for the proposed methodology is presented in Section 3. Section 4 includes a detailed discussion of experimentations and their analysis. The conclusion with the future scope of this paper is placed in Section 5.

2. Related works

Human Activity Recognition (HAR) has been a trending research topic over the past few decades. HAR mainly analyses video sequences and detects spontaneous activity performed by a person. Some of the existing HAR systems are discussed as follows.

Li, Tong and Tang [7] have proposed an action recognition framework by modeling human body posture to achieve multimodal action recognition for monocular videos. The system takes advantage of RGB, optical flow and human pose features to distinguish different actions. The proposed action encoding scheme can handle flexible posture inputs and pose errors as well. Kumar and Bhavani [8] have proposed a human activity recognition system using multimodal egocentric videos. The system segments the region of interest using the watershed algorithm and extracts the pertinent features using HOG, Color and GiST descriptors. Classification is done using random forest classifier with Genetic algorithm for feature reduction.

Rashwan et al. [9] have suggested a novel approach to represent and recognize the actions using Histogram of Optical Flow Co-Occurrence (HOF-CO) and deep convolutional neural network. They have found that HOF-CO can encode relative frequencies of optical flow directions to epitomize the motion dynamics of the human act more precisely. A. Kushwaha, A. Khare and M. Khare [10] also have used optical flow motion vectors with gradients information to encode action dynamics. The proposed approach is view-invariant and robust to lighting variations. Here, a support vector classifier is used for activity classification. Donahue et al. [12] have offered a visual recognition model using a long-term deep recurrent convolutional network. The proposed model is capable of learning spatial-temporal information for action dynamics. Ladjailia et al. [13] have detected human activity over decomposed actions using salient motion information. Optical Flow estimation and motion based descriptor are used here to extract motion information and K Nearest Neighbour (KNN) is used for classification purpose. Jayaswal and Dixit [14] classify different types of violence anomalies using deep neural architecture. The method to analyze human behavior in real-time scenarios uses fine-tuned Xception model for features extraction and an LSTM model for anomaly classification.

The literature, as discussed above uses feature based methods and deep neural architecture to discriminate a person's real-time activities. In contrast, some researchers propose advanced HAR architecture to categorize some complex tasks more precisely, which are discussed as follows:

Arroyo et al. [1] have developed a module to identify various risk situations like entry, exit, loitering, and unattended cash counter detection in stores. The framework detects the human by utilizing color and textural based features extracted using GCH, LBP and HOG descriptors and SVM classifier as a classification scheme. This work involves the loitering event detection for preventing shoplifting in store. However, loitering recognition can reduces shoplifting up to some extends but not completely. Yamato, Fukumoto and Kumazaki [15] have suggested a proposal to build a shoplifting prevention framework for small shops using image analysis and advanced cloud technology. This proposal uses Jubatus system [16] to detect anomalies in real-time scenarios and informs the cloud when the anomaly occurred. Next, the application run on the cloud scrutinizes the item DB in ERP and alerts the security personnel over emailing when the probability of theft is high.

Martínez-Mascorro et al. [17] detect criminal intention at the early stage of suspicious behavior using 3D Convolutional Neural Networks. The work is

validated on video clips taken from the UCF Crime dataset that contains daily actions and shoplifting samples. This model encourages proficient outcomes to detect suspicious behavior as crime prevention. Ansari and Singh [2] have suggested an action recognition model to recognize human shoplifting actions in the indoor environment. The work is practiced with Inception V3 to mine pertinent features from video sequences and LSTM network to differentiate the involved actions. UCF Crime dataset is used for experimentations and achieved up to 75.41% validation accuracy in recognition.

After going through the literature, we found that a cost-effective solution is still needed to detect complex activity like shoplifting in real-time scenarios automatically. Therefore, this paper proposes a dual stream network-based deep learning architecture to detect shoplifting events by analyzing human activities. The details related to the proposed architecture are discussed in the subsequent section.

3. Projected methodology

This section presents a real-time approach to be able to capture the shoplifters by recognizing their stealing actions proficiently. Fig. 1 depicts a block schematic of the suggested method. The proposed method takes two inputs: RGB stream and Composite Stream. The RGB stream represents the appearance information, while the composite stream represents the salient motion information for human acts. The details about the composite stream are discussed in later part of the subsequent paragraph. Here, deep Inception V3 architecture is used to retrieve pertinent features from each frame of both streams, i.e., RGB and Composite streams, separately. The model fuses these features using average fusion and passes them to build a Long Short Term Memory (LSTM) network to classify human acts in any of the two classes (normal and shoplifting).

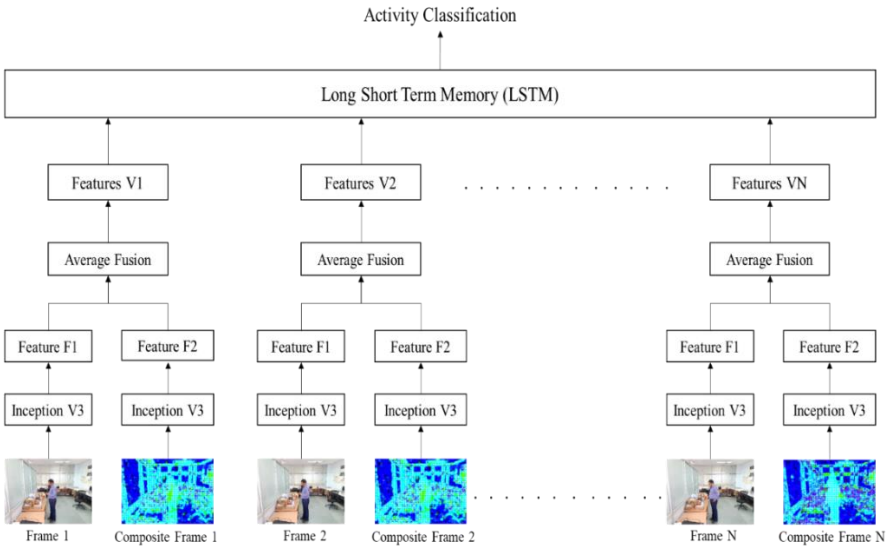

Fig. 1. Block representation of the proposed method

Each frame in the composite stream represents motion evidence of an action. RGB frames are used to generate composite frames using Optical Flow [9, 10] and Histogram of Oriented Gradient (HOG) [1, 18]. The structure of the composite frame is depicted in Fig. 2. The Optical Flow Algorithm takes current and previous RGB frames to evaluate motion vectors like angle and magnitude that represent the moving object information. In contrast, the HOG Algorithm generates gradient vectors that represent gradient directional information for the inputted spatial information. The vectors obtained from optical flow (i.e., angle and magnitude vectors) and HOG (i.e., gradient vector) are represented in a separate plane. Consequently, all these planes are combined to represent the composite frame. The intention behind this concept is to make an image more interpretable and valuable in such a way that it can provide enriched features at the feature extraction phase.

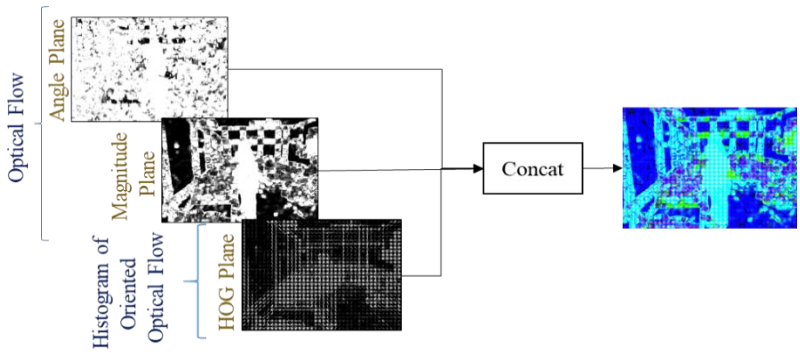

Fig. 2. Structure of composite frame

4. Experimentations and explorations

All experiments are done in the Windows environment on a machine configured with 9th generation i5, 8GB RAM, and 256 GB SSD. CUDA 10.0 with CUDNN 7.5, libraries are used to effectively accelerate the Graphical Processing Unit (GPU) based processing. The dataset used in the experiments is detailed as follows.

4.1. Synthesized shoplifting dataset

We developed the synthesized shoplifting dataset in the laboratory. The videos are recorded using 32 MP camera at a resolution of 480×640. The dataset involves two action classes, namely normal and shoplifting. The normal class involves usual human actions in stores like walking, seeing the product, examining, etc. In contrast, the shoplift class involves human stealing actions like putting the item in clothes or bags. The dataset comprises 175 video clips in which 88 video clips characterize normal behavior of human while remain of clips represents human shoplifting acts. The random sample of 128 clips are used for training and remaining clips are used for testing purpose. To support the current scenario, the clips in the shoplifting dataset involve person's actions with occluded face, partially occluded face and non-occluded face. Around 18%, 24% and 58% of clips represent human acts (i.e., normal

and shoplifting) for covered face, partially covered face and uncovered face. Fig. 3 shows some instances of clips for the same dataset.

Fig. 3. Instances of synthesized shoplifting dataset

4.2. Experimental outcomes

The proposed model is tuned with different cases of modeling parameters to analyze the predictive model's behavior. The parameters used are sequence length, augmentation and batch size. The first parameter, i.e., Sequence Length (SL) represents the number of frames to represent an action. The synthesized shoplifting dataset involves video clips recorded at 30 FPS of length 10 seconds each. Therefore, the value of sequence length is set to 290 in the experimentations. Processing 290 frames of each clip in the synthesized shoplifting dataset is a computationally expensive task. Therefore, the number of frames to represent an action is reduced to half (e.g., set sequence length to 145) in the next phase of experimentations. The next parameter, named augmentation, is used to double clips in the dataset by flipping spatial coordinates of video sequences horizontally. It helps to train the proposed network in deep. Moreover, the parameter, namely batch size, is used to control the stability of the network. The network proposed is trained for dissimilar batch sizes and represents different behavior states of the proposed model. The performance of each modeled case is discussed as follows:

The first case uses sequence length of 290 with no augmentation for the proposed model. It takes approximately 2 hours and 43 minutes in feature extraction.

The experimental results of the proposed model for the first case are presented in Table 1. In experiments, it has been seen that the proposed model for batch size of 16 achieves higher outcomes up to 98.38% training accuracy and 82.97% validation accuracy compared to others. As well as, a lower training loss and moderate validation loss pose stable behavior of the network.

Table 1. Model performance for proposed architecture with sequence length of 290 and no augmentation

Batch size	Iterations	Training accuracy	Training loss	Validation accuracy	Validation loss
4	108	92.19	0.025	78.72	0.742
8	112	96.77	0.081	76.59	0.751
16	123	98.38	0.091	82.97	0.635
32	154	95.96	0.162	76.59	0.795

The second case uses sequence length of 290 with augmented video clips for the proposed model. It spends quite a long time in feature extraction, which is approximately 5 hours and 29 minutes. Table 2 presents the experimental results of the second case of the proposed model. The model for batch size of eight offers up to 93.54% training accuracy and 80.85% validation accuracy. As in the first case, the model presents stable behavior due to low training loss and moderate validation loss.

Table 2. Model performance for proposed architecture with sequence length of 290 and augmentation

Batch size	Iterations	Training accuracy	Training loss	Validation accuracy	Validation loss
4	153	89.51	0.192	76.59	0.684
8	151	93.54	0.161	80.85	0.644
16	140	83.06	0.297	74.46	0.761
32	267	95.96	0.212	78.72	0.589

First and second case use longer sequence length for experiments, and therefore, they require large computation and more space in memory. Accordingly, sequence length is halved here by considering every second frame of the video clip for experiments. This process cuts the computational cost and takes less memory space than the first and second cases. Therefore, the third case uses sequence length of 145 with no augmentation for the proposed model. It takes a lower time in the feature extraction process, which is around 1 hour and 21 minutes. The experimental results of the proposed model for the third case are presented in Table 3, which offers higher accuracy up to 98.79% in training and 87.23% in the validation process for batch size 16. The presented model incurs lower training and validation losses, posing more stable behavior than the first and second cases.

The last case incorporates sequence length of 145 with augmented video clips for the proposed model. Table 4 presents the obtained outcomes of the proposed model for the final case. It takes up to 2 hours and 36 minutes in the feature evaluation process. It provides good accuracy up to 96.77% in training and 90.42% in the validation process for batch size of 32 compared to others. The proposed network for the last case poses stable behavior same as in last three cases.

Table 3. Model performance for proposed architecture with sequence length of 145 and no augmentation

Batch size	Iterations	Training accuracy	Training loss	Validation accuracy	Validation loss
4	97	97.98	0.185	78.72	0.541
8	113	98.79	0.136	78.72	0.413
16	107	98.79	0.043	87.23	0.371
32	172	98.96	0.078	82.97	0.449

Table 4. Model performance for proposed architecture with sequence length of 145 and augmentation

Batch size	Iterations	Training accuracy	Training loss	Validation accuracy	Validation loss
4	186	90.74	0.107	87.23	0.432
8	288	92.77	0.064	82.97	0.764
16	292	96.77	0.086	88.29	0.429
32	188	94.23	0.048	90.42	0.389

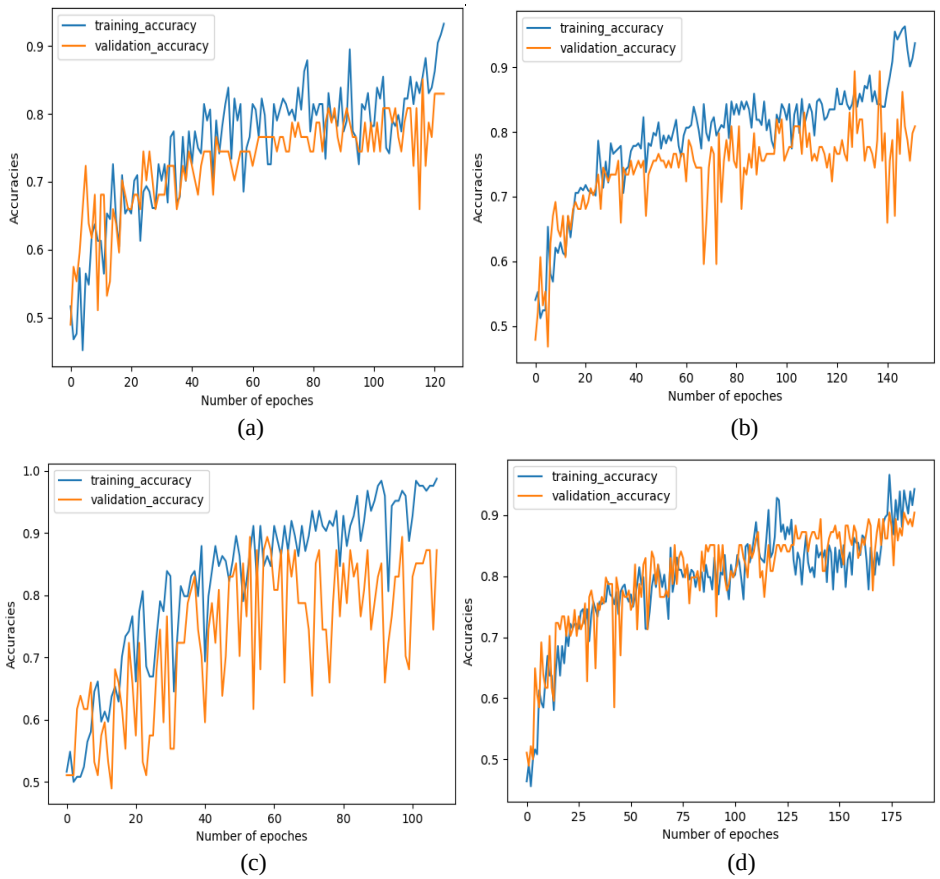

Fig. 4. Trade-off plots for proposed model for: first modeling case (Batch size = 16) (a); second modeling case (Batch size = 8) (b); third modeling case (Batch size = 16) (c); last modeling case (Batch size = 32) (d)

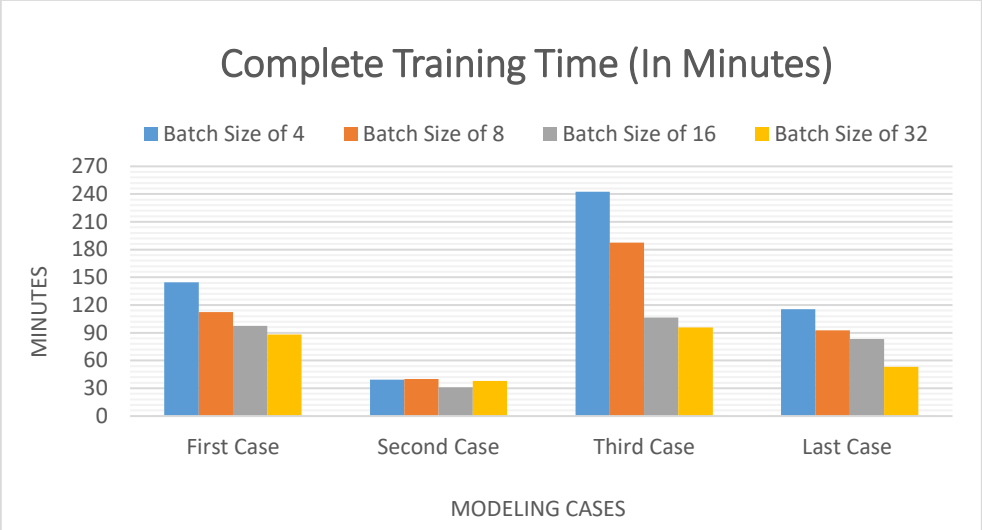

Fig. 5. Training time for each case of classifier modelling

Fig. 4 illustrates the trade-off curves between training and validation for the topmost variant of each case of the proposed model. The proposed model in the second case achieves the lowest accuracy up to 80.85% with batch size of 16. In contrast, the proposed model in the fourth case achieves the highest accuracy of 90.42% with batch size of 32, compared to other cases.

Fig. 5 represents the involved training time of the proposed model for each modeling case. It has been observed that the proposed model trained on sequence length of 290 with augmented video clips (seconds modeling case) takes a long time in complete training. Otherside, the model trained on sequence length of 145 with no augmented clips (third modeling case) spends much lower time in training compared to others. It has also been observed that the model with a shorter batch size spends a longer time in training process, while with larger batch size it comparatively takes less training time.

4.3. Resulting samples

Fig. 6. Resulting image sequences from proposed approach

The proposed model for the afore-mentioned case (Sequence Length = 145, augmentation = Yes and Batch Size = 32) is tested over different video clips. The model generally shows good behavior in activity classification. The resulting samples of tested video clips are presented in Fig. 6.

4.4. Comparison with existing methods

In the end, the proposed method is compared to other existing methods, as depicted in Fig. 7. It is found that the detection accuracy obtained by the proposed method for the last case is about 7.6% higher than [17], 3.6% higher than [14], 2% higher than [12] and 1.2% higher than [2]. Finally, we conclude that the proposed method delivers favorable results compared to other existing methods given in the literature.

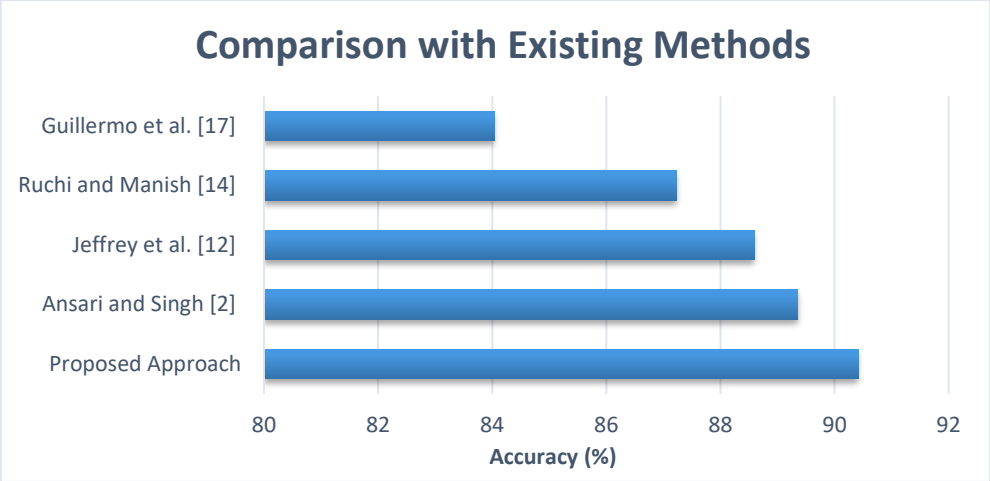

Fig. 7. Comparison with existing methods

5. Conclusion

This paper proposes an Expert Shoplifting Activity Recognition (ESAR) system by combining dual-stream convolutional network architecture with recurrent learning design. Assessing motion dynamics using optical flow and HOG vectors is the foremost task of this work. We extract the pertinent features from both streams (i.e., spatial and motion) using the deep Inception V3 network and use LSTM network to discriminate between normal and shoplifting acts. Our proposed method shows encouraging results with a detection accuracy of up to 90.26% on the synthesized shoplifting inputs. In the future, more stealing actions may be included in the existing shoplifting dataset to make this work more intuitive for identifying different stealing actions. Apart from this, a cloud-based infrastructure may also be incorporated to prevent shoplifting in the stores/shops by spotting the shoplifter using frontal face recognition.

References

1. Arroyo, R., J. J. Yebes, L. M. Bergasa, I. G. Daza, J. Almazán. 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. Ansari, M. A., D. K. Singh. An Expert Eye for Identifying Shoplifters in Mega Stores. – In: Proc. of International Conference on Innovative Computing and Communications (ICICC'21), Vol. **3**, (Vol. **1394**, p. 107), August 2021, Springer Nature.
3. NRF. National Retail Security Survey. National Retail Federation, Washington, DC, USA, 2020.
4. The Global Retail Theft Barometer. 2015.
5. Rankin, G. C. The Indian Penal Code. – LQ Rev., Vol. **60**, 1944, No 37.
6. Singh, D. K. Human Action Recognition in Video. – In: Proc. of International Conference on Advanced Informatics for Computing Research, Singapore, Springer, July 2018, pp. 54-66.
7. Li, C., R. Tong, M. Tang. Modelling Human Body Pose for Action Recognition Using Deep Neural Networks. – Arabian Journal for Science & Engineering (Springer Science & Business Media BV), Vol. **43**, 2018, No 12.
8. Kumar, K. 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.
9. Rashwan, H. A., M. A. Garcia, S. Abdulwahab, D. Puig. Action Representation and Recognition through Temporal Co-Occurrence of Flow Fields and Convolutional Neural Networks. – Multimedia Tools and Applications, Vol. **79**, 2020, No 45, pp. 34141-34158.
10. Kushwaha, 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.
11. 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, pp. 491-500.
12. Donahue, J., L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. – In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625-2634.
13. Ladjailia, A., I. Bouchrika, H. F. Merouani, N. Harrati, Z. Mahfouf. Human Activity Recognition via Optical Flow: Decomposing Activities into Basic Actions. – Neural Computing and Applications, Vol. **32**, 2020, No 21, pp. 16387-16400.
14. Jayaswal, R., M. Dixit. A Framework for Anomaly Classification Using Deep Transfer Learning Approach. – Revue d'Intelligence Artificielle, Vol. **35**, 2021, No 3, pp. 255-263. **https://doi.org/10.18280/ria.350309**
15. Yamato, Y., Y. Fukumoto, H. Kumazaki. Proposal of Shoplifting Prevention Service Using Image Analysis and ERP Check. – IEEJ Transactions on Electrical and Electronic Engineering, Vol. **12**, 2017, pp. S141-S145.
16. Hido, S., S. Tokui, S. Oda. Jubatus: An Open Source Platform for Distributed Online Machine Learning. – In: NIPS 2013 Workshop on Big Learning, Lake Tahoe, December 2013.
17. Martínez-Mascorro, G. A., J. R. Abreu-Pederzini, J. C. Ortiz-Bayliss, A. Garcia-Collantes, H. Terashima-Marín. Criminal Intention Detection at Early Stages of Shoplifting Cases by Using 3D Convolutional Neural Networks. – Computation, Vol. **9**, 2021, No 2.
18. Singh, D. K., S. Paroothi, M. K. Rusia, M. A. Ansari. Human Crowd Detection for City Wide Surveillance. – Procedia Computer Science, Vol. **171**, 2020, pp. 350-359.

Received: 12.10.2021; Second Version: 23.12.2021; Accepted: 21.01.2022