Recognition of Shoplifting Activities in CCTV Footage Using the Combined CNN-RNN Model

Lyudmyla Kirichenko[1,2][0000−0002−2780−7993],
Oksana Pichugina[3,4][0000−0002−7099−8967],
Bohdan Sydorenko[1][0000−0002−5963−5911],
Sergiy Yakovlev[3,5][0000−0001−6736−371X]

[1]*Kharkiv National University of Radio Electronics*
14 Nauki Avenue, 61166 Kharkiv, Ukraine
lyudmyla.kirichenko@nure.ua
[2]*Wroclaw University of Science and Technology*
27 Wyspianskiego, 50-370 Wroclaw, Poland
[3]*National Aerospace University "Kharkiv Aviation Institute"*
17 Chkalova Street, 61070 Kharkiv, Ukraine
o.pichugina@khai.edu
[4]*University of Toronto*
27 King's College Circle, M5S 1A1 Toronto, Canada
[5]*Lodz University of Technology*
Institute of Information Technology
Politechniki 8, 93-590 Łódź, Poland
s.yakovlev@khai.edu

DOI:10.34658/9788366741928.8

Abstract. *The recognition of human activities through surveillance has numerous applications across various fields. This article presents a proposed approach to identify shoplifting in camera-recorded video data using a neural classifier that combines two neural networks, specifically, convolutional and recurrent networks. The hybrid architecture consists of two parallel streams: initial and processed video fragments (histogram of oriented gradients and optical flow). The convolutional network extracts features from each frame of the video fragment, while the recurrent network processes the temporal information from sequences of frames as features to classify the activity.*

Keywords: *human activity recognition, surveillance, shoplifting, convolutional neural network, recurrent neural network, features extraction, histogram of oriented gradients, optical flow*

1. Introduction and Literature Review

Recognition of human actions is an important task in modern video surveillance. Over the years, the number of video cameras in public places has complicated the task of video monitoring. CCTV (Closed Circuit Television) camera networks generate and transmit huge amounts of data, which makes automatically processing all the information crucial.

Video surveillance processing is an important tool for detecting shoplifting, as video analytics can automatically analyze large amounts of video data, detect illegal activities, and send real-time alerts to security guards.

For this, various Machine Learning (ML) algorithms are used. Such ML models are trained on comprehensive datasets of shoplifting, allowing them to identify thief patterns and classify their actions based on certain features.

Convolutional Neural Networks (CNNs) are a powerful tool for image classification and have significantly advanced video processing in recent years. However, it should be noted that video classification requires considering both spatial and temporal characteristics of objects.

In [1], the authors use a pre-trained 3D CNN model to extract video features. Then they use a fully connected neural network to build a regression predicting whether an action was "normal" or "abnormal". The authors tested its model on videos of thefts, fights, and traffic accidents on the UCF-Crime dataset [2]. In [3], the authors presented an approach to real-time anomaly detection using 3D CNN. In the paper [4], the authors used 3D CNN to extract features from video data and classify some events.

Another approach to detecting anomalies in video surveillance is a combination of convolutional and recurrent neural networks (RNNs). It allows the creation of models for extracting spatial and temporal features from a video sequence.

In [5], the authors analyze the existing video classification methods and found that the combination of CNN and RNN works better than methods that use only CNN. For example, in [6], authors use 3D CNN to analyze the presence of violence in surveillance video. To solve this problem, applying only CNN may not be sufficient, so the authors utilize RNN in the model to encode relevant temporal information.

Similarly, in [7], authors use 3D CNN to extract spatial features from video data and LSTM (Long Short-Term Memory) to classify human actions.

Authors of [8] apply a convolutional neural network to analyze typical thief movement features and LSTM for training-derived features.

In [9, 10], a hybrid neural network detects shoplifting. It consists of convolutional and recurrent neural networks. This model uses a CNN to extract important features from video frames. In the recurrent network, gated recurrent units were utilized.

This paper aims to create an effective real-time store theft detection model based on video data processing that utilizes the Combined CNN-RNN Model.

2. Methods and materials

Our study uses the UCF-Crime Dataset [2] as input for our experiments. The dataset includes 1900 videos of varying lengths, totalling 128 hours of actual criminal acts, such as abuse, arrest, arson, assault, traffic accidents, burglary, explosion, fight, robbery, shooting, shoplifting, and vandalism. In particular, the dataset includes 28 videos from a retail store and video surveillance cameras containing shoplifting.

For training our neural network, we artificially increased the number of instances by dividing each video into 32 fragments of 3-second duration. The resulting dataset of 896 video fragments was divided into two classes: 155 videos with shoplifting and 741 videos without shoplifting.

Since video recordings contain information about time and space, both types of information need consideration when analyzing video fragments. We believe convolutional and recurrent neural networks are the best architectures for accomplishing the task. Therefore, we chose a combination of these networks to classify the videos.

Classified video clips of equal duration and many video frames were used as input data. Each frame sequence was marked as either “0” (not shoplifting) or “1” (shoplifting). The marked set of frame sequences was used as a training sample. The features were obtained for each object through a hybrid neural network to train a classifier, which was then used to classify new objects.

In order to improve the quality of the model, we decided to use preprocessing of video fragments. Namely, we used a combination of histograms of oriented gradients and optical flows.

A fast and highly promising way to represent images for classification is by utilizing HOG features. These features were extensively used in pedestrian identification and have remained a reliable technique for feature extraction. HOG features rely on the distribution of gradient angles and magnitudes, making them resistant to minor shifts in lighting and colour variations in visual data [11].

The apparent motion of objects in a visual scene, caused by the motion of a camera or object or both, can be described as optical flow. When a camera records a scene over a certain time, the resulting image sequence can be represented as a function of gray values at the pixel position (x, y) and time t. If the camera or an object within the scene moves, it causes a time-varying shift in the gray values of the image sequence. The optical flow field in the image domain is the resulting two-dimensional pattern of apparent motion [12].

3. Computational Experiment and Results

We develop an algorithm to detect shoplifting, which can be seen as a classification problem. In order to achieve a sufficiently high level of accuracy for our classifier, we conducted model tunning, including extensive research and experiments, including selecting the video classification method, searching for a suitable data set, determining optimal data processing, and configuring neural networks and their parameters.

We describe our main experiment below. Due to the small size of the dataset (only 310 instances) for the non-trivial task of human action classification, we artificially enlarged the dataset. Each video fragment was horizontally mirrored to achieve this, resulting in 620 instances. Additionally, two more copies were generated from each of the 620 fragments, rotated 5 degrees to the left and right, respectively, which resulted in a total of 1860 video fragments.

To enhance the dataset for our purposes, a combination of a histogram of directional gradients and optical flow was applied as preprocessing to the initial set of video fragments. As a result, two sets of 1860 video fragments were obtained: one initial and one preprocessed. The model processed these two sets in parallel, and the results were combined by averaging.

For feature extraction, MobileNetV3Large was used as a convolutional neural network. The ’imagenet’ weights for the model are stored in Keras. The values of calculated accuracy, recall, and F1-scores for each of the classes are presented in Table 1. Thus, the accuracy of the conducted classification is 92%. Since the sample was balanced and considering the presented values of the metrics, this value fully characterizes this result of classification.

Table 1. Accuracy, precision, recall and F1-score values

metric	**precision**	**recall**	**F1-score**
Not Shoplifting	0.90	0.95	0.93
Shoplifting	0.95	0.90	0.92
Accuracy			0.92

4. Conclusions

This study aimed to develop a classifier for identifying shoplifting cases in video data from security cameras. A hybrid neural network classifier involving convolutional and recurrent neural networks was offered to achieve the goal.

The UCF-Crime dataset was chosen as the training dataset, containing videos depicting shoplifting incidents. The major class was under-sampled to address the issue of unbalancing the dataset, while the video data set was artificially enlarged. Additional experiments were conducted using a pre-trained CNN. A neural network with gated recurrent units was utilized for the sequence classification of video clips.

The classifier exhibited a high classification accuracy of 92%, which is several percent higher than the accuracy of models presented in previous relevant studies. Furthermore, our trained classifier demonstrates high performance, enabling its use in real-time applications. Future research will focus on the practical implementation of the proposed model in shopping malls.

Acknowledgment

The work was supported in part by Beethoven Grant No. DFG-NCN 2016/23/G/ST1/04083.

References

[1] Nasaruddin N., Muchtar K., Afdhal A., Dwiyantoro A.P.J., *Deep anomaly detection through visual attention in surveillance videos*, vol. 7, no 1, p. 87, ISSN 2196-1115, doi: 10.1186/s40537-020-00365-y.

[2] Sultani W., Chen. C., Mubarak S., *Papers with code – UCF-crime dataset*. https://paperswithcode.com/dataset/ucf-crime

[3] Sultani W., Chen C., Shah M., *Real-world anomaly detection in surveillance videos*, [In:] *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, doi: 10.1109/CVPR.2018.00678.

[4] Martínez-Mascorro G.A., Abreu-Pederzini J.R., Ortiz-Bayliss J.C., Garcia-Collantes A., Terashima-Marín H., *Criminal intention detection at early stages of shoplifting cases by using 3d convolutional neural networks*, vol. 9, no 2, p. 24, ISSN 2079-3197, doi: 10.3390/computation9020024.

[5] Islam M.S., Sultana S., Roy U.K., Mahmud J.A., *A review on video classification with methods, findings, performance, challenges, limitations and future work*, *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, 2021, vol. 6, no 2, pp. 47–57, doi: 10.26555/jiteki.v6i2.18978.

[6] Li J., Jiang X., Sun T., Xu K., *Efficient violence detection using 3d convolutional neural networks*, [In:] *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–8, doi: 10.1109/AVSS.2019.8909883.

[7] Alfaifi R., Artoli A.M., *Human action prediction with 3d-CNN*, *SN Computer Science*, 2020, vol. 1, no 5, p. 286, doi: 10.1007/s42979-020-00293-x.

[8] Ansari M.A., Singh D.K., *ESAR, an expert shoplifting activity recognition system*, *Cybernetics and Information Technologies*, 2022, vol. 22, no 1, pp. 190–200, doi: 10.2478/cait-2022-0012.

[9] Kirichenko L., Radivilova T., Sydorenko B., Yakovlev S., *Detection of shoplifting on video using a hybrid network*, vol. 10, no 11, p. 199, ISSN 2079-3197, doi: 10.3390/computation10110199.

[10] Kirichenko L., Sydorenko B., Radivilova T., Zinchenko P., *Video surveillance shoplifting recognition based on a hybrid neural network*, [In:] *2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT)*, pp. 44–47, doi: 10.1109/CSIT56902.2022.10000545.

[11] Torrione P.A., Morton K.D., Sakaguchi R., Collins L.M., *Histograms of oriented gradients for landmine detection in ground-penetrating radar data*, *IEEE Transactions on Geoscience and Remote Sensing*, 2014, vol. 52, no 3, pp. 1539–1550, doi: 10.1109/TGRS.2013.2252016.

[12] Shah S.T.H., Xuezhi X., *Traditional and modern strategies for optical flow: an investigation*, *SN Applied Sciences*, 2021, vol. 3, no 3, p. 289, doi: 10.1007/s42452-021-04227-x.