Modern Approaches to Shoplifting Detection: Trends, Limitations, and Future Directions

Introduction

Retailing is the most common form of buying and selling the world over. The mode of retail come in many forms, ranging from small table-top trading to huge malls or hyper-markets. The larger the retail outlet, the higher the revenue potential and the larger the risk, all things being equal. Retail theft - particularly shoplifting – often remains a pervasive problem. A study by the National Association for Shoplifting Prevention indicates that one in every eleven shopper lifts; and that shoplifters are caught about one of every 48 thefts (Chemere, 2018). Hence, the accompanying losses incurred by shop owners due to this menace can be devastating.

To forestall the challenge of shoplifting, retail outlets invest heavily in security systems. Surveillance methods include human security patrols and closed-circuit television (CCTV) monitoring, rely heavily on labour and can be error prone. The activity of reviewing a recorded incident of shoplifting may not be difficult. However, the ability to detect when it occurred to allow the security personnel to review can be daunting. Moreover, the challenge of monitoring is compounded where there are so many shoppers at a given time, particularly in large retail outlets. Video surveillance cameras generate large amount of data and security personnels may not be able to effectively monitor all shopping activities in real time.

In recent times, there has been interest in the use of automatic video surveillance, powered by the application of machine learning and artificial intelligence (Gim et al.,

2020). Obviously, these systems are better at simultaneously monitoring thousands of video feeds, and detecting anomaly behaviours in real time, compared to humans. In this study, I examine modern trends in shoplifting detection and event classification, challenges and future directions.

Current Approaches and Mathematical Models

Convolutional Neural Networks for Spatial Feature Extraction

CNNs are widely employed to extract spatial features from video frames. CNNs are typically feedforward neural networks that learns features using filter optimization. A simple convolutional layer can be represented as follows:

$$y = f(Wx + b),$$

where y is the output feature map, W represents a set of learnable filters, x is the input image, b is the bias component, and $f(.)$ is a non-linear activation function such as ReLU. Some studies leverage on pre-trained models such as Inception V3, ResNet-50, VGG-16 or MobileNetV3Large and apply transfer learning to efficiently extract relevant features from footage.

In recent times, further modification to the CNN architecture can result in better appreciation of object detection and behaviour recognition. For instance, the 3D Convolutional Neural Networks (3D-CNNs) analyse video frames volumetrically by capturing both spatial features and short-term motion patterns. For a given video tensor $X \in \mathbb{R}^{T\times H\times W\times C}$, a 3D convolution operation at layer l is defined as:

$$O_{t,x,y,k} = \text{ReLU}\left(\sum_{i=0}^{d_t-1}\sum_{j=0}^{d_h-1}\sum_{m=0}^{d_w-1} W_{i,j,m,k}\cdot X_{t+i,x+j,y+m,c} + b_k\right)$$

where d_t, d_h, d_w are kernel dimensions across time, height, and with the ReLU activation function.

Recurrent Neural Networks for Temporal Dynamics

RNNs are neural networks that are designed for processing sequential or time series data. In this use case, RNNs capture temporal patterns, required for processing sequential shoplifting behaviours. It is then integrated with CNNs for feature extraction. For instance, a bidirectional LSTM (BiLSTM) processes a sequence of CNN-extracted features $\{x_t\}_{t=1}^{T}$ using forward and backward recurrences. The model can be represented as:

$$\vec{h}_t = f\left(W_x x_t + W_h \vec{h}_{t-1} + b\right),$$

$$\overleftarrow{h}_t = f\left(W_x x_t + W_h \overleftarrow{h}_{t-1} + b\right),$$

The final output is obtained by combining $\vec{h}_t$ and $\overleftarrow{h}_t$.

A typical hybrid model is the CNN-LSTM; which combines CNNs for spatial feature extraction and LSTM for temporal dependencies. In a simple representation of a sequence of N frames, the LSTM cell updates its hidden state h_t as:

$$h_t = \text{LSTM}(h_{t-1}, \text{CNN}(X_t))$$

Vision Transformers (ViTs)

Vision Transformers (ViTs) are another form of deep learning models used for detecting unusual patterns over space and time. Instead of processing the entire images of a video, ViTs split the given inputs into patches to allow the model to focus on local details while maintaining global relationships across patches. The model inherently contains a self-

attention mechanism and for each patch, it computes an attention score in the following form:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where **Queries (Q) r**epresent the patch for which the model wants to compute attention, **Keys (K) r**epresent the patches that are being compared against, and **Values (V)** carry the actual information from the patches. The activation function is softmax, which converts raw scores into a probability distribution. The model then uses these integrated features from the attention scores to detect and classify spatiotemporal anomalies.

Two-Stream Networks

Two-Stream Networks use a dual-pipeline approach to capture both what is visible (spatial features) and how things move (temporal features) in video data. The spatial stream analyzes raw RGB frames, by focusing on visual appearance, and capturing static details such as scenes, objects, and textures. The temporal stream processes optical flow, that is the motion between different sequential frames. The respective features are then combined (F_{fused}) using the weighted sum formula below:

$$F_{\text{fused}} = \alpha F_{\text{spatial}} + (1 - \alpha)F_{\text{temporal}}$$

where F_{spatial} represents feature representation from the spatial stream, F_{temporal} represents feature representation from the temporal stream, and α is a weighting factor (between 0 and 1) that controls the contribution of each stream.

Hybrid Architectures and Anomaly Detection

The combination of CNNs and RNNs in detection is becoming mainstream, as it provides superior performance. For examine, Kirichenko et al. (2022) used CNN for frame-level feature extraction, and RNN, in the form of GRU and LSTM, for temporal sequences. Also, Ansari and Singh (2022) introduce the ESAR system, made up of combined CNN and RNN, that extracts both appearance and motion features in classification. In a related study, Muneer et al. (2023) proposed the CNN-BiLSMT, a hybrid neural network, that achieved over 81% accuracy above traditional 2D and 3D CNNs.

Performance Comparison

3D-CNN: Well-suited for analyzing high-resolution CCTV footage where both spatial and temporal context is important.

CNN-LSTM: Ideal for real-time alert systems due to its ability to extract spatial features (via CNN) and learn time-based patterns (via LSTM).

Vision Transformer (ViT): Effective in complex scenarios such as occlusion, where traditional convolutional models might struggle to maintain context.

Current Trends

Recent literature emphasizes:

- Hybrid Architectures: Integrating spatial and temporal modelling (e.g., CNN–BiLSTM, CNN–GRU) to improve detection accuracy.
- Transfer Learning: Leveraging pre-trained models to mitigate the need for large, annotated datasets.

- Anomaly Detection Framing: Some studies redefine shoplifting detection as an anomaly detection problem, enabling unsupervised learning to flag deviations from typical behaviour.

Limitations

Despite promising advances, several limitations persist:

- **Data Limitation & Bias:** Many studies use small, staged, or biased datasets. Publicly available data on shoplifting are rare, due to factors such as limited annotation, data privacy and protection legislations (eg. GDPR and CCPA). Also, real-world datasets are often based on limited camera angles. Again, models trained on disproportionate datasets flag certain demographics, introducing discriminatory tendencies and bias.
- **Environmental and Operational Variability:** Different lighting conditions, concept drift, cloth or accessories patterns, occlusions and obstructions (such as shelves blocking cameras) negatively impact on the efficiency of movement detections. For instance, feature extractions of same movements across diverse lighting conditions may produce different classification of detection. Again, dense crowds and group interactions within retail space reduces the ability to effectively classify shoplifting.
- **Interpretability:** While classifications of deep networks can be known, the inner workings of these models are often black boxes; complicating the validation and legal admissibility of evidence.

- **Computational Demands:** Real-time detection requires models that are both accurate and computationally efficient, a challenging trade-off. This also means that deployment will use specialized hardware such as GPUs which impact on operational costs.

Proposed Solutions

- **Data Augmentation and New Datasets:** Collecting and publicly releasing real-world, multi-angle surveillance data can improve model generalizability.
- **Explainable AI (XAI):** Incorporating explainability tools to identify which features drive classification decisions can mitigate bias.
- **Optimized Hybrid Models:** Fine-tuning hyperparameters (e.g., batch size, learning rate) in hybrid CNN–RNN architectures improves real-time performance and accuracy.
- **Anomaly Detection Frameworks:** Using unsupervised anomaly detection techniques to learn normal shopping behaviour and flag deviations.

Discussion

The application of AI, machine learning, and deep learning for shoplifting detection has significantly improved the ability of retailers to prevent losses and enhance security measures. Traditional surveillance systems rely heavily on human monitoring, which can be inconsistent and prone to oversight. AI-driven solutions address these limitations by automating the detection process, ensuring continuous surveillance, and minimizing human error.

Machine learning models trained on diverse datasets can recognize suspicious behaviours, such as prolonged loitering, sudden movements, and unauthorized removal of items. These models utilize past incidents to refine their detection capabilities, improving accuracy over time. Additionally, deep learning techniques, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated the ability to analyse video footage with higher precision, identifying complex patterns that may indicate shoplifting attempts.

Despite these advancements, there are challenges associated with AI-powered shoplifting detection. One of the primary concerns is the potential for bias in training data. If datasets are not representative of diverse demographics and behaviours, models may produce inaccurate or unfair classifications. This issue necessitates rigorous data curation and continuous refinement of algorithms to ensure fairness and accuracy.

Privacy concerns also emerge with AI-based surveillance systems. Customers may be uncomfortable with extensive video monitoring, and there are legal and ethical considerations regarding the collection and storage of data. Retailers must balance security improvements with compliance to privacy regulations, ensuring transparency in how AI is implemented within stores.

Another challenge is the cost of deployment and maintenance. Advanced AI models require significant computational resources and infrastructure, which may be inaccessible to smaller retailers. However, cloud-based AI solutions and cost-effective machine learning frameworks can help mitigate this issue by offering scalable and affordable options.

Conclusion

AI, machine learning, and deep learning have transformed shoplifting detection by improving surveillance accuracy, automating security processes, and reducing human oversight errors. These technologies enable proactive loss prevention strategies, helping retailers safeguard inventory and maintain operational efficiency.

However, successful implementation requires addressing challenges such as bias in data, ethical considerations, privacy concerns, and financial constraints. Responsible deployment of AI-based surveillance must include diverse training datasets, clear privacy policies, and cost-effective strategies to ensure accessibility for all retail businesses.

Future research should prioritize creating better datasets, improving interpretability, and designing lightweight models for real-time deployment.

References

1. Gim, U.J.; Lee, J.J.; Kim, J.H.; Park, Y.H.; Nasridinov, A. An Automatic Shoplifting Detection from Surveillance Videos. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY USA, 7–12 2020; Apress: Berkeley, CA, USA, 2020; Volume 34, pp. 13795–13796.

2. Chemere, D.S. Real-time Shoplifting Detection from Surveillance Video. Master Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2018; p. 94.

3. Kirichenko, L.; Radivilova, T.; Sydorenko, B.; Yakovlev, S. Detection of Shoplifting on Video Using a Hybrid Network. Computation 2022, 10(11), 199.

4. Ansari, M.A.; Singh, D.K. ESAR, An Expert Shoplifting Activity Recognition System. Cybernetics and Information Technologies 2022. researchgate.net

5. Muneer, I.; Saddique, M.; Habib, Z.; Mohamed, H.G. Shoplifting Detection Using Hybrid Neural Network CNN-BiLSMT and Development of Benchmark Dataset. Applied Sciences 2023, 13(14), 8341.

6. Fawaz, H.I.; Lucas, B.; Forestier, G.; Pelletier, C.; Schmidt, D.F.; Weber, J.; Webb, G.I.; Idoumghar, L.; Muller, P.-A.; Petitjean, F. InceptionTime: Finding alexnet for Time Series classification. Data Min. Knowl. Discov. 2020, 34, 1936–1962.