AI-Based Real-Time Shoplifting Detection System Using Deep Learning and CCTV Footage

Anuj Gadekar[1], Suraj Khairnar[2], Ayush Gaikwad[3], Shantanu Game[4]
[1,2,3,4]*AVCOE Sangamner,*

Abstract- **Retail theft, specifically shoplifting, results in billions of dollars in losses annually. Traditional surveillance relies heavily on human monitoring, which is prone to fatigue and scalability issues. This paper proposes an automated, real-time shoplifting detection system utilizing a hybrid deep learning architecture. We employ the YOLOv8-Pose model to efficiently extract 17 skeletal keypoints for multi-person tracking and a Long Short-Term Memory (LSTM) network to classify temporal behavioral patterns. The system is designed to distinguish between normal shopping behavior and suspicious gestures, such as concealing items, without relying on facial recognition. Experimental results on a combined dataset of local and UCF Crime videos demonstrate an overall system accuracy of 80%, validating its feasibility as a privacy-preserving and cost-effective surveillance solution.**

***Index Terms*- Computer Vision, Deep Learning, LSTM, Shoplifting Detection, YOLOv8-Pose.**

I. INTRODUCTION

Retail industries lose billions of dollars annually due to inventory shrinkage, with shoplifting being a primary contributor [10]. While Closed-Circuit Television (CCTV) systems are widely deployed in modern stores, their effectiveness is limited by the need for continuous human supervision. A security operator can only monitor a limited number of screens simultaneously, resulting in missed incidents and delayed responses. Moreover, manual monitoring is reactive rather than proactive-theft is typically identified only during post-event audits, long after the perpetrator has left the premises.

To overcome these challenges, intelligent surveillance systems capable of automated abnormal activity recognition are essential [3, 9]. Recent breakthroughs in Deep Learning and Computer Vision have enabled machines to interpret human actions in real time with high precision [7, 8].

In this context, our work proposes a shoplifting detection framework based on action recognition rather than simple object detection. By analyzing skeletal motion patterns instead of relying on facial identity, the system tracks keypoint movements to detect suspicious concealment behaviors — such as hands moving rapidly toward pockets or bags — while ignoring benign customer interactions like browsing or product inspection [1].

II. PROPOSED METHODOLOGY

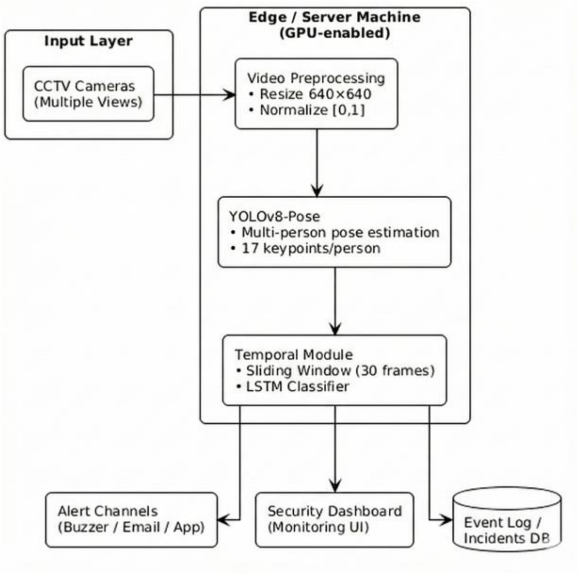
Input Layer
CCTV Cameras (Multiple Views)
Edge / Server Machine (GPU-enabled)
Video Preprocessing
• Resize 640×640
• Normalize [0,1]
YOLOv8-Pose
• Multi-person pose estimation
• 17 keypoints/person
Temporal Module
• Sliding Window (30 frames)
• LSTM Classifier
Alert Channels (Buzzer / Email / App)
Security Dashboard (Monitoring UI)
Event Log / Incidents DB

Fig. 1. System architecture of the proposed shoplifting detection system.

The proposed system architecture is designed as a sequential pipeline comprising three primary stages: Data Acquisition, Multi-Person Feature Extraction, and Temporal Sequence Classification.

A. Data Acquisition

The system accepts video input from standard CCTV surveillance cameras. To ensure compatibility with the deep learning models, the raw video feed is pre-processed in real-time. Frames are resized to a standard resolution (e.g., 640 x 640) to optimize inference speed, and pixel values are normalized to the range [0, 1].

B. Multi-Person Pose Estimation Using YOLOv8-Pose

To accurately analyze customer behavior in retail environments, we employ YOLOv8-Pose, a single-stage, bottom-up pose estimation framework. Unlike conventional top-down pipelines that rely on separate detection and pose estimation stages, YOLOv8-Pose jointly detects individuals and predicts their skeletal keypoints in a unified network. This integration significantly reduces latency and improves tracking efficiency.

The model predicts 17 anatomical keypoints following the COCO format, including the nose, eyes, shoulders, elbows, wrists, hips, knees, and ankles. These keypoints form skeleton structures that are used for subsequent behavioral analysis.

The shift to YOLOv8-Pose offers two major advantages:

- Multi-Person Tracking:
  The bottom-up architecture allows for simultaneous estimation of multiple body poses within crowded scenes while minimizing identity switching across frames.
- Occlusion Resilience:
  Even when some joints are partially hidden (e.g., due to shelving or other customers), the model leverages contextual body structure to infer missing keypoints, preserving temporal continuity.

For each detected person in a frame, the pose representation is defined as a feature vector:

$$V = \{(x_i, y_i, c_i)\}_{i=1}^{17}$$

where (x_i, y_i)denote the normalized pixel coordinates of the i^{th}keypoint and $c_i$is the corresponding confidence score. To ensure robustness, keypoints with $c_i < 0.5$are discarded, preventing noisy or unreliable detections from propagating into the downstream classification module.

TABLE I Pose Keypoints Dataset Table

Frame ID	Nose (x0, y0)	L. Eye (x1, y1)	R. Eye (x2, y2)
0	(0.56, 0.38)	(0.57, 0.36)	(0.54, 0.36)
1	(0.56, 0.37)	(0.58, 0.35)	(0.55, 0.35)
2	(0.57, 0.37)	(0.58, 0.35)	(0.55, 0.35)

As shown in Table I, the raw video frames are converted into numerical vectors. Each row represents a single frame containing the normalized coordinates of the 17 body landmarks.

Algorithm 1: Suspicious Activity Detection
Input:Video Stream V
Output: Alert Signal A

```
Initialize YOLOv8-Pose model Y
Initialize LSTM model M
Initialize buffer B ← []
while V is active do
    Frame_t ← Read(V)
    Keypoints_t ← Y(Frame_t)
    if Keypoints_t is detected then
        Norm_Keypoints ← Normalize(Keypoints_t)
        Append Norm_Keypoints to B
        if Size(B) == 30 then
            Prediction ← M(B)
            if Prediction > Threshold then
                Trigger alert A
                Log timestamp to database
            end if
            Remove oldest entry in B
        end if
    end if
end while
```

This sliding window architecture directly corresponds to the 30-frame buffer used in Algorithm 1, enabling real-time suspicious activity assessment with minimal delay.

C. Sequence Classification (LSTM)

Shoplifting is a temporal activity that cannot be recognized from a single static frame. For example, a person holding a product is harmless, but the sequential motion of placing that product inside a pocket is suspicious. To effectively capture this temporal context, we employ a Long Short-Term Memory (LSTM) network.

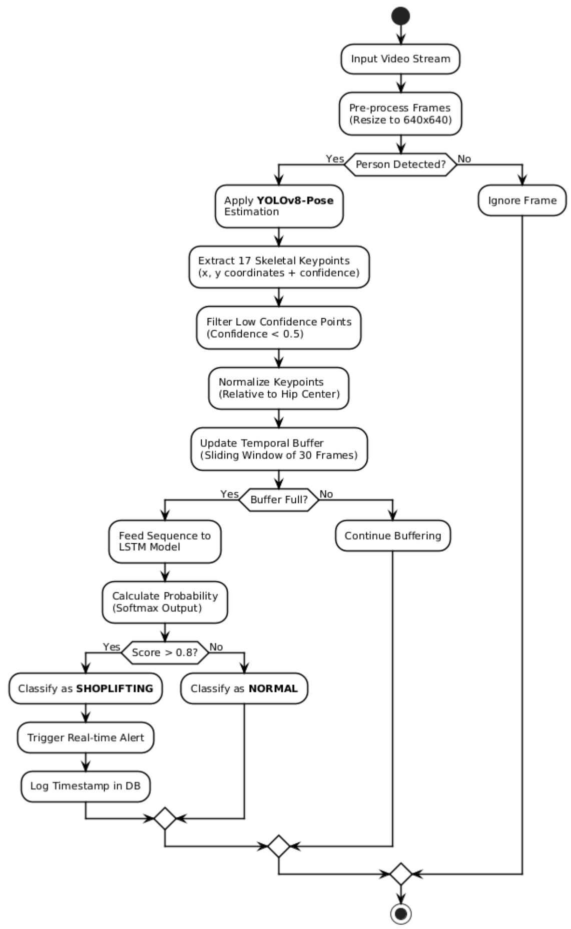
Input Video Stream
Pre-process Frames (Resize to 640x640)
Yes
Person Detected?
No
Apply **YOLOv8-Pose** Estimation
Ignore Frame
Extract 17 Skeletal Keypoints (x, y coordinates + confidence)
Filter Low Confidence Points (Confidence < 0.5)
Normalize Keypoints (Relative to Hip Center)
Update Temporal Buffer (Sliding Window of 30 Frames)
Yes
Buffer Full?
No
Feed Sequence to LSTM Model
Continue Buffering
Calculate Probability (Softmax Output)
Yes
Score > 0.8?
No
Classify as **SHOPLIFTING**
Classify as **NORMAL**
Trigger Real-time Alert
Log Timestamp in DB

Fig. 2. Activity recognition architecture.

III. LSTM-BASED SEQUENCE MODELING

The core of our classification relies on the LSTM cell's ability to maintain a cell state. The forget gate f_t decides what information to discard from the cell state. It is calculated as:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \qquad \textit{(1)}$$

Where $\sigma(\cdot)$denotes the sigmoid activation function, $W_f$represents the learned weight matrix, $h_{t-1}$is the previous hidden state, and $x_t$corresponds to the input vector at time step t. In our YOLOv8-Pose–based implementation, $x_t$is constructed as a flattened feature vector of size 1×34, formed by concatenating the (x, y)coordinate pairs of all 17 detected skeletal keypoints in a frame. This vectorized representation enables the LSTM to learn temporal pose dynamics across consecutive frames.

The output of the network is passed through a Softmax activation function to obtain probability distributions for the classes (Normal vs. Shoplifting):

$$P(y = j \mid x) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \qquad (2)$$

Where z is the output vector from the LSTM and K is the number of classes.

IV. EXPERIMENTAL RESULTS

The system was evaluated on a hybrid dataset consisting of 600 locally captured clips in a simulated retail environment and 400 anonymized videos from the public UCF Crime dataset. The data was split into 70% training, 20% validation, and 10% testing.

A. Training Performance The LSTM network was trained for 50 epochs using the Adam optimizer with a learning rate of 0.001. We utilized the Categorical Cross-Entropy loss function. The model achieved a training accuracy of 94.5% and a validation accuracy of 91.2%, demonstrating strong generalization capabilities without significant overfitting.

B. Person Detection & Pose Estimation (YOLOv8) The initial stage utilizes YOLOv8-Pose to identify individuals and extract skeletal features. As shown in Fig. 3, the model achieves a Mean Average Precision (mAP) of 0.968 at an Intersection over Union (IoU) threshold of 0.5. This high precision is critical, as accurate 17-point skeletal data is the foundation for the subsequent behavioral analysis.

The YOLOv8-Pose model was used in a transfer learning setup. We initialized the network with COCO-pretrained weights and performed lightweight fine-tuning on our simulated retail environment footage for 30 epochs to improve landmark detection under varying camera angles and lighting conditions.

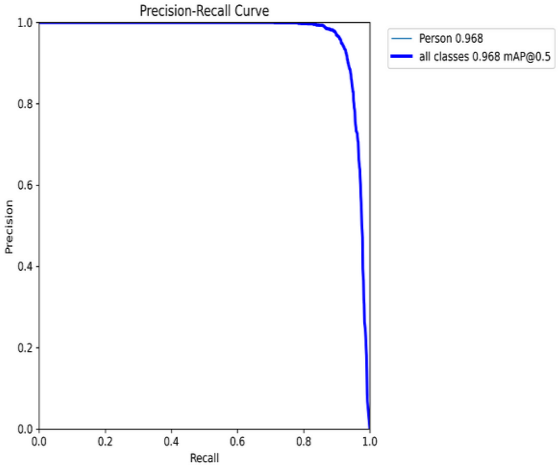

Fig. 3. Precision-Recall curve for Person Detection, achieving 0.968 mAP.

C. System Accuracy & Real-Time Testing During live testing on a dataset of 100 sample clips, the system achieved an Accuracy of 80.00%. The model demonstrated a Recall of 0.90, indicating a strong ability to detect theft attempts when they occur (45 correct detections out of 50 actual thefts). The Precision of 0.75 suggests that while the system is highly sensitive, it occasionally flags benign rapid movements as suspicious. The F1 Score of 0.8182 confirms a balanced performance between precision and recall, validating the system's effectiveness for real-time surveillance. The model operates in real-time at approximately 25 FPS on the hardware listed in Table II, ensuring practical deployment feasibility in live surveillance environments.

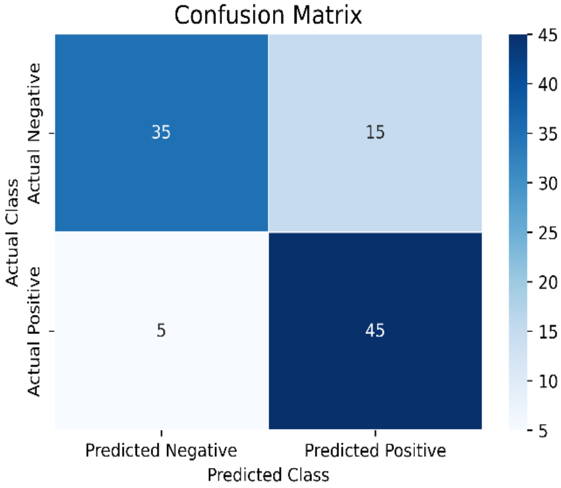

Fig. 4 Confusion Matrix

Metric	Value
Accuracy	0.8000
Precision	0.7500
Recall	0.9000
F1 Score	0.8182

The system achieved an overall accuracy of 80%, which indicates promising initial feasibility for automated surveillance. However, the analysis of false positives highlights a critical need for improved gesture disambiguation, particularly in dense shopping environments where occlusions are frequent.

D. Dataset & Ethical Considerations To ensure robustness, our dataset was curated from diverse sources. It includes 600 locally captured clips recorded in a simulated retail environment with varying lighting conditions and camera angles. Additionally, we incorporated 400 anonymized videos from the public UCF Crime dataset to introduce real-world variability. To strictly adhere to privacy and ethical standards, all facial identities were blurred, and personal details were excluded from the training data.

TABLE II HARDWARE AND SOFTWARE SPECIFICATIONS

Component	Specification
Processor (CPU)	Intel Core i7-10750H (2.60 GHz)
Graphics (GPU)	NVIDIA GeForce RTX 3060 (6GB)
RAM	16 GB DDR4
Operating System	Windows 11 Pro
Programming Language	Python 3.9
Deep Learning Framework	TensorFlow 2.10, Keras
Pose Estimation Lib	YOLOv8

TABLE III COMPARISON WITH EXISTING METHODS

Feature	Traditional CCTV	3D CNN (C3D)	Proposed (Pose+LSTM)
Privacy	Low (Face visible)	Low (RGB input)	High (Skeleton only)
Comp-utation	Manual	High GPU usage	Moderate
Real-time?	No	No	Yes (25+ FPS)
Accuracy	Human-dependent	~79% [7]	~80%

**Note: The C3D result corresponds to performance reported in [7] on its dataset and is not a direct experiment on our dataset.*

V. LIMITATIONS AND CHALLENGES

While the proposed system demonstrates high accuracy in controlled environments, several challenges remain for real-world deployment:

1. Occlusion: If a customer is standing behind a shelf or another person, the camera may lose sight of key body joints (e.g., the wrist). Although YOLOv8 attempts to infer missing points, severe occlusion can still cause temporary tracking failures.
2. Crowd Density: While YOLOv8 significantly improves multi-person tracking compared to older methods, extremely dense crowds in peak hours may still result in "ID Switching," where the system confuses the identity of two overlapping individuals.
3. Ambiguous Gestures: Certain benign actions, such as putting a mobile phone into a pocket or retrieving a wallet for payment, share kinematic similarities with shoplifting. Distinguishing these requires finer-grained finger tracking, which is currently computationally expensive for real-time edge devices.

VI. CONCLUSION & FUTURE WORK

This paper successfully demonstrates the feasibility of an automated, privacy-preserving shoplifting detection system. By integrating YOLOv8-Pose for skeletal feature extraction with an LSTM network for temporal sequence analysis, the proposed model effectively identifies suspicious concealment gestures in real-time. Experimental results indicate an overall accuracy of 80% in simulated retail environments, validating the system's potential to augment traditional surveillance methods. Notably, the reliance on skeletal data rather than facial recognition ensures compliance with ethical privacy standards.

Future research will focus on addressing the limitations of occlusion in crowded scenarios. We aim to implement multi-camera re-identification to track subjects across overlapping fields of view. Additionally, future iterations will integrate a real-time email alert system to instantly notify security personnel of suspicious incidents. We also prioritize integrating object interaction analysis—distinguishing between personal items (e.g., phones) and merchandise-to further minimize false positives and enhance robust deployment. Overall, the proposed method demonstrates the ability to assist existing surveillance infrastructure by providing automated alerts for suspicious concealment actions, reducing reliance on manual monitoring. The system serves as a strong foundation for deployment in small to medium-scale retail stores.

REFERENCES

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no.1, pp. 172186, Jan.2021.doi:10.1109/TPAMI.2019.2929257.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735-1780, Nov.1997. doi:10.1162/ neco. 1997 .9.8.173.

[3] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 6479–6488. doi: 10.1109/ CVPR. 2018.00678.

[4] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," GitHub repository, 2023. Accessed: Dec. 4,2025. [Online]. Available: https://github. com/ultralytics/ultralytics.

[5] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in Proc. IEEE Int. Conf. Comput. Vis.(ICCV), Venice, Italy, Oct. 2017, pp. 2640–2649. doi: 10.1109/ICCV.2017.287.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[7] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, Jan. 2013. doi: 10.1109/TPAMI.2012.59.

[8] K. Simonyan and A. Zisserman,"Two-stream convolutional networks for action recognition in videos," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Montreal, QC, Canada, Dec. 2014, pp. 568–576. doi: 10.5555/2968826.2968890.

[9] T.-H. Cheng, Y. Huang, C.-W. Wang, and S.-Y. Chien, "Video anomaly detection with spatio-temporal dissociation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, Jun. 2022, pp. 14035–14044. doi: 10.1109/CVPR52688.2022.01364.

[10] A. B. M. Musa, M. S. Uddin, and M. A. A. Wadud, "Suspicious activity recognition for shoplifting detection," Int. J. Image, Graph. Signal Process., vol. 8, no. 1, pp. 1–9, 2016. doi: 10.5815/ijigsp.2016.01.01.

[11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, May 2015, pp. 1–15. doi: 10.48550/ arXiv. 1412.6980.