# Person Re-Identification and Crime Scene Classification System

## Technical Write-Up

## 1. Overview and Approach

This system implements a multi-stage pipeline for person re-identification (Re-ID) across video clips and automated crime scene classification. The approach combines computer vision, deep learning, and behavioral heuristics to identify individuals and detect suspicious activities.

### System Architecture

The pipeline consists of two main components:

Part A: Person Re-Identification

1. Detection & Tracking: YOLO11n detects and tracks persons frame-by-frame within each clip
2. Feature Extraction: ResNet50 (pre-trained on ImageNet) extracts 2048-dimensional appearance embeddings from person crops
3. Tracklet Aggregation: Multiple detections of the same person within a clip are aggregated using quality-weighted averaging (50% confidence, 50% bounding box area)
4. Cross-Clip Clustering: Agglomerative clustering with cosine distance (threshold=0.2) groups tracklets across clips to identify unique individuals
5. Visualization: Annotated videos display global person IDs with consistent color coding

Part B: Scene Classification

1. Multi-Model Detection: Three specialized YOLO models run in parallel:
   * General object detection (persons, bags, items)
   * Weapon detection (knives, guns, etc.)
   * Violence detection (fights, aggressive behavior)
2. Behavioral Analysis: State-machine tracking monitors person-item interactions:
   * IDLE → PICKED\_ITEM → BAGGED\_ITEM → (exit without checkout) → CRIME
3. Heuristic Classification: Multiple detection strategies:
   * Object-based: Presence of weapons (conf > 0.45) or violence (conf > 0.5)
   * Behavior-based: Theft patterns (item pickup, bagging, suspicious exit)
   * Spatial reasoning: Checkout zone detection and exit monitoring

## 2. Key Technical Decisions

### Re-ID Pipeline Choices

Why ResNet50 for embeddings?

* Pre-trained on ImageNet provides strong visual features without custom Re-ID training
* 2048-dimensional embeddings offer good discriminative power
* Fast inference (~5ms per crop on GPU)
* Trade-off: Domain-specific Re-ID models (trained on person datasets like Market-1501) would improve accuracy but require additional training data

Why Agglomerative Clustering?

* No need to pre-specify number of persons (distance threshold determines clusters)
* Cosine distance naturally suits normalized embeddings
* Average linkage balances within-cluster compactness and between-cluster separation
* Trade-off: Fixed threshold (0.2) may not generalize across all scenarios; adaptive thresholding would be better

Quality-Weighted Aggregation:

quality\_scores = 0.5 \* normalized\_confidence + 0.5 \* normalized\_area

Rationale: Larger, more confident detections are more reliable for Re-ID. Equal weighting is a simple starting point; could be tuned with validation data.

### Scene Classification Choices

Multi-Model Approach:  
Using three specialized models rather than one general model provides:

* Higher recall (each model targets specific threat types)
* Interpretable detections (clear which model triggered alert)
* Flexibility (can adjust thresholds per model independently)

Confidence Thresholds:

* Weapons: 0.45 (moderate - balance false positives/negatives)
* Violence: 0.5 (stricter - complex to detect accurately)
* These are heuristic values that should be validated on test data

State Machine for Theft:  
Tracks complete behavioral sequence rather than isolated events, reducing false positives from normal shopping behavior (e.g., picking up item to examine).

## 3. Assumptions and Limitations

### Assumptions

1. Video Quality: Reasonable resolution (≥720p), lighting, and camera angles
2. Camera Setup: Relatively stationary cameras with persons visible for multiple frames
3. Appearance Consistency: Persons don't change clothing between clips (same time period)
4. Scene Type: Retail/public spaces with distinguishable checkout areas
5. Detection Models: Pre-trained YOLO models are appropriate for the domain
6. Temporal Scope: All clips are from similar time periods (same day/event)

### Known Limitations

Re-Identification:

* Occlusion: Partially occluded persons may generate poor embeddings
* Pose Variation: Extreme poses (lying down, crouching, sideways, on your back) may hinder matching
* Appearance Changes: Costume changes, adding/removing jackets breaks tracking
* Crowded Scenes: Overlapping persons cause bbox errors and ID switches
* Fixed Threshold: Single clustering threshold (0.2) may over-segment or under-segment
* No Temporal Reasoning: Doesn't use temporal proximity to improve matching

Scene Classification:

* Checkout Zone Detection: Simple heuristic (80% of frame width) may fail in complex layouts
* False Positives: Normal behaviors (trying on items, examining products) may trigger alerts
* False Negatives: Concealed weapons, subtle violence may be missed
* Model Generalization: Custom models trained on specific datasets may not transfer well
* No Context Understanding: Cannot distinguish staged demonstrations from real crimes
* Single-Frame Decisions: Some heuristics (weapon detection) don't consider temporal consistency

System-Wide:

* Computational Cost: Processing 3 YOLO models per frame is expensive (~50ms/frame on GPU)
* Memory Usage: Loading all person data into memory limits scalability
* No Uncertainty Quantification: Binary crime/normal classification with no confidence calibration
* Hard-Coded Parameters: Many thresholds (proximity=50px, min\_frames=5) lack principled tuning

Common-Failure Modes:

* False alarms and missed detections: Imperfect person detection produces ghost tracklets and also fails to detect real people, which pollutes the feature space and contaminates the identity catalogue.
* Track fragmentation and identity splitting: Continuous appearances get broken into multiple short tracklets, causing a single person to be represented by several catalogue entries.
* Identity merging: Distinct people are incorrectly grouped as the same individual, creating false positive matches.
* Identity splitting: One person is assigned multiple global identities across clips, creating false negative matches.
* Clustering failures: Noisy or insufficient embeddings lead to over-clustering or under-clustering, driving both merging and splitting errors.
* Lack of contextual reasoning: The system relies on visual cues alone and cannot infer intent or context, so ambiguous interactions are misclassified based only on proximity or brief contact.

## 4. Future Improvements

### Short-Term Enhancements (1-2 weeks)

Re-ID Improvements:

1. Domain-Specific Model: Fine-tune ResNet or use OSNet/BoT trained on person Re-ID datasets
2. Temporal Smoothing: Use Kalman filtering or optical flow to improve tracklet consistency
3. Adaptive Thresholding: Tune clustering threshold per scene based on embedding distribution
4. Multi-Scale Features: Combine embeddings from multiple crops (full body, upper body, lower body)

Classification Improvements:

1. Temporal Integration: Require sustained detections (e.g., weapon visible for 2+ seconds)
2. Calibrated Confidence: Train a meta-classifier to output calibrated probabilities
3. Action Recognition: Add temporal action models (I3D, SlowFast) for violence detection
4. Scene Understanding: Use scene segmentation to better identify checkout zones

### Medium-Term Enhancements (1-2 months)

1. Active Learning: Collect hard examples where system fails, retrain models
2. Multi-Camera Fusion: When available, use multiple views to improve Re-ID and reduce occlusion
3. Anomaly Detection: Learn normal behavior patterns per scene, flag deviations
4. Tracking Improvements: Replace YOLO tracking with DeepSORT or ByteTrack for better ID consistency
5. Real-Time Optimization: Model quantization, TensorRT compilation for deployment

### Long-Term Research Directions (3+ months)

1. End-to-End Learning: Joint training of detection, Re-ID, and classification modules
2. Weakly-Supervised Learning: Reduce annotation burden by learning from video-level labels
3. Explainable AI: Generate natural language explanations for crime classifications
4. Few-Shot Learning: Quickly adapt to new crime types or environments with minimal data
5. Privacy Preservation: Implement federated learning or differential privacy for sensitive applications

## 5. Validation Strategy

With more time, rigorous evaluation would include:

Re-ID Evaluation:

* Metrics: CMC curves (Rank-1, Rank-5 accuracy), mAP (mean Average Precision)
* Datasets: Test on standard benchmarks (Market-1501, DukeMTMC) and custom dataset
* Error Analysis: Characterize failure modes by occlusion level, pose, lighting

Classification Evaluation:

* Metrics: Precision, Recall, F1-score, ROC-AUC per crime type
* Threshold Tuning: Grid search or Bayesian optimization for confidence thresholds
* Confusion Matrix: Understand misclassification patterns (what normal scenes trigger false alarms?)
* Temporal Consistency: Measure frame-level vs. clip-level accuracy

User Study:

* Security professionals review system outputs for false positive rate acceptability
* Measure time-to-detection compared to manual monitoring

## 6. Deployment Considerations

For production deployment:

* Streaming Architecture: Process video in real-time with low latency (<1s)
* Scalability: Distribute processing across multiple GPUs/servers
* Alert Management: Integrate with security systems, prioritize alerts by severity
* Privacy: Anonymize stored data, comply with surveillance regulations (GDPR, local laws)
* Human-in-the-Loop: Final decisions made by trained personnel, system serves as first-pass filter
* Continuous Monitoring: Track performance degradation, retrain periodically

## Conclusion

This system demonstrates a practical approach to person Re-ID and crime detection using off-the-shelf models and heuristic reasoning. While effective as a proof-of-concept, production deployment would require domain-specific model training, extensive validation, and integration with existing security infrastructure. The modular design allows for incremental improvements to each component independently.