

Person Re-Identification & Scene Classification: Approach & Limits

1 Overview

Cross-clip person re-ID and scene classification for surveillance videos via detection, tracking, embeddings, and action recognition.

2 A. Person Identity Catalogue

2.1 Detection & Tracking

- **Detection:** YOLOv11m (persons).
- **ReID:** OSNet (osnet_a1_x1_0) on person boxes.
- **Tracking:** ByteTrack (IoU-only) per clip \rightarrow frame-level tracklets (local IDs).

2.2 Cross-Clip Re-Identification

The core challenge is associating tracklets across different clips to assign global person IDs.

Tracklet Representation: Each tracklet (clip_id, track_id pair) is represented by a single embedding computed as the L2-normalized mean (or median) of all detection embeddings within that tracklet.

Distance Metric: Cosine distance between normalized tracklet embeddings measures visual similarity.

Constraint-Based Clustering: A custom greedy clustering algorithm merges tracklets across different clips.

ID Assignment: Clusters of size $\geq \text{min_cluster_size}$ (default 2) represent the same person across clips and receive a shared global ID. Singleton tracklets (noise) each receive unique IDs.

3 B. Scene Classification

3.1 Action Recognition

- **Model:** VideoMAE-base fine-tuned on Kinetics-400.
- **Windowing:** Non-overlapping 16-frame segments.
- **Crime keywords:** *fighting, punch, kick, shooting, robbery, stealing, assault*, etc.

3.2 Logic

Each 16-frame segment is classified independently. If any segment matches a crime keyword, the clip is labeled *crime* with the max crime confidence; otherwise *normal*. Output includes detected frame ranges, action labels, and confidences.

4 Limitations & Assumptions

4.1 Person Re-ID

1. **Single-appearance:** Assumes one tracklet per person per clip; split tracks become distinct IDs.
2. **No temporal reasoning:** Ignores timestamps/motion patterns.
3. **Embedding quality:** Degrades with occlusion, extreme poses, blur, poor lighting.

4.2 Scene Classification: Current Limitations

1. **Model-task mismatch 1:** The action-recognition network is ill-suited to this domain, yielding unreliable clip-level labels.
2. **Model-task mismatch 2:** The model does not provide justification for its predictions (e.g., saliency/attention maps or exemplar frames), reducing interpretability and trust.

5 Future Improvements

- **Tracker alternatives:** The current code uses ByteTrack with suboptimal settings, resulting in fragmented tracklets. There is room for improvement through parameter fine-tuning and, if needed, adopting a tracker that incorporates re-ID features.
- **Better tracklet embeddings:** Replace mean/median pooling with quality-weighted or attention pooling over frames; down-weight blurred/occluded detections using sharpness, bbox size, and detector confidence.
- **Fine-tune detection & re-ID models:** Given annotated data, fine-tune the person detector and re-ID backbone on domain-specific samples to improve accuracy for challenging cases (occlusions, lighting, viewpoints).
- **Temporal & geometric constraints:** Requires timestamps & camera calibration.
- **Vector DB at scale:** The current prototype stores embedding vectors in FiftyOne’s MongoDB instance. A more robust approach is to use a dedicated vector database (e.g., Milvus) and run the similarity operations there.
- **Video action model:** The current video action recognition approach is not well suited to this domain; invest time in identifying and evaluating a more appropriate model.