Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Mathematical Informatics (Kazushi Ikeda (Professor))		
Student ID	2411422	Submission date	2026 / 7 / 22
Name	RATURI HIMANSHU		
Thesis title	Towards Real-Time Tracking of Individuals in Crowded Environments via Cross-Modal Attention Fusion and Precision-Gated Association Cascades		

Abstract

Abstract

This research presents a robust, modular computer vision framework designed to enhance multi-object tracking (MOT) and movement analysis in hyper-congested environments. Traditional pipelines regularly experience identity fragmentation in dense crowds due to severe occlusions, target-to-target visual ambiguity, and dynamic camera ego-motion. To address these bottlenecks, we introduce a unified Cross-Modal Attention Fusion pipeline that localizes targets using a fine-tuned YOLOv10x detector and extracts detailed 28-dimensional structural body topologies via an optimized RTMPose-M architecture. Concurrently, deep visual features are isolated using an EfficientNet-B3 backbone with Generalized Mean (GeM) Pooling, while a Bidirectional LSTM models a 16-frame history of higher-order motion derivatives into a 256-dimensional semantic motion token.

Instead of naive feature concatenation, these normalized streams are unified using a 3-layer, 8-head Fusion Transformer Encoder to produce a 512-dimensional Fused Identity Embedding. Data association is governed by a multi-stage matching cascade coupled with an optimized track resurrection buffer that preserves identity clusters for up to 100 frames (dynamically extended to 140 frames under high confidence).

Evaluated on the challenging MOT20 benchmark, the system achieves a strong overall MOTA of 69.86% and an IDF1 score of 69.76%. Crucially, our precision-gated mechanism yields exceptional noise suppression, achieving a False Positive profile of just 10,583-a 36.37% improvement over StrongSORT. Completely bypassing close-range facial recognition, the architecture maps targets exclusively to non-invertible spatial-kinematic embeddings, ensuring strict privacy alignment. Future work details a density-aware adaptive thresholding mechanism and a transition from bounding boxes to instance mask segmentation to ensure absolute token purity.