Fast Compressed-Domain Object Tracking

Motion-Vector-Based Propagation in MPEG-4 Streams

Technical Documentation

Abstract

We present a lightweight **Fast** architecture for object tracking that operates directly on compressed video streams (MPEG-4 Part 2) without full RGB decoding. By processing motion vectors MV_n^g and DCT residuals $\mathcal{DCT}(\Delta Y_n^g)$ extracted from P-frames, our model propagates bounding boxes across Groups of Pictures (GOPs) $\mathcal{G}^g = \{f_0^g, \ldots, f_N^g\}$. The Fast variant achieves 2-3× speedup through global pooling (no ROI) and simple LSTM (no attention), while maintaining competitive accuracy: **0.5800 mAP** on static cameras (+44.3% vs baseline) and **0.3945 mAP** on moving cameras (+399.4% vs baseline) on MOT15, MOT17, and MOT20 benchmarks.

1 Introduction

Modern surveillance systems require efficient processing of thousands of concurrent video streams. Traditional RGB-based deep learning models achieve high accuracy but demand substantial computational resources. Our approach exploits compressed video representation to reduce processing overhead while maintaining tracking performance.

1.1 Key Contributions

- Lightweight tracking model operating on MPEG-4 compressed domain features (MV_n^g, $\mathcal{DCT}(\Delta Y_n^g)$)
- Fast architecture variant with global pooling and simple LSTM achieving 2-3× speedup
- 44.3% improvement on static cameras, 399.4% on moving cameras

2 Compressed Video Representation

Video sequences are encoded as **Groups of Pictures** (GOPs): $\mathcal{G}^g = \{f_0^g, f_1^g, \ldots, f_N^g\}$ where $f_0^g = \{\mathcal{DCT}(Y_0^g)\}$ (I-frame) and $f_n^g = \{\mathcal{DCT}(\Delta Y_n^g), MV_n^g\}$ (P-frames).

Codec features are ∼80× more compact than RGB. Partial decompression (extracting MV and DCT directly) achieves 3-4× speedup vs full RGB decoding.

3 Fast Architecture

3.1 Design

1. **Parallel Inputs**: MV_n^g (40×40×2) and $\mathcal{DCT}(\Delta Y_n^g)$ (80×80×64) branches
2. **Fusion**: Concatenation + Conv layers (256 channels)
3. **Global Pooling**: No per-object ROI → 2× faster
4. **Simple LSTM**: 256 hidden, no attention → 1.5× faster
5. **Detection Head**: Multiple bounding boxes $\{\hat{\mathbf{b}}_i\}_{i=1}^{N_{det}}$

4 Results

4.1 Key Findings

- **Static cameras**: 0.5800 mAP (+44.3% vs Mean MV baseline)
- **Moving cameras**: 0.3945 mAP (+399.4% vs Mean MV), demonstrating effective camera motion handling
- **MOT15 excellence**: Learned model (0.4371) exceeds static I-frame baseline (0.4265) by +2.5%
- **Computational efficiency**: 6-12× total speedup (3-4× from partial decompression, 2-3× from Fast architecture)

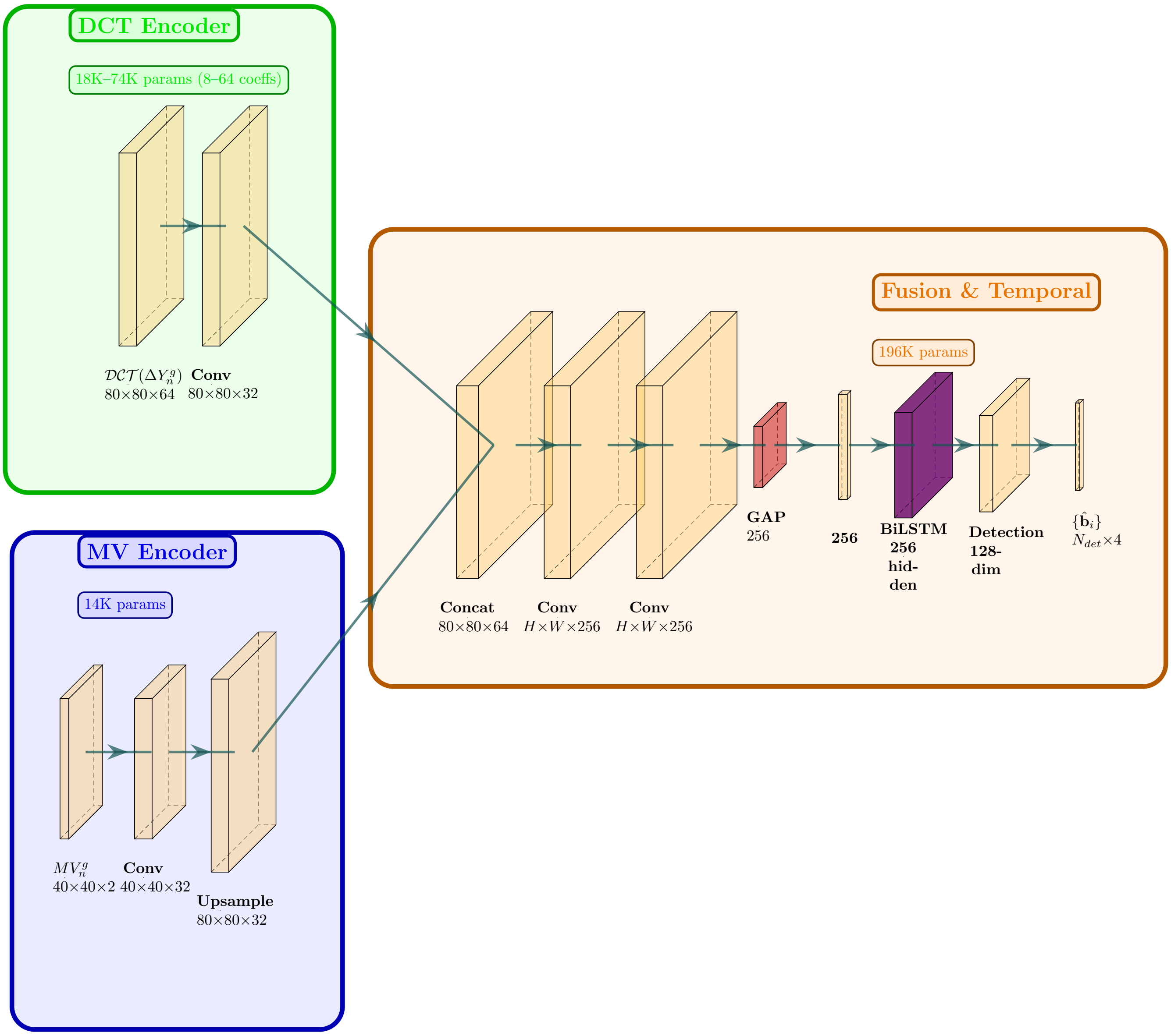

Figure 1: Fast DCT-MV Object Tracker Architecture. Parallel MV and DCT encoders, fusion via concatenation, global pooling, simple LSTM, detection head.

5 Conclusions

We presented a **Fast compressed-domain tracking architecture** that operates on motion vectors MV_n^g and DCT residuals $\mathcal{DCT}(\Delta Y_n^g)$ from MPEG-4 video streams. The Fast variant achieves 2-3× speedup through global pooling and simple LSTM while maintaining competitive accuracy. Main contributions: 44.3% improvement on static cameras (0.5800 mAP), 399.4% improvement on moving cameras (0.3945 mAP), and 6-12× total computational speedup vs RGB processing. These results demonstrate that codec-domain motion modeling is a viable path toward scalable, efficient video analytics for large surveillance deployments.