

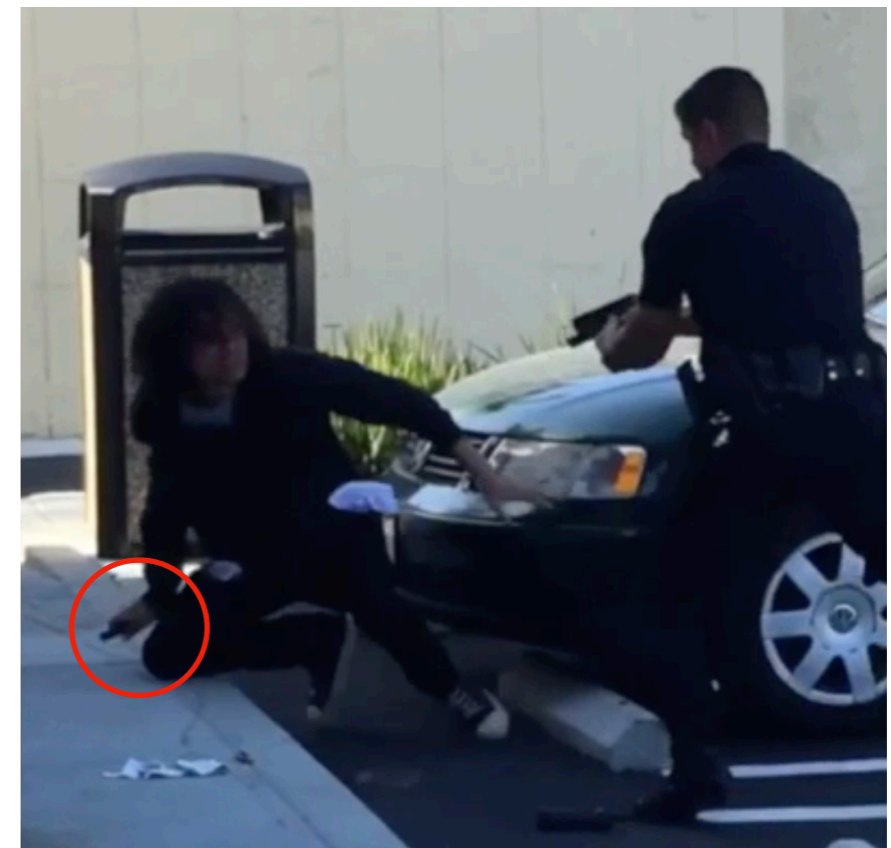
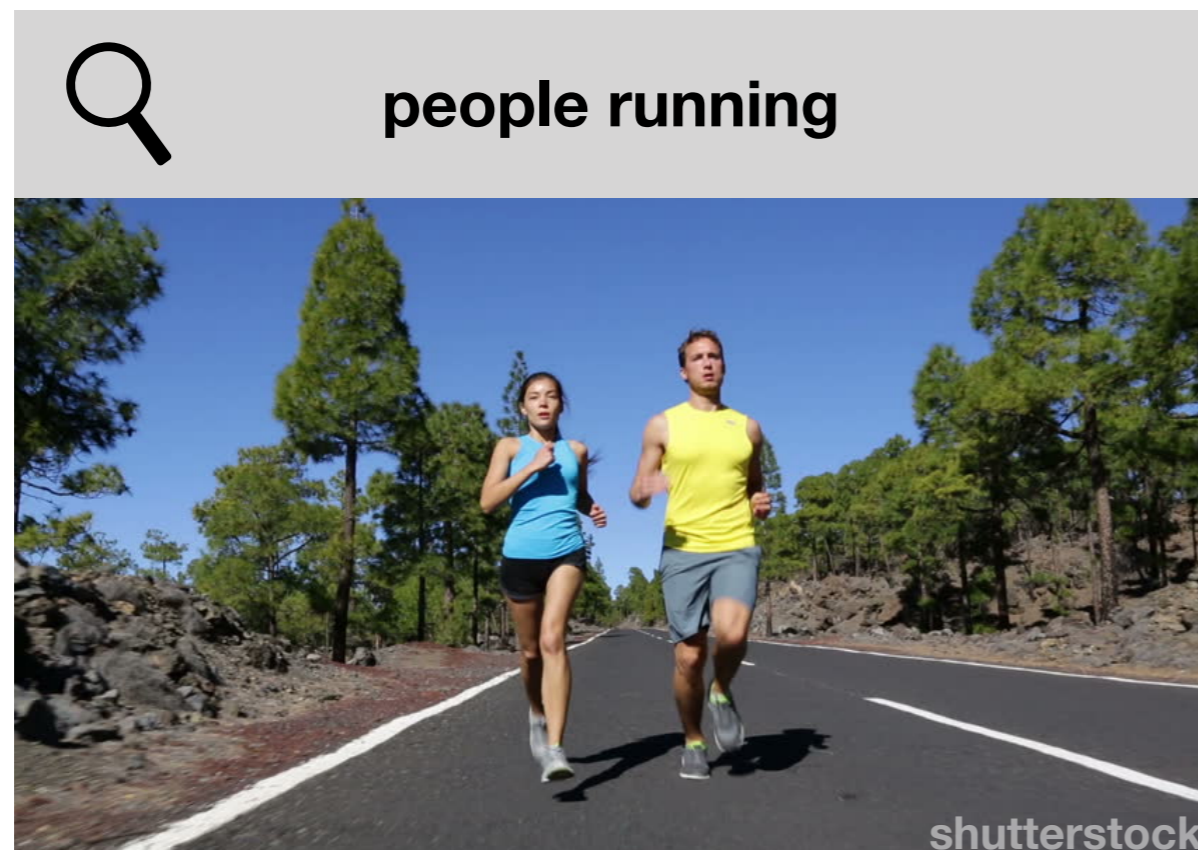
Action Recognition with Improved Trajectories

Heng Wang and Cordelia Schmid
LEAR, INRIA, France

IEEE ICCV 2013

The Problem

- How can we recognize actions in video?
- Applications include gesture recognition, threat detection, media indexing and querying, etc.



Past Approaches

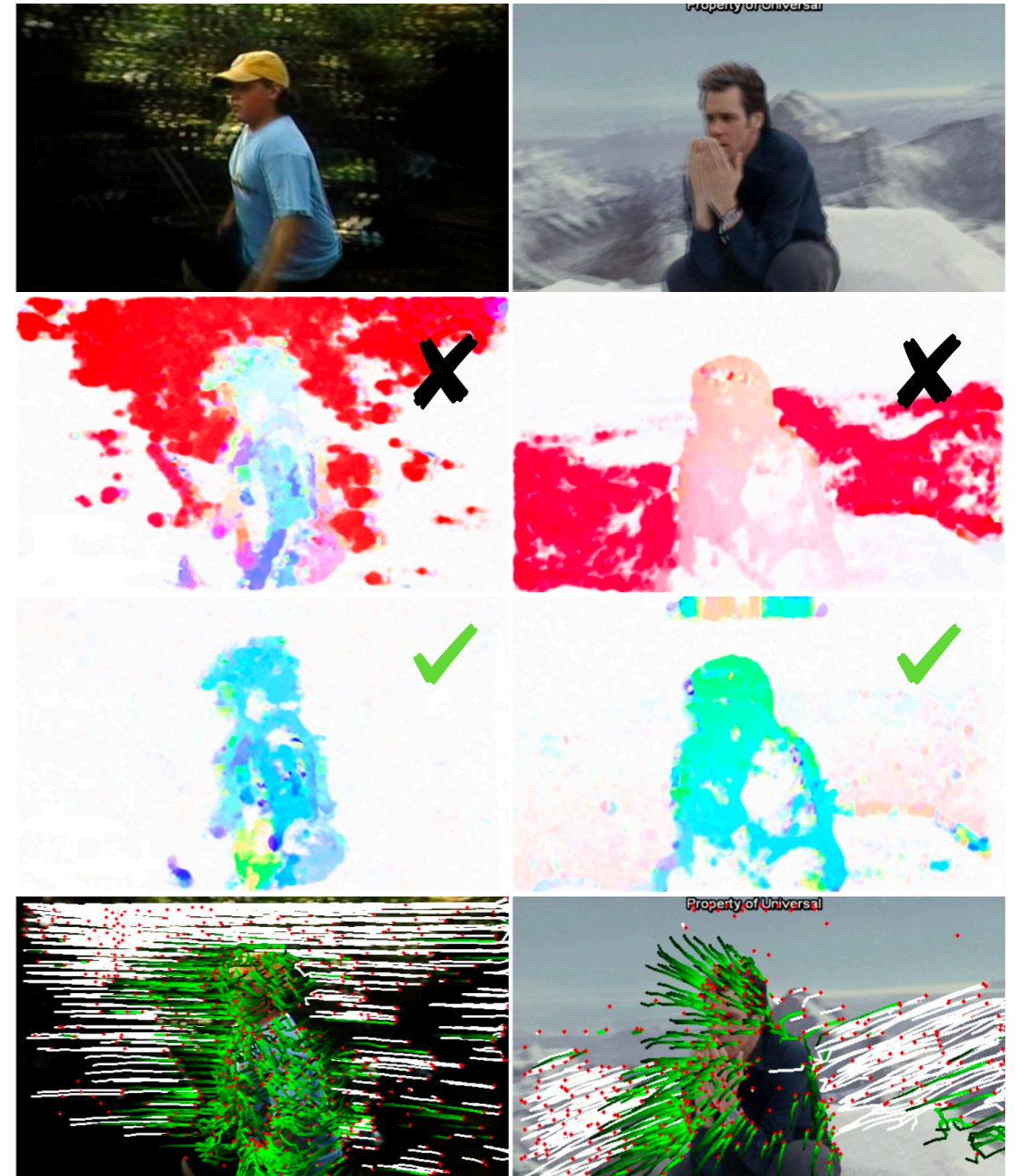
- Image segmentation to separate background and estimate camera motion
- Stabilization using coarse optical flow
- Saliency mapping
- Dense trajectory clustering

Agenda

- The Problem and Past Approaches
- Improved Trajectories
- Experimental Setup
- Results
- Concluding Remarks and Discussion

Action Recognition with Improved Trajectories

- Explicit camera motion estimation
- Corrects optical flow, prunes background
- Leads to better motion descriptor performance



Improved Trajectories



Pipeline Overview

- For consecutive frames:
 - Extract SURF descriptors with nearest-neighbor matching
 - Estimate optical flow, sample by thresholding smallest autocorrelation matrix λ s (optimal sampling for tracking) [35]
- Estimate homography using RANSAC
- Remove camera-induced displacement via thresholding

Features

- SURF works great for detecting blob-like structures
 - (Speeded *[sic]* Up Robust Features)
 - Much faster than SIFT
 - Patented
- Optical flow w/ good-features-to-track [35] great for detecting large gradients (i.e., corners and edges)

Polynomial Expansion Optical Flow Estimation [8]

- Gunnar Farnebäck, 2003
- Estimate displacement \mathbf{d} by modeling pixel neighborhood as a quadratic polynomial

$$f(\mathbf{x}) \sim \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c$$

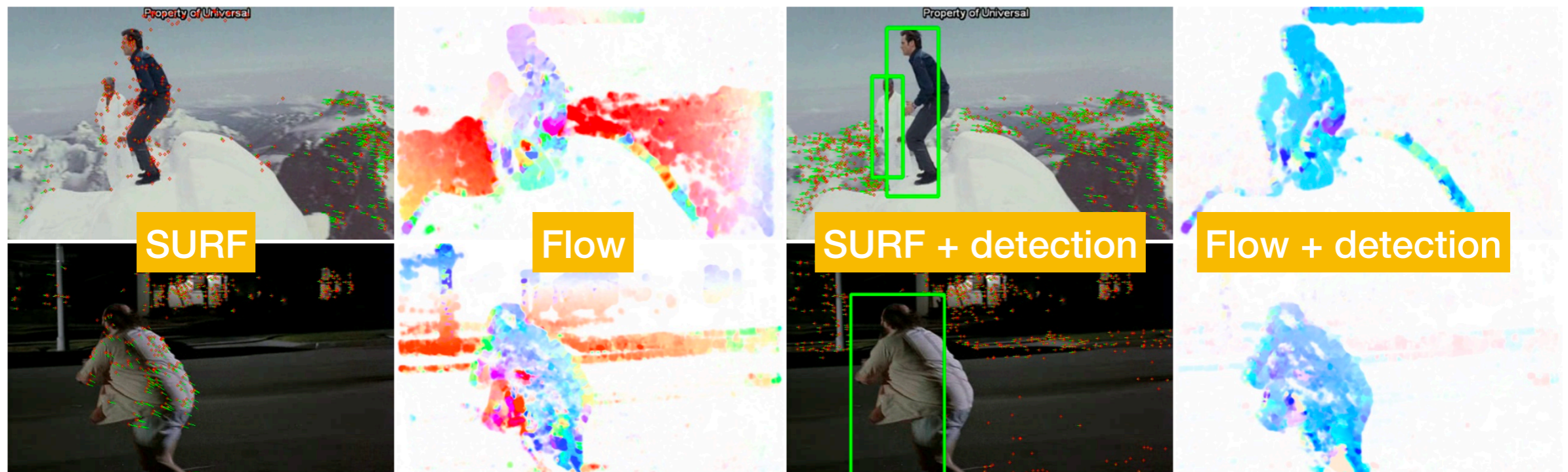
$$f_2(\mathbf{x}) = f_1(\mathbf{x} - \mathbf{d})$$

$$\mathbf{d} = -\frac{1}{2} \mathbf{A}_1^{-1} (\mathbf{b}_2 - \mathbf{b}_1)$$

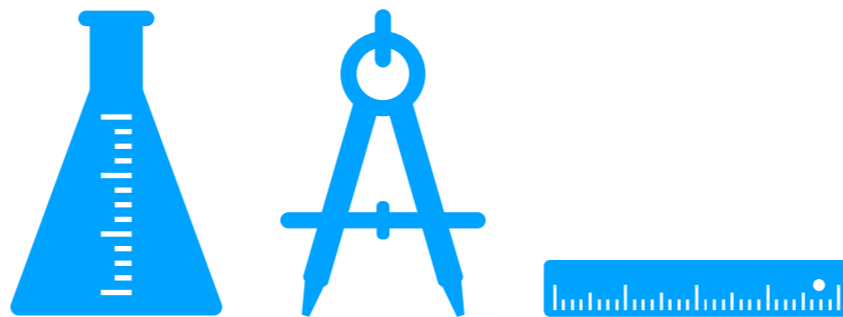
- Assume slowly varying displacement field

Human Detection

- We know humans aren't background *a priori*
- Part-based human detection with tracking, works with occlusion
- Mask away matches from humans when estimating homography



Experimental Setup



Dense Trajectory Features*

- Points densely sampled at different spatial scales
- Points are tracked using in heterogeneous areas (tracked for 15 frames to avoid drift)
- HOG, HOF, MBH, and trajectory (i.e., concatenation of displacement vectors) descriptors are calculated
- Descriptors calculated in space-time volume aligned with trajectory

*** Nothing new, mostly replicating setup in [40]**

Feature Encoding

- Bag of features and Fischer vector (includes 2nd order data)
- 4,000 element codebook build using k-means from 100,000 random features
- Classification:
 - RBF-kernel SVM for bag of features
 - Linear SVM for Fisher vector

Datasets

Hollywood2

HMDB51

Olympic Sports

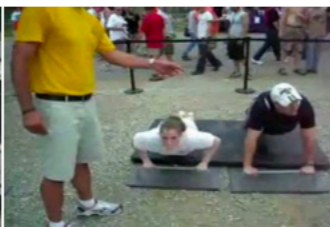
UCF50



(a) AnswerPhone



(a) GetOutCar



(b) Push-Up



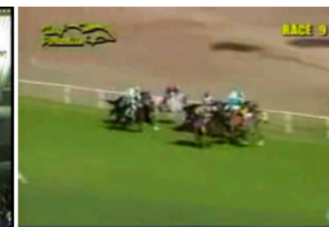
(b) Chew



(c) High-Jump



(c) Springboard



(d) Horse-Race



(d) Playing-Guitar



(a) HandShake



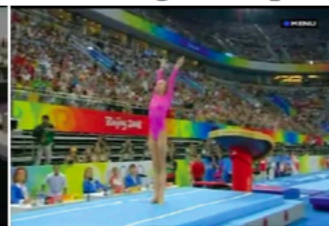
(a) HugPerson



(b) Cartwheel



(b) Pour



(c) Vault



(c) Tennis-Serve



(d) Punch



(d) Ski-Jet

69 movies
12 actions

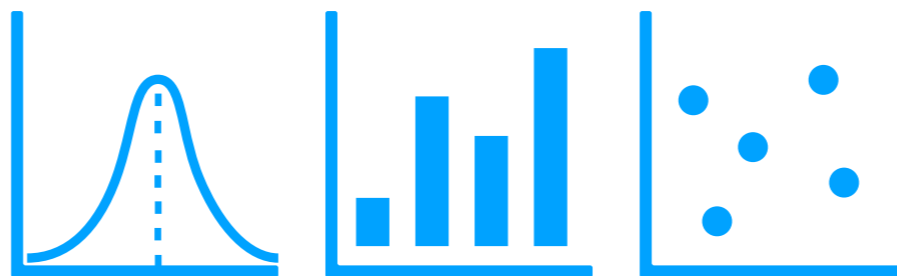
>6k videos
51 actions

783 sequences
16 actions

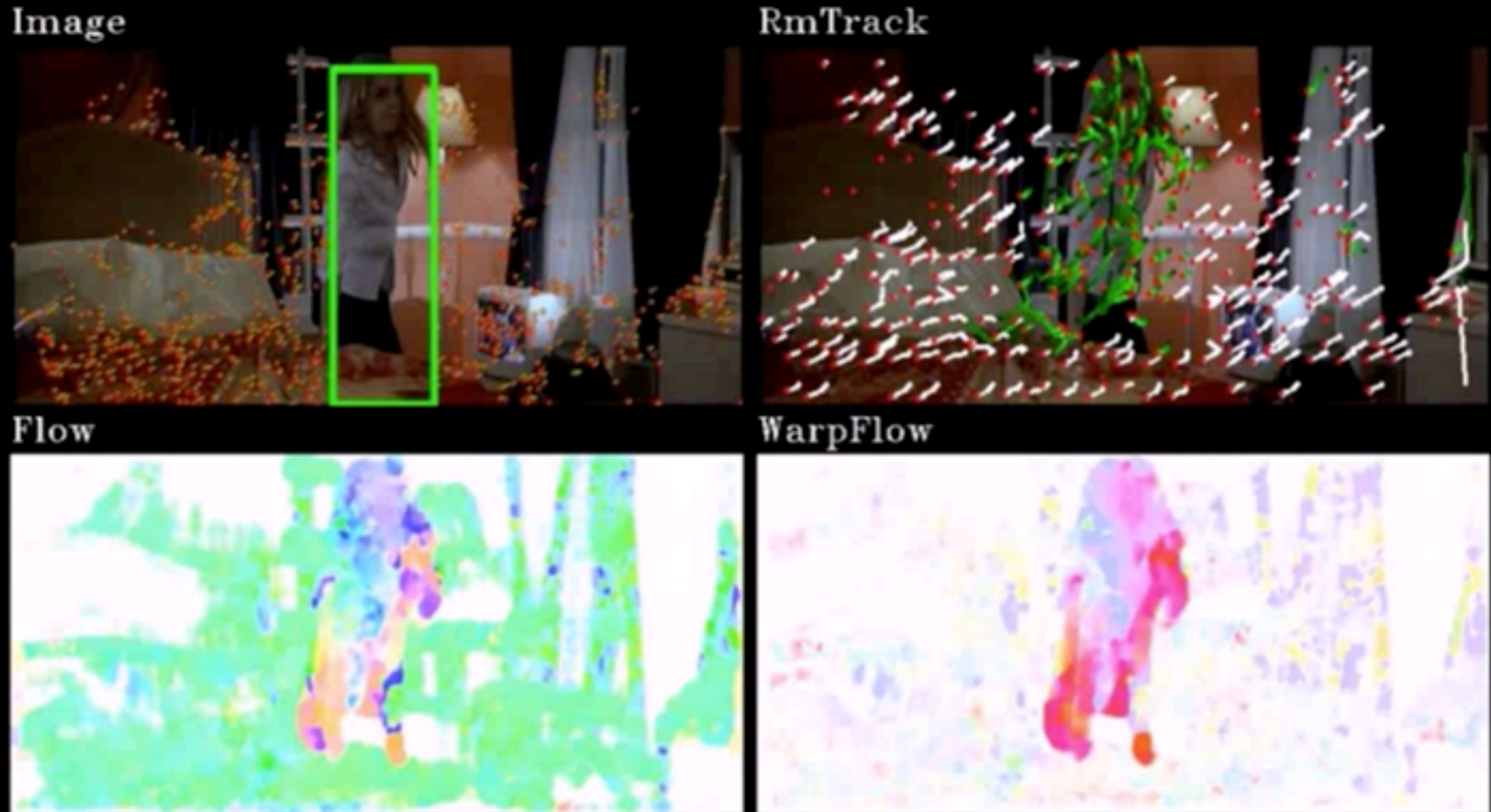
>6k YouTube videos
50 actions

Each dataset has hundreds to thousands of video sequences.

Results



Video Demo



Recognition Accuracy

	Hollywood2			
	Baseline	WarpFlow	RmTrack	Combined
Trajectory	42.2%	47.6%	42.4%	48.5%
HOG	46.9%	46.2%	46.7%	47.1%
HOF	51.4%	58.1%	53.4%	58.8%
MBH	57.4%	60.3%	58.6%	60.5%
HOF+MBH	58.2%	62.3%	59.7%	62.6%
Combined	60.1%	63.6%	61.7%	64.3%

← Use all features

↑
Warping with homography

↑
Background pruning

↑
Warping with homography
and background pruning

Recognition Accuracy

	Hollywood2				HMDB51			
	Baseline	WarpFlow	RmTrack	Combined	Baseline	WarpFlow	RmTrack	Combined
Trajectory	42.2%	47.6%	42.4%	48.5%	25.4%	31.0%	26.9%	32.4%
HOG	46.9%	46.2%	46.7%	47.1%	38.4%	38.7%	39.6%	40.2%
HOF	51.4%	58.1%	53.4%	58.8%	39.5%	48.5%	41.6%	48.9%
MBH	57.4%	60.3%	58.6%	60.5%	49.1%	50.9%	50.8%	52.1%
HOF+MBH	58.2%	62.3%	59.7%	62.6%	49.8%	53.5%	51.0%	54.7%
Combined	60.1%	63.6%	61.7%	64.3%	52.2%	55.6%	53.9%	57.2%

	Olympic Sports				UCF50			
	Baseline	WarpFlow	RmTrack	Combined	Baseline	WarpFlow	RmTrack	Combined
Trajectory	62.4%	73.7%	66.3%	77.2%	65.3%	72.6%	67.8%	75.2%
HOG	77.0%	76.3%	78.7%	78.8%	81.8%	81.6%	82.6%	82.6%
HOF	74.5%	86.2%	77.6%	87.6%	74.3%	85.4%	79.4%	85.1%
MBH	82.4%	87.5%	86.0%	89.1%	86.5%	88.4%	88.0%	88.9%
HOF+MBH	82.1%	88.3%	86.2%	89.7%	87.1%	89.3%	87.5%	89.5%
Combined	84.7%	88.9%	87.0%	91.1%	88.6%	90.9%	88.9%	91.2%

Combined Descriptor Recognition Accuracy

Datasets	Bag of features		Fisher vector	
	DTF	ITF	DTF	ITF
Hollywood2	58.5%	62.2%	60.1%	64.3%
HMDB51	47.2%	52.1%	52.2%	57.2%
Olympic Sports	75.4%	83.3%	84.7%	91.1%
UCF50	84.8%	87.2%	88.6%	91.2%

↑
Dense Trajectory Features

↑
Improved Trajectory Features

Human Detection: Effect on Accuracy

Hollywood2-sub	None	Automatic	Manual
Trajectory	32.3%	35.7%	37.1%
HOG	34.5%	34.9%	34.7%
HOF	43.9%	45.2%	46.7%
MBH	45.8%	47.4%	49.2%
Combined	48.9%	50.7%	51.9%

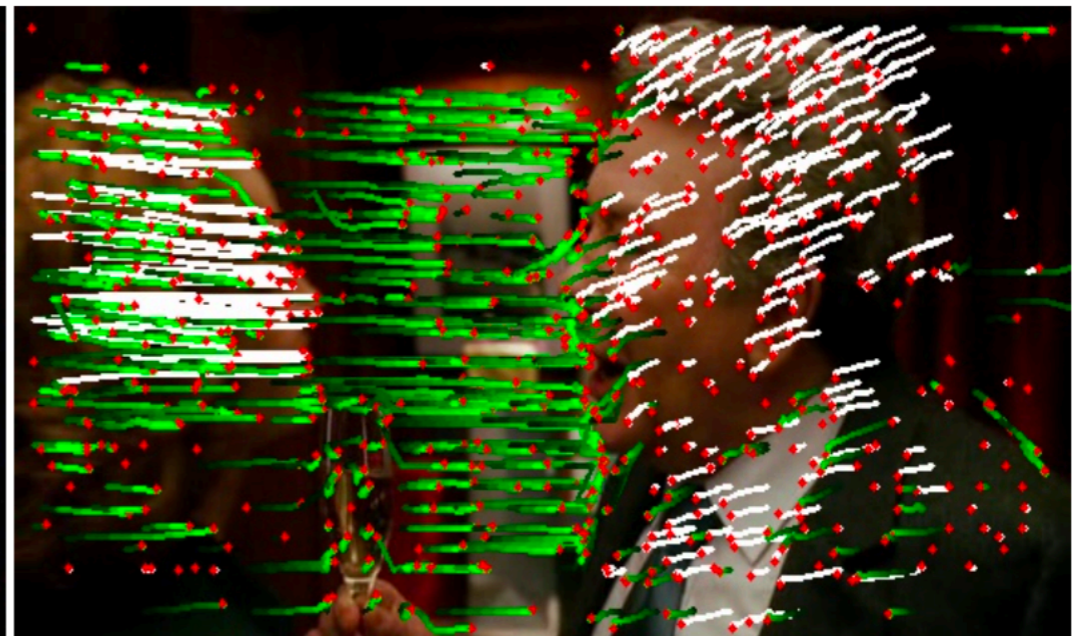
* with Fisher Vector encoding

State of the Art Results

Dataset	State of the Art Accuracy	Improvement Over State of the Art
Hollywood2	62.5%	2%
HMDB51	52.1%	5%
Olympic Sports	83.2%	8%
UCF50	83.3%	8%

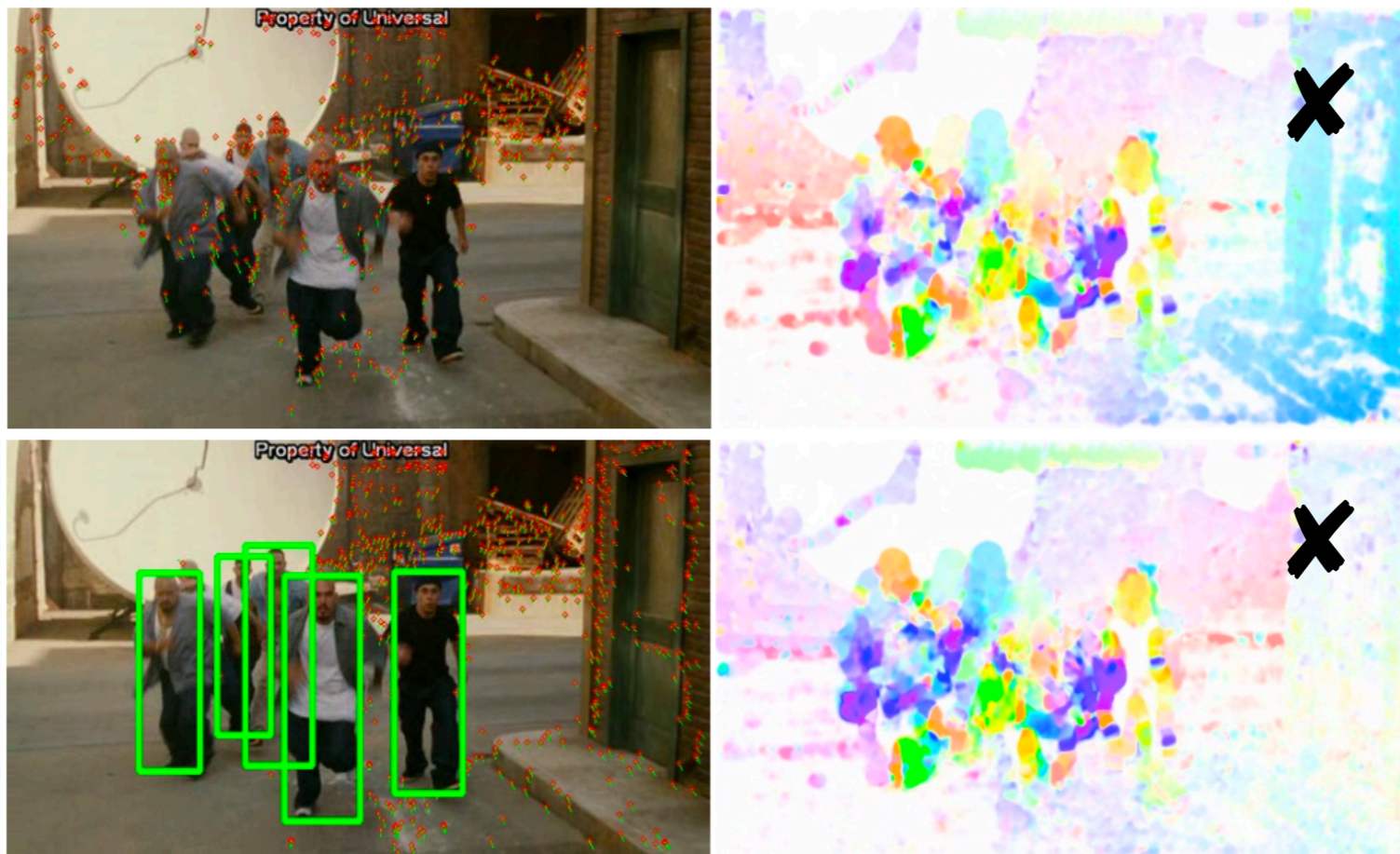
Technique Deficiencies

- Failure cases:
 - Homography is fit to foreground if it dominates the frame
 - Strong motion blur (issue in real-world datasets)



Technique Deficiencies

- Failure cases:
 - Complex mapping from estimated homography to background



Discussion + Q&A

Discussion Points

- How can some of this technique's deficiencies be overcome?
- What other types of *a priori* knowledge can be incorporated?
- The four datasets are all human-centric, how well would this pipeline work for nonhuman agents (e.g., cars)?
- Bag of features and Fischer vectors seem somewhat naïve, would a different encoding work better?