Towards Longer Long-Range Motion Trajectories

Michael Rubinstein, Ce Liu, William T. Freeman

MIT CSAIL Microsoft Research New England

Abstract

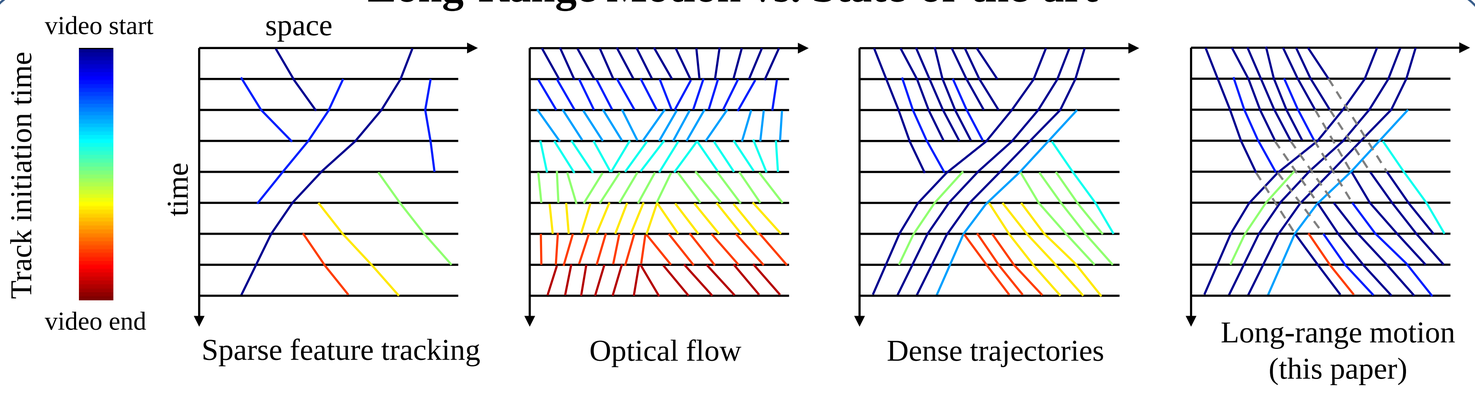
Goal: Finding long-range motion trajectories within videos

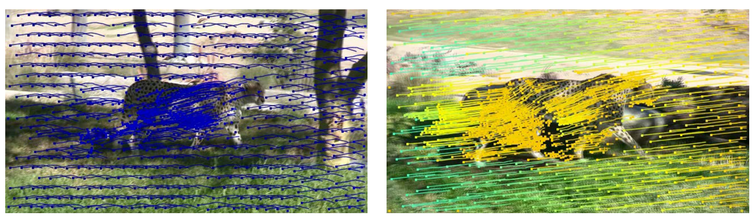
- Generic (low-level) algorithm that associates every scene feature that appears in multiple, not necessarily contiguous, parts of the video with the same track

Our approach:

1. Divide and conquer: compute accurate local (short-range) *tracklets* and combine them into global (long-range) trajectories
2. Combinatorial problem that is solved simultaneously for *all* tracklets in the sequence
3. Quantitative evaluation with/without ground truth data

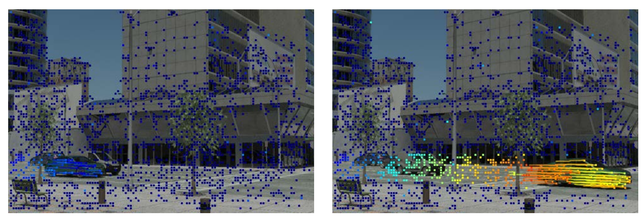
Long-Range Motion vs. State-of-the-art





Frame 15 Frame 170

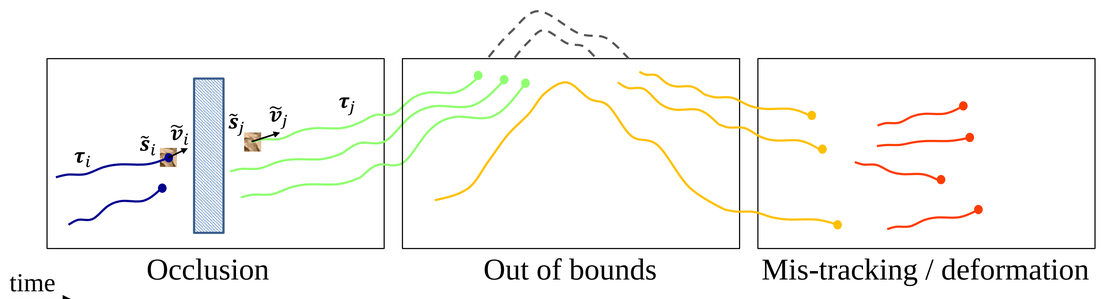
Particle Video (PV) [Sand and Teller 2006]



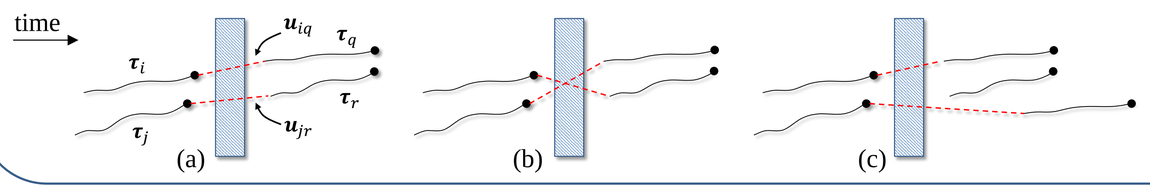
Frame 1 Frame 30

[Sundaram et al. 2010] (LDOF)

Graphical Model



Spatiotemporal regularization: (a) is the linkage with highest ψ_{ij}



Formulation

Tracklet linkage:

(a) Linked tracklets should encode the same scene feature with high probability
(b) Each candidate tracklet should be linked to at most one query tracklet
(c) Spatiotemporally neighboring tracklets should be linked with neighboring candidate tracklets

$$P(L) = \xi(L) \prod_i \phi_i(l_i) \prod_{i,j \in N(i)} \psi_{ij}(l_i, l_j)$$

Exclusion term (b) Track compatibility (a) Link Regularization (c)

Track compatibility

$$\phi_i(l_i) = \phi_i^a(l_i)\phi_i^m(l_i)\phi_i^p(l_i)$$

Link regularization

$$\psi_{ij}(l_i = q, l_j = r) = \exp\left(-\frac{1}{\sigma_a^2}\left\|\boldsymbol{u}_{iq} - \boldsymbol{u}_{jr}\right\|\right)$$

Appearance: $\phi_i^a(l_i = j) = \exp\left(-\frac{1}{\sigma_a^2}\left\|\tilde{\mathbf{s}}_i - \tilde{\mathbf{s}}_j\right\|_{1,d}\right)$

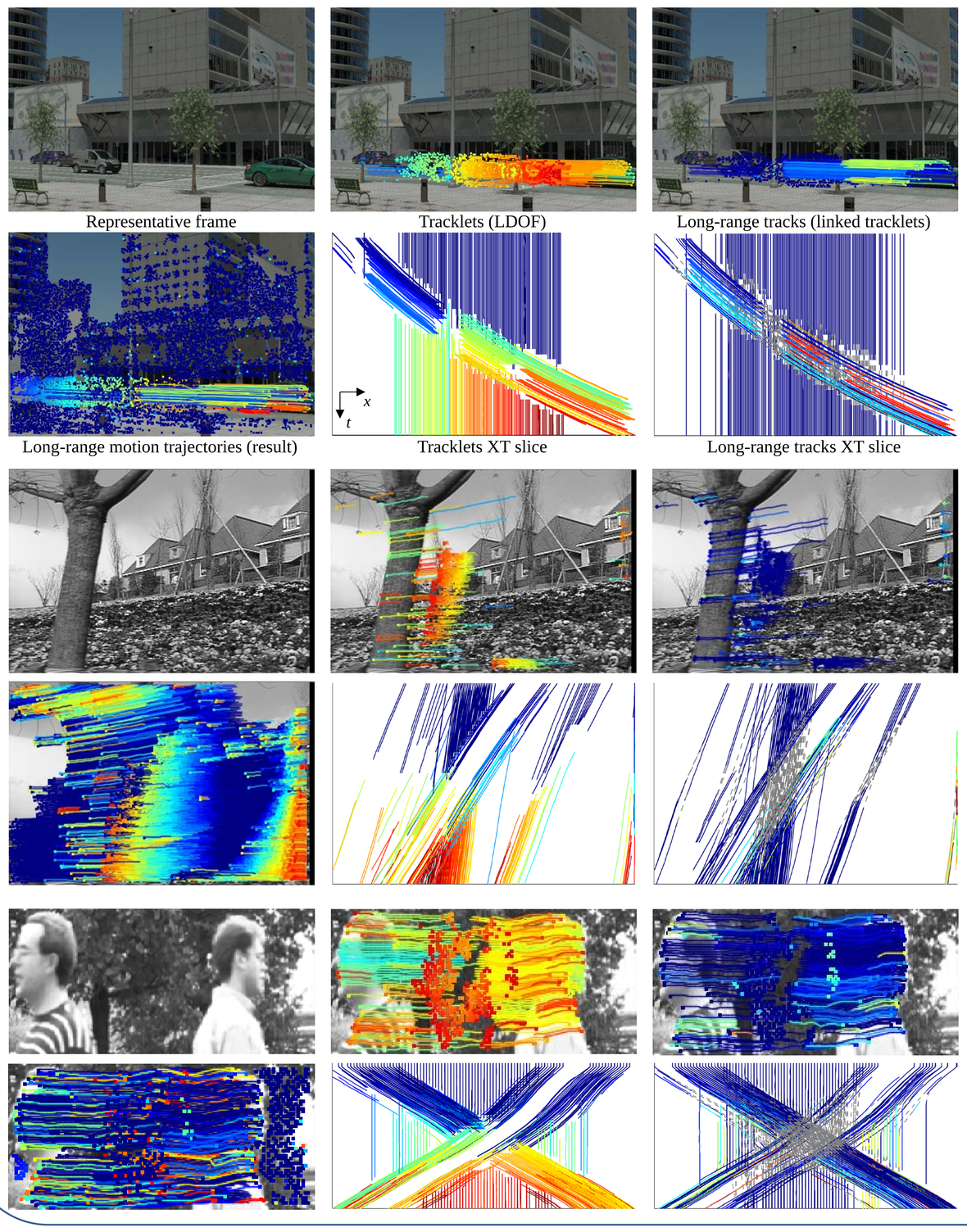
Motion: $\phi_i^m(l_i = j) = \exp\left(-\frac{1}{\sigma_m^2}\left\|\tilde{\boldsymbol{v}}_i - \tilde{\boldsymbol{v}}_j\right\|\right)$

Motion model (prior):

$$\phi_i^p(l_i = j) = \exp\left(-\frac{1}{\sigma_p^2}\left\|\boldsymbol{x}_i(t_i^{end}) - \boldsymbol{x}_j(t_j^{start}) + \tilde{\boldsymbol{v}}_i(t_j^{start} - t_i^{end})\right\|\right)$$

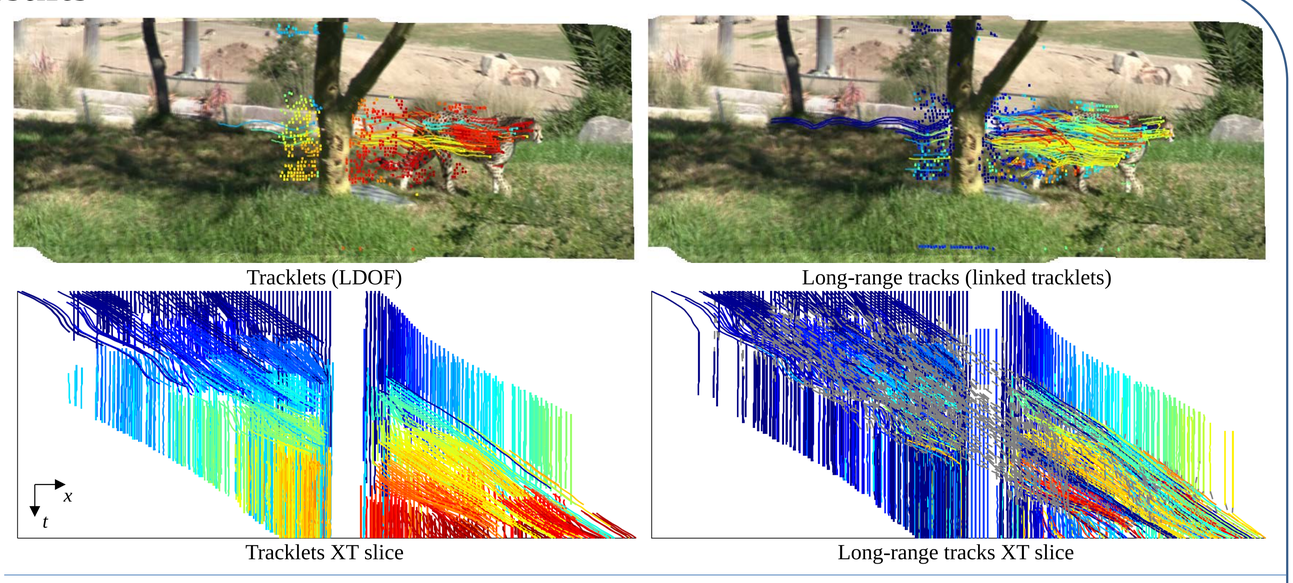
$\boldsymbol{x}_i(t)$	position of track i at time t
t_i^{start}	start time of track i
t_i^{end}	end time of track i
$\tilde{\mathbf{s}}_i$	track i's appearance at t_i^{end}
$\tilde{\boldsymbol{v}}_i$	track i's motion at t_i^{end}
$\tilde{\mathbf{s}}_j$	track j's appearance at t_j^{start}
$\tilde{\boldsymbol{v}}_j$	track j's motion at t_j^{start}
$\boldsymbol{u}_{iq}$	$\boldsymbol{x}_q(t_q^{start}) - \boldsymbol{x}_i(t_i^{end})$

Results



Representative frame Tracklets (LDOF) Long-range tracks (linked tracklets)

Long-range motion trajectories (result) Tracklets XT slice Long-range tracks XT slice



Tracklets (LDOF) Long-range tracks (linked tracklets)

Tracklets XT slice Long-range tracks XT slice

Quantitative evaluation:

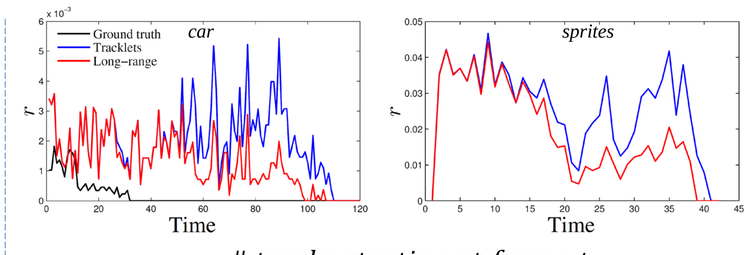
Synthetic data (*car*):

	PV	LDOF	PV + LR	LDOF + LR
r_{obj}	2.58	1.56	1.85	1.23

PV – Particle Video [Sand and Teller 2006]
LDOF – [Sundaram et al. 2010]
LR – Long-range (ours)

$$r_{obj} = \frac{\sum \#\ tracks\ covering\ each\ point}{\#\ points\ covered\ by\ tracks}$$

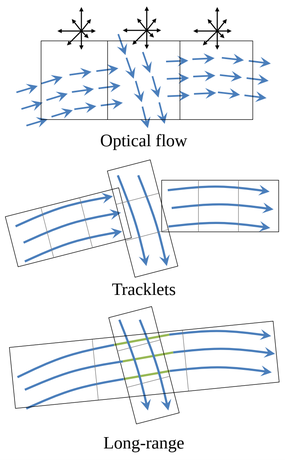
Track refresh number:



$$r = \frac{\#\ tracks\ starting\ at\ frame\ t}{\#\ tracks\ in\ frame\ t}$$

Preliminary results on action recognition:

Motion descriptor:



Results on KTH [Schuldt et al. 2004]: