Track to the Future: Spatio-temporal Video Segmentation with Long-range Motion Cues

José Lezama Karteek Alahari Josef Sivic Ivan Laptev

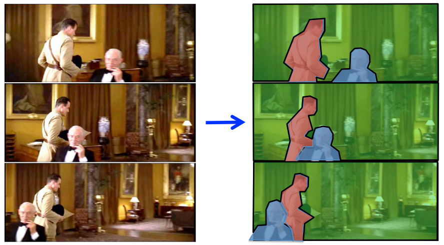
École Normale Supérieure de Cachan INRIA – WILLOW / École Normale Supérieure

Partly supported by Quaero Programme, MSR-INRIA laboratory

The goal

Long-range spatio-temporal video segmentation

Example



Why?

Provide building blocks for

- Object recognition in video (e.g. associate different views of object over time)
- Recognition of long-term object—person interaction
- Human action recognition

How?

Provide over-segmentation which has

- Spatial consistency: Respect object boundaries
- Temporal consistency: Associate object pixels over time

Our Contributions

- Use point-tracks to capture **long-range motion**
- Infer **local depth-ordering** to separate objects

Previous work

- Segment individual frames [Comaniciu & Meer 02, Felzenszwalb & Huttenlocher 04, Shi & Malik 00]

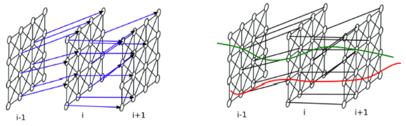
  Not consistent over frames
- Use locally coherent motion (motion-based segmentation) [Shi & Malik 98, Weiss 97, Zitnick et al. 05, Stein et al. 07]

  A small temporal window
- Some work on spatio-temporal segmentation [Dementhon 02, Grundmann et al. 10, Wang et al. 04]

Do not exploit long-range motion constraints

Overview

- Build on graph-based agglomerative segmentation of [Felzenszwalb & Huttenlocher 04, Grundmann et al. 10] and group neighbouring pixels with similar colour and motion

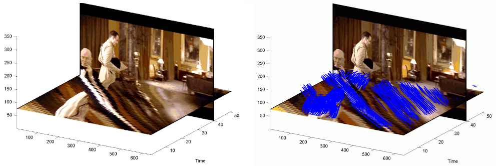


Frames + optical flow Frames + point-tracks

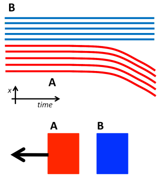
- Introduce point-tracks for long-range support over time
- Encourage all points in a track to belong to the same segment
- Ensure dissimilar tracks are assigned to different segments

How to cluster the tracks?

Find (dis)similarities among point-tracks

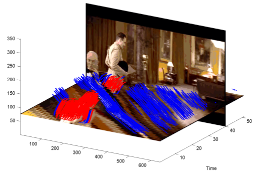


[Tracks from Brox & Malik 10]

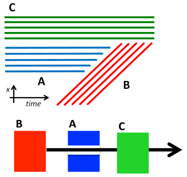


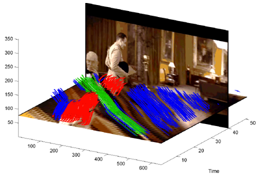
Motion (dis)similarity

[Costeira & Kanade 98, Sivic et al. 06, Brox&Malik 10]



Point-track clusters





Local depth-ordering

Track clustering

- Formulated as an energy minimization problem
- Each variable x_i represents a point-track

Controls the splitting-merging Occlusion cost

$$E(\mathbf{x}) = \sum_{(i,j)\in\mathcal{E}} \Big[\underbrace{\alpha_{ij}\phi_1(x_i,x_j)}_{\text{Merges two tracks}} + \underbrace{(1-\boxed{\alpha_{ij}})\phi_2(x_i,x_j)}_{\text{Separates two tracks}} + \underbrace{\boxed{\gamma_{ij}}\phi_3(x_i,x_j)}_{\text{Orders the tracks}}\Big]$$

$$\phi_1(x_i,x_j)=\begin{cases}0 & \text{if } x_i = x_j,\\ 1 & \text{otherwise.}\end{cases} \quad \phi_2(x_i,x_j)=\begin{cases}1 & \text{if } x_i = x_j,\\ 0 & \text{otherwise.}\end{cases} \quad \phi_3(x_i,x_j)=\begin{cases}1 & \text{if } x_i \ge x_j,\\ 0 & \text{if } x_i < x_j.\end{cases}$$

- Solved using Tree-reweighted message passing (TRW-S) [Kolmogorov 06]

Similarity cost

Similar to [Brox & Malik 10]

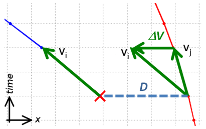
Spatial coordinates Local velocity

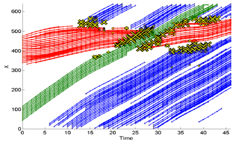
$$\alpha_{ij} = exp\left(-\frac{(1+\overbrace{||\mathbf{a}_i-\mathbf{a}_j||_2}^{\text{Spatial coordinates}})^2\overbrace{||\mathbf{v}_i-\mathbf{v}_j||_2^2}^{\text{Local velocity}}}{\underbrace{2l_{ij}}_{\text{Temporal overlap}}\sigma_s^2}\right)$$

Temporal overlap

Occlusion cost

Measured as a local difference of velocities

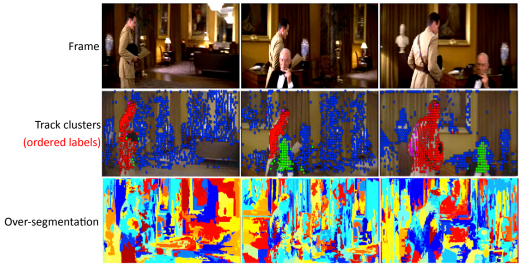




Distance measure

$$\gamma_{ij} = 1 - exp\left(-\frac{\boxed{d}||\mathbf{v}_i-\mathbf{v}_j||_2^2}{\sigma_o^2}\right)$$

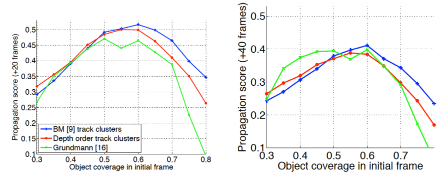
Video segmentation results



Evaluation

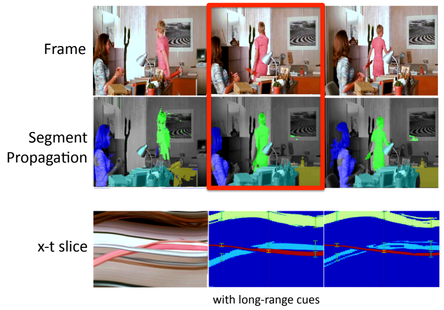
- Video clips selected from Hollywood 2 dataset
- Office scenes with significant motion and (dis-) occlusions
- Ground truth segmentation is labelled for selected frames
- Select a ground truth segmented frame, and propagate the segments over time
- Measure the overlap of segments generated in other ground truth frames

Segment propagation results



(a) +20 frames (b) +40 frames

Manually marked frame



Better propagation

Summary

- Video over-segmentation consistent over frames
- Infer local depth-ordering of point-tracks

Future work

- Object category-level video segmentation
- Long-term object—person interaction
- Parameter learning and optimization methods