**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

---

### The goal
Long-range spatio-temporal video segmentation

### Example
- Frames + optical flow
- Frames + point-tracks

### 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 09]
- Not consistent over frames
- Use locally coherent motion (motion-based segmentation) [Shi & Malik 99, Weiss 99, Zitnick et al. 09, Stain 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
- 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

---

### Track clustering
- Formulated as an energy minimization problem
- Each variable $x_i$ represents a point-track
- Controls the splitting-merging
- Separates two tracks
- Orders the tracks

**Occlusion cost**
Measured as a local difference of velocities

**Similarity cost**
Similar to [Brox & Malik 10]

**Occlusion cost**
- Spatial coordinates
- Local velocity

**Similarity cost**
- $\alpha_{ij} = \exp\left(-\frac{1}{2\sigma^2} \left( |v_i - v_j| \right)^2 \right)$

**Temporal overlap**

---

### 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

---

### 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