Object Segmentation by Long Term Analysis of Point Trajectories

by **Thomas Brox and Jitendra Malik, ECCV 2010**

Introduction

- Contemporary person detectors achieve this goal by learning a classifier and a shape distribution from manually annotated training images.
- Animals or infants are not supplied bounding boxes or segmented training images when they learn to see.

Introduction

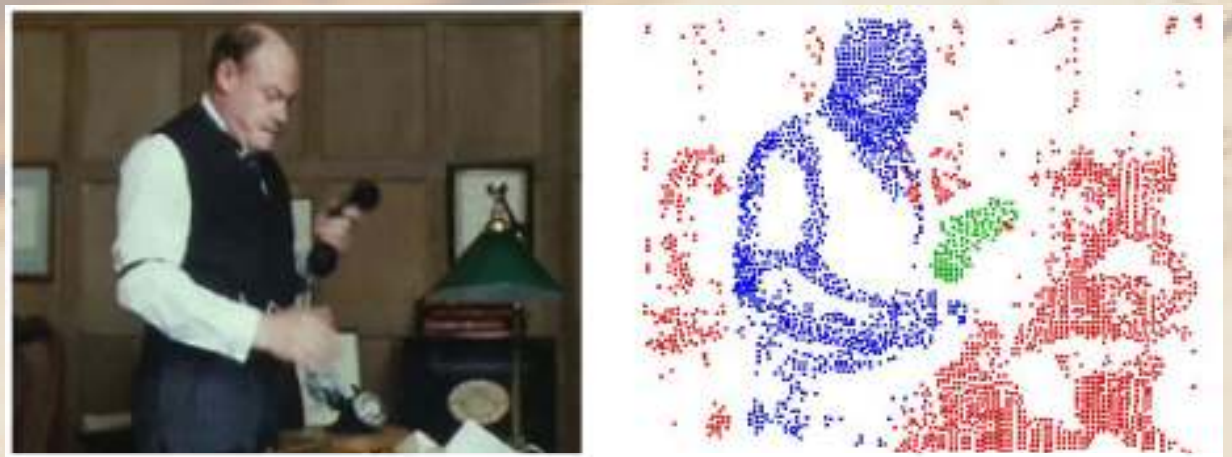


Fig. 1. Left: (a) Bottom-up segmentation from a single input frame is ambiguous. **Right:** (b) Long term motion analysis provides important information for bottom-up object-level segmentation. Only motion information was used to separate the man and even the telephone receiver from the background.

- Biological vision systems learn objects up to a certain degree of accuracy in an unsupervised way by making use of the natural ordering of the images they see.[1]
- Motion provides important information for grouping. (Gestalt - common fate)

[1] Spelke, E.: Principles of object perception. Cognitive Science 14, 29–56 (1990)

Introduction

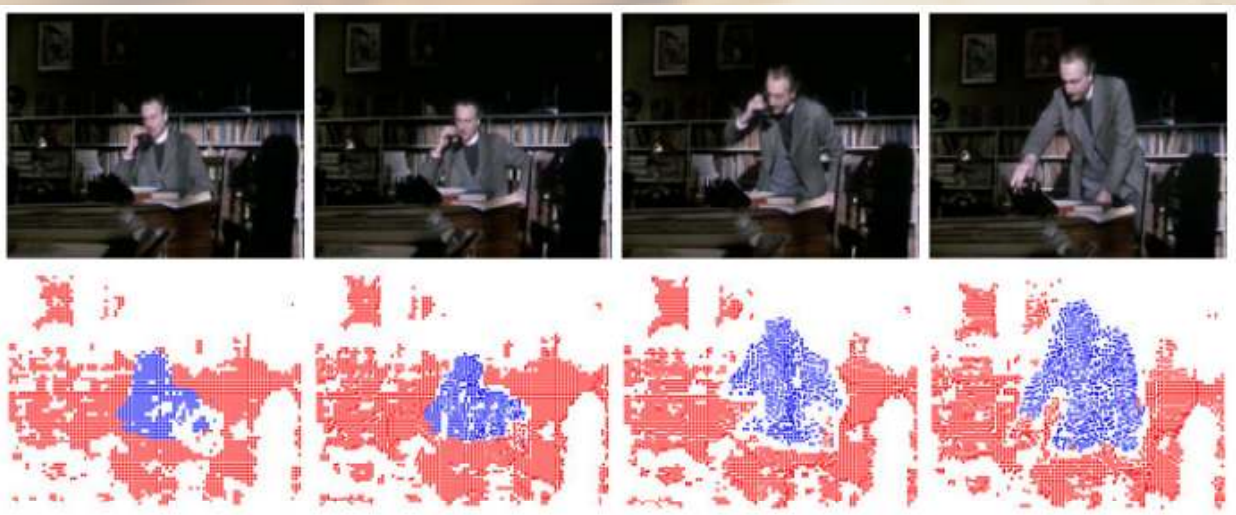


Fig. 2. Frames 0, 30, 50, 80 of a shot from *Miss Marple: Murder at the vicarage*. Up to frame 30, there is hardly any motion as the person is sitting. Most information is provided when the person is sitting up. This is exploited in the present approach. Due to long term tracking, the grouping information is also available at the first frames.

- They argue that temporally consistent clusters over many frames can be obtained best by analyzing long term point trajectories rather than two-frame motion fields.

Introduction

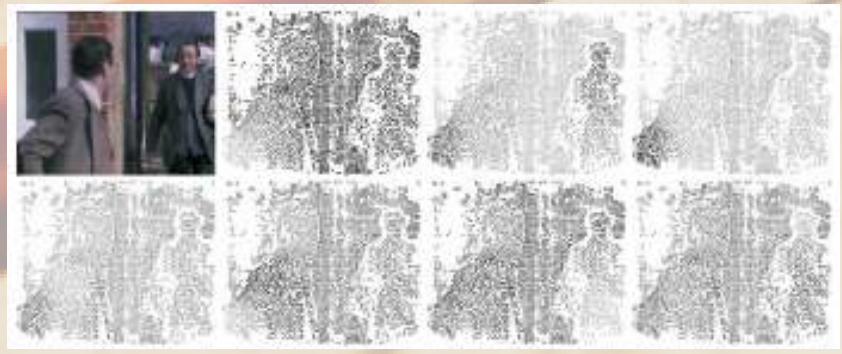
- Segment objects from a video sequence based on appearance and motion
- This is difficult because
  - No constraints on camera motion and object motion
  - Objects need not be rigid
- Required for video editing and video understanding
  - Video editing (object copy/paste etc.)
  - Structure from motion (2D to 3D conversion etc.)
  - Scene context analysis
  - Understanding dynamic object interaction

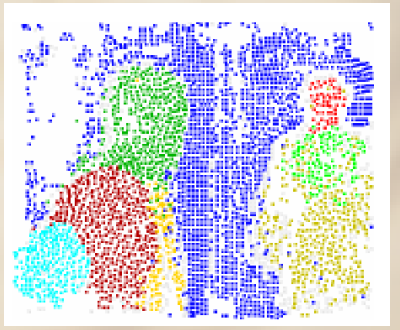
Previous Work

- Factorization based
  - Generalized PCA (GPCA), Local Subspace Analysis (LSA), Agglomerative Subspace Clustering (ALC).
  - Works for rigid objects.
- Multiple Hypothesis video segmentation
  - Uses 2D superpixels at multiple scales.
  - Finds the best hypothesis that connects the superpixels using higher order potentials on a Conditional Random Field (CRF).
- Hierarchical graph segmentation
  - Use agglomerative clustering.
  - Join two supervoxels if their "internal variation" is greater than the edge weight between them.

Method

1. Track points and compute affinities between trajectories.
2. Run spectral clustering with spatial regularity on affinity matrix.





Point Tracking and Affinities between Trajectories

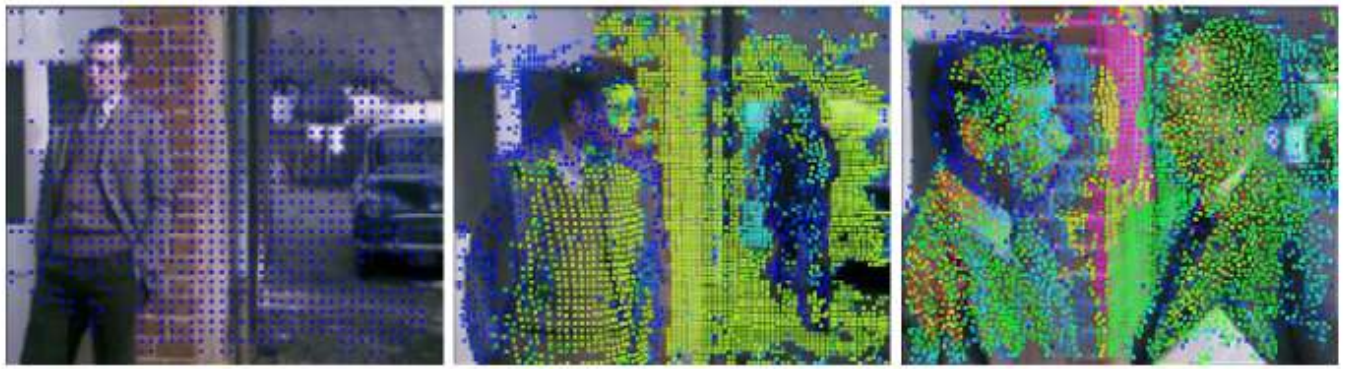


Fig. 3. From left to right: Initial points in the first frame and tracked points in frame 211 and 400. Color indicates the age of the tracks. The scale goes from blue (young) over green, yellow, and red to magenta (oldest). The red points on the right person have been tracked since the person appeared behind the wall. The figure is best viewed in color.

- Obtain point trajectories by running the optical flow based tracker in [2].
- The coverage of the image by tracked points is much denser than with usual keypoint trackers.

[2] Sundaram, N., Brox, T., Keutzer, K.: Dense point trajectories by GPU-accelerated large displacement optical flow. In: European Conf. on Computer Vision. LNCS, Springer, Heidelberg (2010).

Point Tracking and Affinities between Trajectories

$$d_t^2(A,B) = d_{\text{sp}}(A,B)\frac{(u_t^A - u_t^B)^2 + (v_t^A - v_t^B)^2}{5\sigma_t^2}.$$

$u_t := (x_t+5) - x_t$
$v_t := (y_t+5) - y_t$
$d_{sp}(A,B)$, average spatial euclidian distance

$$\sigma_t = \min_{a\in\{A,B\}} \sum_{t'=1}^{5} \sigma(x_{t+t'}^a, y_{t+t'}^a, t+t').$$

σ, local optical flow variance in each frame.

$$w(A,B) = \exp(-\lambda d^2(A,B)) \qquad \lambda = 0.1$$

- The initial points from that time have been tracked to the last frame and are visible as red spots among all the other tracks that were initialized later due to the scaling of the person.
- Define pairwise affiities between all trajectories that share at least one frame
- The actual information is not in the common motion but in motion differences.

Spectral Clustering with Spatial Regularity

$$E(\pi, K) = \sum_{a} \sum_{k=1}^{K} \delta_{\pi^a, k} \|\mathbf{v}^a - \mu_k\|_{\lambda}^2 + \nu \sum_{a} \sum_{b \in \mathcal{N}(a)} \frac{1 - \delta_{\pi^a, \pi^b}}{|\mathbf{v}^a - \mathbf{v}^b|_2^2} \qquad \nu = \frac{1}{2}$$

K, number of clusters

N(a), set of neighboring trajectories based on the average spatial distance of trajectories.

$$\left\| v^a - \mu \right\|_{\lambda} = \sum_i (v_i^a - \mu_i)^2 / \lambda_i$$

- First term, finds smaller and distinct clusters.
- Second term, is penalizing the spatial boundaries between clusters.
  - Boundaries within a smooth area are penalized far more heavily, which avoids splitting clusters at arbitrary locations due to smooth transitions in the eigenvectors.
- Eigenvectors with smaller eigenvalues, separate more distinct clusters.

Experimental Evaluation

- Hopkins 155 dataset
  - Some of the sequences do not correspond to natural scenes.
  - It is much too specialized.
- A new motion segmentation dataset.
  - 26 sequences from detective stories.
  - 10 car and 2 people sequences from Hopkins 155.
  - The annotation is dense in space and sparse in time.

Experimental Evaluation

- Density
  - is the density of the points for which a cluster label is reported.
- Overall clustering error
  - is the number of bad labels over the total number of labels on a per-pixel basis.
- Average clustering error
  - is similar to the overall error but averages across regions after computing the error for each region separately.
- Over-segmentation error
  - the number of clusters merged to fit the ground truth regions

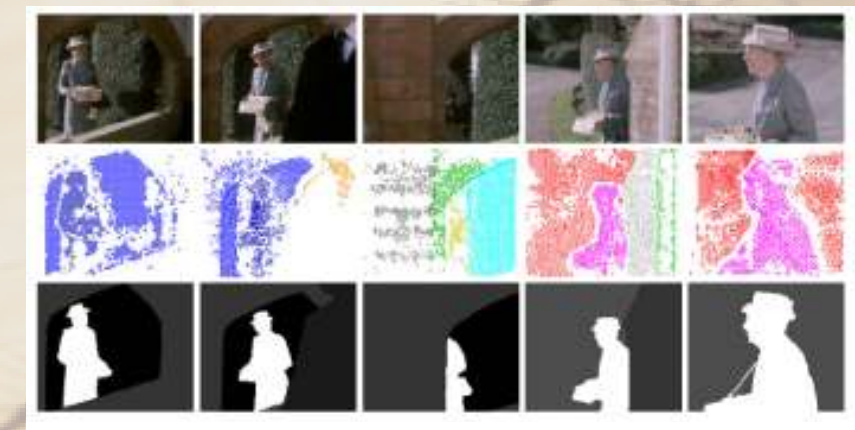


Fig. 6. Frames 1, 110, 135, 170, 250 of a shot from *Miss Marple: Murder at the vicarage* together with our clusters and the ground truth regions. There is much occlusion in this sequence as Miss Marple is occluded by the passing inspector and then by the building. Our approach can link tracks of partially occluded but not of totally occluded objects. A linkage of these clusters is likely to be possible based on the appearance of the clusters and possibly some dynamic model.

Experimental Evaluation

- Generalized PCA, (GPCA)
- Local Subspace Affinity, (LSA)
- Random Sample Consensus, (RANSAC)
- Agglomerative Lossy Compression, (ALC)

	tracks	time
our method	15486	497s
GPCA	12060	2963s
LSA	12060	38614s
RANSAC	12060	15s
ALC	957	22837s

Table 1. Computation times for the people1 sequence of the Hopkins dataset considering only the first 10 frames.

- All these methods ask for the number of regions to be given in advance.
- All these techniques require the tracks to have full length except ALC.
- Tracks are randomly subsampled the tracks for ALC by a factor 16, since it gets intractably slow when considering more than 1000 trajectories.

Experimental Evaluation

	Density	overall error	average error	over-segmentation	extracted objects
First 10 frames (26 sequences)					
our method	3.34%	7.75%	25.01%	0.54	24
GPCA	2.98%	14.28%	29.44%	0.65	12
LSA	2.98%	19.69%	39.76%	0.92	6
RANSAC	2.98%	13.39%	26.11%	0.50	15
ALC corrupted	2.98%	7.88%	24.05%	0.15	26
ALC incomplete	3.34%	11.20%	26.73%	0.54	19
First 50 frames (15 sequences)					
our method	3.27%	7.13%	34.76%	0.53	9
ALC corrupted	1.53%	7.91%	42.13%	0.36	8
ALC incomplete	3.27%	16.42%	49.05%	6.07	2
First 200 frames (7 sequences)					
our method	3.43%	7.64%	31.14%	3.14	7
ALC corrupted	0.20%	0.00%	74.52%	0.40	1
ALC incomplete	3.43%	19.33%	50.98%	54.57	0
All available frames (26 sequences)					
our method	3.31%	6.82%	27.34%	1.77	27
ALC corrupted	0.99%	5.32%	52.76%	0.10	15
ALC incomplete*	3.29%	14.93%	43.14%	18.80	5

- The more traditional methods like GPCA, LSA, and RANSAC do not perform well on this dataset.
- Even when considering only 10 frames, i.e. there is only little occlusion, the error is much larger than for the proposed approach.
- The 10-frame result for ALC with a correction for corrupted tracks is quite good. This is mainly due to the correct number of regions given to ALC.

Experimental Evaluation

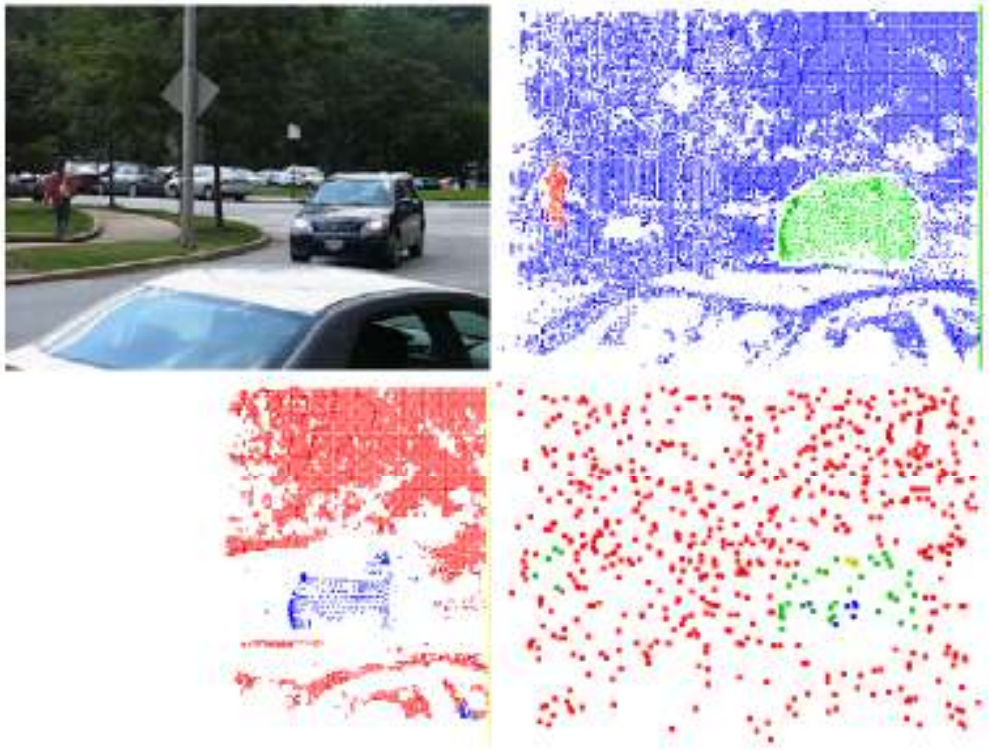


Fig. 7. From left to right: (a) Frame 40 of the cars4 sequence from the Hopkins dataset. (b) The proposed method densely covers the image and extracts both the car and the person correctly. (c) RANSAC (like all traditional factorization methods) can assign labels only to complete tracks. Thus large parts of the image are not covered. (d) ALC with incomplete trajectories [19] densely covers the image, but has problems assigning the right labels.

Conclusion

- They presented a technique for object-level segmentation in a pure bottom-up fashion by exploiting long term motion cues.

- The provided benchmark dataset that comprises a variety of easier and harder sequences, can foster progress in this field.

Application on Action Recognition