

Online Non-rigid Motion and Scene Layer Segmentation

Ali Elqursh

Defense Committee

Ahmed Elgammal (Chair)

Vladimir Pavlovic (Rutgers University)

Dimitris Metaxis (Rutgers University)

Omar Javed (SRI International)

1

Motivation

- Detecting objects of interest is the first step in video analytics

- Object Recognition
- Activity Recognition
- Pose Estimation

- Target: Videos from a moving platform that need online processing

- Existing approaches for object detection and segmentation (not applicable or not effective):

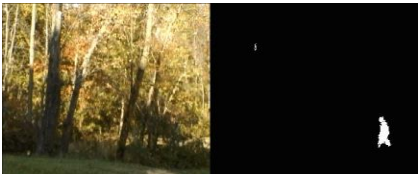
- Background Subtraction
- Object Detectors “Foreground Detection”
- Motion Segmentation
- Video Segmentation


		More General →	
		Available Offline	Online
More General ↓	Static camera	Post processing of surveillance videos	Online surveillance cameras
	Moving camera	Youtube videos	Streaming video sources (Camera Phones, TV broadcast, robotics, Camera mounted on cars, Google Glass,...)

Today's videos are mostly from moving platforms and need online processing. No Effective Solution Exists

2

Background Subtraction	Foreground Object Detection
Accurate segmentation of the foreground (object silhouettes)	Bounding boxes
Assumes stationary camera/scene	Work with moving camera – Designed for single image detection, does not model scene dynamics.
Detect any non-background object	Object specific detectors (Not scalable)
Requires bootstrapping High accuracy / low false alarm	Requires training High false positives/ false negatives <ul style="list-style-type: none"> [Ja11] Face detection algorithms: $\approx 70\%$ accuracy with ≥ 1 FPPI (false positive per image). [Do09] Pedestrian detection: recall $\leq 60\%$ of unoccluded pedestrians with 80 pixel height at a 1 FPPI. Pascal Challenge 2011: precision 51.6% for detecting people and 54.5% for detecting cars.



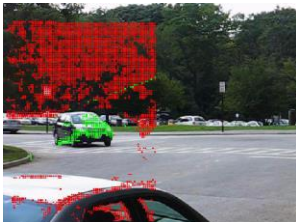



[Felzenszwalb2010]

[Felzenszwalb2010] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan
Object Detection with Discriminatively Trained Part Based Models (PAMI2010)

3

Motion Segmentation	Video Segmentation
Segment a set of point trajectories	Segments coherent regions that have similar color and motion over space and time
Produce sparse segmentation	Produce pixel-level segmentation
Offline process	Offline process
Do not aim at modeling scene background	No concept of background and foreground
<ul style="list-style-type: none"> Trajectory is visible through-out the entire frame sequence Objects are rigid Camera is affine Produces a sparse segmentation 	<ul style="list-style-type: none"> Results in an over segmentation Works on a small-time window Sliding window approach is needed to handle long videos, however doesn't provide consistent solution





[Grundmann2010]

[Grundmann2010] Matthias Grundmann et al. , Efficient Hierarchical Graph Based Video Segmentation (CVPR 2010)

4

Motivation

- Existing approaches
 - Mostly offline
 - Does not learn the scene appearance
 - Either uses long trajectories to capture motion or use two frame optical flow
- Solution:
 - Devise a motion segmentation approach that
 - Is Online
 - Handle trajectories of any length
 - Applicable to perspective cameras
 - Handles nonrigid objects
 - Integrate with a framework for Scene Layer Segmentation that “learns” models of the different layers

5

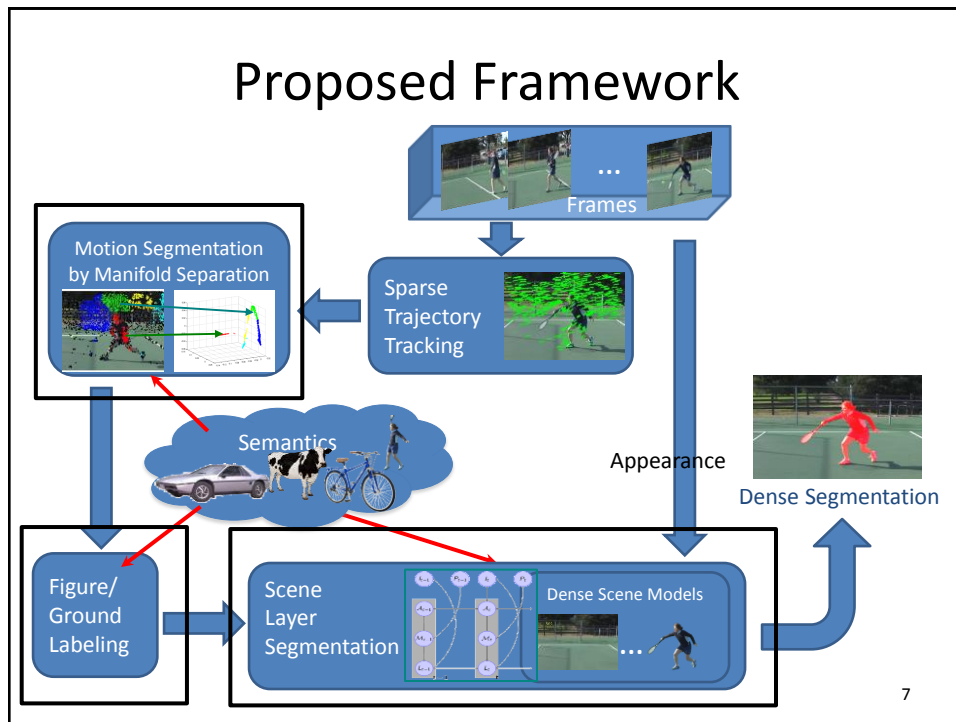
Applications

- 2D to 3D Conversion
 - Manual intensive process! Companies charge up to **\$100,000** per minute of converted footage !
 - Depth effect can be generated by shifting the foreground
 - Proposed framework segments foreground objects
 - Learned background model can be used to fill holes
- Object Recognition from Video data
 - Background clutter severely affect recognition accuracy [Ren2009]
 - Localizing objects in videos also reduces false positives.
- Video search and data mining applications
 - 72 hours of video are uploaded to Youtube every minute! [Youtube]
 - Current search techniques are limited to tags or manual captioning
- Video surveillance from moving cameras
 - Cameras mounted on drones can be used for surveillance
- Car safety and pedestrian detection

[Ren2009] X. Ren and M. Philipose. Egocentric recognition of handled objects: benchmark and analysis. In First Workshop on Egocentric Vision, 2009

[Youtube] <http://www.youtube.com/yt/press/statistics.html>

6

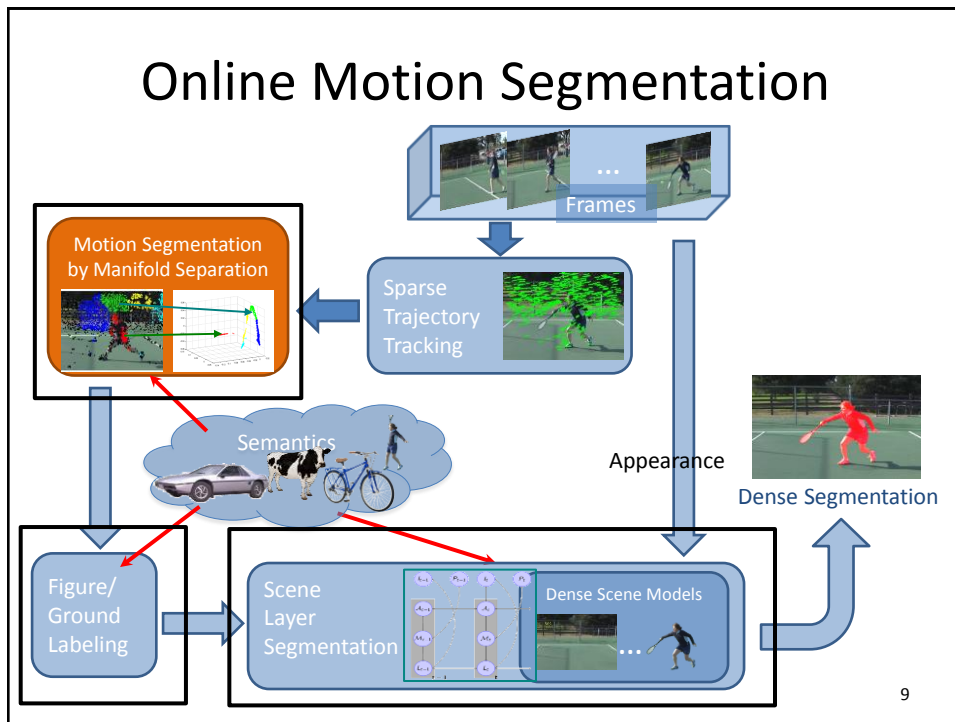


7

Contribution

- Trajectories belonging to a rigid object form a manifold of dimension 3.
 - Cast motion segmentation as manifold segmentation.
- Two methods for online manifold segmentation using explicit and implicit models
- Online method that
 - Learns appearance and motion models of the scene (background and foreground)
 - Produces segmentations of video frames.

8



Motion Segmentation

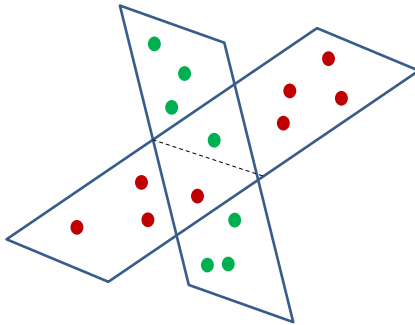
- Motion Segmentation refers to the problem of segmenting different motions in the video.
- Motion Representation
 - Optical Flow
 - Point Trajectories
- Under affine Camera assumption, trajectories generated from rigid motion and under affine projection spans a 4-dimensional subspace [Tomasi1992]

Trajectories \rightarrow $[t_1 \ t_2 \ \dots \ t_N] = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_M \end{bmatrix}_{2M \times 4}$ \leftarrow Camera Matrices
 Rank ≤ 4 \nearrow

$[X_1 \ X_2 \ \dots \ X_N]_{4 \times N}$ \leftarrow 3D Points

[Tomasi1992] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. (IJCV 1992)

Affine Motion Segmentation



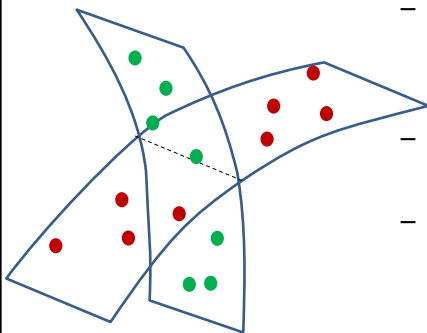
- RANSAC, random sampling and model fitting
- Direct factorization using a shape interaction matrix [Costeria1998]
- Generalize PCA [Vidal2004]

[Costeria1998] J. P. Costeira and T. Kanade: A multi-body factorization method for independently moving objects (IJCV1998)

[Vidal2004] R. Vidal and R. Hartley. Motion segmentation with missing data using power factorization and GPCA. (CVPR 2004).

11

Motion Segmentation



- Spectral Clustering Methods
- Relaxes the assumption of affine subspaces
 - [Yan2006]
 - Neighbors around each trajectory are used to fit a subspace.
 - Affinity matrix built by measuring the angles between subspaces.
 - [Elhamifar2009] Affinity matrix from representing each trajectory as a sparse combination of other trajectories
 - [Brox2010]
 - Affinity matrix capturing similarity in translational motion across all pairs of trajectories.
 - A final grouping step then achieves motion segmentation.
- Offline approaches

[Yan2006] Yan, J. and Pollefeys, M. A General Framework for Motion Segmentation : Independent , Articulated , Rigid , Non-rigid , Degenerate and Non-degenerate. (ECCV2006)

[Elhamifar2009] E. Elhamifar and R. Vidal. Sparse subspace clustering. (CVPR 2009)

[Brox2010] T. Brox and J. Malik. Object Segmentation by Long Term Analysis of Point Trajectories. (ECCV 2010)

12

Motion Segmentation

- New formulation
 - Motion Segmentation via Manifold Separation
- Problem: How can we make the approach online?
- Two approaches
 - Segment by explicitly reconstructing the manifold in a low dimensional representation
 - Implicitly segment the manifolds using label propagation

13

Motion Segmentation via Manifold Separation

- **Claim:** Point trajectories in image space form a 3D manifold
- **Proof:**
 - Trajectories in 3D form a 3D manifold
 - Location at $t = 2, \dots, F$ $f_2(x), \dots, f_F(x)$
 - Since f_2, \dots, f_F are continuous maps
 - Space of trajectories is a manifold of dimension three.
 - Projecting the 3D trajectories into the image coordinates also induces a manifold.
 - Let $g(x) = \frac{f}{z} [x, y]^T$ camera projection function
 - $g(x)$ is continuous at all points except at points with $z=0$
 - Let $\Omega(f) = \Gamma \setminus \{x_1 \dots x_F: z_i \neq 0\}$
 - $G(x_1 \dots, x_F) = (g(x_1), \dots, g(x_F))$ is also a smooth continuous map over $\Omega(f)$
 - $G(\Omega)$ is also a manifold of dimension three.

14

Motion Segmentation via Manifold Separation

- Theoretical justification for why a method for online manifold segmentation is necessary
- Spectral clustering will fail when there exist multiple manifolds that are close or intersecting [Wang2011]
- There is no satisfactory algorithm for online manifold separation

[Wang2011] Yong Wang et al, "Spectral Clustering on Multiple Manifolds," Neural Networks, IEEE Transactions on , 2011

15

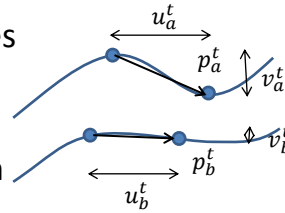
Online Motion Segmentation

- Given point trajectories up to time $t-1$
 - Track end points from $t-1$ to t
 - Update the solution from $t-1$
- General setting
 - Trajectories do not have the same length
 - Constant computation time, i.e. Computation time must be proportional to number of trajectories (Not the length)

16

Online Motion Segmentation

- Define a metric between trajectories that can be computed online
- Similar to [Brox2010]
- Capture similarity in spatial location and motion
- Motion measure



$$- d_M^t(T_a, T_b) = \frac{(u_a^t - u_b^t)^2}{(\sigma_{Mu}^t)^2} + \frac{(v_a^t - v_b^t)^2}{(\sigma_{Mv}^t)^2}$$

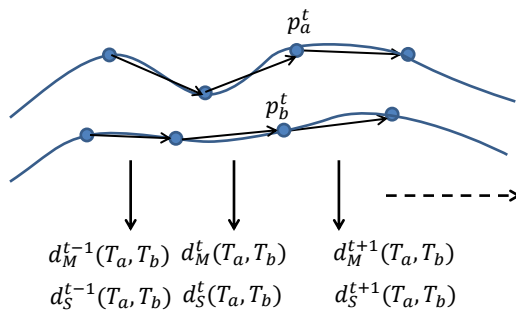
- Spatial Location measure

$$- d_S^t(T_a, T_b) = \|p_a^t - p_b^t\| / \sigma_S^2$$

[Brox2010] T. Brox and J. Malik. Object Segmentation by Long Term Analysis of Point Trajectories. (ECCV 2010)

17

Motion Segmentation Online Distance Computation

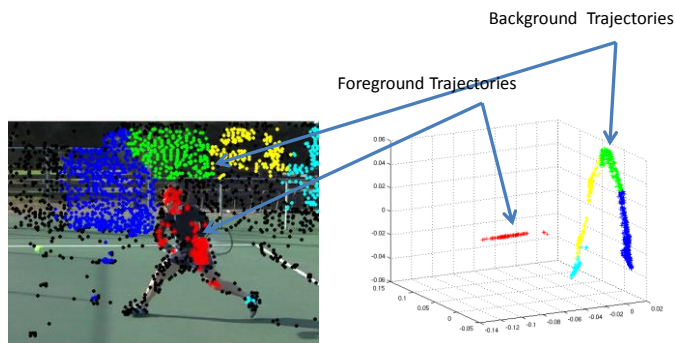


- $d_M^{1:t}(T_a, T_b) = \max(d_M^{1:t-1}, d_M^t)$
- $d_S^{1:t}(T_a, T_b) = \max(d_S^{1:t-1}, d_S^t)$
- $W = \exp(-\frac{D_M}{\lambda_M} + \frac{D_S}{\lambda_S})$

18

Online Motion Segmentation Explicit Modeling

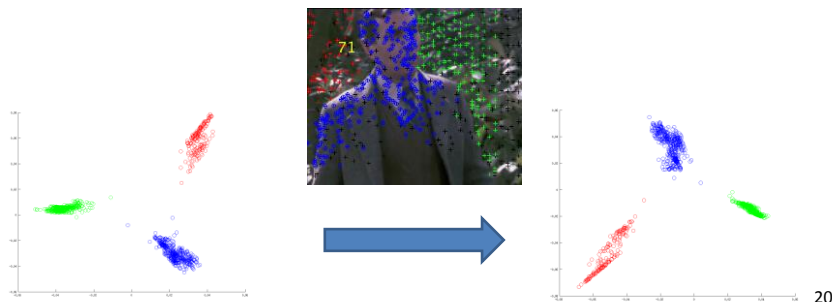
- Compute low dimensional representation using laplacian eigenmaps



19

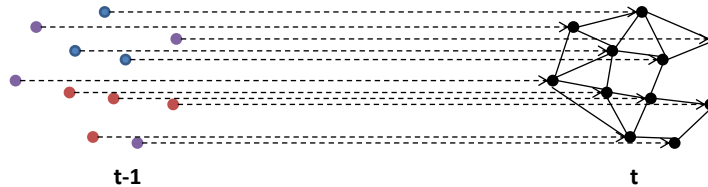
Online Motion Segmentation Explicit Modeling

- Coordinate free clustering
 - Low dimensional representation is in a different coordinate frame each frame
 - (First frame) Fit a Gaussian mixture model of R components
 - (Subsequent) Using previous trajectory assignment to clusters, re-estimate cluster parameters in new coordinate frame



20

Online Motion Segmentation Implicit Modeling



- Manifold structure is implicitly in the graph structure defined by affinity matrix.
- Use label propagation to “propagate” labels
- Minimize the energy function

$$Y^* = \underset{Y}{\operatorname{argmin}} \sum_{ij} w_{ij} \underbrace{(y_i^t - y_j^t)^2}_{\text{Manifold Structure}} + \sum_i \underbrace{(y_i^t - y_i^{t-1})^2}_{\text{Temporal Smoothness}}$$

21

Online Motion Segmentation Implicit Modeling

- In addition to propagating labels, must handle two cases
 - New evidence suggests that two sets of trajectories belong to the same moving object (Merge)
 - Handled automatically by label propagation
 - Strong affinities “links” between two sets causes one label to overcome the other
 - New evidence suggest that one set of trajectory belongs to two differently moving objects (Split)
 - Computing for each cluster the normalized min-cut
 - Split the cluster if the normalized cut cost is below a threshold

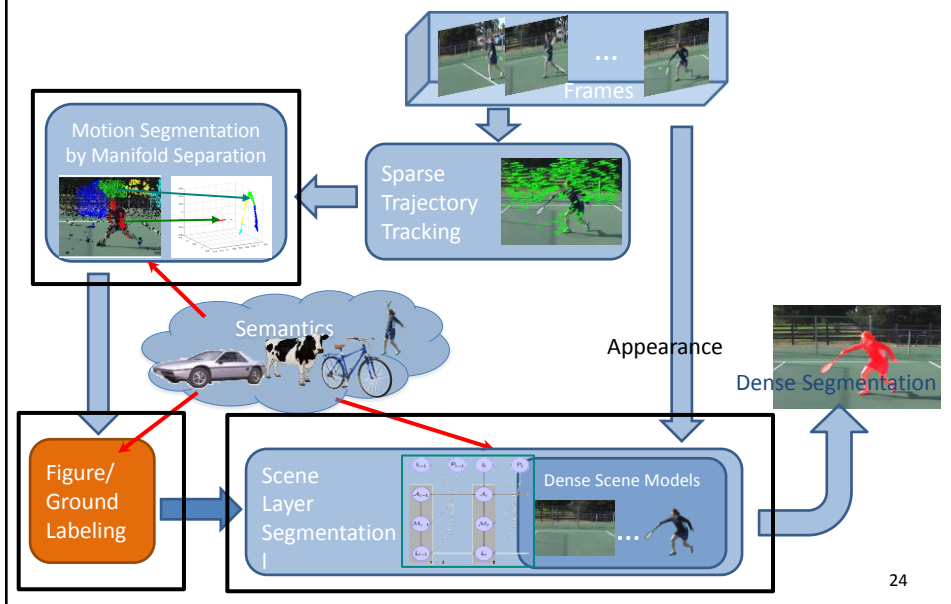
22

Online Motion Segmentation Results



23

Proposed Framework



24

What is Foreground/Background ?

- In stationary camera a common definition for foreground is anything that moves
- On a moving platform, however, everything moves !
- No clear definition for what comprises a foreground object in the moving-camera case
- Motivated by psychological evidence we propose using the following cues
 - Motion discontinuity
 - Appearance discontinuity
 - Surroundedness
 - Occlusion



Is this tree foreground or background ?

25

Figure Ground Labeling

- Segmenting the video into two segments; foreground and background.
- We propose two approaches based on aggregating multiple cues
 - Labeling motion trajectories [Elqursh-ECCV2012]
 - Labeling dense regions foreground/background[Elqursh-ICPR2012]

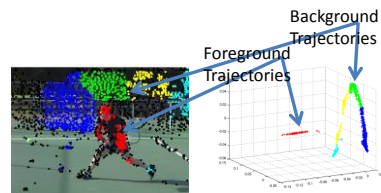
26

Figure Ground Labeling

$$E_l(L) = \alpha_C \sum_i \phi_C(l_i) + \alpha_A \sum_{(i,j)} \phi_A(l_i, l_j) + \alpha_B \sum_{(i,j)} \phi_B(l_i, l_j) + \alpha_S \phi_S(L)$$

Compactness
Affine Motion Compatibility
Boundary in embedding space
Surroundedness

- Energy function over labeling $L = \{l_1, \dots, l_R\}$, where $l_i \in \{0,1\}$ ($0 \equiv fg, 1 \equiv bg$)
- $R < 10$, optimal assignment found by enumerating all possible labels

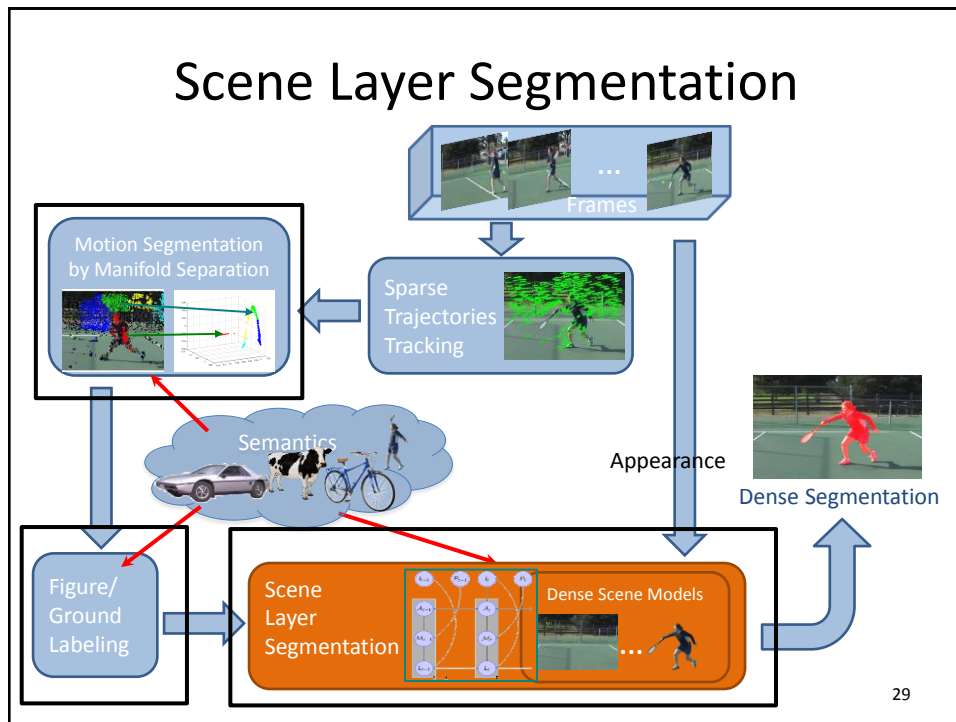


27

Figure Ground Labeling Results



28



Scene Layer Segmentation Related Work

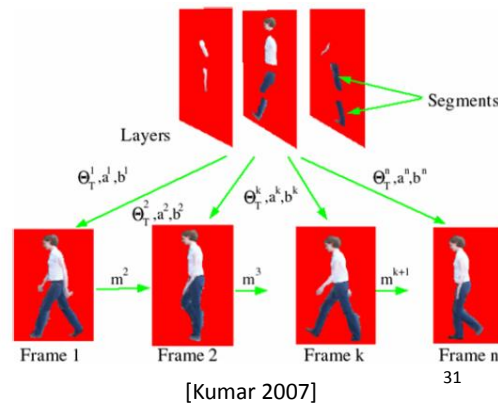
- When using two layers only it is related to
 - Background Subtraction from Moving Camera
- [Sheikh2009]
 - Orthographic motion segmentation over a sliding window to segment a set of trajectories.
 - This is followed by sparse per frame appearance modeling to densely segment images.
- [Kwak2011]
 - Maintains block based appearance models in a Bayesian framework.
 - Update the appearance models by iterating between estimating the motion of the blocks and inferring the labels of the pixels.
 - Once converged, the appearance models are used as priors for the next frame and the process continues.

[Sheikh2009] Sheikh, Y., Javed, O., & Kanade, T. Background Subtraction for Freely moving cameras. (ICCV2009)

[Kwak2011] Kwak, S., Lim, T., Nam, W., Han, B., & Hee, J. Generalized Background Subtraction Based on Hybrid Inference by Belief Propagation and Bayesian Filtering. (ICCV2011) 30

Scene Layer Segmentation Layered Models

- Assume scene is composed of layers
- Each layer has a depth value
- Image is formed by composing the layers
- [Kumar2007] layers composed of segments
- Each layer is transformed to produce the image
- Offline process
- Non-rigidity is approximated by piecewise rigid
- Finds a MAP estimate for most probable configuration
- Semantics: Segment concept
- No object concept



Online Scene Layer Segmentation Generative Model for Videos

- Layer representation allowed to deform over time
 - Enables handling of non-planar layers and non-rigid objects
- Each layer is defined by pixel wise appearance
- Bayesian approach (Maintain distribution over appearance and pixel labels)
- Pixels takes a color value from a distribution instead of only one color value

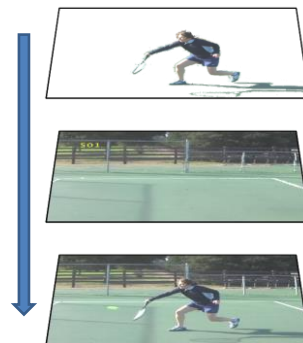
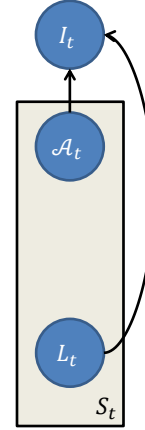


Image formation

- Model scene as two layers; background $A_{b,t}$ and foreground $A_{f,t}$
- Compose Image I_t from $A_{f,t}$ and $A_{b,t}$ according to L_t
- Assume independence in pixel appearance and labels
- Distributions of $a_{k,t}^i$ is represented non-parametrically using N_{KDE} samples
- For two layers, distribution of pixel label l_t^i is Bernoulli



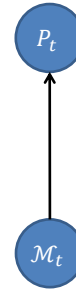
$$I_t = h(\mathcal{A}_t, L_t),$$

$$h_i(\mathcal{A}_t, L_t) = a_{k,t}^i + \epsilon_i, \quad k = l_t^i, \epsilon \sim \mathcal{N}(0, \Sigma_I)$$

33

Trajectory Generation

- Pixel-based motion model $M_{f,t}$ and $M_{b,t}$.
- $m_{k,t}^i = [u_{k,t}^i \ v_{k,t}^i]^T$ is the reverse motion of pixel i (time t to $t-1$)
- Labeled sparse motion vectors at frame t is a set $P_t = \{p_{j,t} : j = 1 \dots M\}$ of tuples $p_{j,t} = (q_{j,t}, w_{j,t}, l_{j,t}^p)$
 - q is the pixel location
 - $w_{j,t} = [u \ v]^T$ denotes the motion vector
 - $l_{j,t}^p = \{f, b\}$ denotes its layer



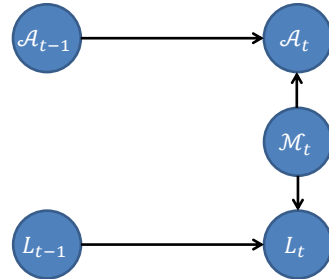
$$P_t = z(\mathcal{M}_t)$$

$$z_j(\mathcal{M}_t) = (j, m_k^j + \epsilon_p, k) \text{ for } j \in 1 \dots N, \quad \epsilon_p \sim \mathcal{N}(0, \Sigma_p)$$

34

Dynamic Bayesian Model

- Generate appearance at time t $A_{f,t}$ from $A_{f,t-1}$ and motion model $M_{f,t}$
- Similarly labels move according to the foreground motion



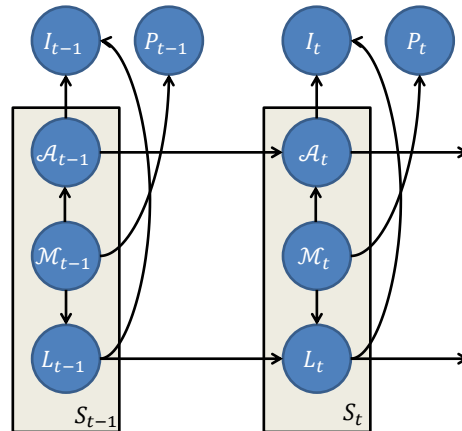
$$L_t = \Omega(M_{f,t}, L_{t-1}), \quad A_t = g(A_{t-1}, M_t),$$

$$\Omega_i(M_{f,t}, L_{t-1}) = l_{t-1}^{j(i,f)}, \quad g_k^i(A_{k,t-1}, M_{k,t}) = a_{k,t-1}^{j(i,k)},$$

35

Online Scene Layer Segmentation Generative Model for Videos

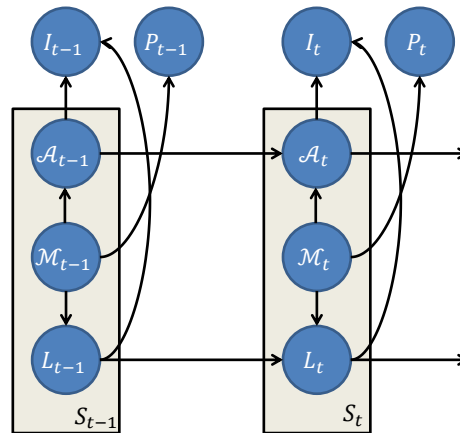
- Factored state S_t
- Observations O_t
 - Image I_t
 - Labeled Trajectories P_t
- Input: Distribution over
 - $A_{t-1} : P(A_{t-1} | O_{1:t-1})$
 - $L_{t-1} : P(L_{t-1} | O_{1:t-1})$
- Goal: Compute
 - $A_t : P(A_t | O_{1:t})$
 - $L_t : P(L_t | O_{1:t})$
- High dimensionality
 - Independence in pixel appearance
 - Multistep approach



36

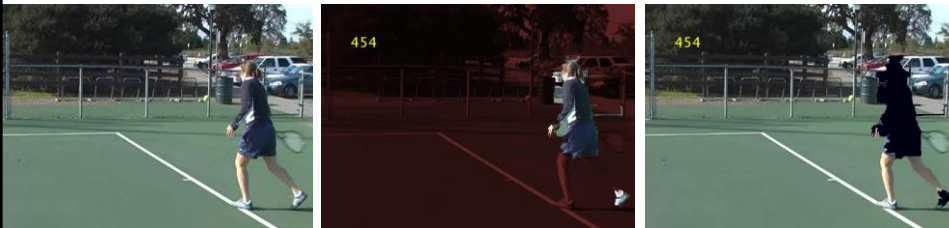
Online Scene Layer Segmentation Approach

- Given observed labeled trajectories P_t , estimate $P(M_t|P_t)$
- Prediction
 - Using M_t and A_{t-1} predict $A_t = P(A_t|P_{1:t}, I_{1:t-1})$
 - Using M_t and L_{t-1} predict $L_t = P(L_t|P_{1:t}, I_{1:t-1})$
- Update
 - Given Image I_t and A_t update L_t
 - $P(A_t|P_{1:t}, I_{1:t})$
 - $P(L_t|P_{1:t}, I_{1:t})$



37

Online Scene Layer Segmentation



38

Experiments

Dataset

- Berkley Motion Segmentation Dataset [Brox2010]
 - Consists of 26 sequences that include rigid and articulated motion.
 - Ground truth provided as frame annotations for 189 frames
 - 5 measures are
 - **Density:** Percentage of labeled trajectories to the total number of pixels.
 - **Overall Error:** Total number of correctly labeled trajectories over the number of labeled trajectories.
 - **Average clustering error :** Average of the ratio of mislabeled trajectories to the number of trajectories for each region.
 - **Over-segmentation error:** Number of segments merged to fit the ground truth regions.
 - **Lt10:** Number of regions covered with less than 10% error. One region subtracted to account for the background

39

Online Motion Segmentation Results

GPCA: [Vidal2004]
LSA: [Yan2006]

		Density	Overall Error	Average Error	Overseg	Lt10
First 10 frames (26 sequences)						
Offline	Ours	3.43%	9.69%	29.93%	0.31	21
	[2]	3.43%	7.49%	25.92%	0.46	20
	RANSAC	3.37%	14.4%	29.87%	0.73	13
	GPCA	3.37%	17.86%	28.64%	0.85	7
	LSA	3.37%	19.69%	39.76%	0.92	6
Frames 50 - 200 frames (7 sequences)						
Online	Ours	3.26%	6.77%	33.44%	2.57	6
	[2]	3.43%	8.32%	37.29%	3.14	6
	RANSAC	2.43%	28.3%	45.46%	1.42	0
All frames (26 sequences)						
	Ours	3.22%	9%	32.89%	2.30	16
	[2]	3.31%	6.82%	27.34%	1.77	27
	RANSAC	2.28%	16.04%	42.6%	1.15	9

Our is online
And more accurate !

Table 1. Evaluation results on the Berkley Dataset

[2] T. Brox and J. Malik. Object Segmentation by Long Term Analysis of Point Trajectories. (ECCV 2010)

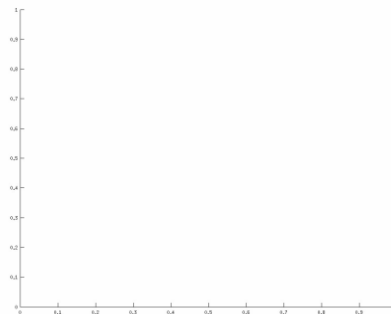
40

Online Motion Segmentation Results



41

Figure-Ground Labeling Explicit Modeling



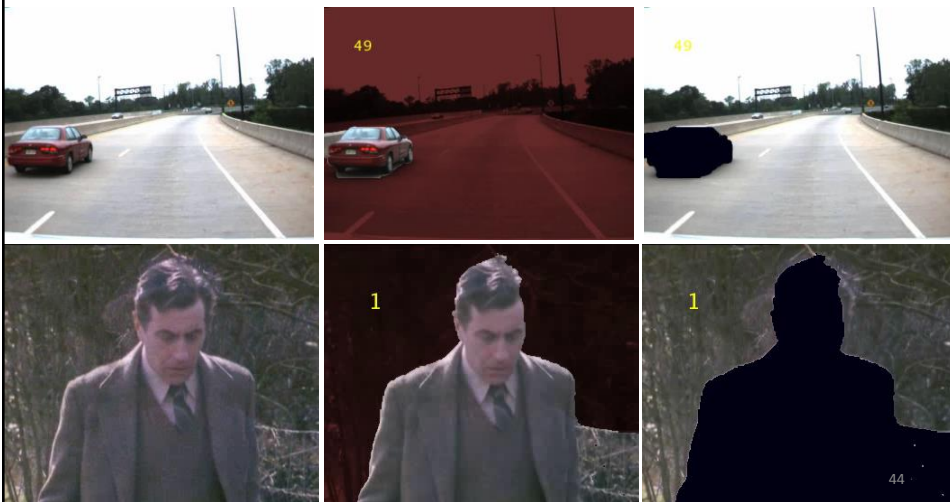
42

Scene Layer Segmentation Evaluation

- 6 sequences
 - 5 sequences from Berkley Motion Segmentation Dataset (cars1, people1, people2, tennis)
 - 1 driving sequence (drive1)
- Own implementation of [Sheikh2009]
- Comparison with results from [Kwak2011]

43

Scene Layer Segmentation Results





Scene Layer Segmentation Results

	cars1			people1			people2			tennis			drive		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Ours-1	0.84	0.99	0.91	0.94	0.85	0.89	0.69	0.88	0.77	0.86	0.92	0.89	0.55	0.96	0.70
Ours-2	0.85	0.97	0.90	0.97	0.88	0.92	0.87	0.88	0.88	0.90	0.81	0.85	0.60	0.67	0.63
[15]	0.63	0.99	0.77	0.78	0.63	0.70	0.73	0.83	0.78	0.27	0.83	0.40	0.02	0.66	0.04
[9]	0.92	0.84	0.88	0.95	0.93	0.94	0.85	0.89	0.86	-	-	-	-	-	-

- Ours-1 : Without using label prior
- Ours-2 : With label prior

[15] Sheikh, Y., Javed, O., & Kanade, T. Background Subtraction for Freely moving cameras. (ICCV2009)

[9] Kwak, S., Lim, T., Nam, W., Han, B., & Hee, J. Generalized Background Subtraction Based on Hybrid Inference by Belief Propagation and Bayesian Filtering. (ICCV2011)

46

Conclusion

- Motion Segmentation can be cast as Manifold Separation Problem
- Two approaches for Online Motion Segmentation via Manifold Separation
- Framework for Scene Layer Segmentation
 - Learns the appearance of different layers
 - Segments the video frame
- Show how multiple cues can be integrated in our Bayesian framework to achieve Figure Ground Labeling

47

Publications

- Ali Elqursh and Ahmed Elgammal. Online Motion Segmentation via Label Propagation (ICCV2013) Under Review
- Ali Elqursh and Ahmed Elgammal. Online Moving Camera Background Subtraction (ECCV2012)
- Ali Elqursh and Ahmed Elgammal. Video Figure Ground Labeling (ICPR2012)
- Ali Elqursh and Ahmed Elgammal. Single Axis Relative Rotation from Orthogonal Lines (ICPR2012)
- Ali Elqursh and Ahmed Elgammal. Line-Based Relative Pose Estimation (CVPR 2011)

48

Thank you !

49