JHU vision lab

Semantic (less) Motion and Video Segmentation

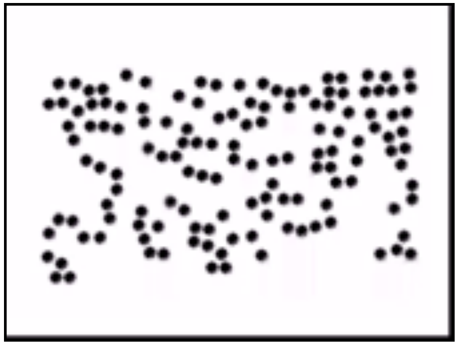
René Vidal
Johns Hopkins University

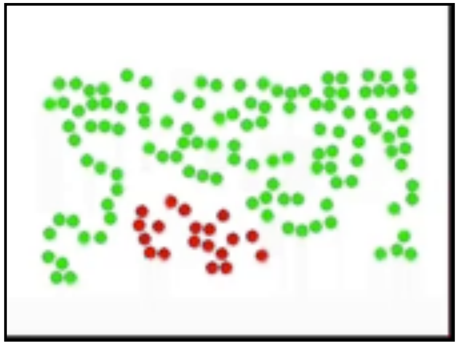
THE DEPARTMENT OF BIOMEDICAL ENGINEERING
The Whitaker Institute at Johns Hopkins

Center for Imaging Science

Talk Outline

- Semantic-less Motion Segmentation (Vidal et al., ECCV02, IJCV06; Vidal, Ma and Sastry CVPR03, PAMI05; Vidal and Sastry CVPR03; Vidal and Ma ECCV04, JMIV06; Vidal and Hartley, CVPR04; Tron and Vidal, CVPR07; Li et al. CVPR07; Goh and Vidal CVPR07; Vidal and Hartley, PAMI08; Vidal et al. IJCV08; Rao et al. CVPR 08, PAMI 09; Elhamifar and Vidal, CVPR 09)





- Coarse-to-Fine Semantic Video Segmentation (Jain et al. ICCV 2013)

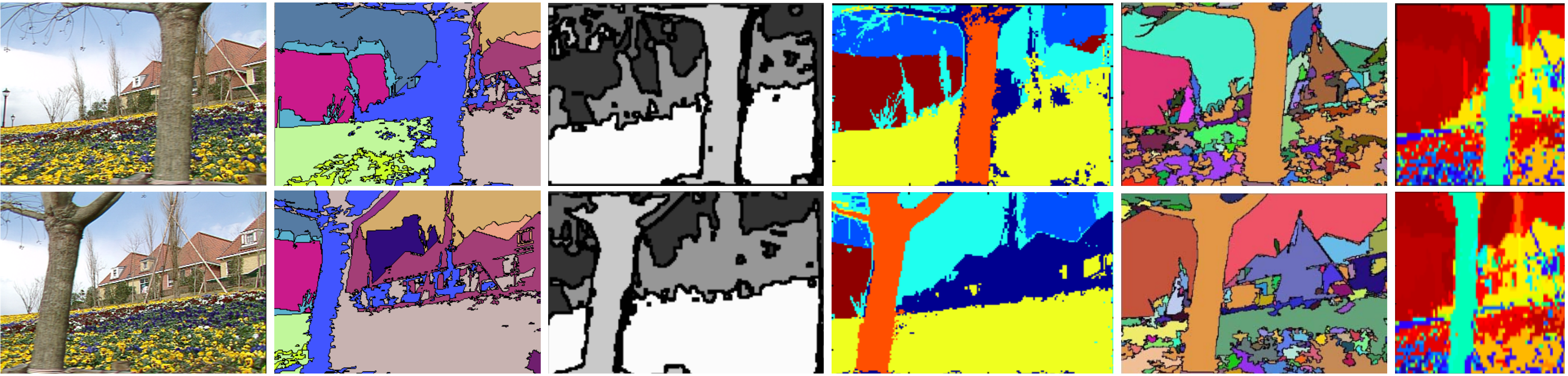
Part I
Semantic-less Motion Segmentation

E. Elhamifar, A. Goh, R.Tron, S. Rao, R. Hartley, Y. Ma, S. Soatto, S. Sastry
René Vidal
Johns Hopkins University

2D Motion Segmentation Problem

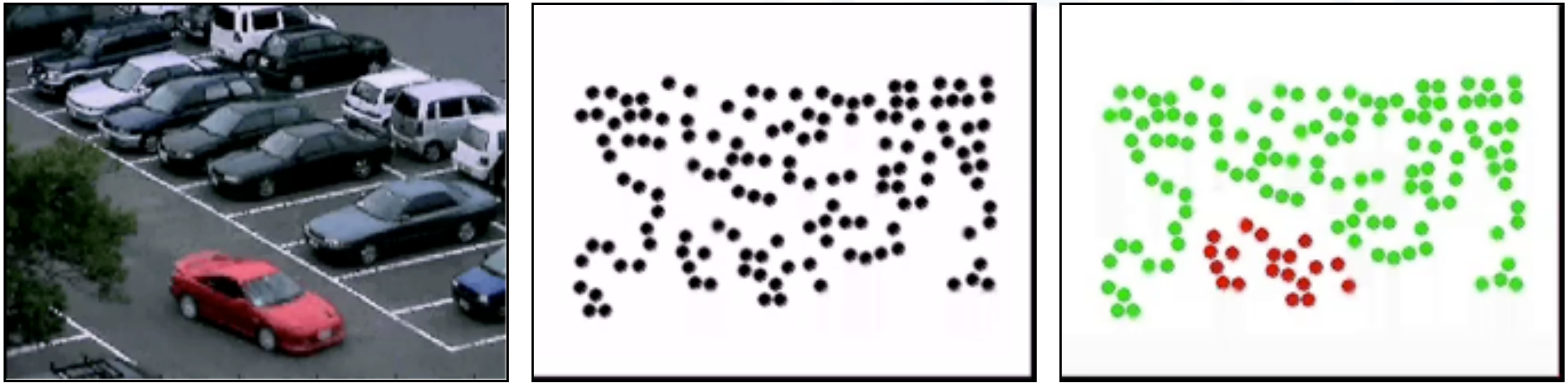
Prior Work on 2D Motion Segmentation

- Cluster locally estimated models (Wang-Adelson '93-'94)
- Fit one dominant motion at a time (Irani-Peleg '92)
- Fit a mixture model (Jepson-Black'93, Ayer-Sawhney '95, Darrel-Pentland'95, Weiss-Adelson'96, Weiss'97, Torr-Szeliski-Anandan '99, Khan-Sha'01)
- Apply normalized cuts to motion profile (Shi-Malik '98)



Original Grundman '10 Wang-Adelson'94 Khan-Shah'01 Brendel'09 Dementhon'02

3D Motion Segmentation Problem

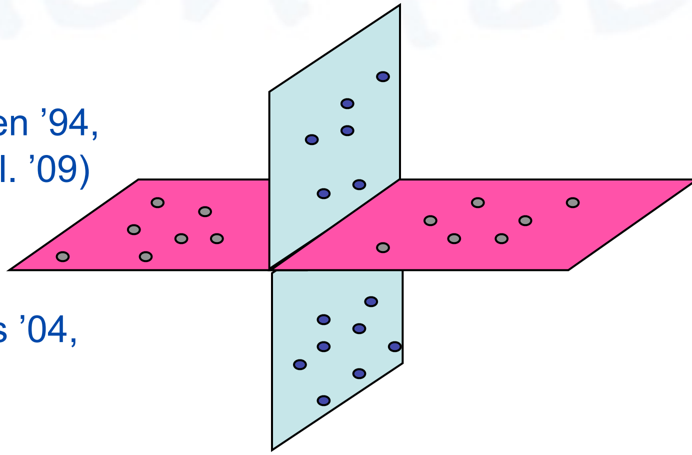


- Motion of a rigid-body lives in 3D affine subspace
  (Boult and Brown '91, Tomasi and Kanade '92)
  - P = #points
  - F = #frames

$$\mathtt{W} = \mathtt{M}\,\mathtt{S}^T$$

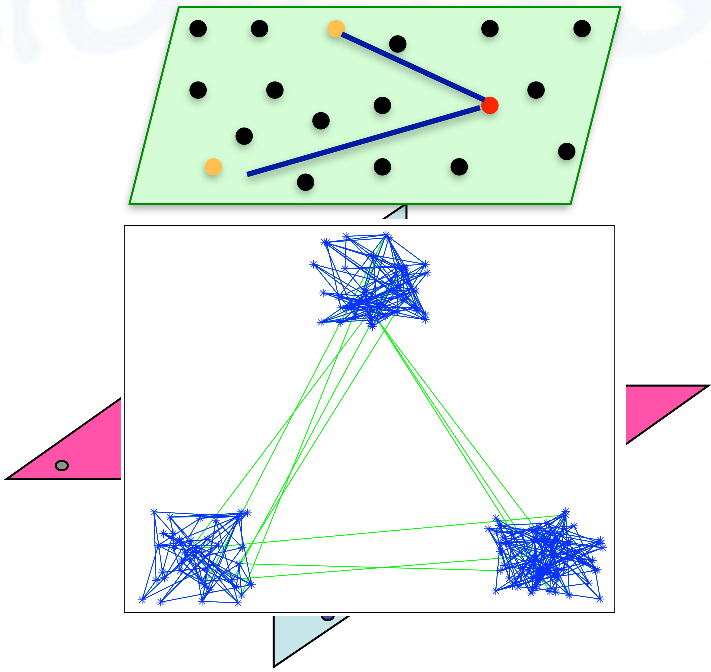
$$\underbrace{\begin{bmatrix} x_{11} \cdots x_{1P} \\ \vdots \qquad \vdots \\ x_{F1} \cdots x_{FP} \end{bmatrix}}_{2F\times P} = \underbrace{\begin{bmatrix} \mathtt{A}_1 \\ \vdots \\ \mathtt{A}_F \end{bmatrix}}_{2F\times 4} \underbrace{\begin{bmatrix} X_1 \cdots X_P \end{bmatrix}}_{4\times P}$$

Prior Work on 3D Motion Segmentation



- Iterative methods
  - K-subspaces (Bradley-Mangasarian ’00, Kambhatla-Leen ’94, Tseng’00, Agarwal-Mustafa ’04, Zhang et al. ’09, Aldroubi et al. ’09)
- Probabilistic methods
  - Mixtures of PPCA (Tipping-Bishop ’99, Grubber-Weiss ’04, Kanatani ’04, Archambeau et al. ’08, Chen ’11)
  - Agglomerative Lossy Compression (Ma et al. ’07, Rao et al. ’08)
  - RANSAC (Leonardis et al.’02, Yang et al. ’06, Haralik-Harpaz ’07)
- Algebraic methods
  - Factorization (Boult-Brown’91, Costeira-Kanade’98, Gear’98, Kanatani et al.’01, Wu et al.’01)
  - Generalized PCA: (Shizawa-Maze ’91, Vidal et al. ’03 ’04 ’05, Huang et al. ’05, Yang et al. ’05, Derksen ’07, Ma et al. ’08, Ozay et al. ‘10)
- Spectral clustering-based methods (Zelnik-Manor ’03, Yan-Pollefeys ’06, Govindu ’05, Agarwal et al. ’05, Fan-Wu ’06, Goh-Vidal ’07, Chen-Lerman ’08, Elhamifar-Vidal ’09 ’10, Lauer-Schnorr ’09, Zhang et al. ’10, Liu et al. ’10, Favaro et al. ’11, Candes ’12)

How to Define a Good Subspace Affinity?



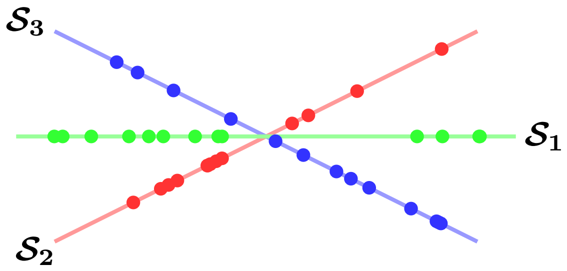
- Spectral clustering
  - Represent points as nodes in graph G
  - Connect points i and j with weight c_{ij}
  - Infer clusters from Laplacian of G
- Good affinity matrix C for subspaces?
  - $c_{i,j} = \exp(-d^2(\boldsymbol{y}_i, \boldsymbol{y}_j))$
  - Points in the same subspace: $c_{ij} \neq 0$
  - Points in different subspaces: $c_{ij} = 0$
- Challenge: cannot define a pairwise affinity
- Multiway affinity based on d+1 or d+2 points (Chen-Lerman '08)
- Affinity based on angles between local subspaces (Yan-Pollefeys '06)

Sparse Subspace Clustering (SSC)

- Data in a union of subspaces are self-expressive

$$\boldsymbol{y}_i = \sum_{j=1}^{N} c_{ji} \boldsymbol{y}_j \implies \boldsymbol{y}_j = Y \boldsymbol{c}_i \implies Y = YC$$

- Data in a union of subspaces admit a subspace-sparse representation



- The affinity can be constructed using L1 minimization

$$P_1 : \min \|\boldsymbol{c}_i\|_1 \quad \text{s.t.} \quad \boldsymbol{y}_i = Y \boldsymbol{c}_i, \quad c_{ii} = 0$$

E. Elhamifar and R. Vidal. Sparse Subspace Clustering. CVPR 2009.
E. Elhamifar and R. Vidal. Clustering Disjoint Subspaces via Sparse Representation. ICASSP 2010.
E. Elhamifar and R. Vidal. Sparse Subspace Clustering: Algorithm, Theory and Applications. TPAMI 2013.

Hopkins 155 motion segmentation database

- Collected 155 sequences (Tron-Vidal ‘07)
  - 120 with 2 motions
  - 35 with 3 motions
- Types of sequences
  - Checkerboard sequences: mostly full dimensional and independent motions
  - Traffic sequences: mostly degenerate (linear, planar) and partially dependent motions
  - Articulated sequences: mostly full dimensional and partially dependent motions
- Point correspondences
  - In few cases, provided by Kanatani & Pollefeys
  - In most cases, extracted semi-automatically with OpenCV

R. Tron and R. Vidal. A Benchmark for the Comparison of 3-D Motion Segmentation Algorithms. CVPR 2007.

Results on the Hopkins 155 database

- 2 motions, 120 sequences, 266 points, 30 frames

	GPCA	LLMC	LSA	RANSAC	MSL	SCC	ALC	SSC
Checkerboard	6.09	3.96	2.57	6.52	4.46	1.30	1.55	**1.12**
Traffic	1.41	3.53	5.43	2.55	2.23	1.07	1.59	**0.02**
Articulated	2.88	6.48	4.10	7.25	7.23	3.68	10.70	**0.62**
All	4.59	4.08	3.45	5.56	4.14	1.46	2.40	**0.82**

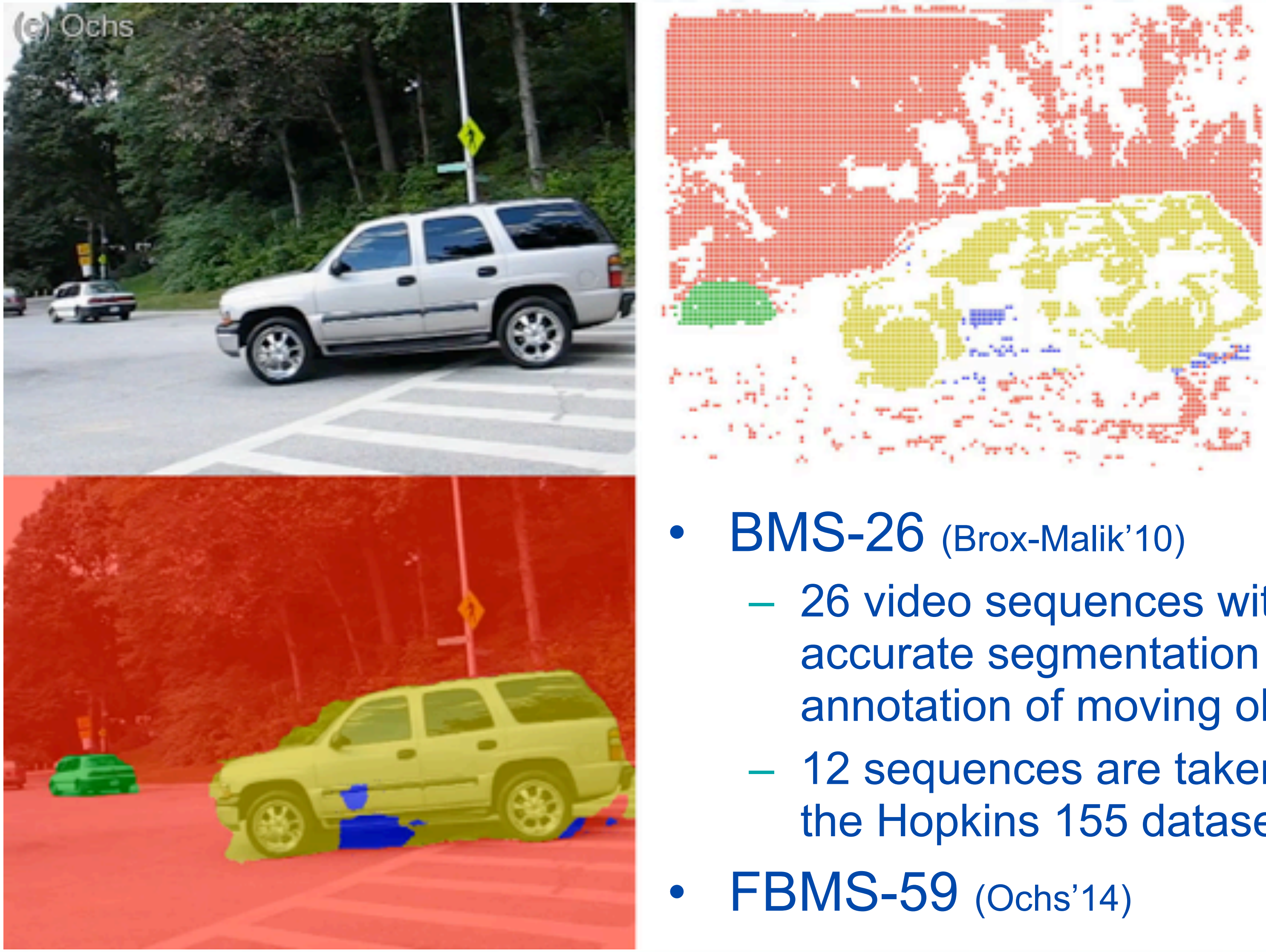
- 3 motions, 35 sequences, 398 points, 29 frames

	GPCA	LLMC	LSA	RANSAC	MSL	SCC	ALC	SSC
Checkerboard	31.95	8.48	5.80	25.78	10.38	5.68	5.20	**2.97**
Traffic	19.83	6.04	25.07	12.83	1.80	2.35	7.75	**0.58**
Articulated	16.85	9.38	7.25	21.38	2.71	10.94	21.08	**1.42**
All	28.66	8.04	9.73	22.94	8.23	5.31	6.69	**2.45**

- All

	GPCA	LLMC	LSA	RANSAC	MSL	SCC	ALC	LRR	LRSC	SSC
All	10.34	4.97	4.94	9.76	5.03	2.33	3.37	3.16	3.28	**1.24**

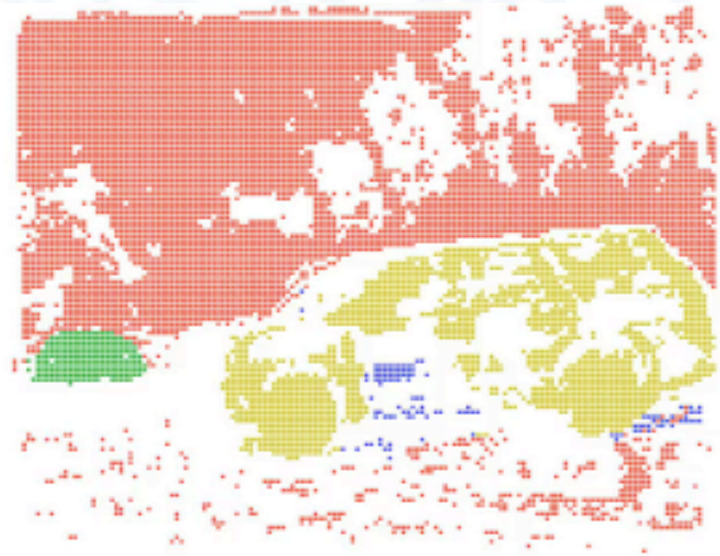
Dense 3D Motion Segmentation



- BMS-26 (Brox-Malik'10)
  - 26 video sequences with pixel-accurate segmentation annotation of moving objects
  - 12 sequences are taken from the Hopkins 155 dataset
- FBMS-59 (Ochs'14)

T. Brox, J. Malik Object segmentation by long term analysis of point trajectories, ECCV 2010
P. Ochs and T. Brox. Higher Order Motion Models and Spectral Clustering. CVPR, 2012
P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis, PAMI 2014

Dense 3D Motion Segmentation



- Sparse trajectory clustering:
  - Spectral clustering based on pairwise motion affinities
- Dense segmentation
  - Variational approach based on color, texture, etc.

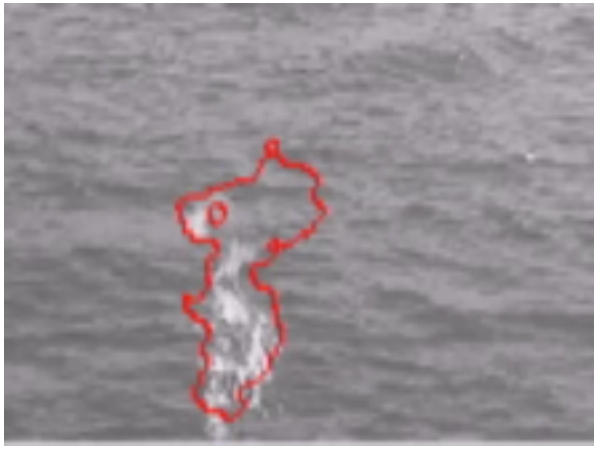
T. Brox, J. Malik Object segmentation by long term analysis of point trajectories, ECCV 2010
P. Ochs and T. Brox. Higher Order Motion Models and Spectral Clustering. CVPR, 2012
P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis, PAMI 2013

Future Vistas in 3D Motion Segmentation

- Good progress in the last decades
  - Sparse trajectories
  - Complete trajectories
  - Short videos
  - Affine cameras

- Ongoing and future directions
  - Dense trajectories
  - Incomplete and corrupted trajectories
  - Appearing and disappearing objects
  - Longer videos
  - Static objects
  - Deformable objects
  - Strong perspective effects

(Torr et al. '98, Shashua et al. '00, '01, '02, Vidal et al. '02, '06, '07)



(Doretto'03, Chan'05, '09, Ghoreyshi-Vidal'06)

Coarse-to-fine Semantic Video Segmentation Using Supervoxel Trees

Aastha Jain
LinkedIn

Shaunak Chatterjee
UC Berkeley

René Vidal
Johns Hopkins

Semantic Video Segmentation Problem

- Given a video sequence, assign a class label to each pixel

SUNY Dataset. Chen et al. Propagating multi-call pixel labels throughout video frames, WNYIPW 2010

Center for IMAGING SCIENCE

Computational Challenges

$$\left.\begin{aligned} V &= \text{number of supervoxels} \\ L &= \text{number of labels} \end{aligned}\right\} O(L^V) \text{ possible segmentations}$$

- Existing energy minimization approaches trade-off accuracy for efficiency by finding an approximate solution
  - Graph cuts [Boykov et al. TPAMI01]
  - Belief propagation [Felzenszwalb-Huttenlocher IJCV06]
  - Hierarchical graph cuts [Kumar UIA09]

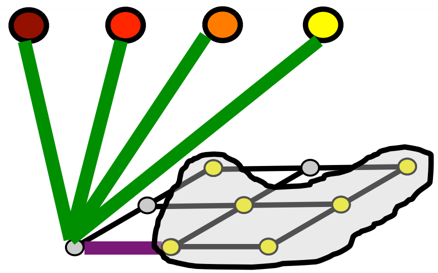
- While successful for many tasks in image segmentation, these approximate methods continue to be very slow for applications in video segmentation

- How to perform efficient semantic video segmentation?

Proposed Approach

- Observations
  - Real videos are spatially and temporally coherent
  - Set of coherent labelings is much smaller than the set of all labelings
- Approach
  - Construct a hierarchy of supervoxels
  - Propose a coarse-to-fine energy minimization strategy
- Advantages
  - Exact: it gives the same solution as minimizing over the finest graph
  - General: it can be used with any supervoxel hierarchy and any energy minimization algorithm to minimize any energy function
  - Efficient: it gives 2x-10x speedup for several datasets with varying degrees of spatio-temporal coherence

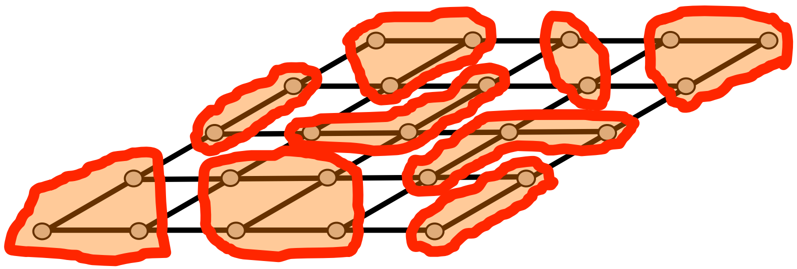
Energy Minimization Problem



object categories

$$l \in \mathcal{L} = \{1, \ldots, L\}$$

labels: $x_i \in \mathcal{L}$



supervoxels

$$E(x) = \lambda_U \sum_{v_i \in \mathcal{V}} \psi_i^U(x_i, V) + \lambda_P \sum_{e_{ij} \in \mathcal{E}} \psi_{i,j}^P(x_i, x_j, V) + \lambda_H \sum_{c \in \mathcal{C}} \psi_c^H(x_c, V)$$

- $\psi_i^U(l, I)$: cost of assigning label l to *supervoxel* i
- $\psi_{ij}^P(l_1, l_2, I)$: cost of assigning labels l_1 and l_2 to *supervoxels* i and j
- $\psi_c^H(x_c, I)$: label consistency cost for clique $c \in \mathcal{C}$

Superpixel computation: Ren CVPR03, Felzenszwalb IJCV04, Levinshtein TPAMI09, Vedaldi ECCV08, Veksler ECCV10, Achanta TPAMI12

Energy design: Winn CVPR06, Shotton CVPR08, Shotton IJCV09, Rabinovich CVPR07, Fulkerson ICCV09, Micusik ICCVW09, Ladicky ICCV09, Russell ECCV10, Vijayanarasimhan POCV09, Larlus CVPR08, Verbeek NIPS08, Gould NIPS08, Yang CVPR10

Energy minimization: Boros DAM02, Boykov TPAMI01, Kolmogorov TPAMI04, Kohli CVPR08

Hierarchy of Supervoxels

- Supervoxel Based Methods [Xu and Corso CVPR12]
  - SWA [Sharon CVPR00], Graph Based [Felzenszwalb IJCV04], Hierarchical [Grundmann CVPR10], Mean Shift [Paris CVPR07], Nystom [Fowlkes TPAMI04]

Original image

Level 5(coarsest)

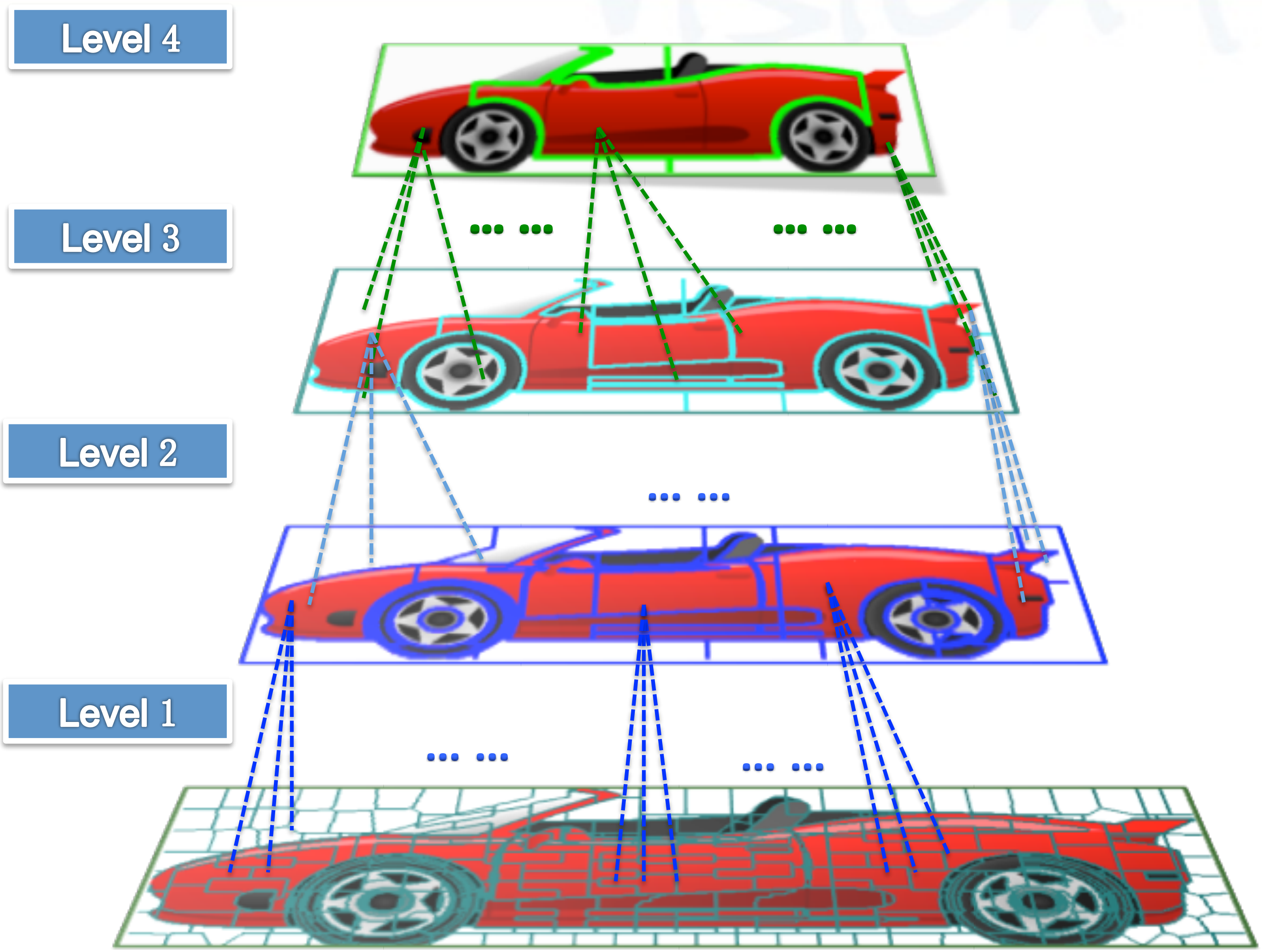
Level 4

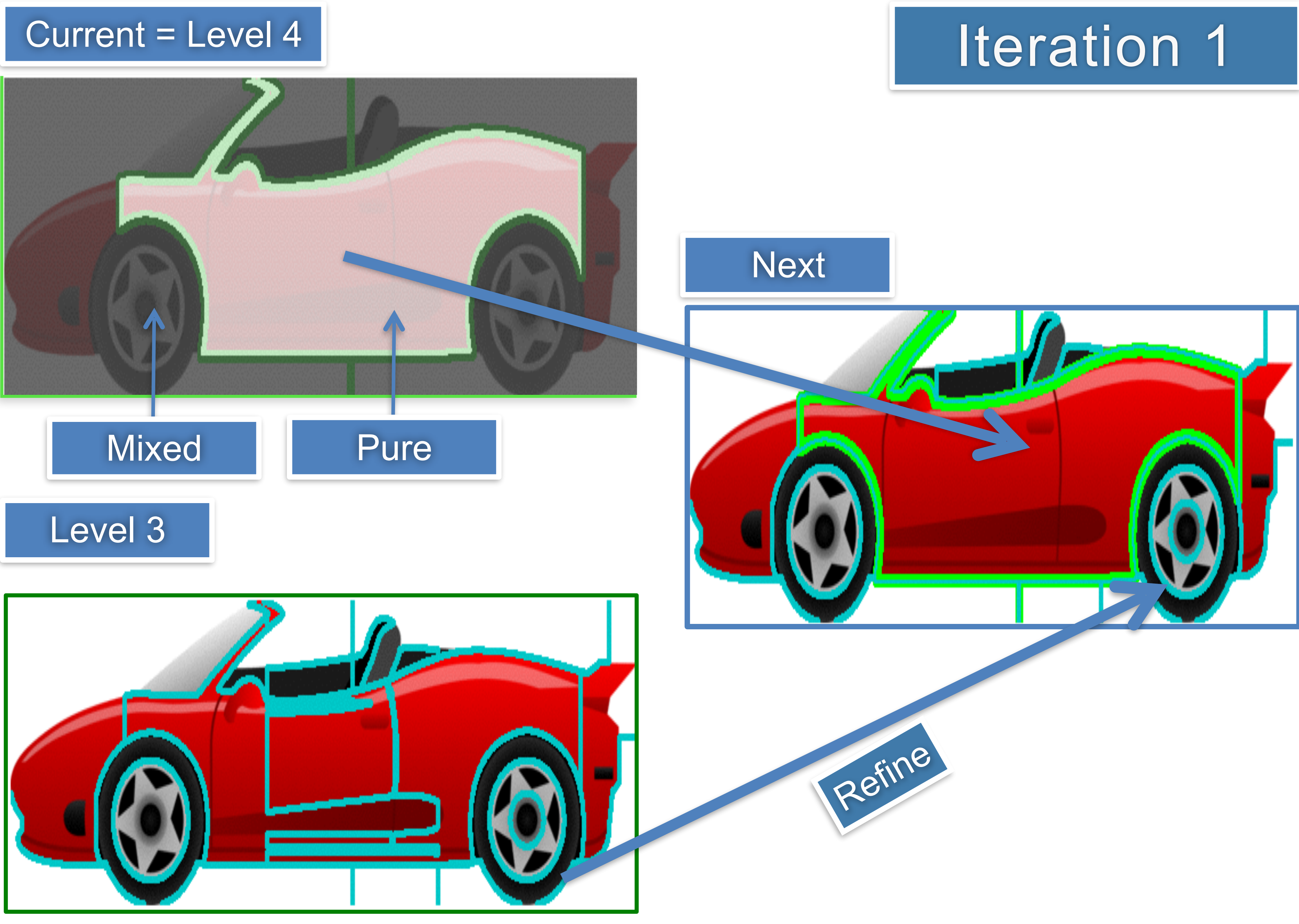
Level 3

Level 2

Level 1 (finest)

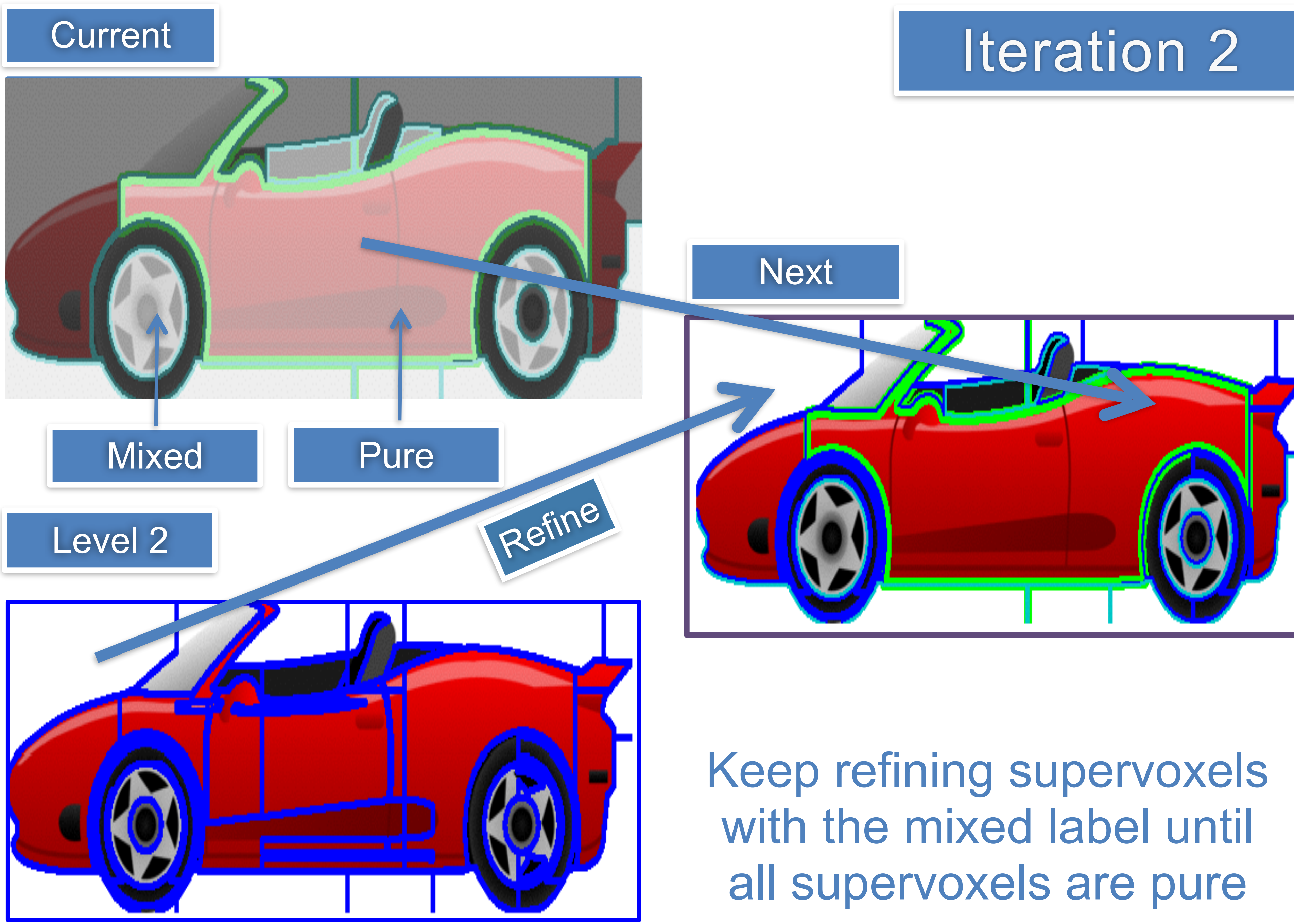
Coarse-to-Fine Energy Minimization





Current = Level 4
Iteration 1
Next
Mixed
Pure
Level 3
Refine

Iteration 2



Keep refining supervoxels with the mixed label until all supervoxels are pure

Exactness of the Coarse-to-Fine Solution

Algorithm 1 Coarse-to-fine Inference Algorithm $(\mathcal{V}^{1:m}, \psi)$

1: $\mathcal{V}^{curr} \leftarrow \mathcal{V}^m$
2: **repeat**
3: Find $x_{\mathcal{V}^{curr}}$ which minimizes $E_{\mathcal{V}^{curr}}$
4: **for all** ${v_i}^j \in \mathcal{V}^{curr}$ such that ${x_i}^j = L+1$ **do**
5: Refine ${v_i}^j$
6: $\mathcal{V}^{curr} \leftarrow \mathcal{V}^{curr} \cup \mathcal{R}(i,j,j-1) \setminus {v_i}^j$
7: **end for**
8: **until** $L+1 \notin x_{\mathcal{V}^{curr}}$
9: **return** $x_{\mathcal{V}^{curr}}$

- **Theorem.** If the coarse potentials in $E_{\mathcal{V}^{curr}}$ are lower bounds of their constituent exact potentials, the set of minimizers of the coarse-to-fine procedure (with algorithm *A* in step 3) is the same as that of running algorithm *A* at the finest level

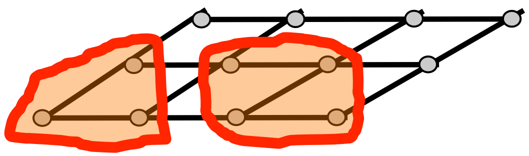
Chatterjee and Russel. A temporally abstracted Viterbi algorithm, UAI11.
Finley and Joachims Training Structural SVMs when Exact Inference is Intractable, 2008.

Construction of the Coarse Potentials

- Consider the energy at the finest level (level 1)

$$E(x) = \lambda_U \sum_{v_i \in \mathcal{V}} \psi_i^U(x_i, V) + \lambda_P \sum_{e_{ij} \in \mathcal{E}} \psi_{i,j}^P(x_i, x_j, V) + \lambda_H \sum_{c \in \mathcal{C}} \psi_c^H(x_c, V)$$

- Unary cost for a *coarse supervoxel* at level *j*
  - Pure label: sum of the unary costs of constituent supervoxels at level 1

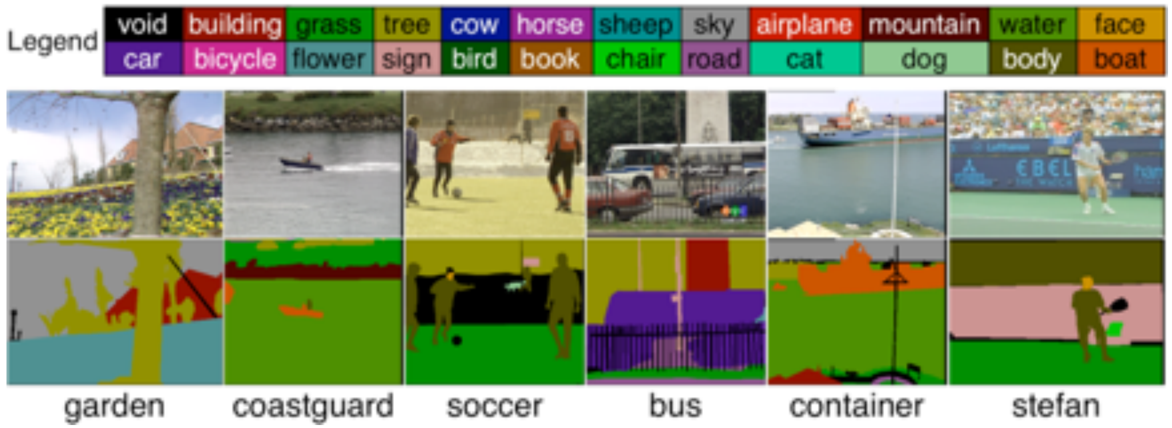


  - Mixed label: minimum cost over constituent supervoxels at level 1 subject to all the constituent supervoxels not getting the same label

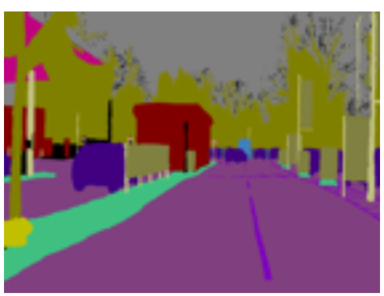
- Pairwise cost
  - Pure label: sum of the pairwise costs of the edges connecting the constituent supervoxels
  - Mixed label: zero

Experiments: Datasets

- SUNY
  - 24 classes, 2 in each video, 70 training frames, 100 testing frames



- CamVid
  - 11 classes, 100 training frames, 100 testing frames



Void	Building	Wall	Tree	VegetationMisc	Fence
Sidewalk	ParkingBlock	Column_Pole	TrafficCone	Bridge	SignSymbol
Misc_Text	TrafficLight	Sky	Tunnel	Archway	Road
RoadShoulder	LaneMkgsDriv	LaneMkgsNonDriv	Animal	Pedestrian	Child
CartLuggagePram	Bicyclist	MotorcycleScooter	Car	SUVPickupTruck	Truck_Bus
Train	OtherMoving				

Experiments: Quantitative Results

- Time taken by the different inference algorithms (in minutes)

Algorithm		CamVid					SUNY		
		CamVid1	CamVid2	CamVid3	CamVid4	CamVid5	Bus	Football	Ice
GC	Flat	130.1	137.3	117.6	145.1	140.1	35.3	25.0	32.7
	Coarse-to-fine	32.7	40.9	27.3	43.8	29.4	6.5	2.3	5.3
BP	Flat	256.0	270.1	258.3	307.0	319.2	50.3	34.7	50.9
	Coarse-to-fine	50.5	79.1	61.5	107.7	90.5	9.3	4.1	8.3

- Computational speedup
  - CamVid: 3x-5x (2x-4x with time to compute hierarchy)
  - SUNY: 7x-10x (5x-6x with time to compute hierarchy)
- Percentage of time spent on bound computation
  - Graph cut: 40-50%
  - Belief propagation: 20-25%

Experiments: Qualitative Results

- Reduced problem size

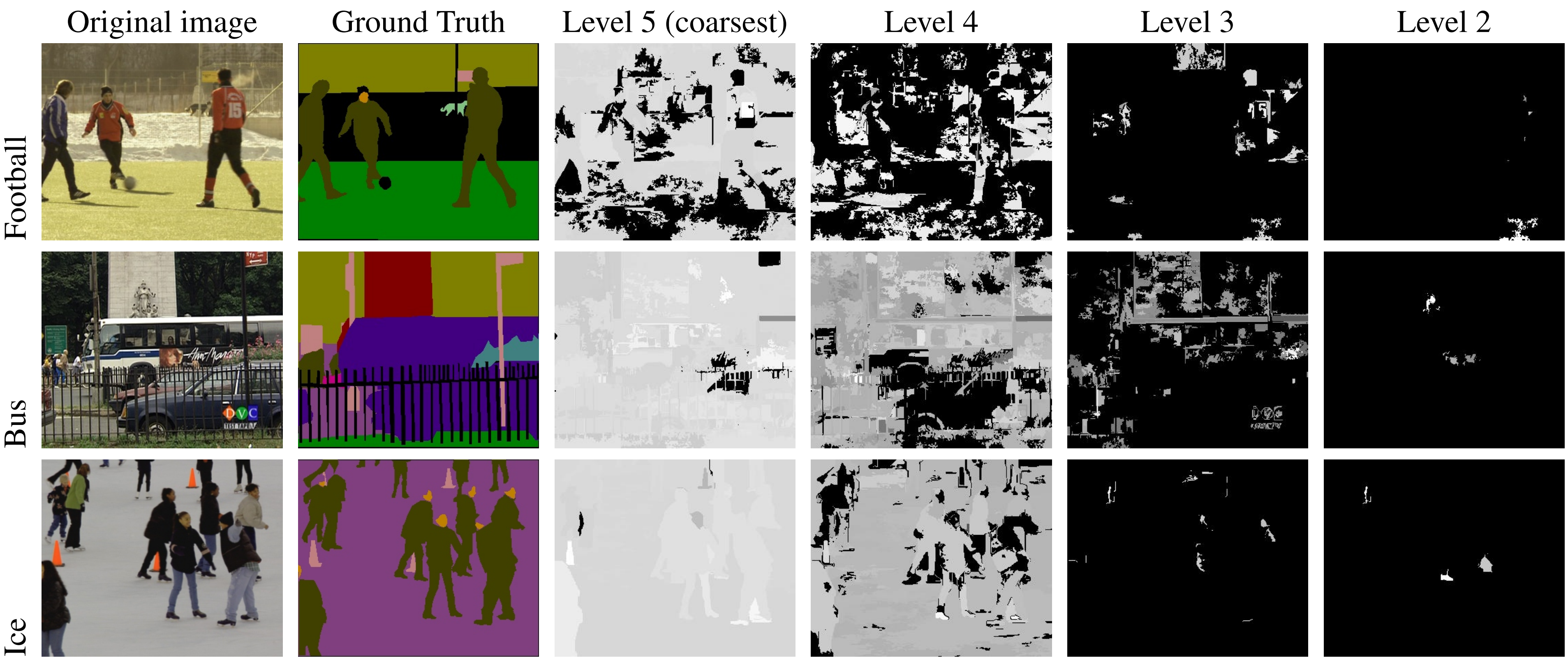


Figure 2. Explored portions of the supervoxel tree. The blacked out portions in each superpixel level denotes the patch of superpixels which were never refined during inference. The top row shows results from the "football" video, the middle row from the "bus" video and the bottom row from the "ice" video (all from the SUNY dataset).

Experiments: Qualitative Results

- Segmentation accuracy versus number of refinement cycles

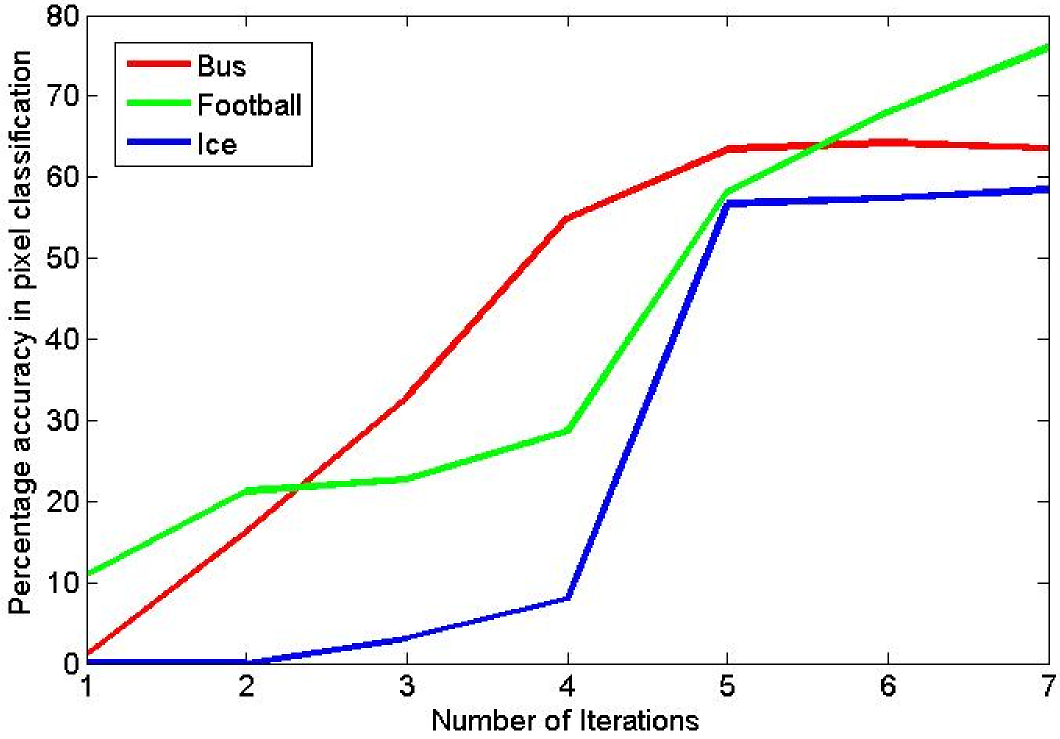


Figure 3. Percentage of correctly classified supervoxels after every iteration of the coarse-to-fine belief propagation algorithm.

Discussion

- An exact, general and efficient coarse-to-fine energy minimization strategy for semantic video segmentation
  - It produces the same set of solutions as minimizing over the finest graph
  - It can be used with several energy minimization and hierarchy construction algorithms
  - It gives a 2x-10x speedup relative to flat algorithm
- Advances in energy minimization or hierarchy construction algorithms will only improve the efficiency of our framework

Thank You!

Vision Lab @ Johns Hopkins University

http://www.vision.jhu.edu

Center for Imaging Science @ Johns Hopkins University

http://www.cis.jhu.edu