Correspondence Estimation in Images: New Techniques and Applications

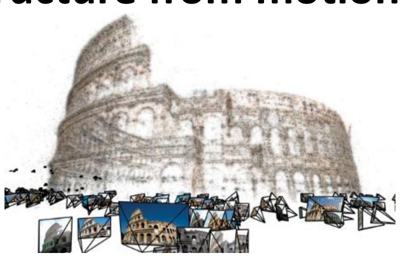
Sudipta N. Sinha

Microsoft Research

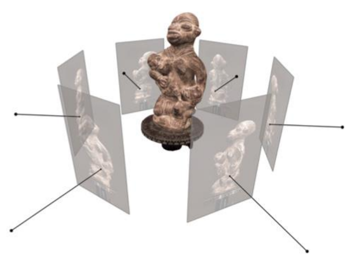
Correspondence Estimation in Computer Vision

2-view		n-view
rigid	vs.	non-rigid
sparse		dense

Structure from motion

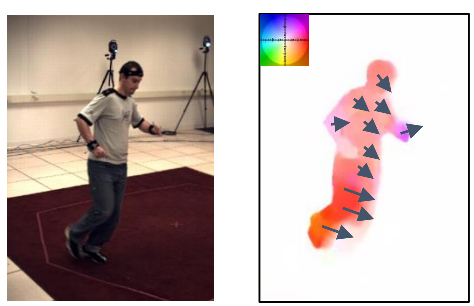


Multiview stereo



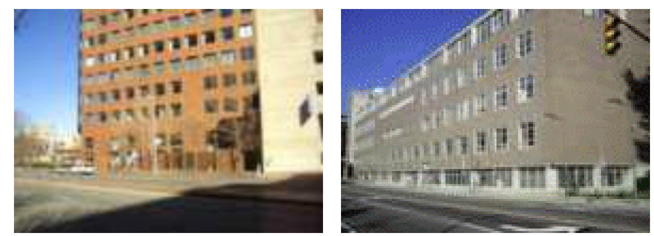
Binocular Stereo

Optical flow



<u>Different scenes</u>

SIFT Flow (Liu et al. 2008)

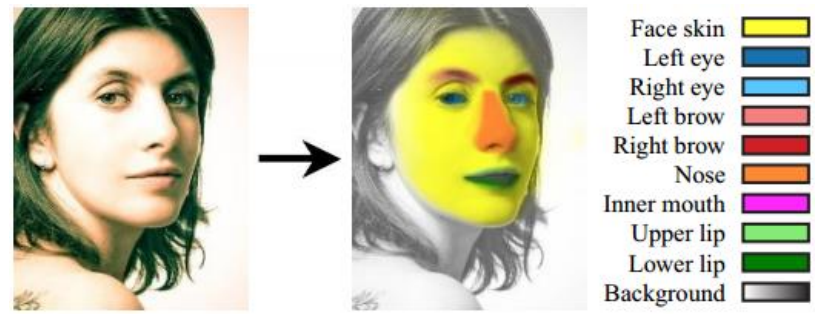


Deformable Spatial Pyramid Matching (Kim et al. 2013)

More applications

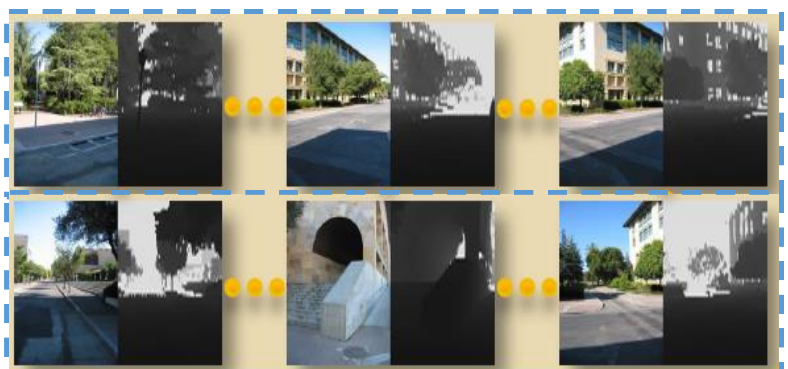
Label transfer (Face parsing) [Smith et al 2013]

Labeled images

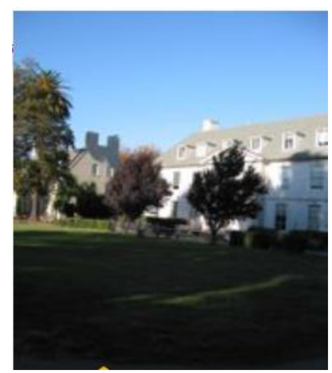


Input Output Label Types

Depth transfer [Karsch et al. 2012]



RGB-D database



Query

Predicted Depth

Overview

- Dense Correspondence Estimation
  - Surface Stereo
  - High Resolution Stereo Matching
- Joint Correspondence and Cosegmentation
  - Align different object instances
- Sparse Correspondences and Applications
  - Improved place recognition
  - Color Consistency in Photo Collections

Surface-based stereo

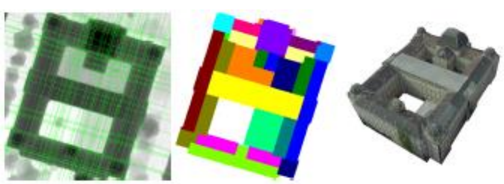
- Piecewise planar stereo

Birchfield and Tomasi 2001

Furukawa et al. 2008

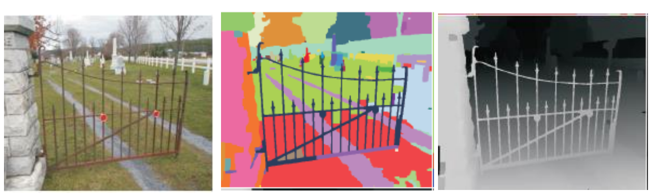
Sinha et al. 2009

- Surface stereo



Zebedin et al. 2008

Gallup et al. 2010

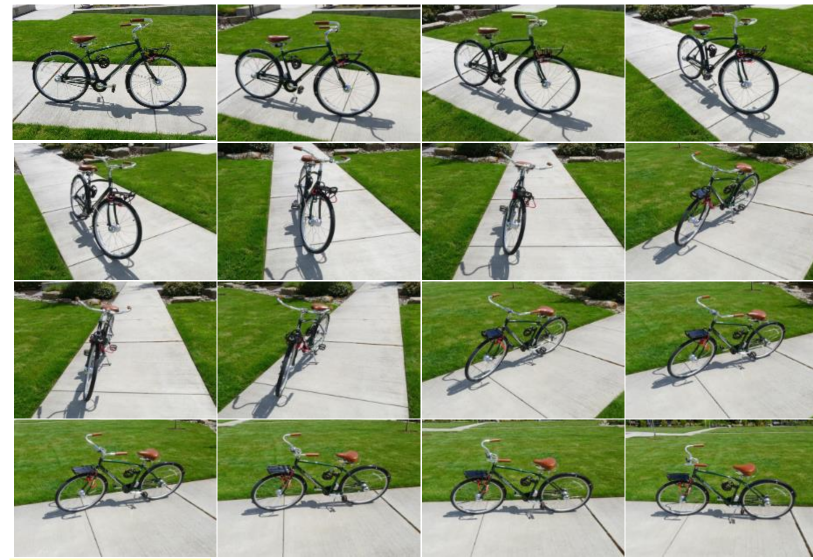


Bleyer et al. 2010, 2011

Multiple View Object Cosegmentation using Appearance and Stereo Cues

Kowdle, Sinha and Szeliski (ECCV 2012)

Input Images

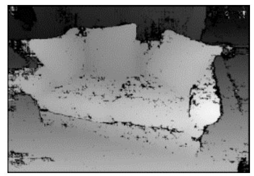


Final Results

Multiple View Object Cosegmentation using Appearance and Stereo Cues

Kowdle, Sinha and Szeliski (ECCV 2012)

Input images

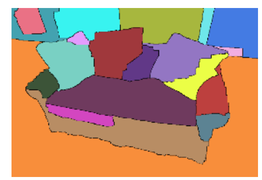


Stereo matching

Piecewise planar stereo

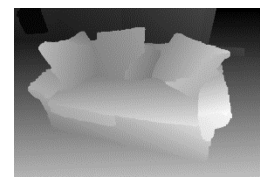
- Learn per-plane local color models
- Combine stereo and color cues
- Accurate occlusion boundaries

Plane hypotheses



Plane labels

+



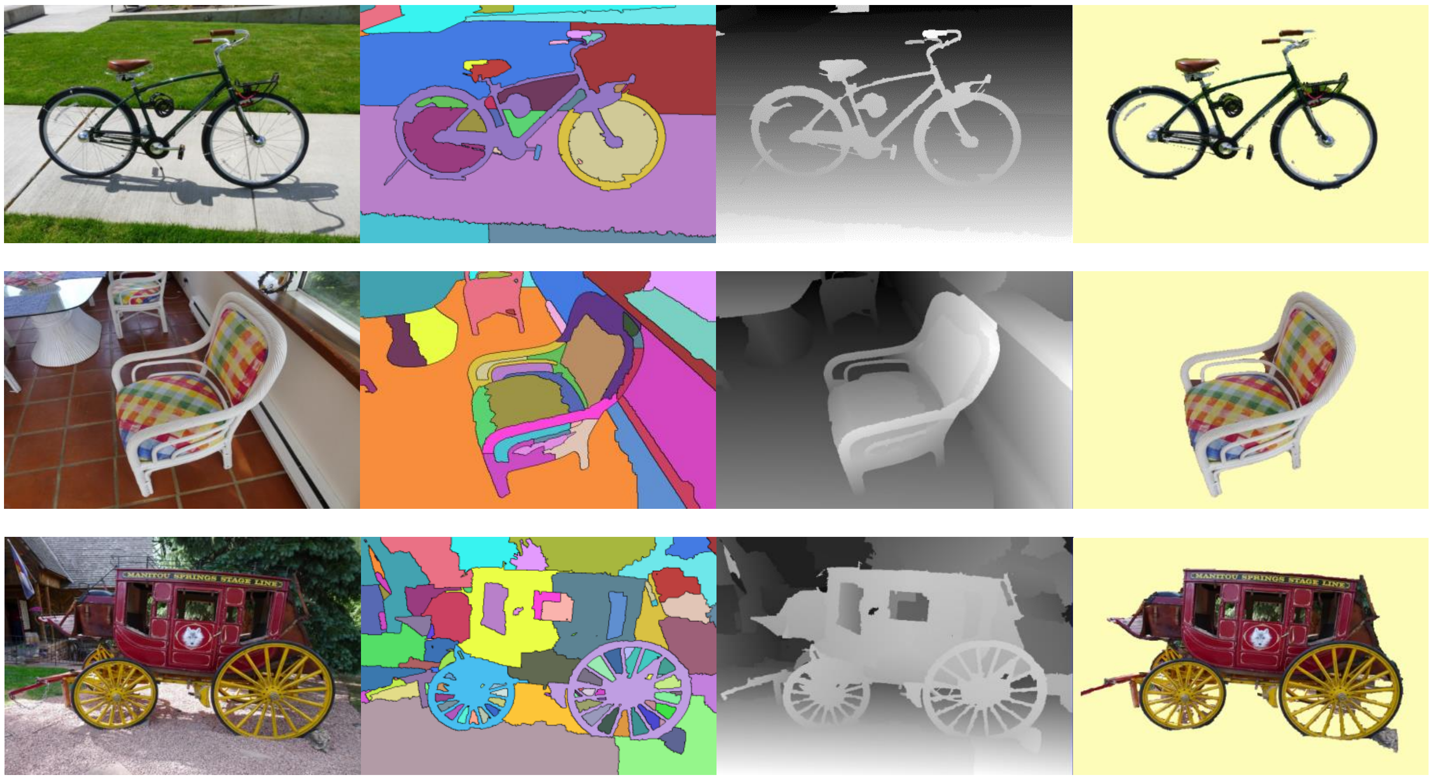
Depth map

Multiview Foreground Object Segmentation

- Infer what constitutes the foreground *object*
- Soft segmentation consistency in multiple-views

Multiple View Object Cosegmentation using Appearance and Stereo Cues

Kowdle, Sinha and Szeliski (ECCV 2012)



Overview

- Dense Correspondence Estimation
  - Surface Stereo
  - High Resolution Stereo Matching
- Joint Correspondence and Cosegmentation
  - Align different object instances
- Sparse Correspondences and Applications
  - Improved place recognition
  - Color Consistency in Photo Collections

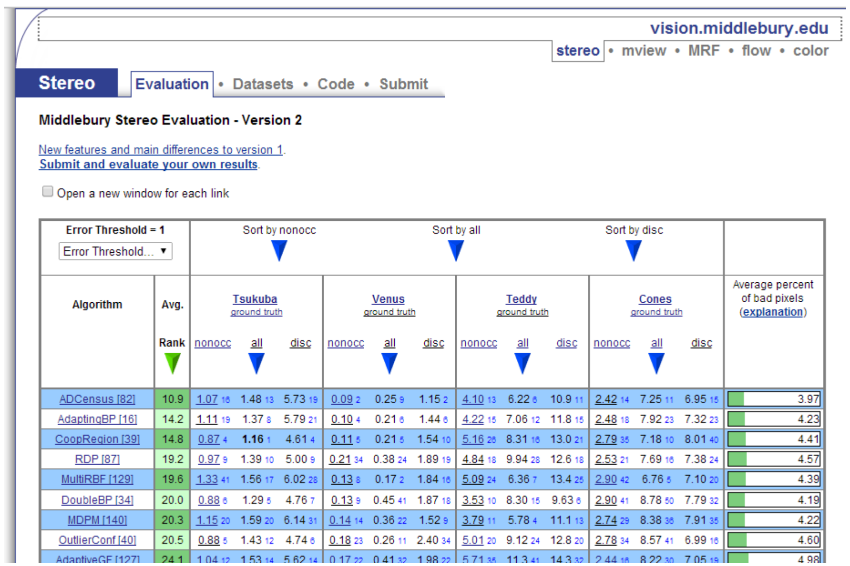
Overview

- Dense Correspondence Estimation
  - Surface Stereo
  - High Resolution Stereo Matching
- Joint Correspondence and Cosegmentation
  - Align different object instances
- Sparse Correspondences and Applications
  - Improved place recognition
  - Color Consistency in Photo Collections

Stereo benchmarks

Middlebury
(v2 now offline)

KITTI


vision.middlebury.edu
stereo • mview • MRF • flow • color
Stereo | Evaluation • Datasets • Code • Submit

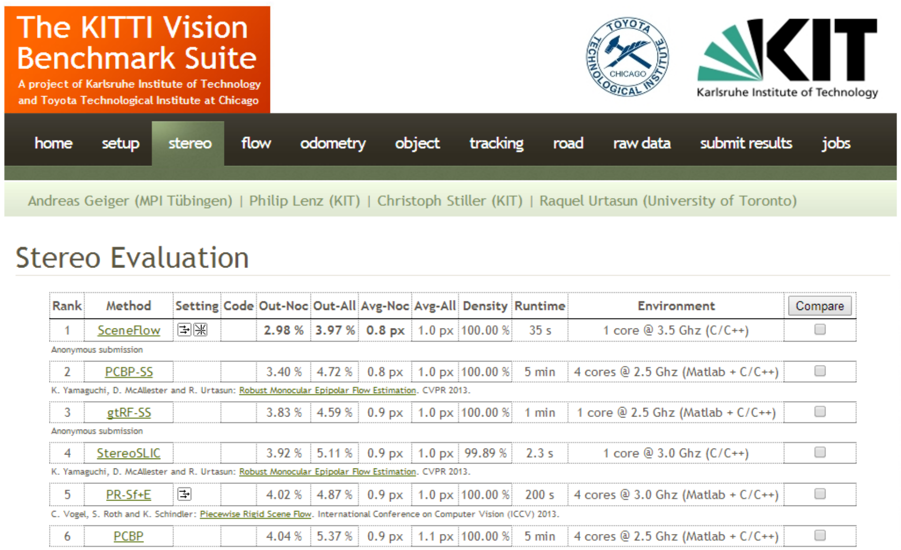
Middlebury Stereo Evaluation - Version 2

New features and main differences to version 1.
Submit and evaluate your own results.

Open a new window for each link

Error Threshold = 1 | Sort by nonocc | Sort by all | Sort by disc

Algorithm	Avg. Rank	Tsukuba nonocc	Tsukuba all	Tsukuba disc	Venus nonocc	Venus all	Venus disc	Teddy nonocc	Teddy all	Teddy disc	Cones nonocc	Cones all	Cones disc	Average percent of bad pixels (explanation)
ADCensus [82]	10.9	1.07 16	1.48 13	5.73 19	0.09 2	0.25 9	1.15 2	4.10 13	6.22 6	10.9 11	2.42 14	7.25 11	6.95 15	3.97
AdaptingBP [16]	14.2	1.11 19	1.37 8	5.79 21	0.10 4	0.21 6	1.44 6	4.22 15	7.06 12	11.8 15	2.48 18	7.92 23	7.32 23	4.23
CoopRegion [39]	14.8	0.87 4	**1.16** 1	4.61 4	0.11 5	0.21 5	1.54 10	5.16 26	8.31 16	13.0 21	2.79 35	7.18 10	8.01 40	4.41
RDP [87]	19.2	0.97 9	1.39 10	5.00 9	0.21 34	0.38 24	1.89 19	4.84 18	9.94 28	12.6 18	2.53 21	7.69 16	7.38 24	4.57
MultiRBF [129]	19.6	1.33 41	1.56 17	6.02 28	0.13 8	0.17 2	1.84 16	5.09 24	6.36 7	13.4 25	2.90 42	6.76 5	7.10 20	4.39
DoubleBP [34]	20.0	0.88 6	1.29 5	4.76 7	0.13 9	0.45 41	1.87 18	3.53 10	8.30 15	9.63 6	2.90 41	8.78 50	7.79 32	4.19
MDPM [140]	20.3	1.15 20	1.59 20	6.14 31	0.14 14	0.36 22	1.52 9	3.79 11	5.78 4	11.1 13	2.74 29	8.38 36	7.91 35	4.22
OutlierConf [40]	20.5	0.88 5	1.43 12	4.74 6	0.18 23	0.26 11	2.40 34	5.01 20	9.12 24	12.8 20	2.78 34	8.57 41	6.99 16	4.60
AdaptiveGF [127]	24.1	1.04 12	1.53 14	5.62 14	0.17 22	0.41 32	1.98 22	5.71 35	11.3 41	14.3 32	2.44 16	8.22 30	7.05 19	4.98


The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago
Karlsruhe Institute of Technology
home | setup | stereo | flow | odometry | object | tracking | road | raw data | submit results | jobs
Andreas Geiger (MPI Tübingen) | Philip Lenz (KIT) | Christoph Stiller (KIT) | Raquel Urtasun (University of Toronto)

Stereo Evaluation

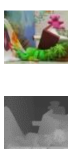
Rank	Method	Setting	Code	Out-Noc	Out-All	Avg-Noc	Avg-All	Density	Runtime	Environment	Compare
1	SceneFlow			**2.98 %**	**3.97 %**	**0.8 px**	1.0 px	100.00 %	35 s	1 core @ 3.5 Ghz (C/C++)	
Anonymous submission											
2	PCBP-SS			3.40 %	4.72 %	0.8 px	1.0 px	100.00 %	5 min	4 cores @ 2.5 Ghz (Matlab + C/C++)	
K. Yamaguchi, D. McAllester and R. Urtasun: Robust Monocular Epipolar Flow Estimation. CVPR 2013.											
3	gtRF-SS			3.83 %	4.59 %	0.9 px	1.0 px	100.00 %	1 min	1 core @ 2.5 Ghz (Matlab + C/C++)	
Anonymous submission											
4	StereoSLIC			3.92 %	5.11 %	0.9 px	1.0 px	99.89 %	2.3 s	1 core @ 3.0 Ghz (C/C++)	
K. Yamaguchi, D. McAllester and R. Urtasun: Robust Monocular Epipolar Flow Estimation. CVPR 2013.											
5	PR-Sf+E			4.02 %	4.87 %	0.9 px	1.0 px	100.00 %	200 s	4 cores @ 3.0 Ghz (Matlab + C/C++)	
C. Vogel, S. Roth and K. Schindler: Piecewise Rigid Scene Flow. International Conference on Computer Vision (ICCV) 2013.											
6	PCBP			4.04 %	5.37 %	0.9 px	1.1 px	100.00 %	5 min	4 cores @ 2.5 Ghz (Matlab + C/C++)	

450 x 376 pixels
D ≈ 16…60

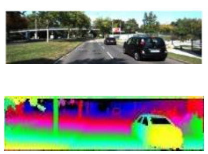
1241 x 376 pixels
D ≈ 70…150

Kim et al. 2013 (Disney Research)

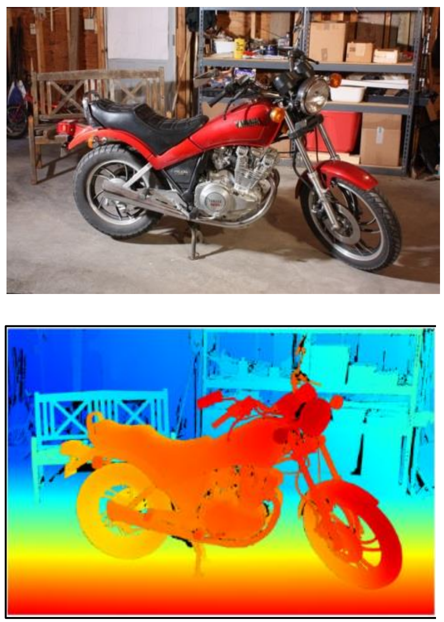
Middlebury Cones/Teddy



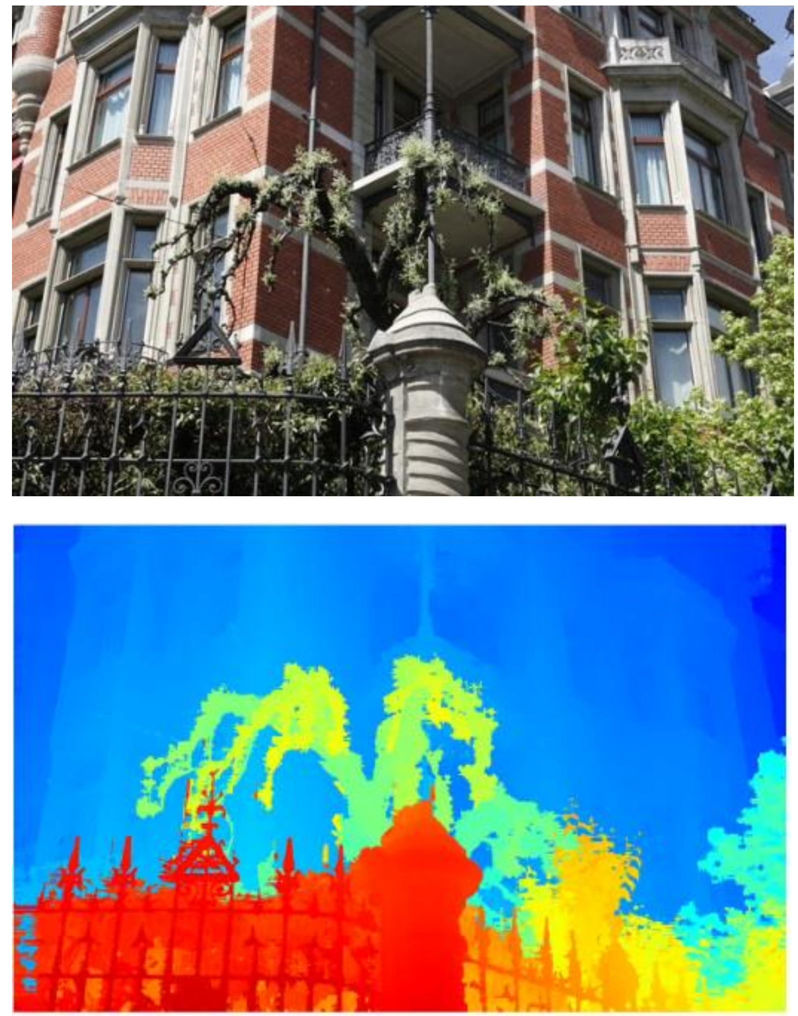
KITTI



Middlebury v3



5 – 6 MPixels



~20 MPixels

Disparity Search Space

Middlebury (old)	KITTI
10 Mdisp.	40 Mdisp.
Middlebury New	**Disney Mansion**
1.5 Gdisp.	**20 Gdisp.**

O(**P*****D**) = O($\mathbf{s}^3$) for resolution **s,** **P**: pixels,

D: disparities

(Most) methods are O(**P*****D**), or O(**P***$\mathbf{D}^2$); they do not scale

Our Goals:

- Ideally, want O(P)
- Avoid enumerating all disparities
- Optimization should scale

Related Work

- Efficient approximate energy minimization
  - Semi-global Matching (SGM) [Hirschmüller 2005]
- Disparity Refinement [Ma 2013, …]
- Avoid exploring the whole DSI
  - Coarse-to-fine [long tradition]
  - Seed & Grow [Cech & Sara 2007, …]
  - PatchMatch stereo [Bleyer et al. 2011]
  - ELAS [Geiger et al. 2010]
  - Bilateral space edge-aware stereo [Barron et al. 2015]
  - Tunable Stereo [Pillai et al. 2016]

Efficient High-Resolution Stereo Matching using Local Plane Sweeps

CVPR 2014

Sudipta Sinha
Microsoft Research

Daniel Scharstein
Middlebury College

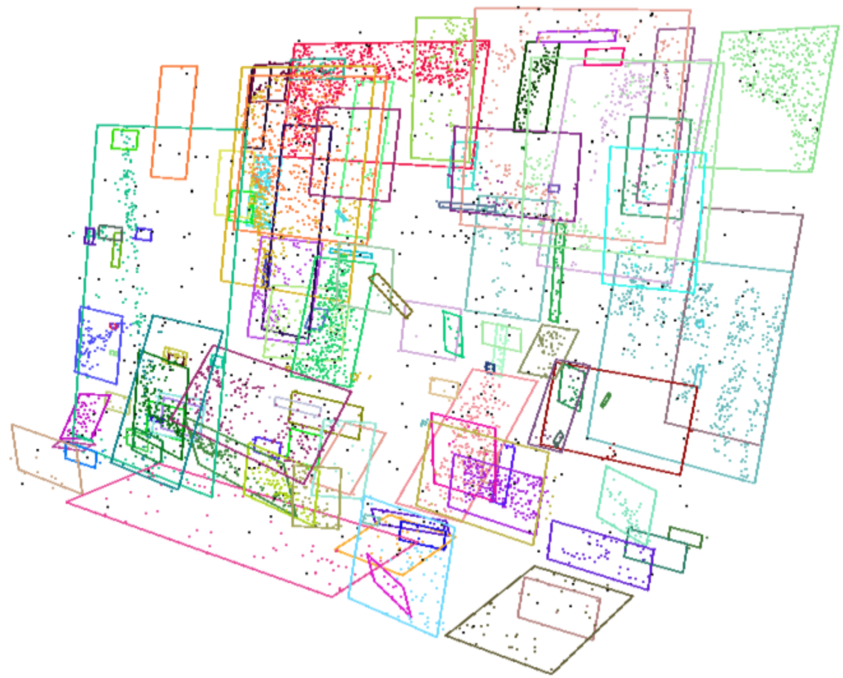
Richard Szeliski*
Facebook

* while at Microsoft Research

Local Plane Sweep Stereo

- Sparse feature matching; refine vertical disparities
- Generate plane hypotheses (with unknown extents)

Plane Hypothesis Generation



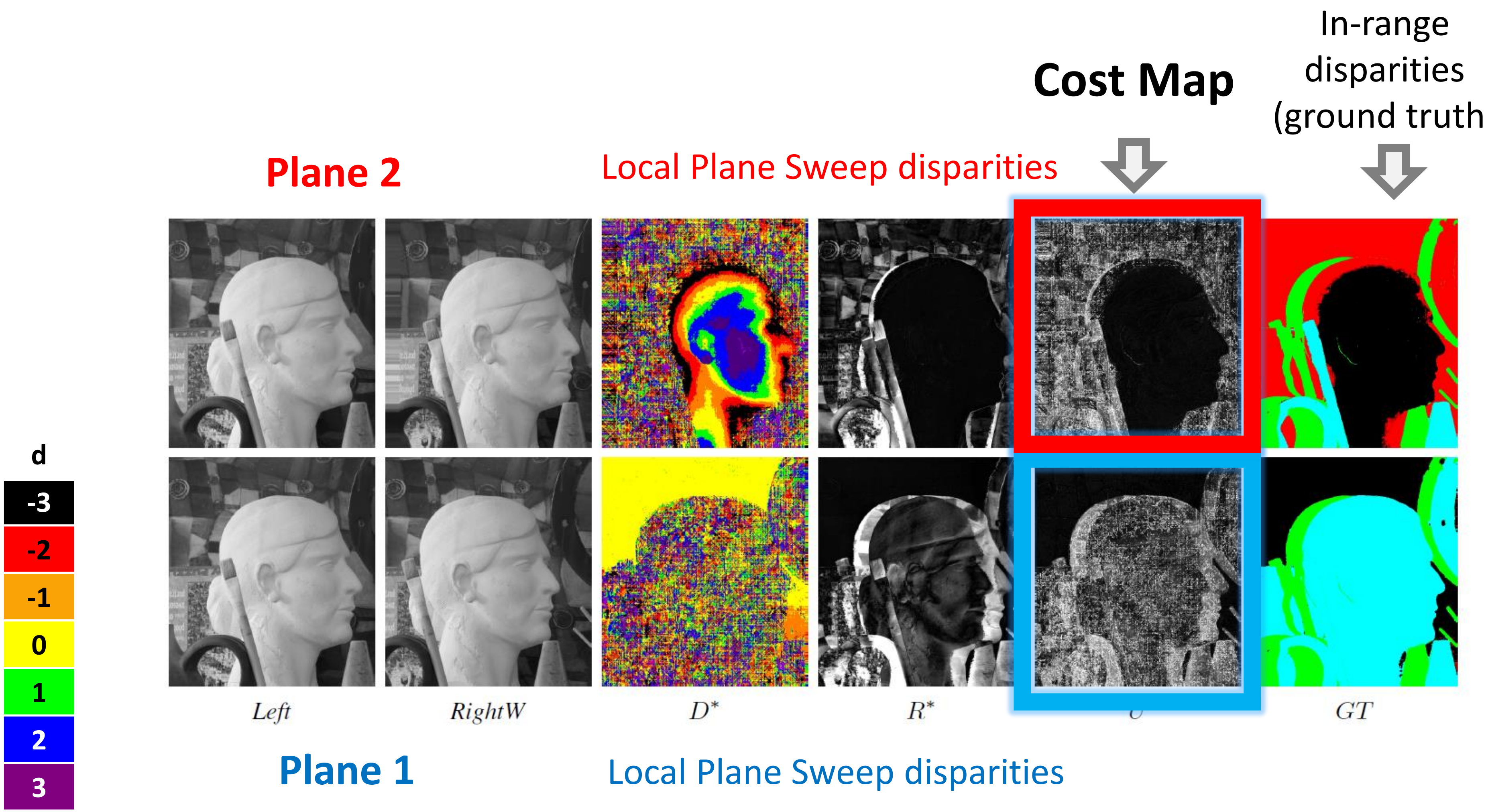
Sparse Correspondences
Plane Labels
Left Image

Local Plane Sweep Stereo

- Sparse feature matching; refine vertical disparities
- Generate plane hypotheses (with unknown extents)
- Perform local plane sweeps (LPS) around planes

  - narrow disparity range; SGM optimization

Local Plane Sweeps



Local Plane Sweep Stereo

- Sparse feature matching; refine vertical disparities
- Generate plane hypotheses (with unknown extents)
- Perform local plane sweeps (LPS) around planes

 - narrow disparity range; SGM optimization

Impose Tile structure

 - Perform LPS on tiles and propagate planes to adjoining tiles

Local Plane Sweep Stereo

- Sparse feature matching; refine vertical disparities
- Generate plane hypotheses (with unknown extents)
- Perform local plane sweeps (LPS) around planes

 - narrow disparity range; SGM optimization

Impose Tile structure

 - Perform LPS on tiles and propagate planes to adjoining tiles

- Global optimization

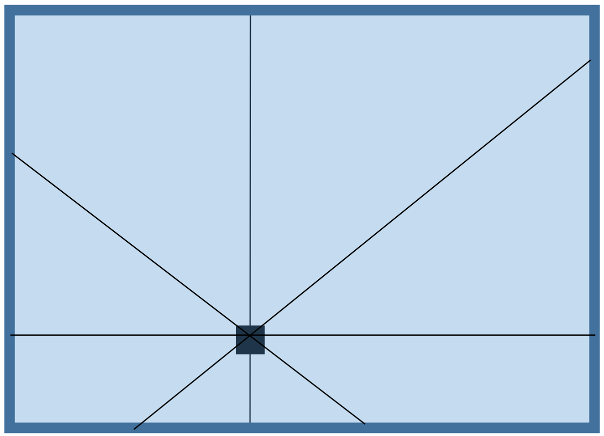
 - Assign pixels to surface proposals

 - Approximate energy minimization (via SGM)

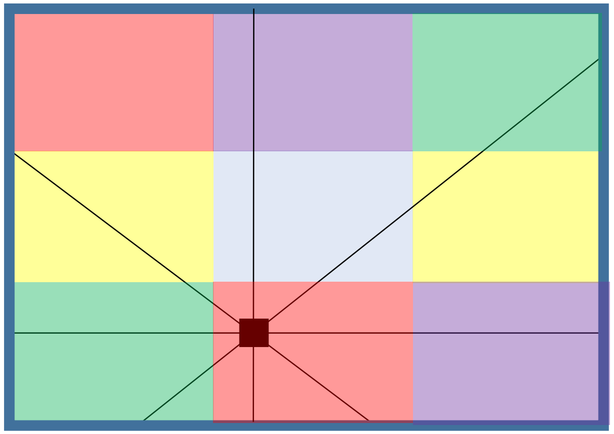
 - Extend SGM to exploit tile structure and sparse label sets

Global Optimization (via SGM)

- Message passing on 1D paths (8 directions)



SGM
- same labels at all pixels



LPS – SGM
- *Label sets vary tile to tile*
- *Needs book-keeping at tile boundaries*

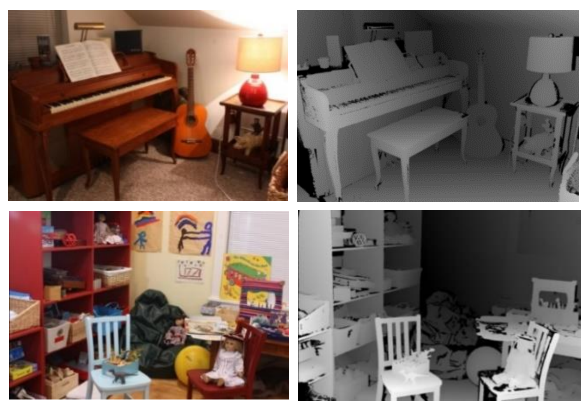
Datasets

Midd9

2003-2006 Middlebury
(1.4 – 2.7 MP)

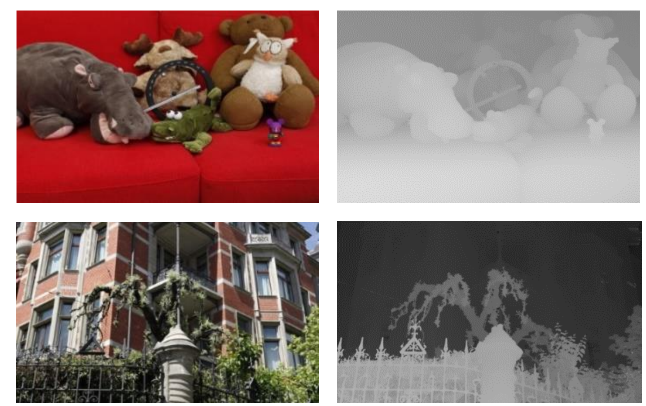
New7

2011-2014 Middlebury
(5.1 – 6.0 MP)



Disney

Kim et al. 2013, in SIGGRAPH
(4.5 – 20 MP)

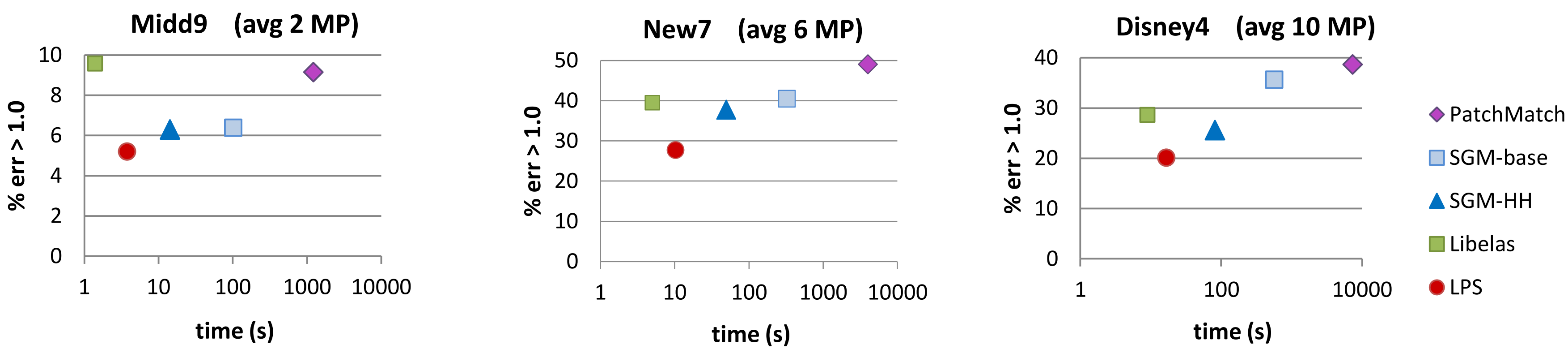


Experiments

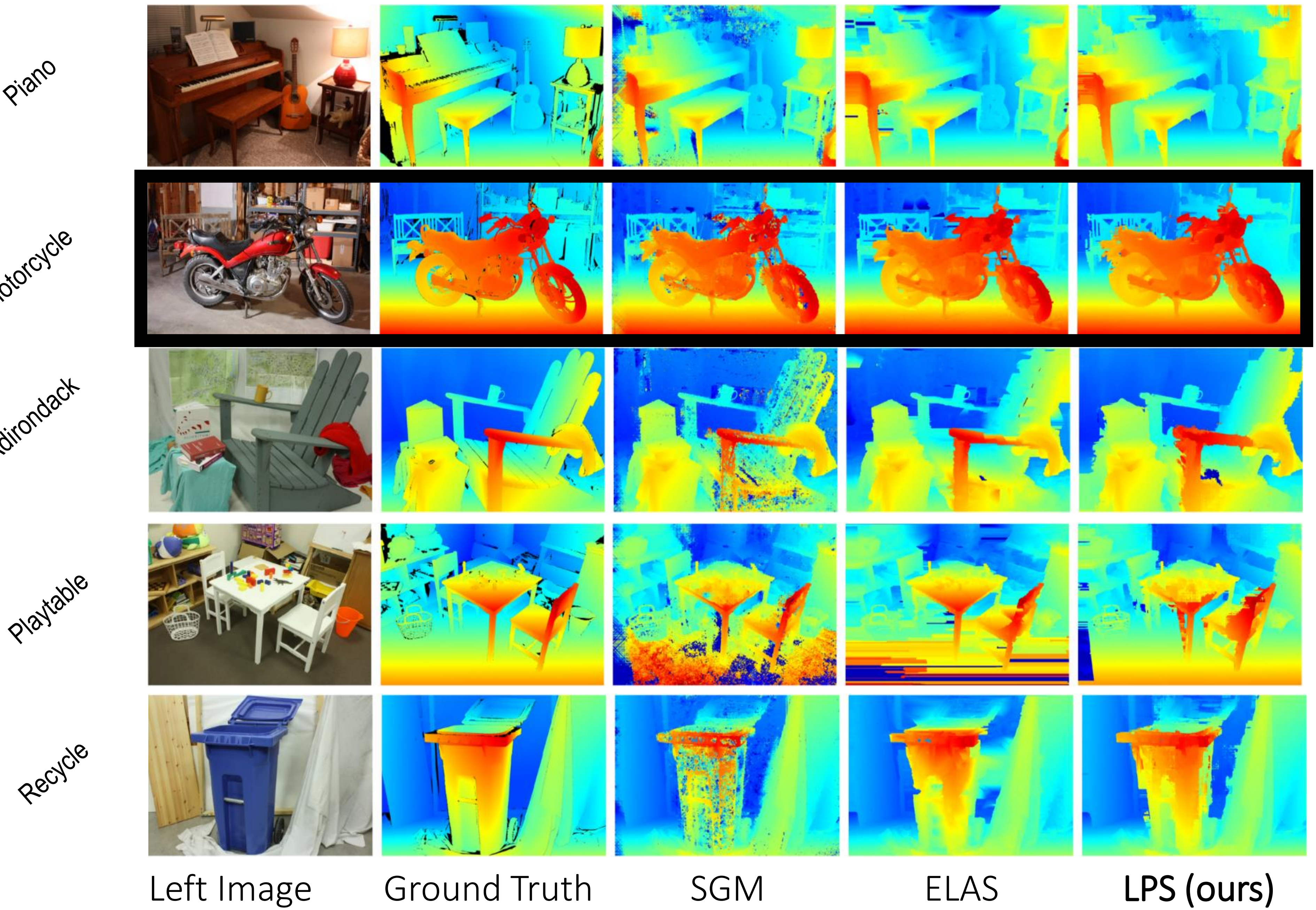
- Evaluation:
  - PatchMatch Stereo [Bleyer et al. 2011]
  - SGM (our impl.)
  - SGM-HH [Hirschmüller 2005]
  - ELAS [Geiger et al. 2010]
  - LPS
- Metric:
  - 1 pixel disparity error at non-occluded pixels.

Results – Accuracy vs. Runtime

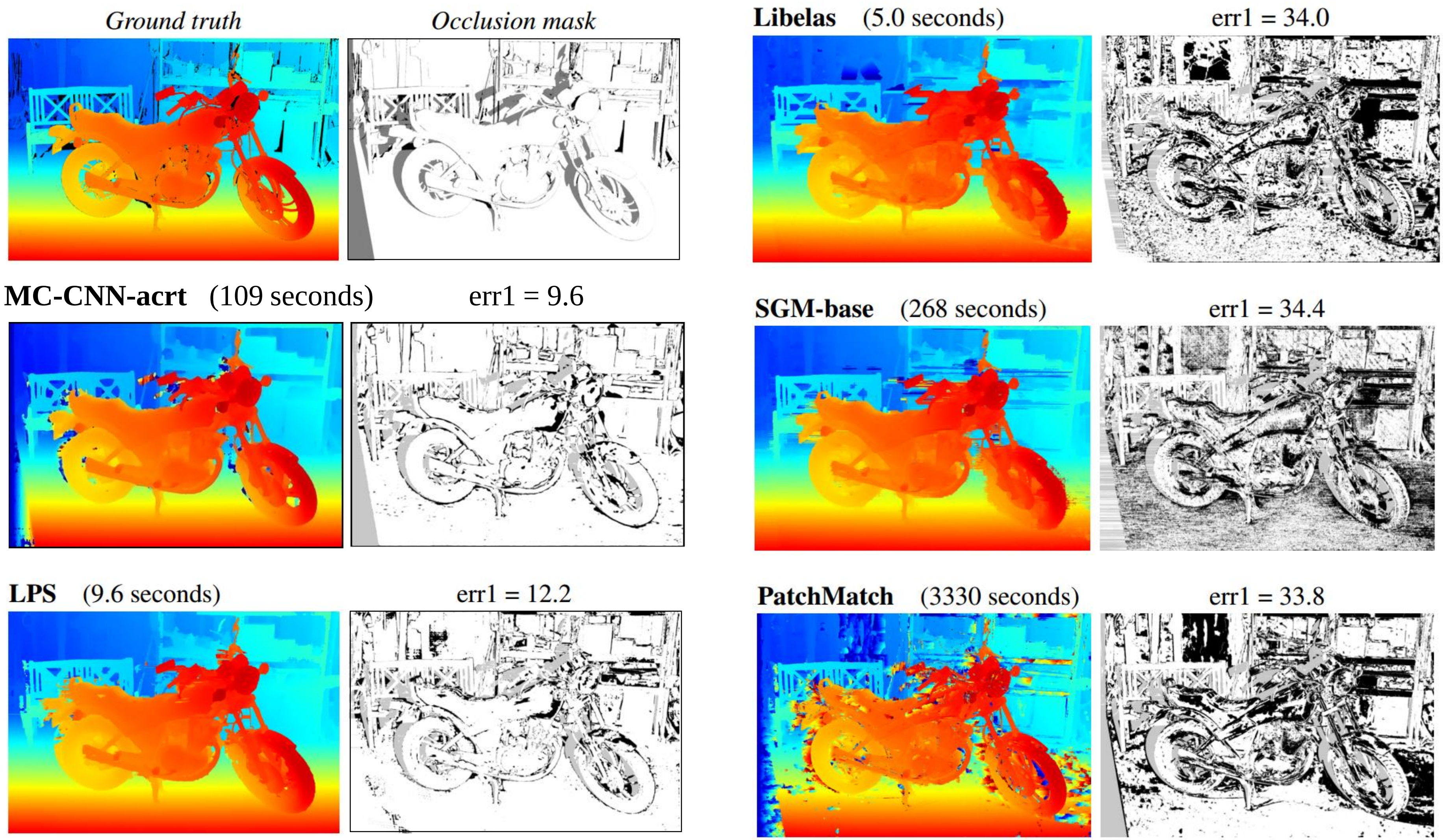
Error vs. runtime, 1.0 pixel threshold



- LPS is the most accurate followed by SGM
- ELAS is the fastest, LPS is 2nd.
  - no GPUs were used.


Piano
otorcycle
dirondack
Playtable
Recycle
Left Image
Ground Truth
SGM
ELAS
LPS (ours)

Motorcycle (1 pixel error maps)



Summary

Advantages

- Avoid exploring full search space
- Runtime independent of disparity range
- Handles weakly textured slanted surfaces

Limitations

- Can miss surfaces not among initial proposals
- No good “stopping criterion” for proposal generation
- Unclear how to incorporate coarse to fine reasoning

Promising Directions

- Avoid monolithic optimization
- Residual analysis to guide efficient search

Overview

- Dense Correspondence Estimation
  - Surface Stereo
  - High Resolution Stereo Matching
- Joint Correspondence and Cosegmentation
  - Align different object instances
- Sparse Correspondences and Applications
  - Improved place recognition
  - Color Consistency in Photo Collections

Overview

- Dense Correspondence Estimation
  - Surface Stereo
  - High Resolution Stereo Matching
- **Joint Correspondence and Cosegmentation**
  - Align different object instances
- Sparse Correspondences and Applications
  - Improved place recognition
  - Color Consistency in Photo Collections

Joint Cosegmentation and Dense Correspondence Estimation

CVPR 2016 (to appear)

Tatsunori Taniai
Univ. of Tokyo

Sudipta Sinha
Microsoft Research

Yoichi Sato
Univ. of Tokyo

Problem

<u>Input</u>

Image pair containing semantically related objects but different instances

Source

Target

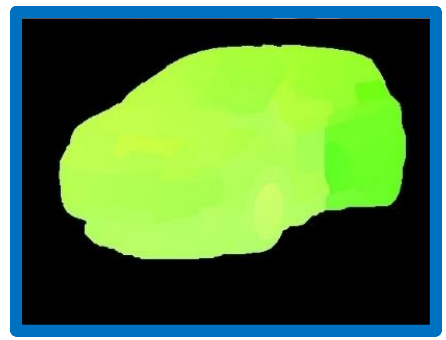
Problem

<u>Input</u>

Image pair containing semantically related objects but different instances


Source


Target


Mask + Flow


Warped Source Image

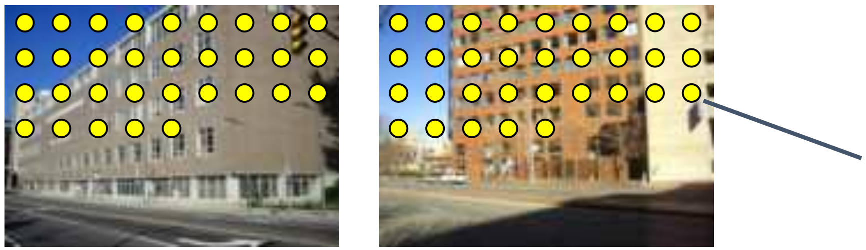
<u>Output</u>

Find the common region i.e. ***foreground (binary) mask***

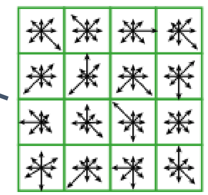
Find the dense ***flow map*** *associated with foreground.*

SIFT Flow [Liu+ 2008]

Robust visual-similarity matching using dense SIFT



Dense SIFT descriptors



- At every pixel
- Same scale
- No rotation

Efficient coarse to fine inference [Felzenszwalb and Huttenlocher 2004]

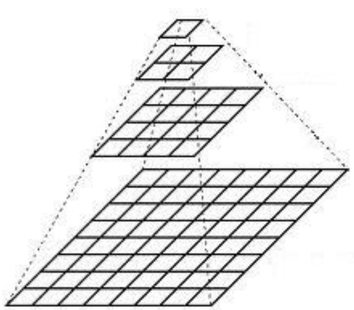
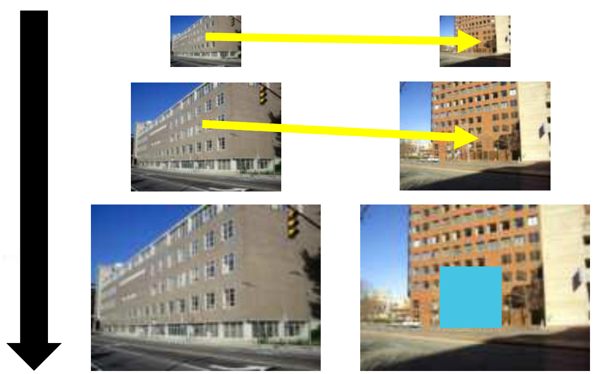


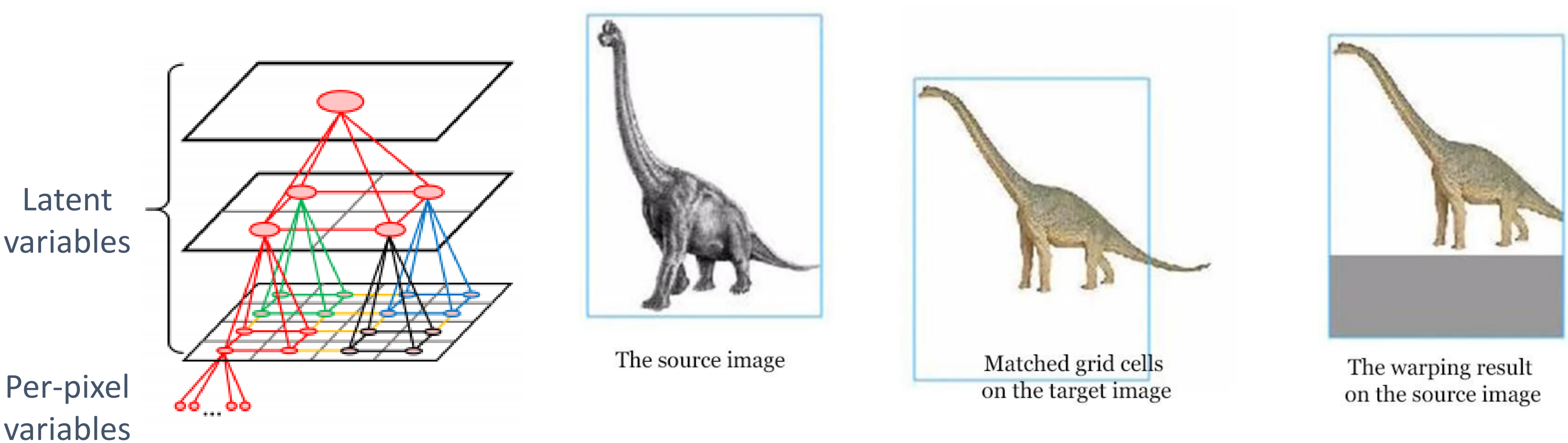
Image pyramid



- Estimate and propagate from coarse to fine levels
- Search in a limited range from propagated points

Generalized Deformable Spatial Pyramids

[Hur+15, Kim+13]



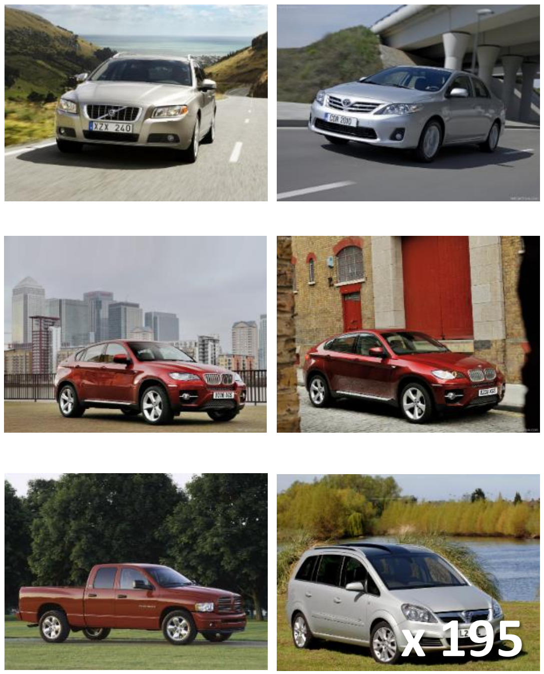
- Powerful yet flexible regularization
- Hierarchy is not segmentation/flow aware

Image Co-segmentation

- Common region shares similar statistics
- Pixel correspondence in common region not modeled

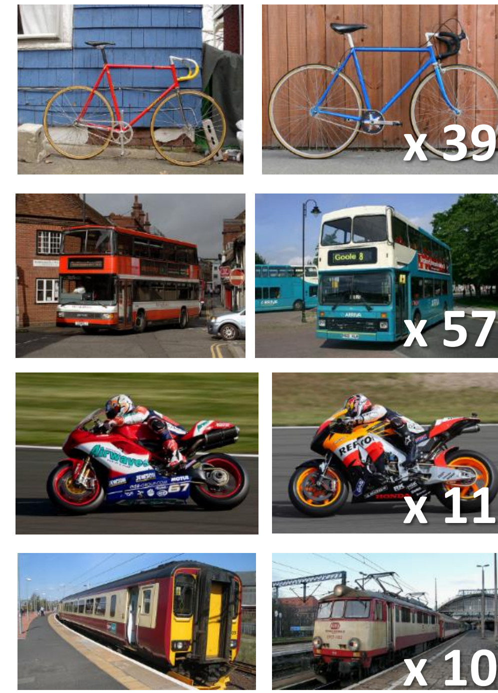
New Dataset

FG3DCar [1]



JODS [2]

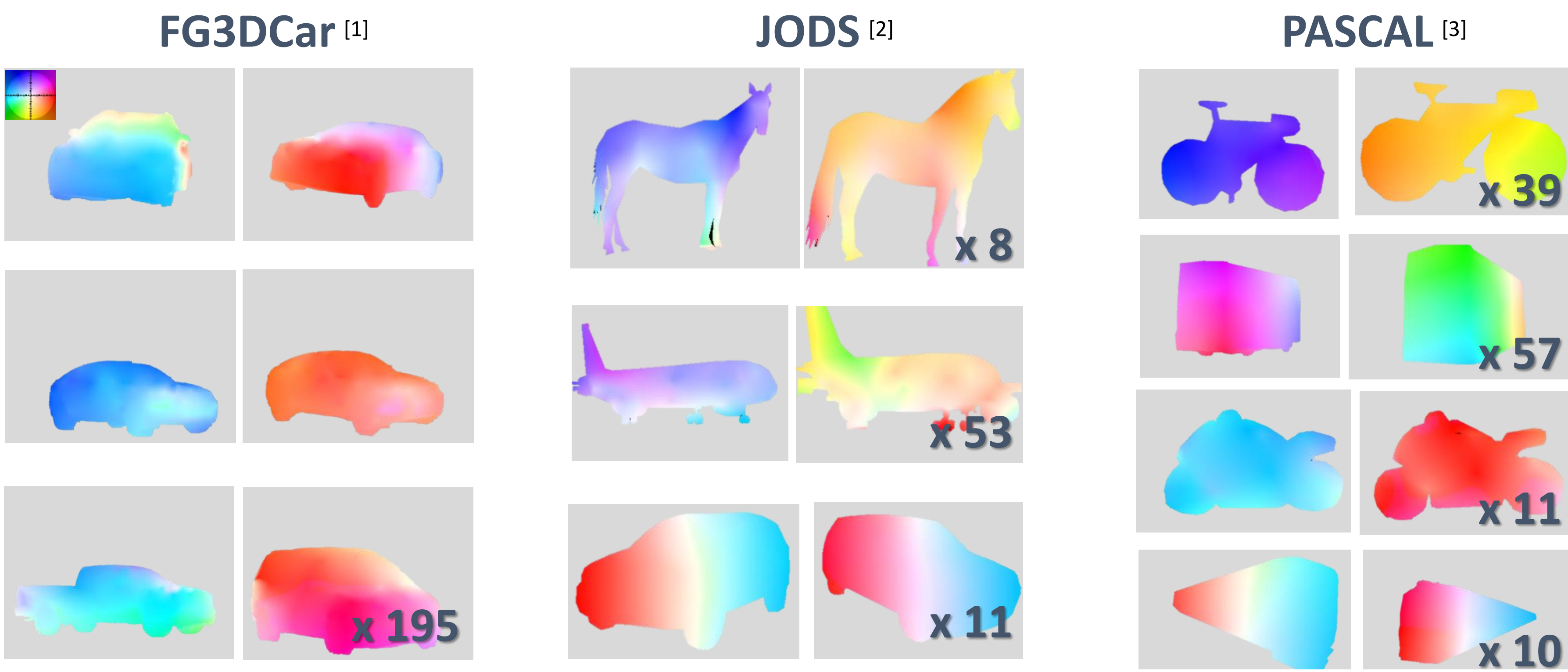
PASCAL [3]



[1] Lin et al 2014, "Jointly Optimizing 3D Model Fitting and Fine-Grained Classification"
[2] Rubinstein et al 2013 "Unsupervised Joint Object Discovery and Segmentation in Internet Images"
[3] Hariharan et al 2011, "PASCAL segmentation dataset"

400 pairs

New Dataset



Dense ground truth correspondence obtained by interpolating sparse key-point matches (annotated by a user).

400 pairs

Challenges

- Objects are unknown
- Appearance, shape similarity cues are weak
- Viewpoints, backgrounds differ

Towards ..

unsupervised visual object discovery

Our Approach

- Jointly recover flow and segmentation
- Hierarchical model

  (Structure) Layered graph of nested image regions

  (Continuous Label Space)

  - binary (segmentation)
  - 2D similarity transform (flow) (4-dof)

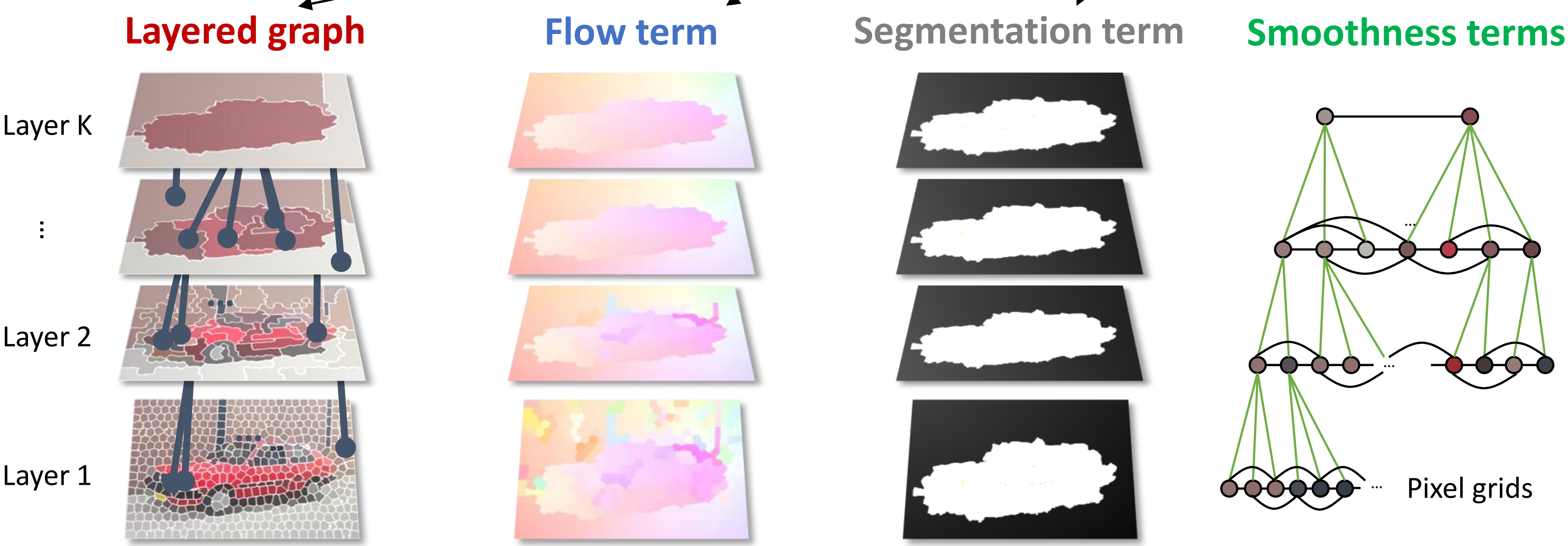
  (Spatial regularization)

  - between neighbors
  - between parent-child nodes.
- Energy minimization/Inference

  Local alpha expansions (graph cuts) [Taniai et al. 2014]

Hierarchical Model

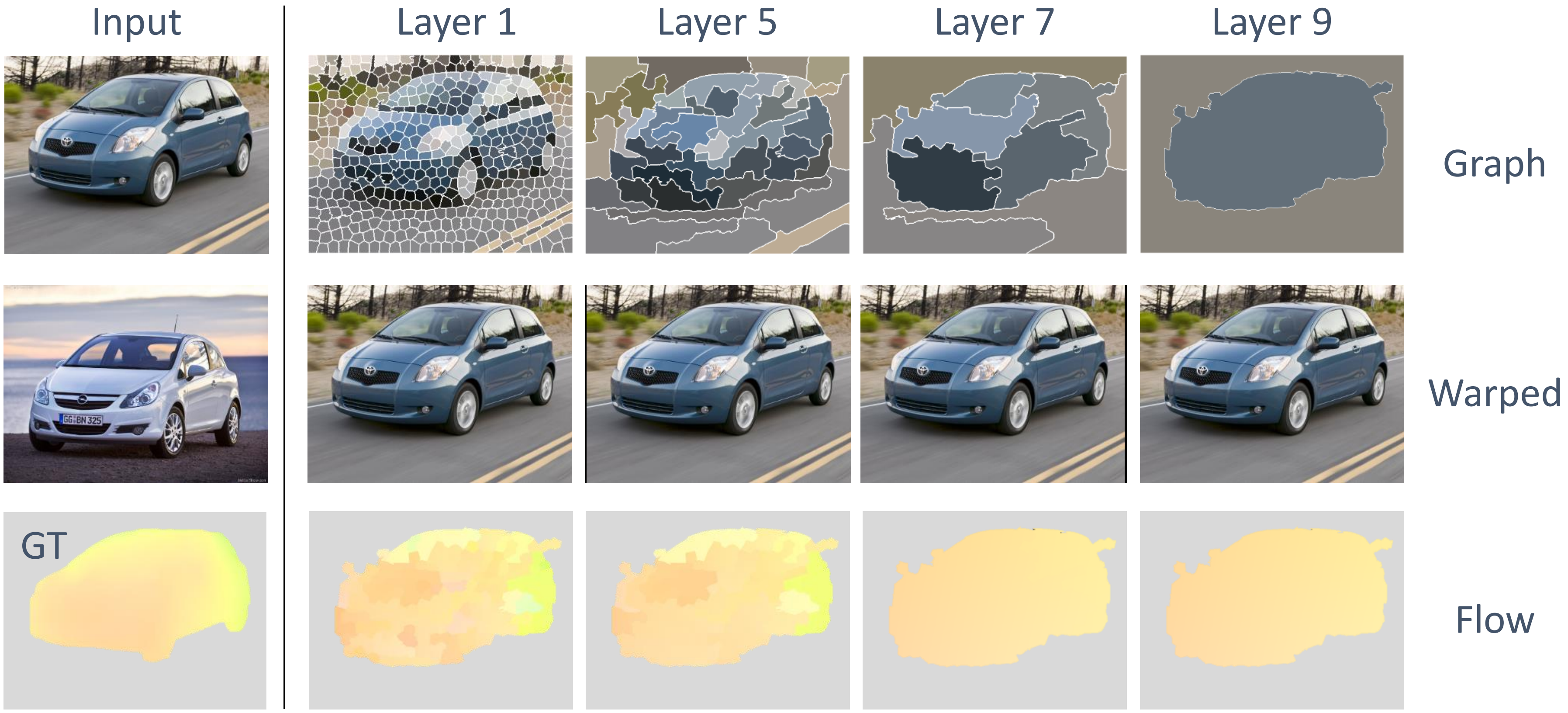
$$F(G, f, \alpha) = E_{graph}(G) + E_{flow}(f|G) + E_{seg}(\alpha|G) + E_{reg}(f, \alpha|G)$$



- Spatial neighbor edges
- Parent child edges

- structure inferred one layer at a time
- Patch matching with HOG descriptors
- FG/BG color likelihoods

Hierarchical flow visualization



Foreground / Background cues

Foreground patches are likely to have a good match (low cost)

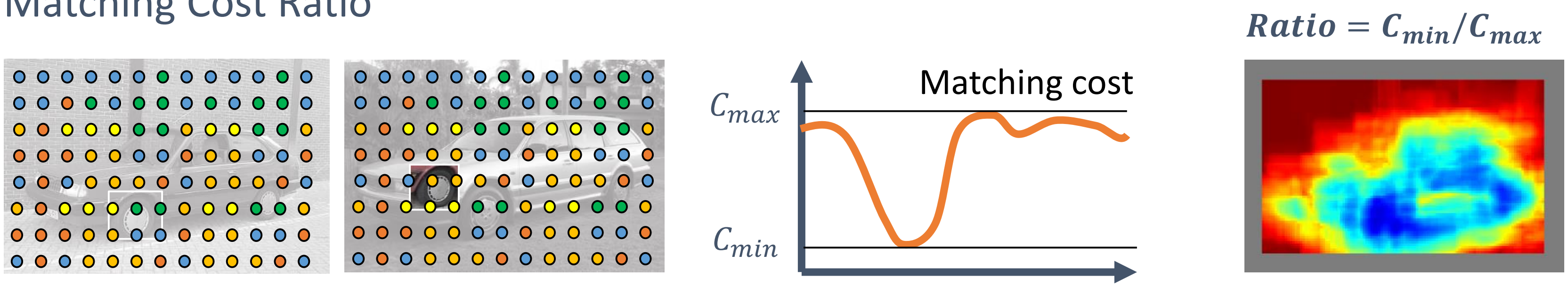
More discriminative

Background patches will have random matches (usually high cost)

Less discriminative

Foreground / Background cues

Matching Cost Ratio



Geodesic distance from image boundary

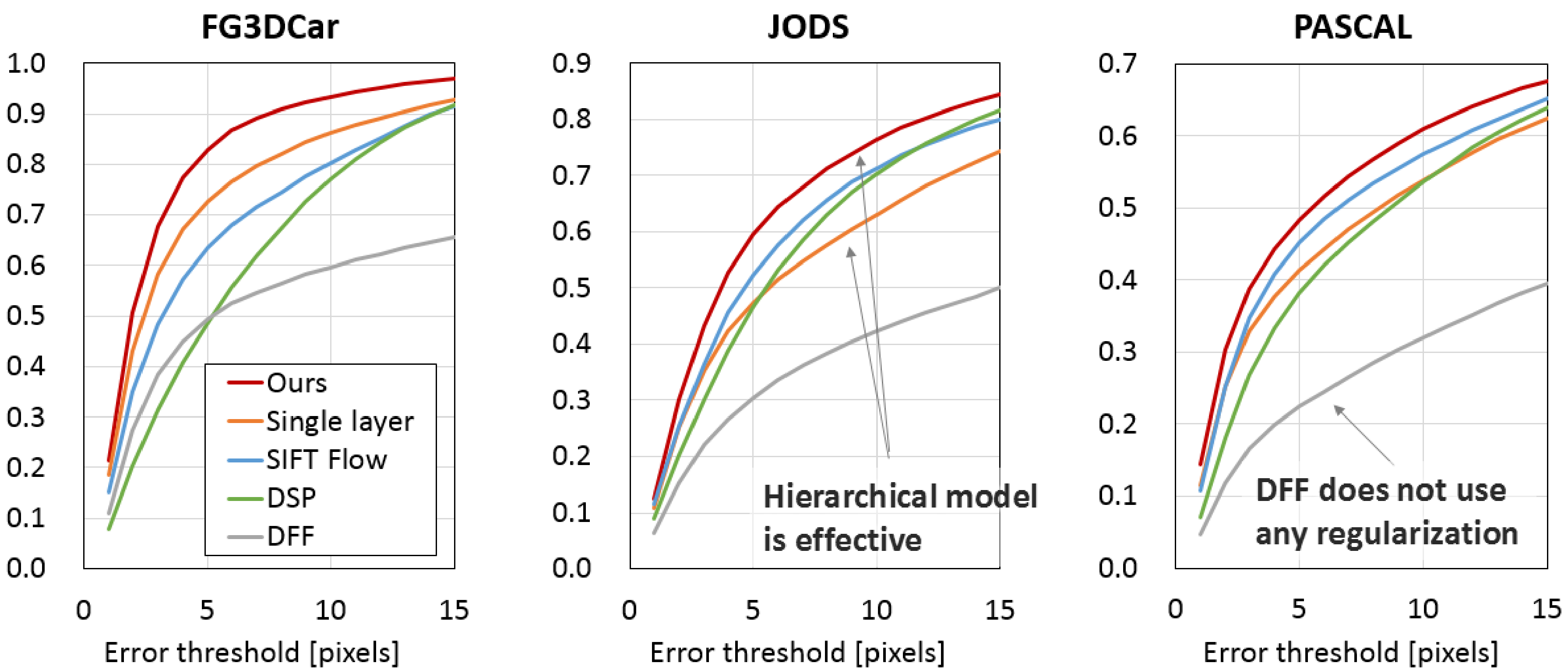
- Construct seeds and initial mask for GrabCut [Rother+04]
- Learn FG/BG color models for each image

Evaluation

Baselines

- Our single layer model
- SIFT Flow [Liu et al. 2008]
- Deformable spatial pyramids (DSP) [Kim et al. 2013]
- DAISY filter flow (DFF) [Yang et al. 2014]
- Cosegmentation by composition [Faktor and Irani 2013]
- Discriminative Clustering [Joulin et al. 2010, 2012]
- NRDC … [HaCohen et al. 2011]
- Cosegmentation by Co-sketch [Dai et al. 2013]

Flow Accuracy

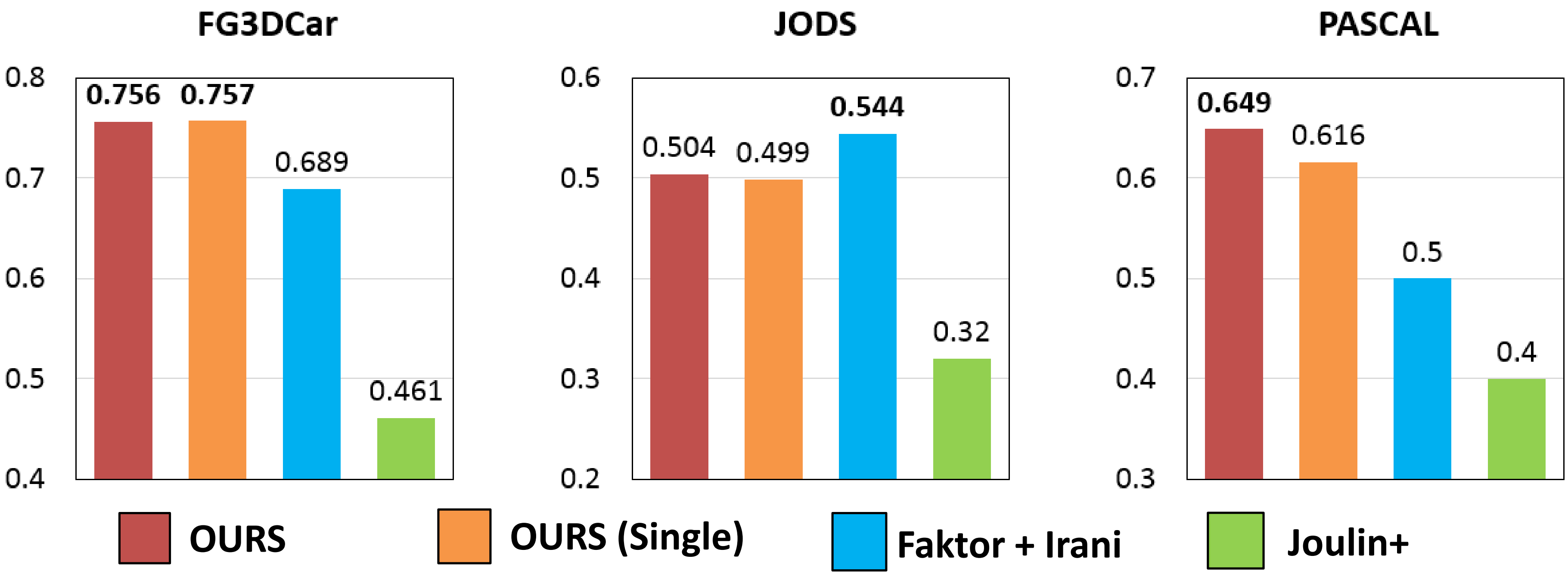


Accuracy Metric

- Percentage of flow errors above a threshold (2d distance)

Our method consistently outperforms all the baselines

Cosegmentation Accuracy



Accuracy Metric

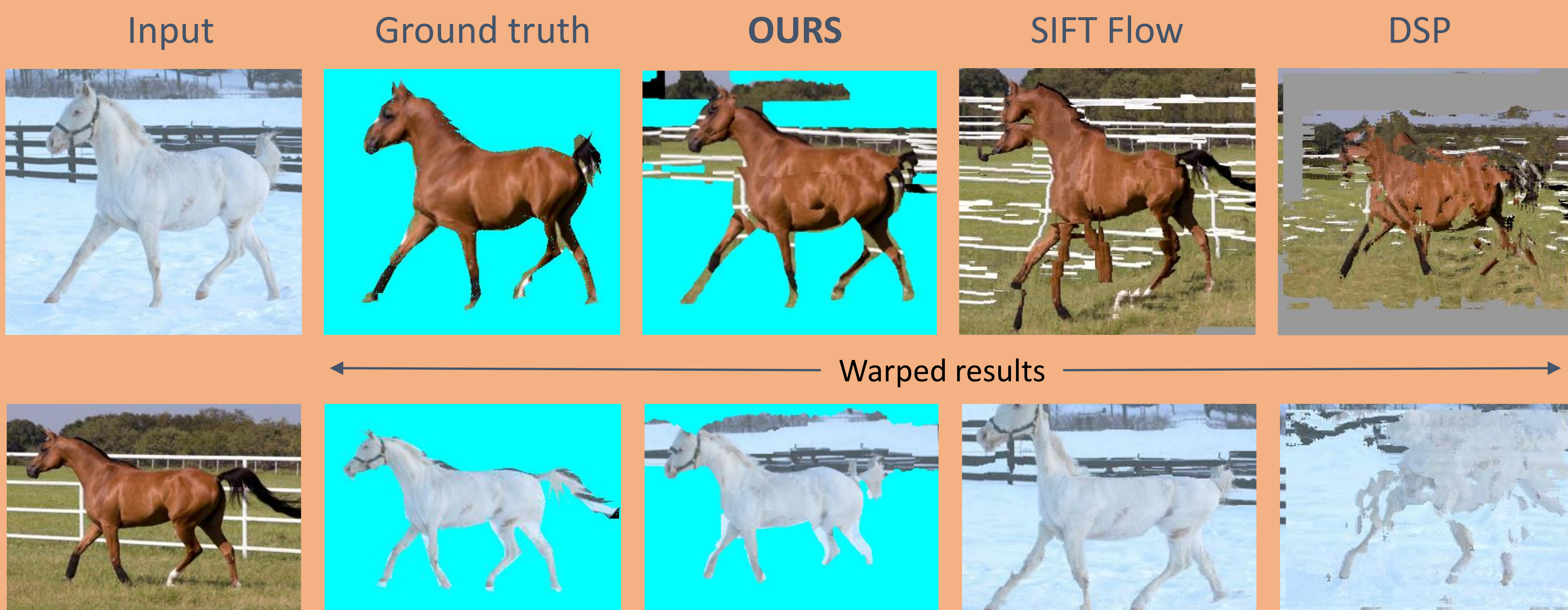
- Intersection-over-union ratio

Our method achieves comparable or better accuracies

Alignment Results

Source: FG3DCar

Alignment Results



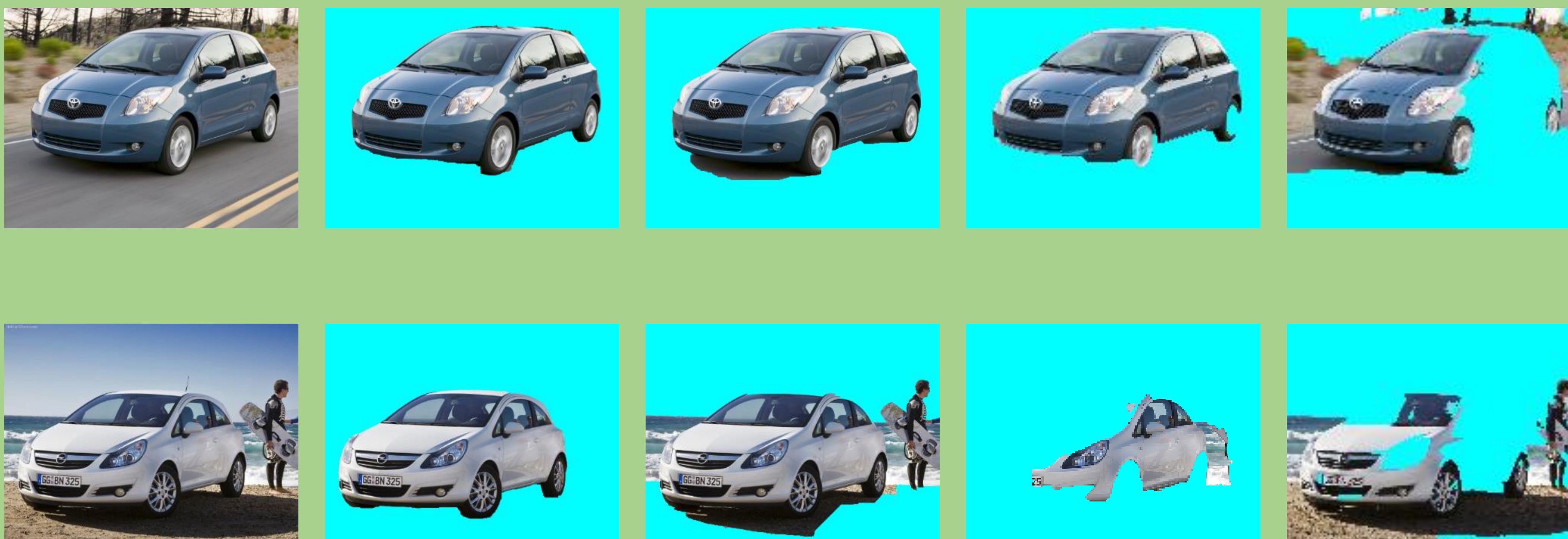
Source: JODS

Alignment Results

Source: PASCAL

Cosegmentation Results

Input | Ground truth | **OURS** | Faktor & Irani | Joulin et al.



Source: FG3DCar

Cosegmentation Results

Input | Ground truth | **OURS** | Faktor & Irani | Joulin et al.

Source: JODS

Cosegmentation Results

Input | Ground truth | **OURS** | Faktor & Irani | Joulin et al.

Source: PASCAL

Future Work

- Try pre-trained ConvNet features
- Add bi-directional flow consistency
- Use multiple images, add cycle-consistency

Overview

- Dense Correspondence Estimation
  - Surface Stereo
  - High Resolution Stereo Matching
- Joint Correspondence and Cosegmentation
  - Align different object instances
- Sparse Correspondences and Applications
  - Improved place recognition
  - Color Consistency in Photo Collections

Overview

- Dense Correspondence Estimation
  - Surface Stereo
  - High Resolution Stereo Matching
- Joint Correspondence and Cosegmentation
  - Align different object instances
- Sparse Correspondences and Applications
  - **Improved place recognition**
  - Color Consistency in Photo Collections

Leveraging Structure from Motion to Learn Discriminative Codebooks for Scalable Landmark Classification

CVPR 2013

Alessandro Bergamo*

Amazon

Lorenzo Torresani

Dartmouth College

* while at Dartmouth College

SfM from Internet photos

TREVI FOUNTAIN (3201 IMGS) 2337 cams, 57K pts (3 comps)
TYN CHURCH (3307 IMGS) 3307 cams, 298K pts (10 comps)
PIAZZA NAVONA (3013 IMGS) 1004 cams, 182K pts (8 comps)
CHICHEN ITZA (3434 IMGS) 955 cams, 216K pts (19 comps)

Problem

Goal

A single image from one of **k** locations. Recognize the location.

Approach

Image categorization (BoW/VLAD/Fisher → linear SVM)

Train a binary classifier for each location

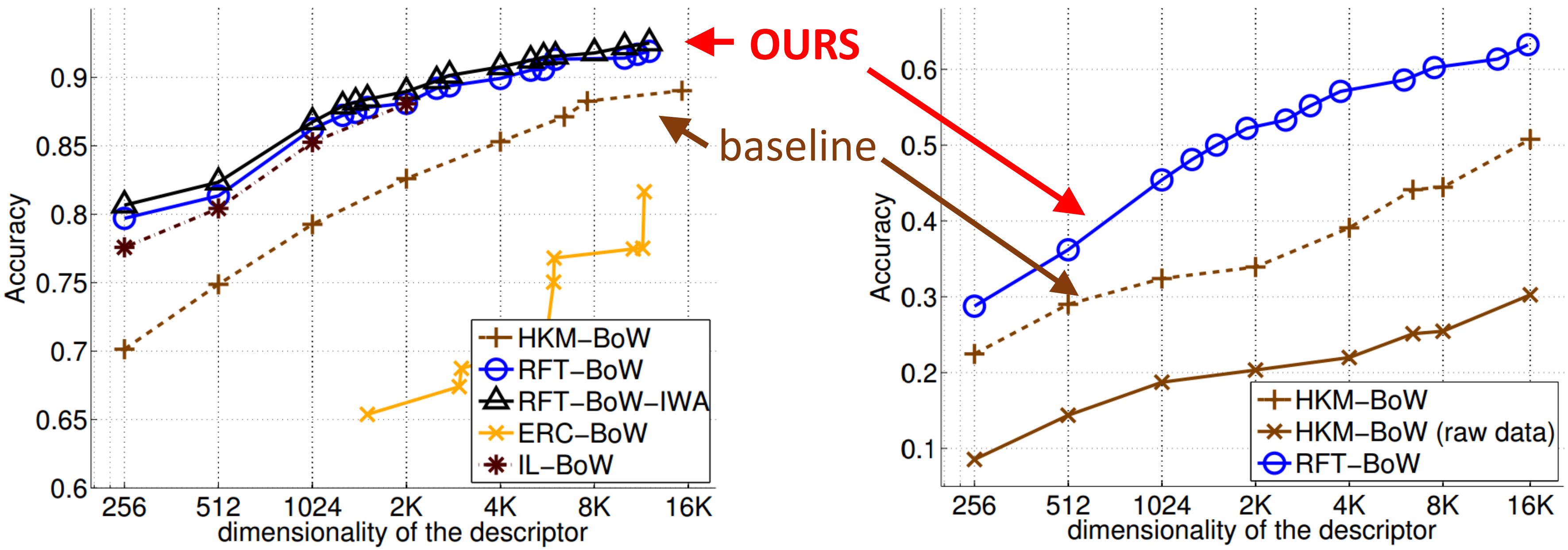
Idea (Discriminative Codebook Learning)

- Each track (n-view correspondence) is a unique class.
- Train a discriminative random forest.
- Use it to quantize/aggregate local descriptors.

Results: top-1 classification accuracy

<u>Landmark3D</u>
(25 places, 5K test images)

<u>Landmark-620</u>
(620 places, 62K test images)



- More accurate than k-means (SIFT/DAISY)
- for both BoW/VLAD representations

Overview

- Dense Correspondence Estimation
  - Surface Stereo
  - High Resolution Stereo Matching
- Joint Correspondence and Cosegmentation
  - Align different object instances
- Sparse Correspondences and Applications
  - Improved place recognition
  - **Color Consistency in Photo Collections**

Efficient and Robust Color Consistency for Community Photo Collections

CVPR 2016 (to appear)



Jaesik Park*
Intel Labs

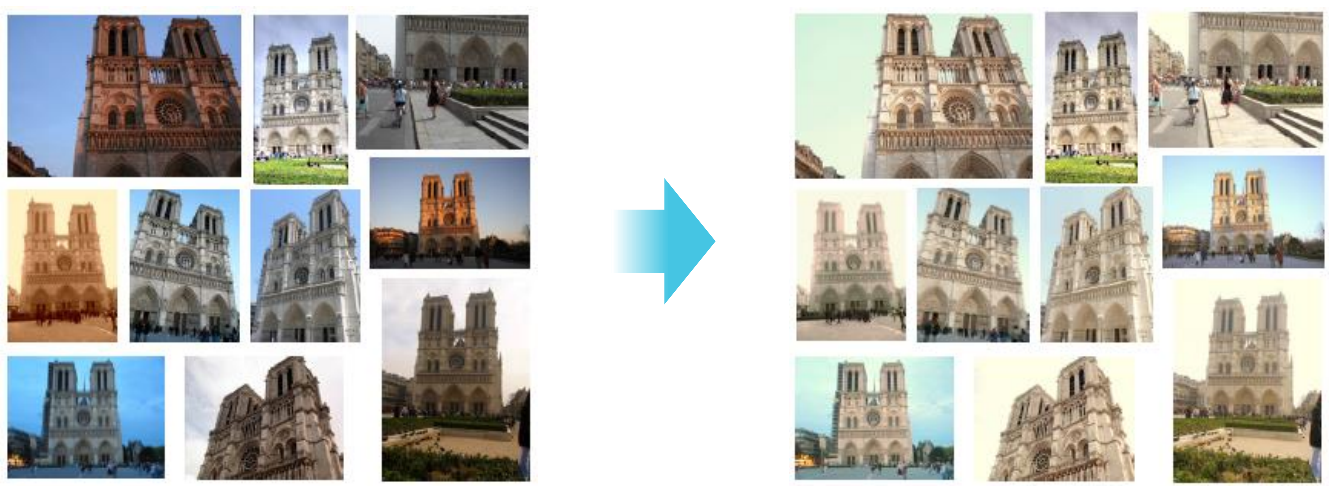
Yu-Wing Tai*
SenseTime

In So Kweon
KAIST

* while at KAIST

Goal

Improve the color consistency of images in a photo collection



Rigid Scenes

Feature matching + Structure from Motion (SfM)

Non-rigid scenes

Feature matching in image pairs

Construct a match graph

Compute maximal cliques [Bron-Kerbosch algorithm (1977)]

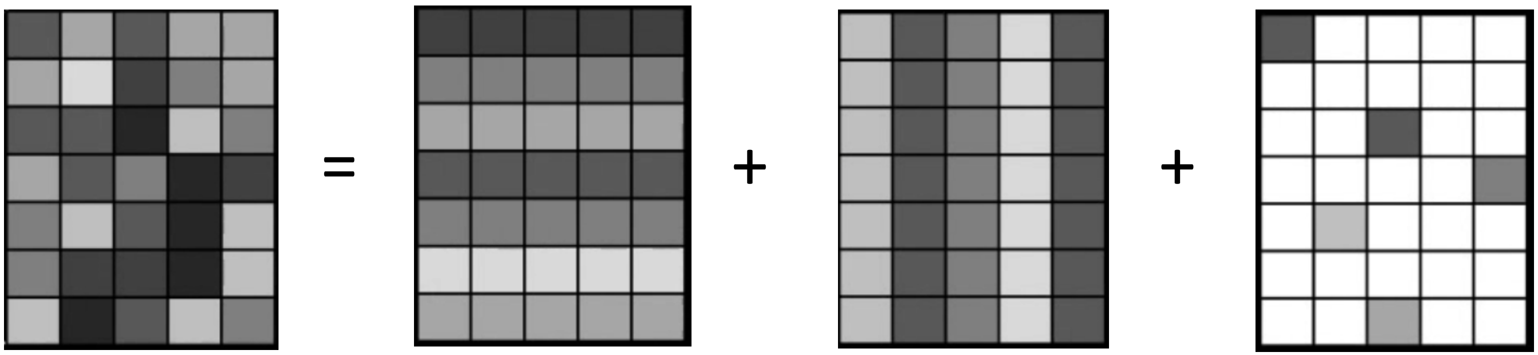
Main Idea

Color Correction Model: $I' = (cI)^{\gamma}$

Constraints from sparse correspondences ..

$$I_i(x_{ij}) = (c_i a_j e_{ij})^{\gamma_i}$$

Low-rank Matrix Factorization formulation



Low-Rank Matrix Decomposition Technique

(Cabral et al. 2013, in ICCV)

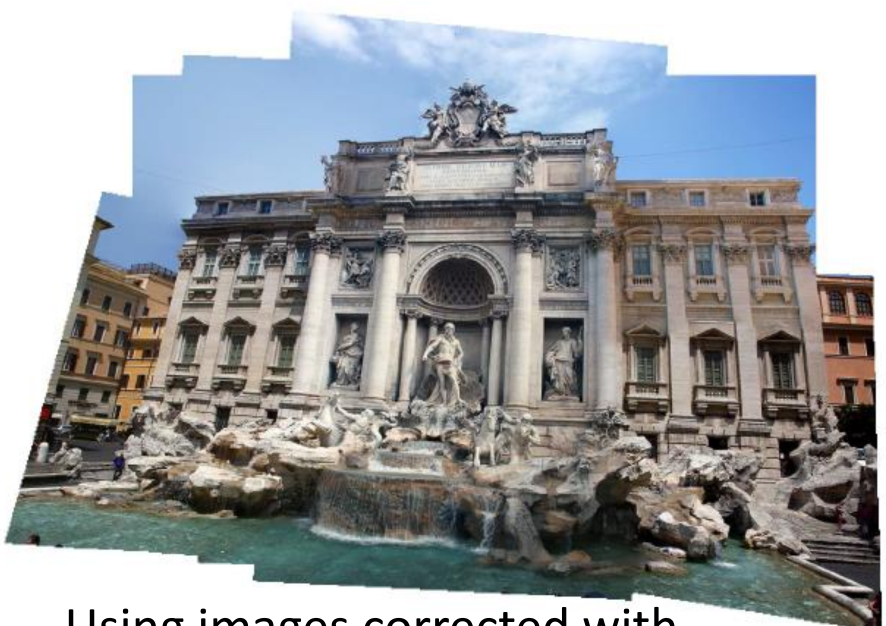
Results – ICE SKATER (36 images)

- Our method is faster than [HaCohen et al. 2012] which requires dense correspondence.
- Robust formulation; resilient to outliers.

Image Stitching

Input Images for Microsoft ICE (stitcher)

Using original images



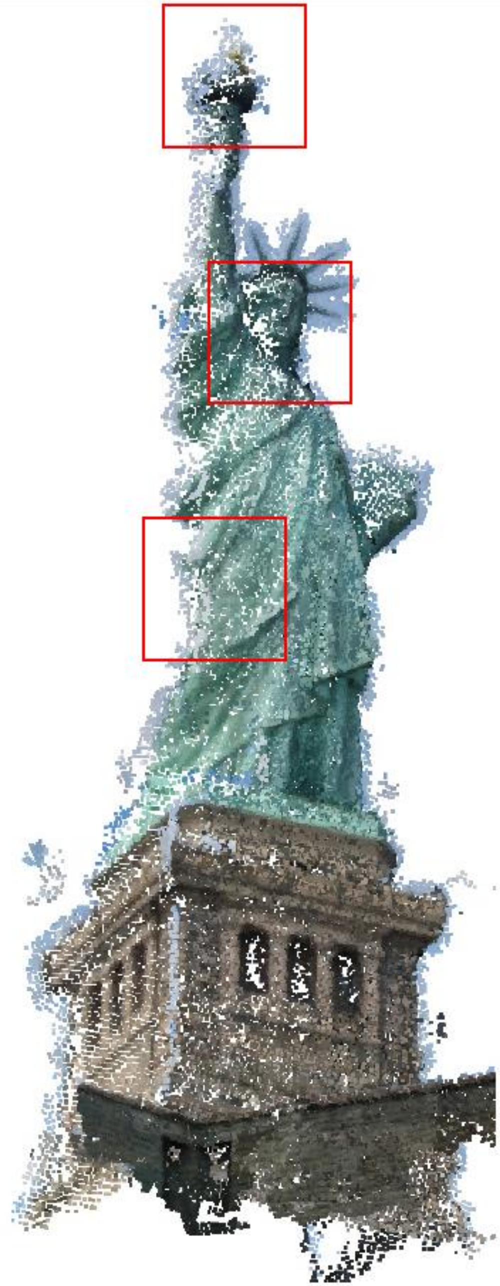
Using images corrected with
Photoshop CS6

Using our corrected images

Multi-view Stereo

original images

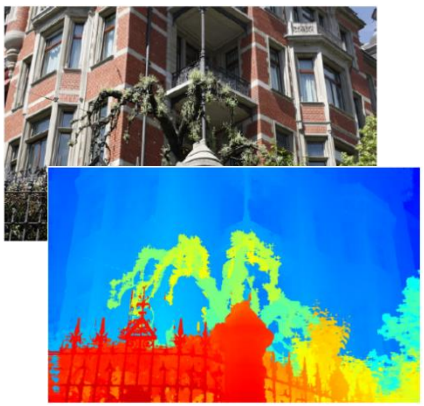
corrected images



Using original images

Using corrected images

Conclusions



High resolution Stereo

- Local plane sweeps
- Reduce search space
- SGM optimization

Flow + Cosegmentation

- Joint formulation
- Hierarchical MRF model
- Continuous labels
- Graph cuts

Color Consistency in Photo Collections

- Uses sparse feature matches
- Robust matrix factorization
- Efficient color transfer