Efficient Dense 3D Rigid-Body Motion Segmentation in RGB-D Video

Jörg Stückler and Sven Behnke

Autonomous Intelligent Systems Group
Computer Science Institute VI
University of Bonn, Germany

Objective: segment RGB-D images into moving rigid parts and estimate motion of parts efficiently

I_{ref}

$I_{seg}, \mathcal{L}$

Objective: segment RGB-D images into moving rigid parts and estimate motion of parts efficiently

I_{ref}

$I_{seg}, \mathcal{L}$

Difficulty

- abitrary camera/object motion
- unknown number of segments

Objective: segment RGB-D images into moving rigid parts and estimate motion of parts efficiently

I_{ref}

$I_{seg}, \mathcal{L}$

Difficulty

- abitrary camera/object motion
- unknown number of segments

Existing methods

- sparse: require texture
- dense: computationally demanding, do not exploit RGB-D

Approach: Expectation-Maximization

$$\arg\max_{\Theta} \sum_{\mathcal{L}} \underbrace{p(\mathcal{L} \mid I_{seg}, \Theta, I_{ref})}_{\text{labelling likelihood}} \underbrace{\ln p(I_{seg} \mid \Theta, I_{ref}, \mathcal{L})}_{\substack{\text{observation likelihood given} \\ \text{labelling } \mathcal{L} \text{ and rigid-body motions } \Theta}}$$

Approach: Expectation-Maximization

$$\arg\max_{\Theta} \sum_{\mathcal{L}} \underbrace{p(\mathcal{L} \mid I_{seg}, \Theta, I_{ref})}_{\text{labelling likelihood}} \underbrace{\ln p(I_{seg} \mid \Theta, I_{ref}, \mathcal{L})}_{\substack{\text{observation likelihood given} \\ \text{labelling } \mathcal{L} \text{ and rigid-body motions } \Theta}}$$

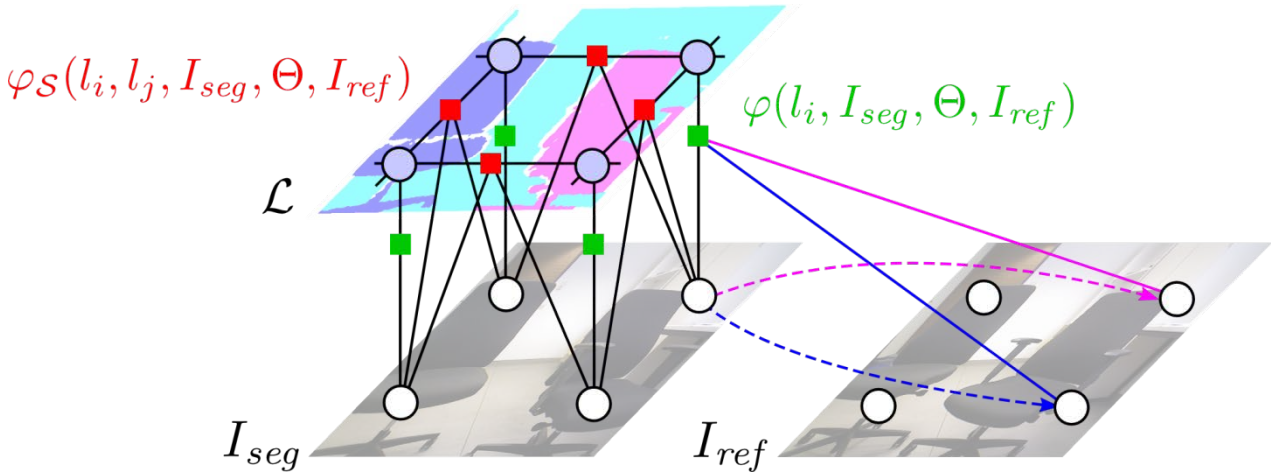
E-step: „soft“ labelling into segments of coherent rigid-body motion

Approach: Expectation-Maximization

$$\arg\max_{\Theta} \sum_{\mathcal{L}} \underbrace{p(\mathcal{L} \mid I_{seg}, \Theta, I_{ref})}_{\text{labelling likelihood}} \ln \underbrace{p(I_{seg} \mid \Theta, I_{ref}, \mathcal{L})}_{\substack{\text{observation likelihood given} \\ \text{labelling } \mathcal{L} \text{ and rigid-body motions } \Theta}}$$

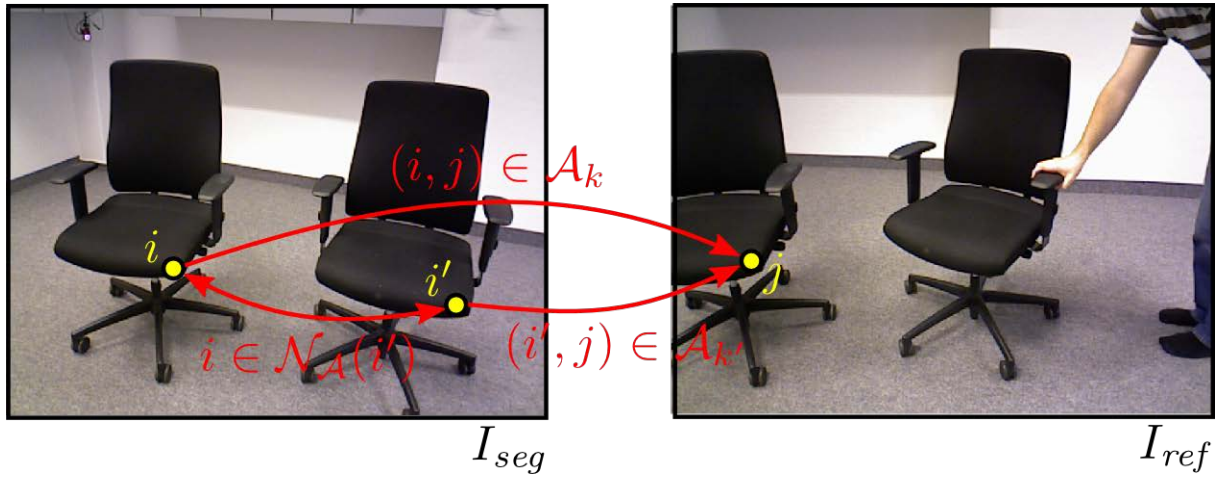
E-step: „soft“ labelling into segments of coherent rigid-body motion

M-step: optimize for rigid-body motions of segments



$$p(\mathcal{L} \mid I_{seg}, \Theta, I_{ref}) \propto \exp\left(\sum_{l_i \in \mathcal{C}_U} \ln \varphi(l_i, I_{seg}, \Theta, I_{ref}) + \sum_{(l_i,l_j) \in \mathcal{C}_P} \ln \varphi(l_i, l_j, I_{seg}, \Theta, I_{ref})\right)$$

- unary potentials: observation likelihood under motion
- pair-wise potentials:
  - spatial coherence, contrast-/curvature-sensitive Potts model
  - unique associations with sites in reference image



Discount labelling-pairs that associate with same site in reference image:

$$\ln \varphi_{\mathcal{A}}(l_i, l_j, I_{seg}, \Theta, I_{ref}) := \begin{cases} -\alpha & \text{if } l_i = k \wedge l_j = k' \wedge k \neq k' \\ 0 & \text{otherwise} \end{cases}$$

Back to the EM-Objective

$$\arg\max_{\Theta} \sum_{\mathcal{L}} \underbrace{p(\mathcal{L} \mid I_{seg}, \Theta, I_{ref})}_{\text{labelling likelihood}} \underbrace{\ln p(I_{seg} \mid \Theta, I_{ref}, \mathcal{L})}_{\text{observation likelihood given labelling } \mathcal{L} \text{ and rigid-body motions } \Theta}$$

Considering the summation over all labellings would be intractable!

Factored Observations

$$\arg\max_{\Theta} \sum_{\mathcal{L}} \quad p(\mathcal{L} \mid I_{seg}, \Theta, I_{ref}) \qquad \sum_i \ln p(x_i \mid \theta_{l_i}, I_{ref})$$

Assume stochastic independence between image site observations

Variational Approximation

$$\arg\max_{\Theta} \sum_{\mathcal{L}} \prod_i q(l_i \mid I_{seg}, \theta_{l_i}, I_{ref}) \qquad \sum_i \ln p(x_i \mid \theta_{l_i}, I_{ref})$$

Mean-field approximation of labelling likelihood:

$$p(\mathcal{L} \mid I_{seg}, \Theta, I_{ref}) \approx \prod_i q(l_i \mid I_{seg}, \theta_{l_i}, I_{ref})$$

Variational Approximation

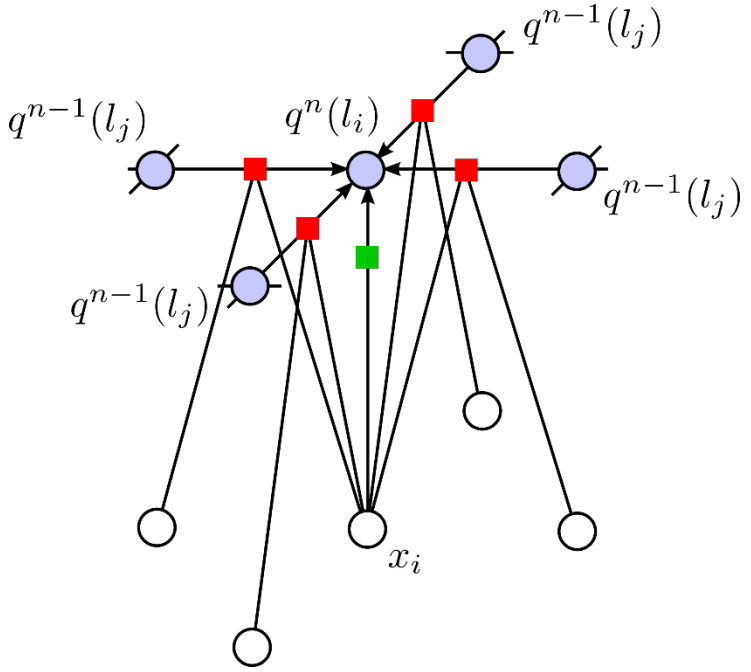
$$\arg\max_{\Theta} \sum_k \sum_i \; q(l_i = k \mid I_{seg}, \theta_{l_i}, I_{ref}) \qquad \ln p(x_i \mid \theta_{l_i}, I_{ref})$$

Mean-field approximation of labelling likelihood:

$$p(\mathcal{L} \mid I_{seg}, \Theta, I_{ref}) \approx \prod_i q(l_i \mid I_{seg}, \theta_{l_i}, I_{ref})$$

→ motion can be optimized for each label separately!

Minimizing KL-divergence of mean-field approximation yields **local updates**



$$\ln q(l_i) =$$

$$const. + \ln \varphi(l_i, I_{seg}, \Theta, I_{ref}) + \sum_{j \in \mathcal{C}_P(i)} \sum_{l_j} q(l_j) \ln \varphi(l_i, l_j, I_{seg}, \Theta, I_{ref})$$

with $\sum_i q(l_i) = 1$ and $q(l_i) := q(l_i \mid I_{seg}, \theta_{l_i}, I_{ref})$

Initialize mean-field iterations

with graph-cuts ML solution

$$\mathcal{L}_{ML} := \arg\max_{\mathcal{L}} p(\mathcal{L} \mid I_{seg}, \Theta, I_{ref})$$

$$q^0(l_i) := \begin{cases} 1 & \text{if } l_i = l_{ML,i} \\ 0 & \text{otherwise} \end{cases}$$

Initialize mean-field iterations

with graph-cuts ML solution

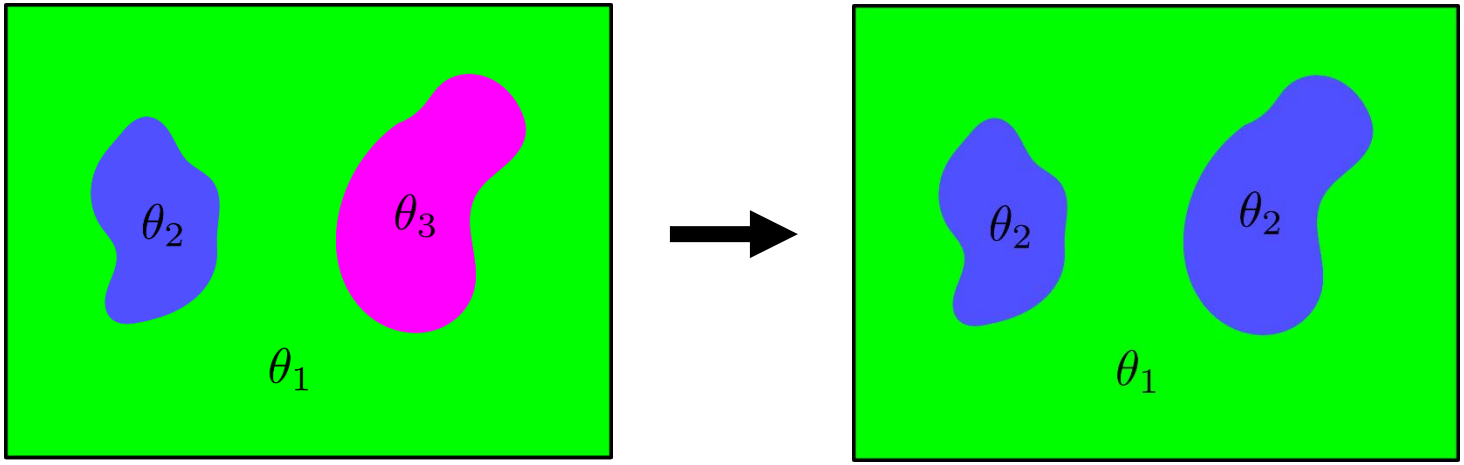
$$\mathcal{L}_{ML} := \arg\max_{\mathcal{L}} p(\mathcal{L} \mid I_{seg}, \Theta, I_{ref})$$

$$q^0(l_i) := \begin{cases} 1 & \text{if } l_i = l_{ML,i} \\ 0 & \text{otherwise} \end{cases}$$

One step update:

$$q^1(l_i) = \eta\, \varphi(l_i, I_{seg}, \Theta, I_{ref}) \prod_{j \in \mathcal{C}_P(i)} \varphi(l_i, l_{ML,j}, I_{seg}, \Theta, I_{ref})$$

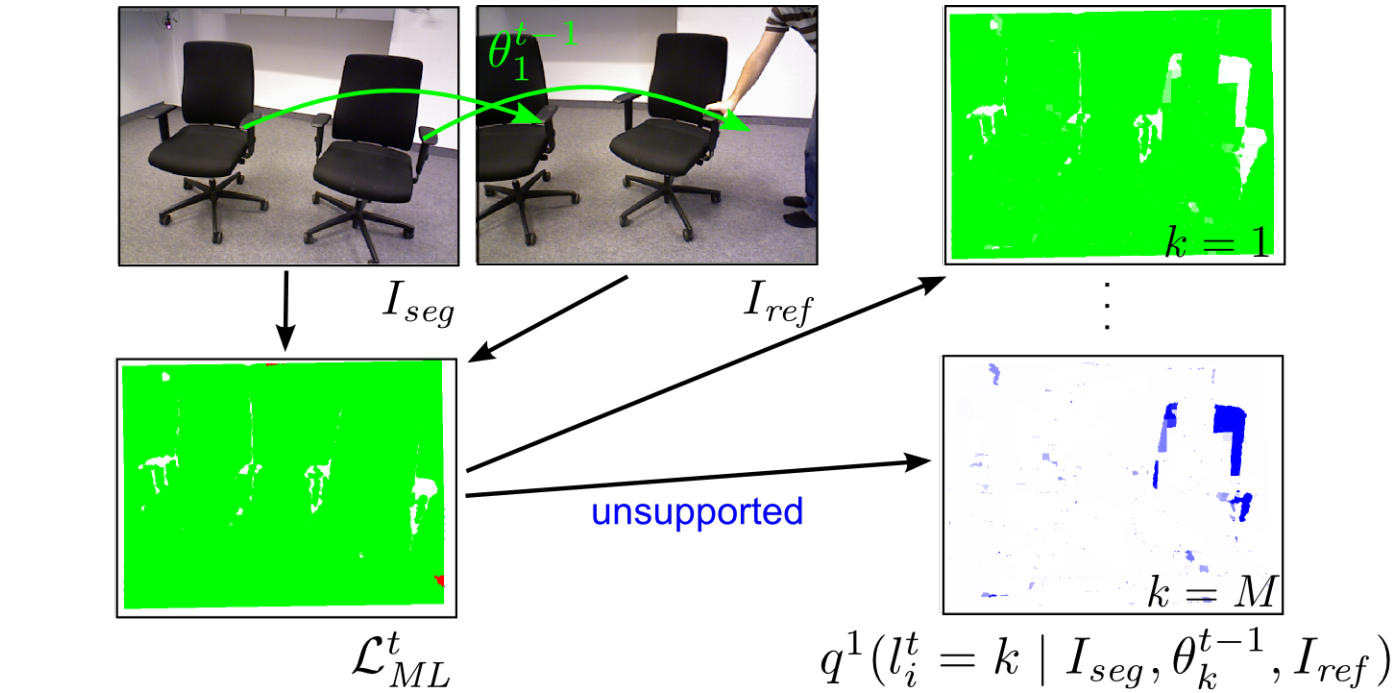
→ local conditionals on ML solution!

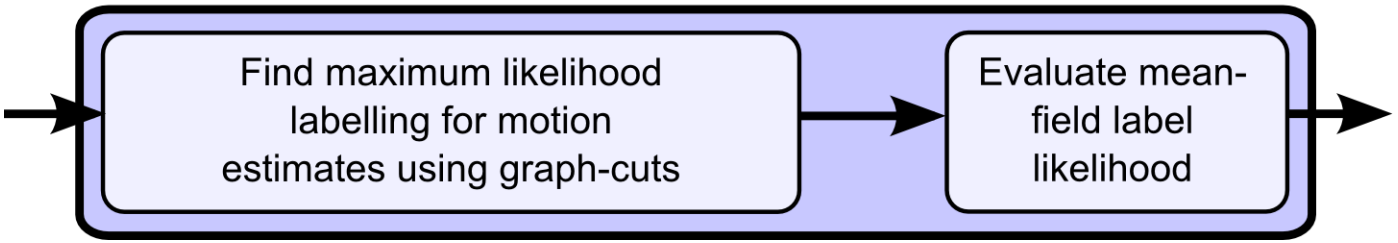


$\theta_2 \approx \theta_3$

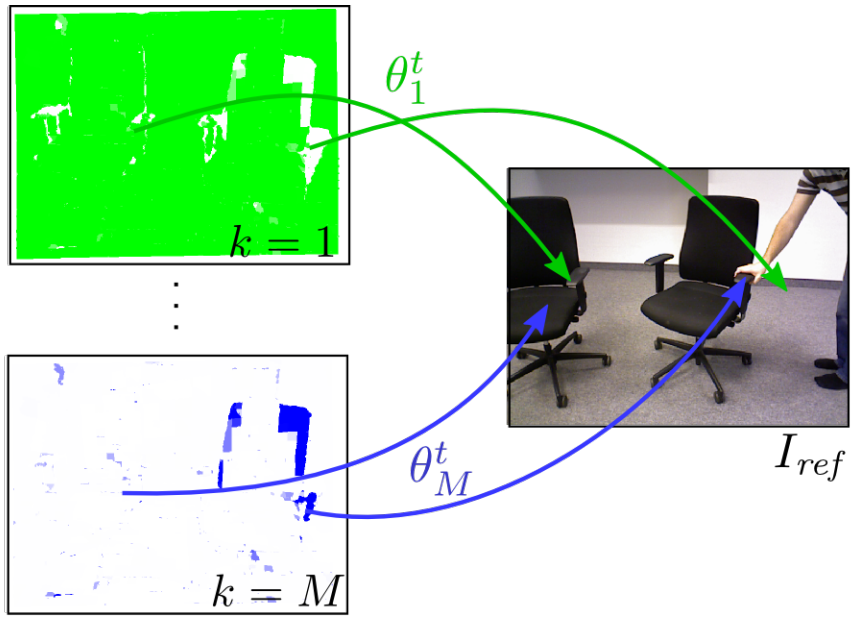
- Trade model complexity via label costs (Delong et al., 2012)
- Inject one new label per EM-iteration, remove unused labels

Efficient EM-Algorithm: E-Step





Efficient EM-Algorithm: M-Step

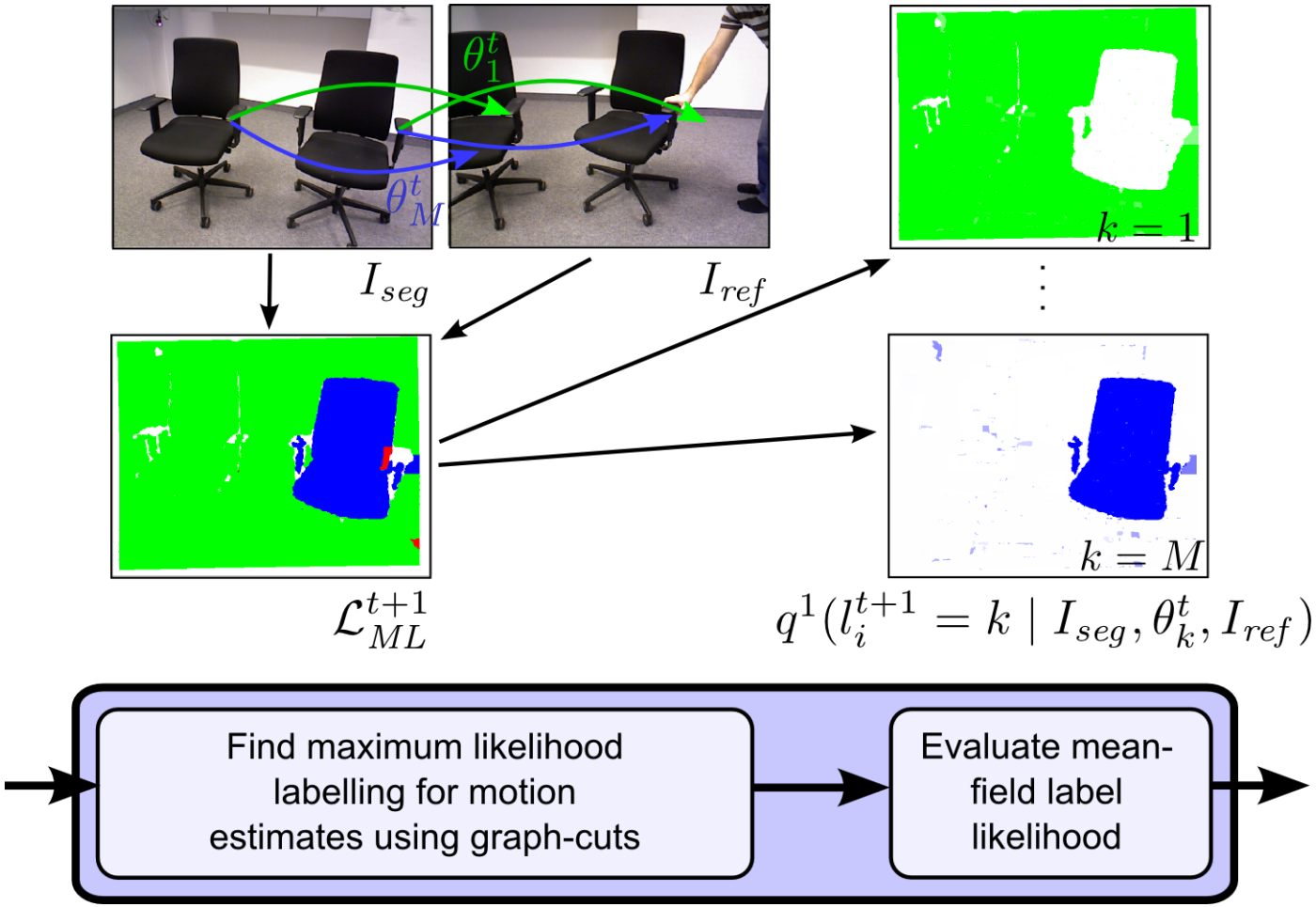


$$q^1(l_i^t = k \mid I_{seg}, \theta_k^{t-1}, I_{ref})$$

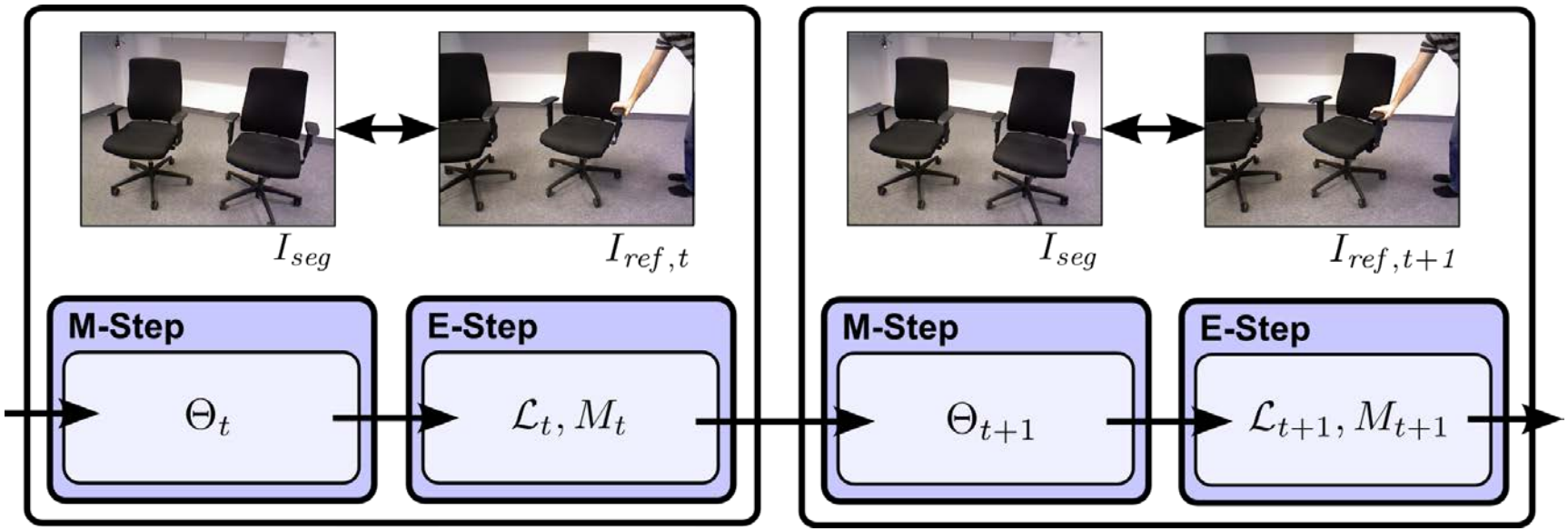
Determine best motion estimates
given mean-field labelling likelihood

$$\arg\max_{\theta_k} \sum_i q^1(l_i^t = k) \ln p(x_i \mid \theta_k, I_{ref})$$

Efficient EM-Algorithm: E-Step

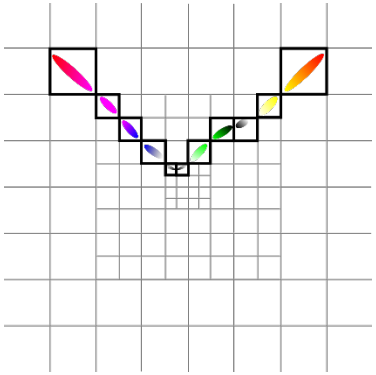


Sequential EM



Efficient Image Representation

- Represent RGB-D images compactly in multi-resolution surfel maps
- Shape and color distribution in octree
- Local multi-resolution
- From 640x480 pixels to several 1,000 voxels
- Fast aggregation and registration on CPU



5cm

2.5cm

Experiments: Setup

- RGB-D sequences, 30Hz, 640x480
- Moving objects of varying sizes/shape/texture
- Moving camera
- Ground truth:
  - motion capture
  - manual labelling of frames at every 5 sec
- Evaluation of pose and segmentation accuracy
- Run-time evaluated on Intel Core-i7-4770K QuadCore CPU (3.50GHz)

Experiments: Results

all frames

Run-time (msec): 200.2 ± 42.3
Error in #segments: 0.05 ± 0.29
Avg. seg. acc.: 0.95
Median lin. error (m): 0.012
Median ang. error (rad): 0.047

Run-time (msec): 213.1 ± 54.7
Error in #segments: 0.11 ± 0.43
Avg. seg. acc.: 0.94
Median lin. error (m): 0.018
Median ang. error (rad): 0.029

Run-time (msec): 138.7 ± 37.5
Error in #segments: -0.58 ± 1.01
Avg. seg. acc.: 0.63
Median lin. error (m): 0.034
Median ang. error (rad): 0.049

Experiments: Results

all frames (real-time)

Run-time (msec): 200.2 ± 42.3
Error in #segments: 0.05 ± 0.29 (-0.09 ± 0.35)
Avg. seg. acc.: 0.95 (0.91)
Median lin. error (m): 0.012 (0.013)
Median ang. error (rad): 0.047 (0.045)

Run-time (msec): 213.1 ± 54.7
Error in #segments: 0.11 ± 0.43 (0.04 ± 0.45)
Avg. seg. acc.: 0.94 (0.91)
Median lin. error (m): 0.018 (0.020)
Median ang. error (rad): 0.029 (0.030)

Run-time (msec): 138.7 ± 37.5
Error in #segments: -0.58 ± 1.01 (-0.43 ± 0.92)
Avg. seg. acc.: 0.63 (0.65)
Median lin. error (m): 0.034 (0.030)
Median ang. error (rad): 0.049 (0.048)

Summary

- Efficient EM for rigid-body motion segmentation in RGB-D video
- Good accuracy in segmentation and pose estimate
- Limitations: 3D aperture, local optima in registration

Future Work:

- Other image representations/registration methods, GPU
- Further cues (e.g., IMU) to handle 3D aperture problems

- Multi-resolution surfel maps: http://code.google.com/p/mrsmap
- Dataset: http://www.ais.uni-bonn.de/download/rigidmultibody

Thank You!