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

Jörg Stückler
http://www.ais.uni-bonn.de/~stueckler
Sven Behnke
http://www.ais.uni-bonn.de/behnke

Computer Science Institute VI
University of Bonn
Bonn, Germany

Common motion is a fundamental grouping cue in video sequences. While for monocular and stereo image sequences, several approaches to motion segmentation have been investigated, it still remains a research problem to compute dense 3D motion segmentation efficiently. Many approaches match images sparsely at interest points and infer the groups of points with common 3D rigid-body motion (e.g., [1, 5]). Methods for dense 3D motion segmentation are still far from real-time performance (e.g., [6, 8]).

We propose in this paper an efficient approach to dense 3D motion segmentation in RGB-D video. We formulate an EM framework (see Fig. 1) that recovers a segmentation $\mathcal{L} = \{l_i = k\}_{i=1}^{N}$ into motion segments $\mathcal{M} = \{m_k\}_{k=1}^{M}$, estimates their 3D rigid-body motions $\Theta = \{\theta_k\}_{k=1}^{M}$, and also finds the number of segments M by optimizing

$$\arg\max_{\Theta} \sum_{\mathcal{L}} p(\mathcal{L} \mid I_{seg}, \overline{\Theta}, I_{ref}) \; \ln p(I_{seg} \mid \Theta, I_{ref}, \mathcal{L}), \tag{1}$$

Our approach makes no difference between background and foreground objects and copes with camera motion as well as multiple moving objects.

We model the likelihood of a labelling $\mathcal{L}$ in a random field

$$p(\mathcal{L} \mid I_{seg}, \Theta, I_{ref}) \propto \prod_{i} p(z_i \mid \theta_{l_i}, I_{ref}) \prod_{j \in \mathcal{N}(i)} p(l_i, l_j \mid I_{seg}) \tag{2}$$

that incorporates the likelihood of the data at each site and pair-wise interaction terms between neighbors $\mathcal{N}(i)$ of site i. The data likelihood $p(z_i \mid \theta_{l_i}, I_{ref})$ quantifies the likelihood of the observation $z_i \in I_{seg}$ at a site under its label's motion estimate θ_{l_i}. For the pair-wise interaction terms we use a contrast-sensitive Potts model [2]. We incorporate additional pair-wise couplings to avoid multiple data associations of a site in the segmented image to the same site in the reference image.

Efficient graph-cuts [3] are then applied find a maximum likelihood labelling $\mathcal{L}_{ML} = \arg\max_{\mathcal{L}} p(\mathcal{L} \mid I_{seg}, \overline{\Theta}, I_{ref})$. To estimate motion, we apply approximations to arrive at

$$\arg\max_{\Theta} \sum_{i} \sum_{l_i} p(l_i \mid \mathcal{L}_{ML} \setminus \{l_i\}, I_{seg}, \overline{\Theta}, I_{ref}) \ln p(z_i \mid \theta_{l_i}, I_{ref}). \tag{3}$$

The weight of a site intuitively is the likelihood of belonging to a segment.

Model complexity is controlled using label costs [4]. We start with one segment and perform EM with one additional segment in each iteration until the number of segments keeps constant and EM has converged. For the sequential segmentation of an image towards a video we propose to initialize EM in each frame with the result from the previous frame. This way, the EM algorithm requires only few iterations.

The performance of our EM approach strongly depends on the underlying image representation. In principle, any representation is suitable that defines data likelihood $p(z_i \mid \theta_{l_i}, I_{ref})$, image site neighborhood $\mathcal{N}_{\mathcal{S}}(i)$, and dissimilarity for the pair-wise interaction terms. To solve Eq. (3), an image registration technique is required that allows to incorporate individual weights for the image sites. Instead of segmenting the large number of pixels in the image, we represent RGB-D images compactly as multi-resolution surfel maps [7]. These maps capture noise characteristics in a local multi-resolution structure in which the maximum resolution adapts to the distance of the measurements. In effect, the content of an RGB-D image is compressed from 640×480 pixels to only several thousand voxels, making dense inference of labels in the map efficient.

In experiments on test sequences with ground-truth segmentations and rigid-body motions, we demonstrate that our approach efficiently identifies moving segments with high accuracy and recovers 3D rigid-body motion of the segments at good accuracy (see Fig. 2 for examples), also at real-time operation on a CPU.

[1] M. Agrawal, K. Konolige, and L. Iocchi. Real-time detection of independent motion using stereo. In *Proc. of the IEEE Workshop on Motion*, 2005.

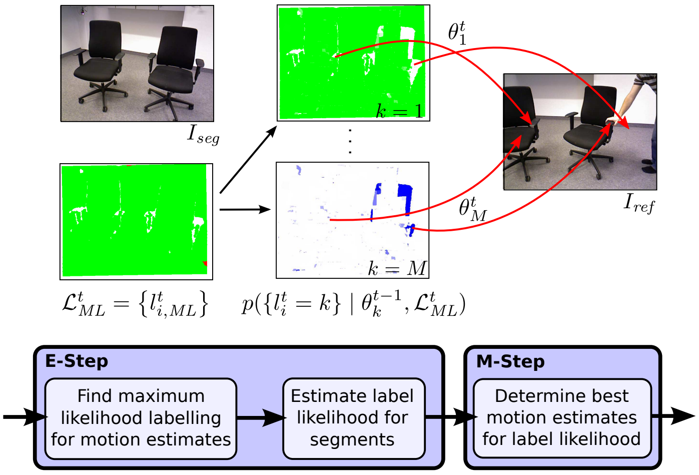
θ_1^t
I_{seg}
$k=1$
θ_M^t
$k=M$
I_{ref}
$\mathcal{L}_{ML}^t = \{l_{i,ML}^t\}$
$p(\{l_i^t = k\} \mid \theta_k^{t-1}, \mathcal{L}_{ML}^t)$
E-Step
Find maximum likelihood labelling for motion estimates
Estimate label likelihood for segments
M-Step
Determine best motion estimates for label likelihood

Figure 1: We segment motion in an RGB-D image I_{seg} towards a reference image I_{ref} in an efficient EM framework. In the E-step, we evaluate the likelihood of image site labels l_i under the latest motion estimates θ_k. Efficient graph cuts yield a maximum likelihood labelling $\mathcal{L}_{ML}$ given the motion estimates, which is then used to approximate the label likelihoods. In the M-step, motion is reestimated taking into account label likelihoods.

[2] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. In *Proc. of the IEEE Int. Conf. on Computer Vision*, 2001.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23:2001, 2001.

[4] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. *Int. J. of Computer Vision*, 96 (1):1–27, 2012.

[5] A. Gruber and Y. Weiss. Multibody factorization with uncertainty and missing data using the EM algorithm. In *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[6] A. Roussos, C. Russell, R. Garg, and L. de Agapito. Dense multibody motion estimation and reconstruction from a handheld camera. In *Proc. of the IEEE Int. Symp. on Mixed and Augmented Reality (ISMAR)*, 2012.

[7] J. Stückler and S. Behnke. Multi-resolution surfel maps for efficient dense 3D modeling and tracking. *J. of Visual Communication and Image Representation*, 2013.

[8] G. Zhang, J. Jia, and H. Bao. Simultaneous multi-body stereo and segmentation. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, 2011.

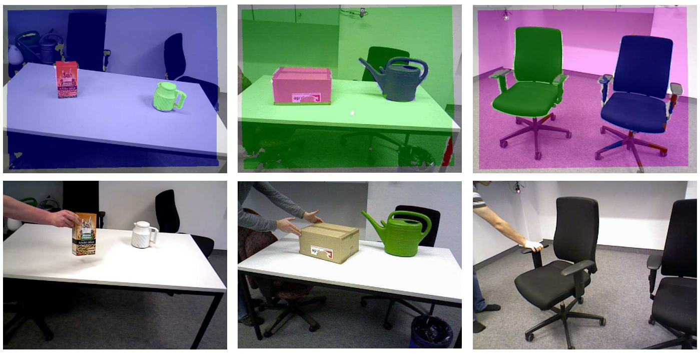

Figure 2: Example segmentations (outliers dark red).