In Proceedings of 24th British Machine Vision Conference (BMVC), Bristol, UK, 2013.

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

Abstract

Motion is a fundamental segmentation cue in video. Many current approaches segment 3D motion in monocular or stereo image sequences, mostly relying on sparse interest points or being dense but computationally demanding. We propose an efficient expectation-maximization (EM) framework for dense 3D segmentation of moving rigid parts in RGB-D video. Our approach segments two images into pixel regions that undergo coherent 3D rigid-body motion. Our formulation treats background and foreground objects equally and poses no further assumptions on the motion of the camera or the objects than rigidness. While our EM-formulation is not restricted to a specific image representation, we supplement it with efficient image representation and registration for rapid segmentation of RGB-D video. In experiments we demonstrate that our approach recovers segmentation and 3D motion at good precision.

1 Introduction

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 [1, 9, 12, 13, 15]. Methods for dense 3D motion segmentation are still far from real-time performance [14, 16, 23, 25].

In this paper, we propose an efficient approach to dense 3D motion segmentation. We formulate an expectation-maximization framework (see Fig. 1) that recovers motion segments, estimates their 3D rigid-body motion, and also finds the number of segments in the scene. Our formulation makes no difference between background and foreground objects and, hence, copes well with camera motion and multiple moving objects in the scene. We exploit dense depth information from RGB-D cameras and utilize highly efficient probabilistic image representation and registration techniques to obtain a rapid segmentation method. Instead of segmenting the large number of pixels in the image, we represent RGB-D images compactly as point distributions in 3D voxels at multiple resolutions. These maps capture the noise characteristics of the sensor in a local multi-resolution structure in which the maximum resolution in the map 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,

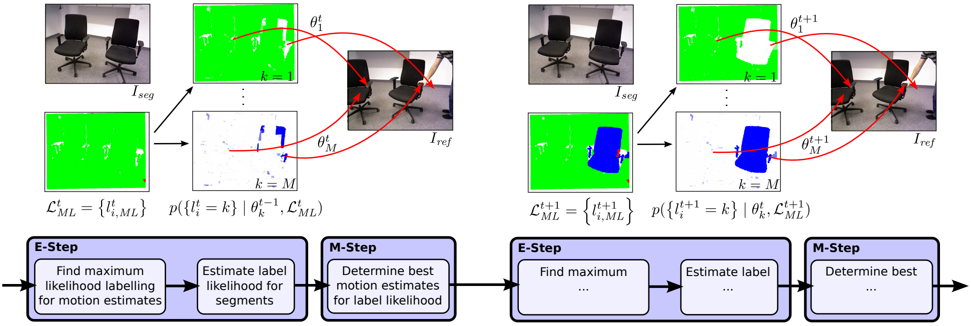
I_{seg}
θ_1^t
$k=1$
I_{ref}
θ_M^t
$k=M$
$\mathcal{L}_{ML}^t = \{l_{i,ML}^t\}$
$p(\{l_i^t = k\} \mid \theta_k^{t-1}, \mathcal{L}_{ML}^t)$
I_{seg}
θ_1^{t+1}
$k=1$
I_{ref}
θ_M^{t+1}
$k=M$
$\mathcal{L}_{ML}^{t+1} = \{l_{i,ML}^{t+1}\}$
$p(\{l_i^{t+1} = k\} \mid \theta_k^t, \mathcal{L}_{ML}^{t+1})$
E-Step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

Figure 1: We segment motion in an RGB-D image I_{seg} towards a reference image I_{ref} in an efficient expectation-maximization 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, new motion estimates for each segment are found through image registration which takes the soft assignment of sites to labels into account.

making dense inference of labels in the map efficient. In experiments, we demonstrate that our approach efficiently identifies moving segments with high accuracy and recovers 3D rigid-body motion of the segments at good precision.

2 Related Work

Several approaches to 3D motion segmentation have been proposed that represent images sparsely through image features. Multi-body factorization methods [24] find groups of points with common 3D rigid-body motion through factorization of the measurement matrix. These approaches have been extended to also cope with outliers and noisy observations [9, 13, 15]. Exploiting depth measurements for interest points from a calibrated stereo camera, Agrawal et al. [1] propose a real-time capable framework for 3D motion segmentation based on RANSAC and structure-from-motion. These approaches, however, do not provide dense segmentations. Some approaches segment 2D image motion densely based on optical flow. Cremers and Soatto [5] propose motion competition, a variational framework for dense motion segmentation of monocular image sequences. They estimate the 2D parametric motion of multiple motion segments. Occlusions and multiple data associations are explicitly modelled in the variational framework of Unger et al. [19], but the method is far from real-time performance. In our approach, we also handle multiple data associations as additional pairwise labelling constraints during graph-cut optimization of the motion segmentation. Kumar et al. [10] segment scenes into 2D motion layers using a conditional random field model that incorporates occlusions and lighting conditions. The work by Ayvaci and Soatto [2] defines an energy functional on a superpixel graph which is optimized using efficient graph cuts. While these methods yield impressive results, they estimate motion of 2D layers in the image and do not necessarily provide segments with consistent 3D rigid-body motion.

Superpixel segmentation can also be formulated based on color, stereo depth, and stereo 3D flow simultaneously [21]. This approach operates at about 2 Hz using a GPU for optical flow computation and is not designed to find coherent segments of rigid-body motion. With a stereo camera, Zhang et al. [25] propose dense 3D multibody structure-from-motion using an energy minimization framework. The approach relies on plane fitting to make the segmentation robust and is reported to require ca. 10 min per frame. Wang et al. [23] transfer the approach of Cremers and Soatto [5] to 3D time-of-flight images. They formulate a 3D optical flow constraint, and optimize for the 3D motion segmentation using level sets, but do not report on computational load. Recently, a variational framework has been proposed that integrates rigid-body motion segmentation with dense 3D reconstruction [14] from monocular image sequences. The batch method requires about 8 to 9 sec per frame on a GPU. We make efficient use of dense depth in RGB-D images for 3D motion segmentation—also integrating texture cues. The frame-rate of our approach is between 2 to 10 Hz on a CPU.

In simultaneous localization mapping and moving object tracking (SLAMMOT, [22]), dynamic objects are segmented in laser scans through distance comparisons, and subsequently tracked while concurrently mapping the environment statics in a SLAM framework. Van de Ven et al. [20] recently proposed a graphical model that integrates CRF-Matching [11] and CRF-Clustering [18] within a single framework for 2D scan-matching, moving object detection, and motion estimation. They infer associations, motion segmentation, and 2D rigid-body motion through inference in the model using max-product loopy belief propagation. We formulate dense 3D motion segmentation of RGB-D images using expectation-maximization and perform fast approximate inference using graph cuts.

In summary, the contributions of our work are a general expectation-maximization framework for dense sequential 3D rigid-body motion segmentation in RGB-D video with tractable efficient approximations, and an efficient implementation based on a compact image representation and fast probabilistic registration techniques.

3 Dense 3D Motion Segmentation of Rigid Parts

Our approach segments moving rigid parts between two RGB-D frames, i.e., it determines the number of rigid parts, their 3D rigid-body motion, and the image regions that map the parts. We assume that an image I is partitioned into a set of discrete sites $I = \{z_i\}_{i=1}^N$ such as pixels or map elements in a 3D representation. Let $\mathcal{L} = \{l_i\}_{i=1}^N$ be the labelling of the image sites. The labelling $l_i = k$, $k \in \{O, 1, \ldots, M\}$ denotes the membership of site i in one of the distinct motion segments $\mathcal{M} = \{m_k\}_{k=1}^M$ or in the set of outliers O. All sites within a segment move with a common six degree-of-freedom (6-DoF) rigid-body motion θ_k between the segmented image I_{seg} and a reference image I_{ref}.

3.1 Expectation-Maximization Framework

We explain the segmented image by the rigid-body motion of segments towards the reference image, i. e., we seek rigid-body motions $\Theta = \{\theta_k\}_{k=1}^M$ that maximize the observation likelihood of the segmented image in the reference image $\arg\max_\Theta p(I_{seg} \mid \Theta, I_{ref})$. In our formulation, the labelling of the image sites is a latent variable that we estimate jointly with the rigid-body motions of the segments using expectation-maximization (EM) [7],

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

where $\overline{\Theta}$ is the latest motion estimate of the segments from the previous iteration of the EM algorithm, and $p(\mathcal{L} \mid I_{seg}, \overline{\Theta}, I_{ref})$ is the posterior distribution of the image labelling. The EM

algorithm alternates the following two steps in several iterations until convergence, or until a maximum number of iterations is reached:
E-step: Determine the posterior distribution of the image labelling given the latest motion estimates $\overline{\Theta}$ to form the conditional expectation in (1).
M-step: Find new motion estimates Θ by maximizing the conditional expectation (1), given the posterior distribution of the image labelling.

3.2 Image Labelling Posterior

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 outlier label $l_i = O$, we set the data likelihood to a constant p_O. In our concrete implementation, an image site i is transformed into the frame of the reference image I_{ref} given the motion estimate for its labelling. Subsequently, the site is associated with a closest site in the reference image. The data likelihood for site i is retrieved from this matching. For the pair-wise interaction terms we use a contrast-sensitive Potts model [3]

$$\ln p_{\mathcal{S}}(l_i, l_j \mid I_{seg}) = -\gamma(z_i, z_j)\, \delta(l_i, l_j), \text{ where } \delta(l_i, l_j) := \begin{cases} 0 & , \text{if } l_i = l_j, \\ 1 & , \text{if } l_i \neq l_j, \end{cases}, \tag{3}$$

and $\gamma(z_i, z_j) > 0$ controls the strength of the coupling.

3.3 Efficient Solution of the Expectation-Maximization Formulation

We propose an efficient solution to the EM formulation. Firstly, we see that the matching likelihood between image segments towards the reference image given motion estimates and labelling, $p(I_{seg} \mid \Theta, I_{ref}, \mathcal{L})$, factorizes into the matching likelihood of the individual observations since we assume stochastic independence between the observations and each site is associated to exactly one segment given a concrete labelling, i.e., $p(I_{seg} \mid \Theta, I_{ref}, \mathcal{L}) = \prod_i p(z_i \mid \theta_{l_i}, I_{ref})$. By this, Eq. (1) becomes $\arg\max_\Theta \sum_{\mathcal{L}} p(\mathcal{L} \mid I_{seg}, \overline{\Theta}, I_{ref}) \sum_i \ln p(z_i \mid \theta_{l_i}, I_{ref})$. Note that each term of the inner sum only depends on one of the image labels.

Since exact inference of the joint label likelihood $p(\mathcal{L} \mid I_{seg}, \overline{\Theta}, I_{ref})$ is not tractable, we need to resort to approximations. One possible approach is to use inference algorithms such as loopy belief propagation to infer the marginal distribution over site labellings $p(l_i \mid I_{seg}, \overline{\Theta}, I_{ref})$, and to optimize $\arg\max_\Theta \sum_{\mathcal{L}} \sum_i p(l_i \mid I_{seg}, \overline{\Theta}, I_{ref}) \ln p(z_i \mid \theta_{l_i}, I_{ref})$.

We take a more efficient approach by using graph-cuts [4] to find an approximate maximum likelihood labelling $\mathcal{L}_{ML} = \arg\max_{\mathcal{L}} p(\mathcal{L} \mid I_{seg}, \overline{\Theta}, I_{ref})$. Next, we apply a mean field approximation to the joint label likelihood $p(\mathcal{L} \mid I_{seg}, \overline{\Theta}, I_{ref})$ to write

$$\begin{aligned} \arg\max_\Theta \sum_{l_1} p(l_1 \mid \mathcal{L}_{ML} \setminus \{l_1\}, I_{seg}, \overline{\Theta}, I_{ref}) \dots \\ \sum_{l_N} p(l_N \mid \mathcal{L}_{ML} \setminus \{l_N\}, I_{seg}, \overline{\Theta}, I_{ref}) \sum_i \ln p(z_i \mid \theta_{l_i}, I_{ref}), \end{aligned} \tag{4}$$

where $p(l_i \mid \mathcal{L} \setminus \{l_i\}, I_{seg}, \overline{\Theta}, I_{ref}) = \eta \; p(z_i \mid \overline{\theta}_{l_i}, I_{ref}) \prod_{j \in \mathcal{N}(i)} p(l_i, l_j \mid \mathcal{L} \setminus \{l_i\}, I_{seg})$ and η is a normalization constant such that $\sum_k p\left(l_i = k \mid \mathcal{L}_{ML} \setminus \{l_i\}, I_{seg}, \overline{\Theta}, I_{ref}\right) = 1$. That is, for each image site i, we set the labelling of the neighboring sites constant according to the maximum likelihood labelling $\mathcal{L}_{ML}$, and evaluate the local conditional likelihood of the site labelling l_i.

By rearranging the sums and exploiting the normalization, we arrive at

$$\arg\max_{\Theta} \sum_i \sum_{l_i} p(l_i \mid \mathcal{L}_{ML} \setminus \{l_i\}, I_{seg}, \overline{\Theta}, I_{ref}) \ln p(z_i \mid \theta_{l_i}, I_{ref}). \tag{5}$$

Each image site i is assigned a weight for the reestimation of the rigid-body motion θ_k. The weight intuitively is the likelihood that site i belongs to the segment.

3.4 Resolving Ambiguous Data Associations

Our approach also needs to avoid multiple associations of image sites in the segmented image with the same image site in the reference image. Otherwise, the approach could explain different parts of the segmented image with the same part in the reference image, e.g. at missing image overlap or in occluded regions. For sites i and j in the segmented image that map to the same site in the reference image for different motion segments k and k', repectively, we additionally model the pair-wise labelling probability

$$\ln p_{\mathcal{A}}(l_i, l_j) := \begin{cases} -\alpha & \text{, if } l_i = k \wedge l_j = k', \\ 0 & \text{, otherwise,} \end{cases} \tag{6}$$

where α sets the strength of the couplings.

3.5 Model Complexity

The pair-wise interaction terms prefer large motion segments and naturally control the number of segments to be small. In the case that a single 3D motion segment occurs as multiple unconnected image segments in the image, our approach so far may still use different but redundant motion segments for the image segments. To control model complexity, we enhance the graph-cut optimization in Sec. 3.3 with label costs [6].

We initialize the EM algorithm with a guess of the number of motion segments ($M = 1$ in our experiments). To let our approach possibly increase the number of segments, we append one additional, yet unsupported segment before the M-step. All sites in segments that are yet unsupported in the image are assigned the outlier data likelihood p_O. By this, our EM algorithm prefers to explain sites that misalign with the already existing segments with new motion segments. We define a motion segment to be supported if it labels sites in the image and reject very small segments as outliers. Unsupported segments (eventually the additional segment) are dicarded after the E-step.

3.6 Sequential Segmentation

While our EM formulation may in principle segment motion between arbitrary images, we augment it to perform efficiently on image sequences. We segment the first image I_{seg} in a sequence iteratively towards subsequent images $I_{ref,t}$. At each new image at time t, our approach estimates the number of segments M_t, a new segmentation $\mathcal{L}_t$, and new motion estimates Θ_t. Instead of starting our EM procedure all over for each new image, we initialize the approach with the estimates from the last image $I_{ref,t-1}$. This way, the EM algorithm requires significantly less iterations per image to converge (typically one iteration suffices).

4 Image Representation

The performance of our EM algorithm in Sec. 3 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 $\gamma(z_i, z_j)$ for the pair-wise interaction terms. To solve for the motion estimates of the segments in Eq. (5), an image registration technique is required that allows to incorporate individual weights for the image sites.

Instead of processing the RGB-D image pixel-wise, we choose to represent the image content in compact multi-resolution 3D surfel maps (MRSMaps, [17]). This image representation respects the noise characteristics of the sensor, provides a probabilistic representation of the data, and supports efficient weighted registration. It stores the joint color and shape statistics of points within 3D voxels at multiple resolutions sparsely in an octree. The maximum resolution at a point is limited proportional to its squared depth in order to capture the disparity-dependent noise of the RGB-D camera. In effect, the map exhibits a local multi-resolution structure which well reflects the accuracy of the measurements and compresses the image from 640×480 pixels into only a few thousand voxels. Our MRSMap implementation is available open-source from http://code.google.com/p/mrsmap/ .

4.1 Data Likelihood in Multi-Resolution Surfel Maps

Each voxel in a MRSMap contains a surfel z_i which is defined by mean $\mu_i \in \mathbb{R}^6$ and covariance $\Sigma_i \in \mathbb{R}^{6\times 6}$ of the colored points falling into the voxel. Given the labelling l_i of the surfel, the surfel z_i^{seg} is observed at a corresponding surfel z_j^{ref} under the label's rigid-body motion θ_{l_i}, i.e.,

$$p(z_i^{seg} | z_j^{ref}, \theta_{l_i}) = \mathcal{N}\left(d_{i,j}(\theta_{l_i}); 0, \Sigma_{i,j}(\theta_{l_i})\right),$$
$$d_{i,j}(\theta_{l_i}) := \mu_j^{ref} - T(\theta_{l_i})\mu_i^{seg}, \; \Sigma_{i,j}(\theta_{l_i}) := \Sigma_j^{ref} + R(\theta_{l_i})\Sigma_i^{seg} R(\theta_{l_i})^T, \quad (7)$$

where $T(\theta_{l_i})$ is the transformation matrix for the pose estimate θ_{l_i} and $R(\theta_{l_i})$ is its rotation matrix. Note, that our data likelihood takes spatial as well as color information into account.

The evaluation of the data likelihood involves the association $(i, j) \in \mathcal{A}$ of the surfel z_i^{seg} with a surfel z_j^{ref} from the reference image. The mean position of the surfel z_i^{seg} is transformed to the reference image according to the motion estimate θ_{l_i}. We then search for a matching surfel in the reference image from coarse to fine resolutions. We adapt the search radius proportional to the resolution and find the association on the finest resolution possible.

Special care needs to be taken at image borders, background at depth discontinuities, and occlusions. We assign the last observed data likelihood at such borders and in occlusions.

4.2 Smoothness Terms in Multi-Resolution Surfel Maps

We establish pair-wise terms between all six direct neighbors of a voxel in the 3D grid. In addition, we couple a voxel with its children and its parent voxel within the octree. In this way, spatial coherence can be enforced despite the sparseness of the representation and

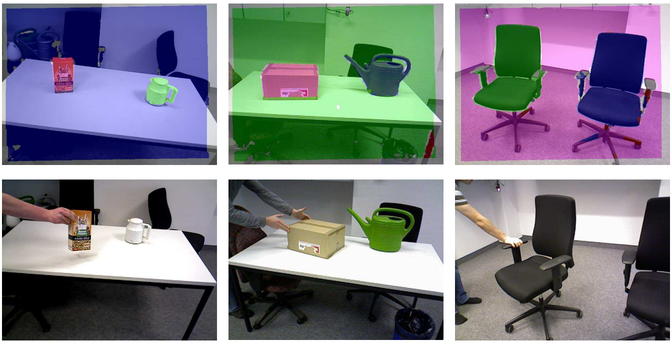

Figure 2: Example segmentations (top, outliers dark red) towards a reference image (bottom) from test sequences (left: small, middle: medium, right: large).

across the discrete changes of the depth-dependent resolution limit. We lessen pair-wise couplings between nodes at highly curved or textured spots,

$$\gamma(z_i, z_j) := \max\{\gamma_n n_i^T n_j, \gamma_L \left|\mu_{L,i} - \mu_{L,j}\right|, \gamma_\alpha \left|\mu_{\alpha,i} - \mu_{\alpha,j}\right|, \gamma_\beta \left|\mu_{\beta,i} - \mu_{\beta,j}\right|\}, \tag{8}$$

where $n_\cdot$ are the surface normals, $\mu_{L,\cdot}$, $\mu_{\alpha,\cdot}$ and $\mu_{\beta,\cdot}$ are the color means of the surfels in the $L\alpha\beta$ color space [17], and $\gamma_\cdot$ are weighting factors.

4.3 Motion Estimation between Multi-Resolution Surfel Maps

The MRSMaps are registered in an iterative dual refinement procedure similar to the iterative closest points algorithm [17]. The algorithm alternates between efficient pose and data association refinement steps. Assuming the current pose estimate θ fixed, new surfel associations $\mathcal{A}$ are estimated in an efficient multi-resolution procedure. Given these associations, a new pose is estimated by maximizing the observation likelihood of the associated surfels $\widehat{\theta} = \arg\max_\theta \sum_{(i,j)\in\mathcal{A}} \ln p(z_i^{seg} | \theta, z_j^{ref})$, marginalized on the spatial dimensions. We augment this algorithm to incorporate the weighting in our EM objective function (Eq. 5) through

$$\underset{\theta_{l_i}}{\arg\max} \sum_{(i,j)\in\mathcal{A}} p(l_i \mid \mathcal{L}_{ML} \setminus \{l_i\}, I_{seg}, \overline{\Theta}, I_{ref}) \ln p(z_i^{seg} \mid \theta_{l_i}, z_j^{ref}). \tag{9}$$

5 Experiments

We evaluate segmentation and motion estimation accuracy of our approach on three RGB-D video sequences with ground-truth information. We recorded two large objects (chairs), two medium sized objects (a watering can and a box), and two small objects (a cereal box and a tea can) (see Fig. 2). The objects as well as the camera have been moved during the recordings. The sequences contain 1,100 frames at 640×480 VGA resolution and at full 30 Hz frame-rate. Ground truth of the 3D rigid-body motion has been obtained with a motion capture system. We attached infrared reflective markers to the backside of the objects. While recording the data, we took care that the reflective markers were not visible for the RGB-D camera. For frames at every 5 seconds, we manually annotated the individual object parts that move throughout the sequences. Invalid depth readings or non-rigid objects like arms

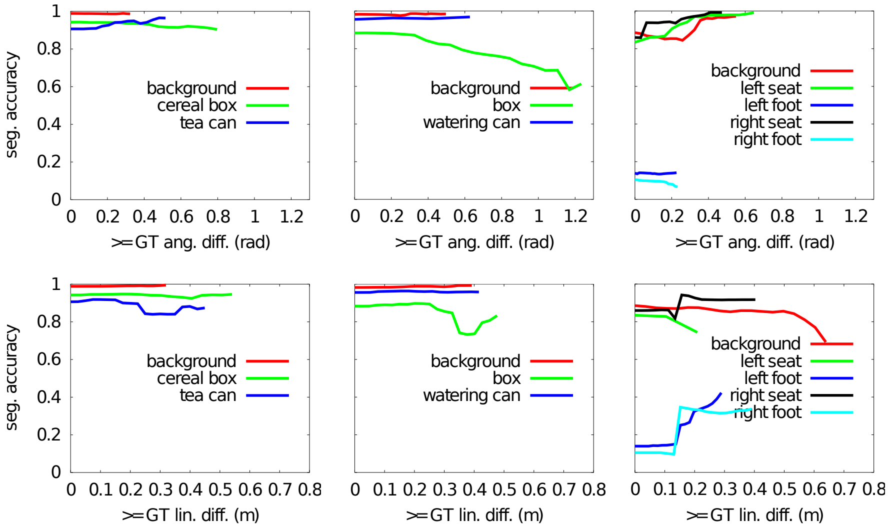

Figure 3: Average segmentation accuracy vs. angular (top) and linear (bottom) ground-truth object motion (left: small, middle: medium, right: large objects). The mean is determined for segment motion greater or equal the value on the x-axis.

sequence	small	medium	large
run-time in msec	200.2±42.3	213.1±54.7	138.7±37.5
error in M	0.05±0.29 (-0.09±0.35)	0.11±0.43 (0.04±0.45)	-0.58±1.01 (-0.43±0.92)
avg. seg. acc.	0.95 (0.91)	0.94 (0.91)	0.63 (0.65)
median lin. acc. in m	0.012 (0.013)	0.018 (0.020)	0.034 (0.030)
median ang. acc. in rad	0.047 (0.045)	0.029 (0.030)	0.049 (0.048)

Table 1: Mean ± standard deviation of run-time and error in the number of segments, segmentation and motion estimate accuracy over all frames (in brackets: real-time mode).

and legs of persons are annotated with dont-care labels. Additionally, we set pixels to dont care in the ground truth that project outside the reference image due to camera motion. Not all annotated segments move between a ground-truth frame and an arbitrary frame in the sequence. We thus automatically determine groups of objects that moved jointly between the frames (0.12 rad angular and 0.05 m linear thresholds) and merge their segments.

The sequences are processed sequentially, starting from each ground-truth labelled image as the image to be segmented. If not stated otherwise, the sequences are processed frame-by-frame. In real-time mode, we drop frames if they would arrive during processing. The experiments have been run on an Intel Core i7-4770K CPU@3.50 GHz. We quantify the segmentation accuracy of the ground-truth segments with the measure proposed in [8], $\sigma = \frac{\textit{true positives}}{\textit{true pos.+false pos.+false negatives}}$, for which we back-project the resulting motion segmentation into images. We also measure angular and linear errors between ground-truth and estimated motion. We determined the parameters of our approach empirically, while for the MRSMaps we use a maximum resolution of 0.0125 m at a factor of 0.007 on the squared measurement distance. The run-time of our approach is given in Table 1. It segments images fast at a frame rate of about 2 to 10 Hz which depends on the number of segments and distance to surfaces.

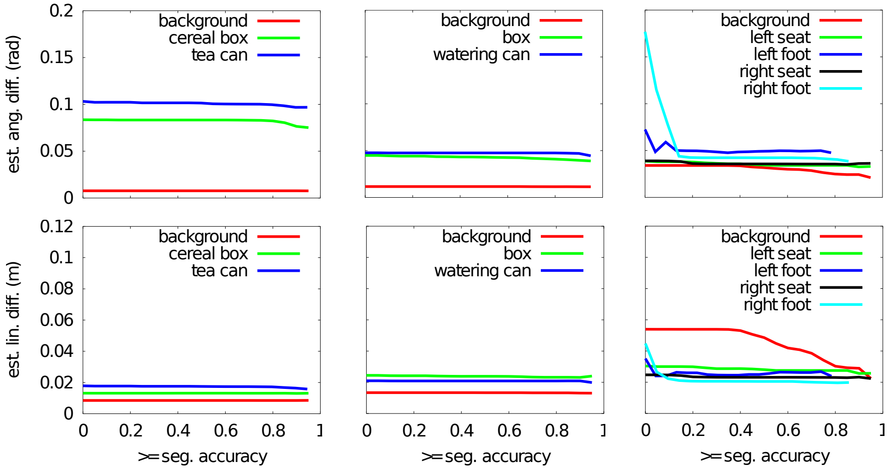

Figure 4: Median angular (top) and linear (bottom) error of camera motion estimate vs. object segmentation accuracy (left: small, middle: medium, right: large objects). The median is determined for seg. accuracies greater or equal the value on the x-axis.

5.1 Segmentation Accuracy

In Fig. 3, we show average segmentation accuracy in dependency on the actual linear and angular motion of the objects. To visualize the effect of different degrees of object motion onto the segment accuracy, we vary a threshold for the linear and angular motion and determine the avg. segmentation accuracy for those results for which the motion is above the threshold. Most objects and the background in the sequences can be very well segmented. The box-shaped objects show a continuous drop in segmentation accuracy with rotation since sides of the boxes become occluded. For the chairs (bottom row) it can be seen that object motion improves segmentation accuracy. This is explained by the distant hence noisy, structureless, and untextured background which allows only coarse misalignments to be detected. The chair feet cannot be reliably segmented because of their thin and rotationally repetitive structure. Besides this, our approach recovers the number of segments well and achieves good overall accuracies in segmentation and motion estimates (see Table 1). Notably, if frames are dropped to operate in real-time, we obtain similar performance.

5.2 Motion Estimate Accuracy

The results in Fig. 4 demonstrate that our approach yields accurate motion estimates of the camera relative to the objects. Here, we determine the median pose accuracy for all results above the varied segmentation accuracy threshold. While for many objects motion accuracy increases with segmentation accuracy, the motion is well estimated also for low segmentation accuracies. This indicates that segmentation accuracy is mostly low for small displacements. Only for the small objects, or for the background at low segmentation accuracy, the pose estimates are slightly off. The small objects are difficult to track in angle with our depth-based registration method due to measurement noise and hands of persons that touch the object to move it. If the background is undersegmented, the registration arbitrates between the background and a foreground object until motion is sufficiently large to split the segment.

6 Conclusions

In this paper we presented an efficient motion segmentation approach for RGB-D image sequences. We employ expectation-maximization to infer image labelling and motion estimates, and propose efficient approximations based on graph-cuts. Our approach recovers the number of motion segments and is suited for online operation in real-time. An efficient probabilistic image representation that supports rapid registration of RGB-D images facilitates fast performance.

In our experiments, we demonstrated high accuracy of our method with regards to segmentation and motion estimates. The performance of our motion segmentation approach strongly depends on the underlying image representation. In order to improve the segmentation of fine-detailed structure and to increase the accuracy of motion estimation for small objects, we will integrate point features into our dense segmentation approach. It could also be useful to adapt an oversegmentation of the image such as superpixels or supervoxels to our approach. While we handle degrading image overlap, segmentation evidence from multiple view points would be beneficial to increase overlap.

References

[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.

[2] A. Ayvaci and S. Soatto. Motion segmentation with occlusions on the superpixel graph. In *Proc. of the IEEE ICCV Workshops*, 2009.

[3] 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.

[4] 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.

[5] D. Cremers and S. Soatto. Motion competition: A variational approach to piecewise parametric motion segmentation. *Int. J. of Computer Vision*, 62:249–265, 2005.

[6] 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.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Stat. Society, Series B*, 39(1):1–38, 1977.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *Int. J. of Computer Vision*, 88(2), 2010.

[9] 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.

[10] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Learning layered motion segmentations of video. In *Proc. of the Int. Conf. on Computer Vision (ICCV)*, 2005.

[11] F. Ramos, D. Fox, and H. Durrant-Whyte. CRF-Matching: Conditional random fields for feature-based scan matching. In *Proc. of Robotics: Science and Systems (RSS)*, 2007.

[12] D. Ross, D. Tarlow, and R. Zemel. Learning articulated structure and motion. *Int. J. of Computer Vision*, 88:214–237, 2010.

[13] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. Segmenting, modeling, and matching video clips containing multiple moving objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 477–491, 2007.

[14] 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.

[15] K. Schindler and D. Suter. Two-view multibody structure-and-motion with outliers through model selection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28:983–995, 2006. ISSN 0162-8828.

[16] H. Sekkati and A. Mitiche. Concurrent 3-D motion segmentation and 3-D interpretation of temporal sequences of monocular images. *IEEE Trans. on Image Processing*, 15(3): 641–653, 2006.

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

[18] G. D. Tipaldi and F. Ramos. Motion clustering and estimation with conditional random fields. In *Proc. of the IEEE/RSJ Int. Conf. on IROS*, 2009.

[19] M. Unger, M. Werlberger, T. Pock, and H. Bischof. Joint motion estimation and segmentation of complex scenes with label costs and occlusion modeling. In *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1878 –1885, 2012.

[20] J. van de Ven, F. Ramos, and G.D. Tipaldi. An integrated probabilistic model for scan-matching, moving object detection and motion estimation. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2010.

[21] M. Van den Bergh and L. van Gool. Real-time stereo and flow-based video segmentation with superpixels. In *IEEE WS on App. of Computer Vision (WACV)*, 2012.

[22] C. Wang, C. Thorpe, M. Hebert, S. Thrun, and H. Durrant-whyte. Simultaneous localization, mapping and moving object tracking. *International Journal of Robotics Research*, 2004.

[23] S. Wang, H. Yu, and R. Hu. 3d video based segmentation and motion estimation with active surface evolution. *Journal of Signal Processing Systems*, pages 1–14, 2012.

[24] L. Zelnik-Manor, M. Machline, and M. Irani. Multi-body factorization with uncertainty: Revisiting motion consistency. *Int. J. of Computer Vision*, 68(1), 2006.

[25] 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.