In Proceedings of 23rd Joint Conference on Artificial Intelligence (IJCAI), Beijing, China, 2013.

Hierarchical Object Discovery and Dense Modelling From Motion Cues in RGB-D Video

Jörg Stückler and Sven Behnke
Autonomous Intelligent Systems, University of Bonn
53113 Bonn, Germany
stueckler@ais.uni-bonn.de, behnke@cs.uni-bonn.de

Abstract

In this paper, we propose a novel method for object discovery and dense modelling in RGB-D image sequences using motion cues. We develop our method as a building block for active object perception, such that robots can learn about the environment through perceiving the effects of actions. Our approach simultaneously segments rigid-body motion within key views, and discovers objects and hierarchical relations between object parts. The poses of the key views are optimized in a graph of spatial relations to recover the rigid-body motion trajectories of the camera with respect to the objects. In experiments, we demonstrate that our approach finds moving objects, aligns partial views on the objects, and retrieves hierarchical relations between the objects.

1 Introduction

Motion is an important cue in object perception. While many static cues such as color or shape can be used to generate object hypotheses, common motion is a further fundamental grouping cue that is especially useful for active perception by robots. For novel, previously unseen objects, motion provides a clear segmentation hint for the constituent parts of an object. Some approaches to the unsupervised learning of object models have been proposed in the robotics community that exploit motion (e.g., [Fitzpatrick, 2003; Kenney *et al.*, 2009; Sturm *et al.*, 2011; Katz *et al.*, 2012; Herbst *et al.*, 2011]). In this paper, we propose a novel approach that simultaneously segments motion in image sequences, and builds dense 3D models of the moving segments. Our approach also reasons on the hierarchy of object parts on-the-fly.

We segment motion between images densely within an expectation-maximization framework using an efficient registration method for RGB-D images. This motion segmentation approach is integrated in a simultaneous localization and mapping framework to incrementally build maps of moving objects and the background. In this process of simultaneous motion segmentation, localization, and mapping (SMOSLAM), we incrementally extract key views and segment motion in these views. From the motion segments, we generate objects, optimize poses of partial views onto the objects, and deduce a hierarchy of object parts from the relations of the motion segments throughout the sequence.

We demonstrate in experiments that our approach is capable of finding moving objects, aligns partial views on the objects, and infers hierarchical relations between the objects.

2 Related Work

Bottom-up cues for single-image segmentation such as texture [Cremers *et al.*, 2007; Delong *et al.*, 2012] or 3D-shape [Holz and Behnke, 2012; Silberman *et al.*, 2012] often do not suffice to find segment borders that coincide with the boundaries of objects. Thus, they are frequently combined with top-down cues to integrate spatial and semantic context (e.g., [Carreira and Sminchisescu, 2012]). Motion is a further important bottom-up cue that can be utilized in image sequences. In contrast to texture and shape, common motion provides unambiguous segmentation hints for the constituent parts of a rigid object.

Many approaches to motion segmentation employ point features in multi-body structure-from-motion [Zelnik-Manor *et al.*, 2006; Gruber and Weiss, 2004; Schindler and Suter, 2006; Rothganger *et al.*, 2007; Agrawal *et al.*, 2005; Ross *et al.*, 2010; Katz *et al.*, 2012], but these methods do not provide a dense segmentation of objects like ours. Several dense methods [Cremers and Soatto, 2005; Unger *et al.*, 2012; Kumar *et al.*, 2005; Ayvaci and Soatto, 2009; Zhang *et al.*, 2011; Wang *et al.*, 2012; Roussos *et al.*, 2012] have been proposed that demonstrate impressive results. These methods are, however, either computationally demanding and yet far from real-time performance, or they do not extract segments with common 3D rigid-body motion. Our motion segmentation approach makes use of dense depth available in RGB-D images to retrieve 3D rigid-body motion segments efficiently. We also take motion segmentation a step further by integrating it with simultaneous localization and mapping (SLAM) and deducing the hierarchical relations between the moving parts.

The mapping of static as well as dynamic parts of environments is an actively researched topic in the robotics community. Early work focused on 2D mapping using laser scanners. Anguelov et al. [2002] learned templates and object classes of non-stationary parts of an environment in a two-level hierarchical model. Haehnel et al. [2003] proposed an EM algorithm that filters dynamic parts of the environment in order to make the 2D occupancy mapping of the static environment parts robust. They then extract 3D models of the dynamic parts by stitching the laser measurements. In simultaneous localization mapping and moving object tracking (SLAMMOT, [Wang *et al.*, 2004]), dynamic objects are segmented in 2D laser scans through distance comparisons, and subsequently tracked while concurrently mapping the environment statics in a SLAM framework. We integrate 3D motion segmentation in a SLAM framework that also reasons about hierarchical relations between object parts. Van de Ven et al. [2010] recently proposed a graphical model that integrates CRF-matching [Ramos *et al.*, 2007] and CRF-clustering [Tipaldi and Ramos, 2009] 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 efficient expectation-maximization and perform fast approximate inference of the motion segmentation using graph cuts.

3 Multi-View Motion Segmentation

We assume that an image I is partitioned into a set of discrete sites $\mathcal{S} = \{s_i\}_{i=1}^N$ such as pixels or map elements in a 3D representation (see Fig. 1). Let $\mathcal{L} = \{l_i\}_{i=1}^N$ be the labelling of the image sites. The labels denote the membership in distinct motion segments $\mathcal{M} = \{m_k\}_{k=1}^M$. All sites within a segment move with a common six degree of freedom (6-DoF) rigid-body motion θ_k between the image I_{seg} to be segmented and a reference image I_{ref}.

Our goal is to explain the segmented image by the rigid-body motion of segments into the reference image, i. e., we seek rigid-body motions $\Theta = (\theta_1, \ldots, \theta_M)$ that maximize the observation likelihood of the segmented image in the reference image:

$$\arg\max_{\Theta} p(I_{seg} \mid \Theta, I_{ref}). \tag{1}$$

The labelling of the image sites is a latent variable that we estimate with the rigid-body motions of the segments using an expectation-maximization (EM) algorithm,

$$\arg\max_{\Theta} \sum_{\mathcal{L}} p(\mathcal{L} \mid I_{seg}, \overline{\Theta}, I_{ref}) \; \ln p(I_{seg} \mid \Theta, I_{ref}, \mathcal{L}), \tag{2}$$

where $\overline{\Theta}$ is the recent motion estimate of the segments from the previous iteration of the EM algorithm.

We model the likelihood of the labelling in a conditional random field

$$p(\mathcal{L} \mid I_{seg}, \overline{\Theta}, I_{ref}) \propto \exp\left(\sum_i \ln p(z_i \mid l_i, \overline{\Theta}, I_{ref}) \sum_{(i,j)\in\mathcal{N}} \ln p(l_i, l_j \mid I_{seg}) \right) \tag{3}$$

that incorporates the likelihood of the labelling given the observations z_i at each site and the label's corresponding

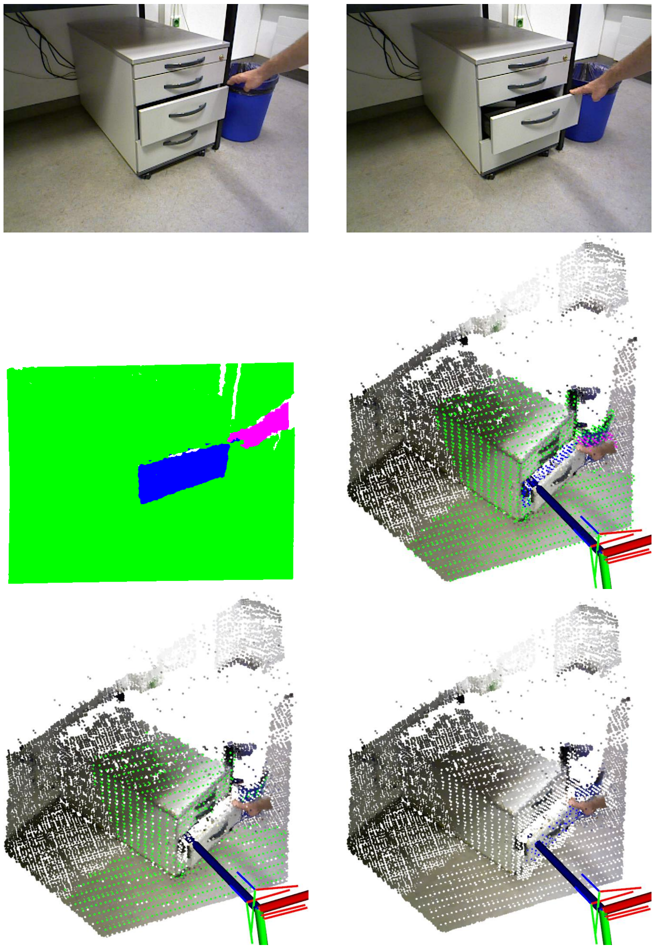

Figure 1: Dense motion segmentation. Top: Image to be segmented (left) and reference (right). Middle: Max. likelihood labelling in image (left) and in 3D at 5 cm res. (right). Bottom: Label likelihood for two segments (white: low; green/blue: high, best viewed in color). The reference image is overlayed and transformed with the motion estimates.

motion estimate. Pairwise interaction terms between direct neighbors in the image representation enforce spatial coherence within motion segments using a contrast-sensitive Potts model $p(l_i, l_j \mid I_{seg}) = \gamma(s_i, s_j)\, \delta(l_i, l_j)$, where

$$\delta(l_i, l_j) := \begin{cases} l_i = l_j, & 0 \\ l_i \neq l_j, & 1 \end{cases} \tag{4}$$

and $\gamma(s_i, s_j)$ measures the dissimilarity of the image sites.

It is, however, intractable to compute the joint image labelling posterior exactly. Instead, we approximate it through the maximum likelihood labelling $\mathcal{L}_{ML}$, given the motion estimates $\overline{\Theta}$. We find this labelling using efficient graph-cut optimization [Boykov *et al.*, 2001]

$$\mathcal{L}_{ML} = \arg\max_{\mathcal{L}} p(\mathcal{L} \mid I_{seg}, \overline{\Theta}, I_{ref}). \tag{5}$$

Given this labelling, we determine the pseudo-likelihood of a site label l_i:

$$p(l_i = k \mid I_{seg}, \overline{\Theta}, I_{ref}, \mathcal{L}_{ML}) = \\ \eta \, p(z_{seg,i} \mid \overline{\Theta}, I_{ref}, l_i = k) \prod_{j \in \mathcal{N}(i)} p(l_i = k, l_{ML,j}), \quad (6)$$

where η normalizes the probability over label values of l_i.

3.1 Resolving Ambiguous Data Associations

The image site labellings decide for an association of sites between both images. In order to prevent the graph-cut optimization from establishing labellings that would associate multiple times to a site in the reference image, we introduce additional pair-wise couplings. 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 log likelihood

$$\ln p_{\mathcal{A}}(l_i, l_j) := \begin{cases} -\alpha & \text{, if } l_i = k \wedge l_j = k', \\ 0 & \text{, otherwise,} \end{cases} \quad (7)$$

where α sets the strength of the couplings.

3.2 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 appears as multiple unconnected image segments in the image, our approach may still use different but redundant motion segments for the image segments. Furthermore, it may not be desirable to set the number of motion segments in advance. We thus adapt the number of motion segments to the data similar to the approach in [van de Ven *et al.*, 2010].

We initialize the EM algorithm with a guess of the number of motion segments ($M = 1$ in our experiments). While this guess influences the number of required iterations, we found that it has only little effect on finding the correct number of segments. Initially, we label all sites to belong to the first segment. 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.

To let our approach possibly increase the number of segments, we append one additional, yet unsupported segment before the M-step. After the E-step, we reduce the number of motion segments in two ways. We discard unsupported segments (eventually also the additional segment) and merge redundant segments. We measure the redundancy of a segment k using the average effective number of segments of the image sites

$$\overline{N}_{\mathit{eff}}(k) := \frac{\sum_{i:l_i=k} N_{\mathit{eff}}(i)}{|\{i : l_i = k\}|} \quad (8)$$

within the segment [van de Ven *et al.*, 2010]. It takes the average over the effective number of segments that the image sites within the segment are matching, i.e.,

$$N_{\mathit{eff}}(i) := \sum_k \frac{1}{p(z_i \mid \theta_k, I_{ref})^2}. \quad (9)$$

Ideally, the image sites labelled as belonging to a motion segment are mainly explained by the segment and, hence, $\overline{N}_{\mathit{eff}}(k) \approx 1$. We classify a motion segment k to be redundant, if $\overline{N}_{\mathit{eff}}(k)$ is larger than a threshold (set to 1.8 in our experiments), and subsequently merge it with the most similar segment.

3.3 Segmentation Towards Multiple Reference Images

Our formulation supports the segmentation of an image in reference to multiple other images $\mathcal{I}_{ref} = \{I_{ref,r}\}_{r=1}^{R}$ which allows to include motion hints from multiple perspectives into the segmentation of an image. Each motion segment in I_{seg} then moves with a rigid-body motion θ_k^r to the reference image $I_{ref,r}$. We denote the set of motions towards one reference image r by θ^r. We may either optimize the segmentation for all reference images concurrently, i. e.,

$$\operatorname*{arg\,max}_{\theta^1,\ldots,\theta^R} p(I_{seg} \mid \theta^1, \ldots, \theta^R, \mathcal{I}_{ref}), \quad (10)$$

or we may only estimate the motion estimate towards a specific reference image r while keeping the motion estimates of the other images fixed, i. e.,

$$\operatorname*{arg\,max}_{\theta^r} p(I_{seg} \mid \theta^1, \ldots, \theta^r, \ldots, \theta^R, \mathcal{I}_{ref}). \quad (11)$$

In both cases, the label likelihood given the observations at a site is the product of the observation likelihoods in the individual reference images,

$$p(l_i \mid z_{seg,1,i}, \ldots, z_{seg,R,i}, \overline{\Theta}, I_{ref}) \\ \propto \prod_r p(z_{seg,r,i} \mid l_i, \overline{\Theta}, I_{ref})^{\alpha}. \quad (12)$$

To keep the data term in balance with the pairwise interaction terms, we normalize the combined observation likelihood to the effect of a single image using $\alpha := 1/R$.

3.4 Image Representation

In principle, any image representation is suitable for our motion segmentation method that defines data likelihood $p(z_i \mid \theta_{l_i}, I_{ref})$, image site neighborhood $\mathcal{N}(i)$, and dissimilarity $\gamma(s_i, s_j)$ for the pair-wise interaction terms. To solve for the motion estimates of the segments in Eq. (2), an image registration technique is required that allows to incorporate individual weights for the image sites.

Instead of labelling the RGB-D image pixel-wise, we choose to represent the image content in a multi-resolution 3D representation to gain efficiency. These multi-resolution surfel maps [Stückler and Behnke, 2012] respect the noise characteristics of the sensor, provide a probabilistic representation of the data, and support efficient registration of motion segments. They store the joint color and shape statistics of points within 3D voxels at multiple resolutions in an octree. The maximum resolution at a point is limited proportional to its squared distance 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. RGB-D images can be efficiently mapped into this representation and registered. We use an open-source implementation of multi-resolution surfel maps.[1]

Observation Likelihood

Each voxel s_i in a multi-resolution surfel map contains a surfel observation z_i specified by mean and covariance of the points falling into the voxel. The observation likelihood of a site $s_{seg,i}$ given label $l_i = k$ and reference image I_{ref} is the matching likelihood of the surfel $z_{seg,i}$ under the rigid-body motion θ_k for the labelling,

$$\begin{aligned} p(z_{seg,i}|z_{ref,j}, \theta_k, l_i = k) &= \mathcal{N}\left(d_{i,j}(\theta_k); 0, \Sigma_{i,j}(\theta_k)\right), \\ d_{i,j}(\theta_k) &:= \mu_{ref,j} - T(\theta_k)\mu_{seg,i}, \\ \Sigma_{i,j}(\theta_k) &:= \Sigma_{ref,j} + R(\theta_k)\Sigma_{seg,i}R(\theta_k)^T, \end{aligned} \tag{13}$$

where $T(\theta_k)$ is the transformation matrix for the pose estimate θ_k and $R(\theta_k)$ is its rotation matrix. We only use the spatial components of the surfels.

Spatial Coupling and Local Dissimilarity Measure

Each voxel in the map is connected to its eight direct neighbors in the 3D grid. In addition, we connect each voxel to its parent and childs. For the contrast-sensitive Potts model in Eq. (4), we measure local dissimilarity through $\gamma(s_i, s_j) := 1 - \min(1, \max(\xi(s_i), \xi(s_j)))$, where $\xi(s)$ is a linear combination of the principal shape curvature [Pauly *et al.*, 2003] and the trace of the color covariance within voxel s.

Motion Estimation

In order to maximize our EM objective function for the individual motion segments k, we augment our registration approach [Stückler and Behnke, 2012] with the weighting through the label likelihoods, i.e.,

$$\begin{aligned} \arg\max_{\theta_k} \sum_{(i,j)\in\mathcal{A}_k} & p(l_i = k \mid I_{seg}, \overline{\Theta}, I_{ref}, \mathcal{L}_{ML}) \\ & \ln p(z_{seg,i} \mid \theta_k, z_{ref,j}, l_i = k). \end{aligned} \tag{14}$$

Exemplary label likelihoods in two motion segments are visualized in Fig. 1.

Border and Occlusion Handling

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

4 Hierarchical Object Modelling through Simultaneous Motion Segmentation, Localization, and Mapping

We extend our view-based SLAM approach [Stückler and Behnke, 2012] towards simultaneous motion segmentation, localization, and mapping (SMOSLAM) of objects $\mathcal{O} = \{o_i\}_{i=1}^{O}$. For each object, we maintain a graph of view poses for those key views that contain the object. The view poses are connected through edges that represent spatial relations which we estimate through motion segmentation between the key views. The poses of the pose graphs are optimized using the g2o graph optimization framework [Kuemmerle *et al.*, 2011].

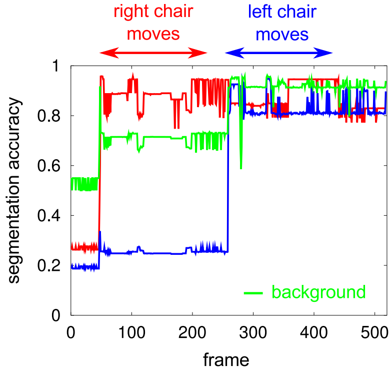

Figure 5: Segmentation accuracy for incremental segmentation of the first frame to all other images in the chairs sequence. The segmentation accuracies of the chairs rapidly reach high values when the objects start to move. The achievable segmentation accuracy is limited due to sensor noise, occlusions, and limited field-of-view. This is handled through motion segmentation towards multiple key views in later processing stages.

Not all objects may be visible in a single key view, or objects may split into sub-parts between different key views. This property enables us to learn part-of relations between the objects in a hierarchical object map. We infer these relations from the overlap of motion segments between key views.

We perform SMOSLAM incrementally, working sequentially on the images in a RGB-D video sequence. The current image is segmented towards the latest key view in the map in order to track the relative motion of the camera towards the objects in the reference key view. If one of the objects moved a specific distance or angle, we generate a new key view and resume tracking towards this view. We initialize the motion segmentation of the new key view with the segmentation of the previous one, and add the previous key view as a reference image to the segmentation. By this, the segmentation of the previous key view acts as a regularizing prior while the segmentation is further optimized with respect to the current image. Most importantly, we associate the motion segments of the previous key view with the object SLAM graph to discover object relations.

4.1 Discovering Objects and Hierarchical Relations

We analyze relations of motion segments between key views to associate motion segments with objects and to infer hierarchical relations between objects. We quantify the image overlap of the motion segments to set equivalence and part-of relations between segments, i.e., we define a motion segment m to be part of another segment m', if m overlaps m' at

[1] http://code.google.com/p/mrsmap/

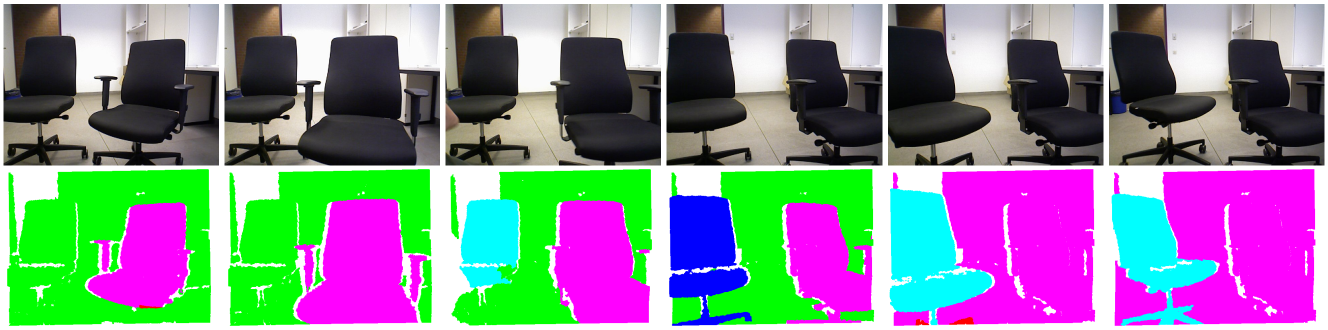

Figure 2: Key views (top) and motion segmentations (bottom) estimated with our approach in the chairs sequence. Each key view is concurrently segmented in reference to its predecessor and successor.

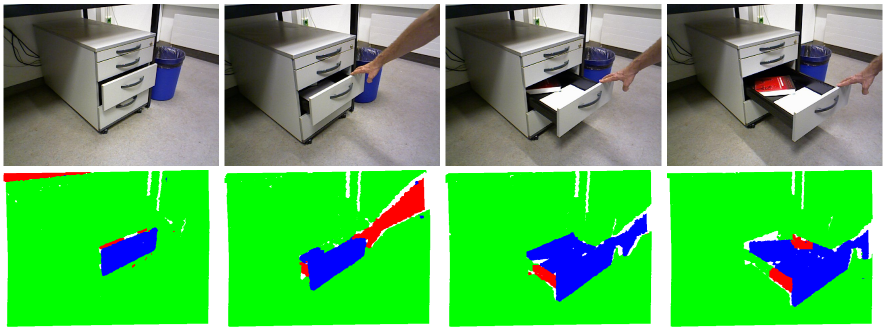

Figure 3: Key views (top) and motion segmentations (bottom) estimated with our approach in the static container sequence. Each key view is concurrently segmented in reference to its predecessor and successor.

least by some threshold (e.g., set to 80% in our experiments) within the image:

$$\text{overlap}(m, m') \Rightarrow \text{part-of}(m, m'). \tag{15}$$

The segments are defined equivalent, if both segments overlap each other significantly:

$$\text{overlap}(m, m') \wedge \text{overlap}(m', m) \Rightarrow \text{equivalent}(m, m'). \tag{16}$$

From these relations, we infer new objects, or equivalence and part-of relations of motion segments towards existing objects. If a motion segment m in a key view is equivalent part of an object o, i. e., equivalent(m, o), all motion segments m' that are equivalent to m are also equivalent part of o:

$$\begin{aligned} &\text{equivalent}(m, o) \wedge \text{equivalent}(m', m) \\ &\qquad \Rightarrow \text{part-of}(m', o) \wedge \text{equivalent}(m', o). \end{aligned} \tag{17}$$

Analogeously, for all motion segments m that are part of an object o, motion segments m' from other key views in part-of relations to m are also contained in the same object,

$$\begin{aligned} &\text{part-of}(m, o) \wedge \text{part-of}(m', m) \\ &\qquad \Rightarrow \text{part-of}(m', o). \end{aligned} \tag{18}$$

A motion segment that is not equivalent part of an object creates a new part. It is then equivalent part of the new object. For each pair of key views for which motion segments are part of the same object, we create an edge between the key views in the SLAM graph of the object.

Finally, we deduce hierarchical relations between objects from the relations between key views and objects. An object o is contained within another object o', if a motion segment m exists in a key view for which this segment is equivalent to the contained object o, but only part of the containing object o':

$$\begin{aligned} &\exists m : \text{equivalent}(m, o) \wedge \neg\, \text{equivalent}(m, o') \\ &\qquad \wedge \text{part-of}(m, o') \Rightarrow \text{part-of}(o', o). \end{aligned} \tag{19}$$

Isolated smallest objects have no further parts.

5 Results

We demonstrate our approach in three RGB-D image sequences that contain 30 images per second at VGA (640×480) resolution. Note that our current implementation does not process the sequences in real-time, but achieves about 80 to 800 msec per frame on a notebook PC with an Intel Core i7 3610QM 2.3 GHz (max. 3.3 GHz) QuadCore CPU. The first sequence shows two moving chairs from a static camera position (see Fig. 2). In the second sequence, a drawer of a container at a fixed position is opened while the camera is moving (see Fig. 3). The third sequence involves the same container which is now first moved before it is opened (Fig. 4). Since ground truth motion estimates are not available, we visualize resulting segmentations of key

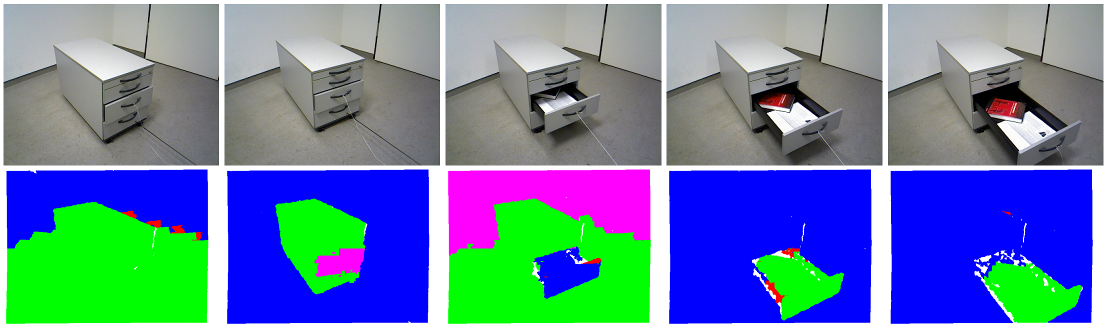

Figure 4: Key views (top) and motion segmentations (bottom) estimated with our approach in the moving container sequence. Each key view is concurrently segmented in reference to its predecessor and successor.

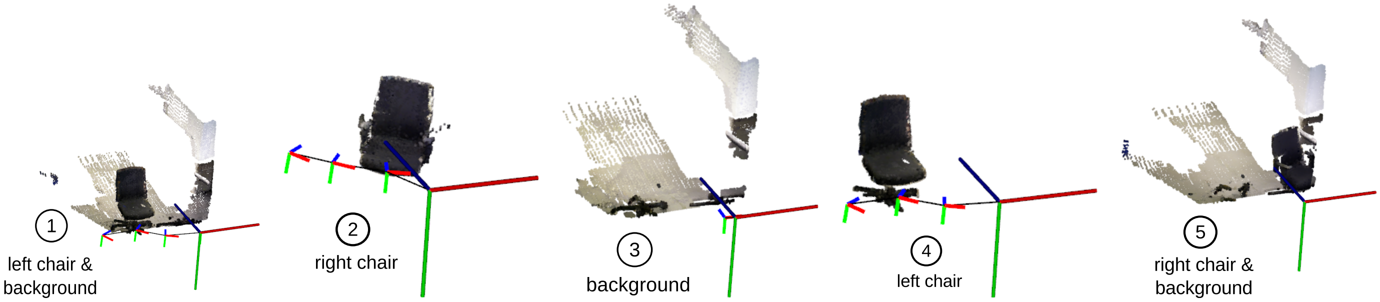

Figure 6: Objects found with our approach in the chairs sequence. The numbering specifies the temporal order of retrieval. The key view poses within the object SLAM graphs are visualized by coordinate frames. Black lines depict spatial relations between the key views. We overlay the motion segments that have been associated with the objects in the estimated key view poses relative to the object.

views, object relation graphs, and object SLAM graphs for the sequences.

5.1 Chairs Sequence

In the first sequence, two chairs are moved in the horizontal plane and rotated around the vertical axis while the camera is static. Both chairs move separately, starting with the rightmost chair (see Fig. 2).

We first demonstrate the performance of our dense motion segmentation approach. We compare the resulting labelling with a manual ground truth image through $\sigma = \mathit{true\ positives}/(\mathit{true\ pos.} + \mathit{false\ pos.} + \mathit{false\ neg.})$ as a measure of segmentation accuracy [Everingham *et al.*, 2010]. The initial key view is sequentially segmented in reference to the images in the sequence, performing one iteration of our EM algorithm per frame. We project the 3D segmentation into the image to compare the image segmentation with the ground truth labelling, and associate each found segment to the ground truth segment with best overlap. The results in Fig. 5 show that our approach is well capable of segmenting the moving objects with high accuracy. The segmentation accuracy of the chairs rapidly reaches high values when the objects start to move. Note that due to sensor noise, occlusions, and parts of the objects leaving the field-of-view, the achievable segmentation accuracy is limited. This is handled through motion segmentation of key views towards both preceding and subsequent key views in our SMOSLAM approach.

Fig. 2 shows the key views extracted by our approach and the estimated motion segmentation of the views. The segments correspond well to the actual objects in the images. The objects and the poses of the key views in the object SLAM graphs are shown in Fig. 7. The trajectory of the camera with respect to the objects has been well recovered, such that the corresponding motion segments of the objects accurately align. Both chairs and the background segment are smallest parts in the object hierarchy. Since between the first two key views and their neighbors, the left chair does not move, left chair and background are found within a single segment. The object relation graph reflects the containment of both parts (left chair and background segment) within this combined object, which is inferred through the split of the segment in the third key view. Similar arguments apply for the left chair and the background in the last two key views.

5.2 Static Container Sequence

In the second sequence, a drawer is moved open by a person, while the container is kept fixed and the camera is slowly moved. Our algorithm succeeds in segmenting the drawer with good accuracy (see Fig. 3). From the object SLAM graphs it can be seen that the relative pose of the camera towards the objects is recovered and the motion segments accurately overlap for the estimated key view poses. Remarkably, the drawer segment makes the inside of the container explicit.

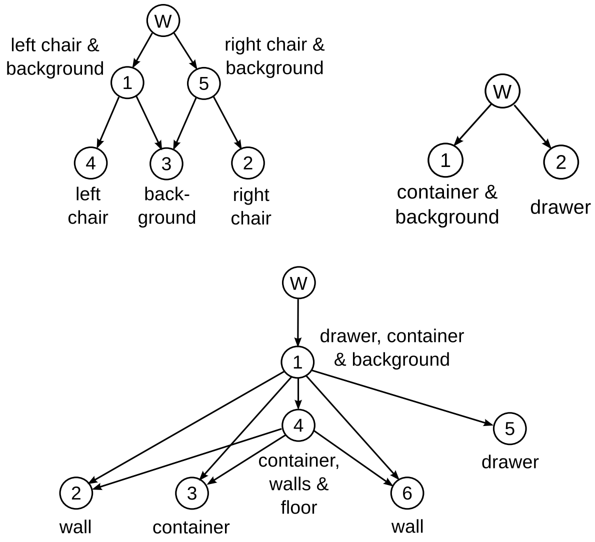

Figure 7: Object hierarchies deduced by our approach from the chairs (top left), the static (top right), and the moving container sequence (bottom). The object numbering specifies the temporal order of retrieval. Each node corresponds to an object. Arrows depict part-of relations (pointing from containing object to the part).

5.3 Moving Container Sequence

In the third sequence, our algorithm partially succeeds in segmenting the container from the background (see Fig. 4). Since the container moves parallel to the ground plane, our approach cannot well distinguish if the ground plane itself is moving or not. As a consequence, the drawer is part of the object combined from container, drawer, and background. The wall part is found twice, since it is not detected in the second key view.

6 Conclusions

In this paper, we introduced a novel method for learning object maps with hierarchical part relations from motion cues. Motion segmentation between RGB-D key views finds the rigid parts in images and estimates their motion. It is based on an efficient expectation-maximization algorithm and employs a compact local multi-resolution 3D representation of RGB-D images to process images efficiently.

We integrate our motion segmentation method with SLAM into a framework for simultaneous motion segmentation, localization, and mapping. Our mapping approach extracts moving objects from key views and aligns the parts by optimizing a graph of spatial relations. From the overlap of motion segments, we deduce a hierarchy of object parts.

In experiments, we demonstrated that our approach is capable of extracting motion segments and aligning multiple views on objects. In each of the sequences, our approach deduces hierarchical object relations.

The robustness and accuracy of our motion estimates and segmentation strongly depend on the underlying registration method. We currently work on including point features into the registration to further improve the range of applications of our approach. This would allow for tracking smaller objects, or reduce aperture problems along planar surfaces. One limitation of our method is that we currently only establish relations between segments in temporal sequence. In future work, we will also establish equivalence and part-of relations between parts that interrupt their motion. The robustness of our approach could further be increased by reasoning on the uncertainty of segmentation decisions and hierarchical relations. Our overlap measure between segments could be enhanced by tracking correspondences through time between key views. In order to scale our approach to larger scenes, graph pruning and map merging needs to be incorporated. We also plan to extract articulation models from the hierarchical object relations and the relative object trajectories. Finally, we will pursue the application of our approach for interactive perception of objects by robots.

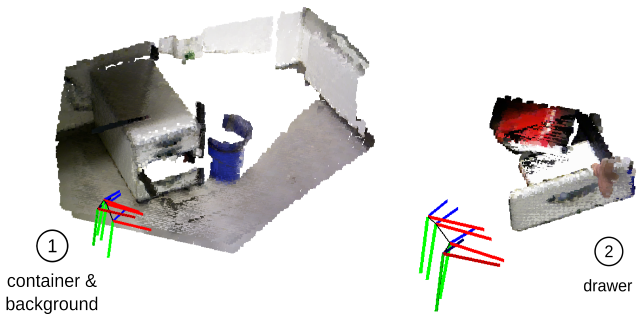

Figure 8: Object parts found with our approach in the static container sequence. The numbering specifies the temporal order of retrieval. The key view poses within the object SLAM graphs are visualized by coordinate frames. Black lines depict spatial relations between the key views. We overlay the motion segments that have been associated with the objects in the estimated key view poses relative to the object.

References

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

[Anguelov *et al.*, 2002] D. Anguelov, R. Biswas, D. Koller, B. Limketkai, S. Sanner, and S. Thrun. Learning hierarchical object maps of non-stationary environments with mobile robots. In *Proc. of UAI*, 2002.

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

[Boykov *et al.*, 2001] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on Pattern Analysis and Mach. Intell.*, 2001.

[Carreira and Sminchisescu, 2012] J. Carreira and C. Sminchisescu. CPMC: Automatic object segmentation using constrained parametric min-cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2012.

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

[Cremers *et al.*, 2007] D. Cremers, M. Rousson, and R. Deriche. A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape. *Int. J. of Computer Vision*, 72:195–215, 2007.

[Delong *et al.*, 2012] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. *International Journal of Computer Vision*, 2012.

[Everingham *et al.*, 2010] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[Fitzpatrick, 2003] P. Fitzpatrick. First contact: an active vision approach to segmentation. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2003.

[Gruber and Weiss, 2004] 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.

[Haehnel *et al.*, 2003] D. Haehnel, R. Triebel, W. Burgard, and S. Thrun. Map building with mobile robots in dynamic environments. In *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2003.

[Herbst *et al.*, 2011] E. Herbst, X. Ren, and D. Fox. RGB-D object discovery via multi-scene analysis. In *Proc. of the IEEE Int. Conf. on Robots and Systems (IROS)*, 2011.

[Holz and Behnke, 2012] D. Holz and S. Behnke. Fast range image segmentation and smoothing using approximate surface reconstruction and region growing. In *Proc. of the Int. Conf. on Intelligent Autonomous Systems (IAS)*, 2012.

[Katz *et al.*, 2012] D. Katz, M. Kazemi, J. A. Bagnell, and A. Stentz. Interactive segmentation, tracking, and kinematic modeling of unknown articulated objects. Technical report, Carnegie Mellon Robotics Institute, 2012.

[Kenney *et al.*, 2009] J. Kenney, T. Buckley, and O. Brock. Interactive segmentation for manipulation in unstructured environments. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2009.

[Kuemmerle *et al.*, 2011] R. Kuemmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. g2o: A general framework for graph optimization. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2011.

[Kumar *et al.*, 2005] 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*, 2005.

[Pauly *et al.*, 2003] M. Pauly, R. Keiser, and M. Gross. Multi-scale feature extraction on point-sampled surfaces. In *Eurographics*, 2003.

[Ramos *et al.*, 2007] 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.

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

[Rothganger *et al.*, 2007] 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 Mach. Intell.*, 2007.

[Roussos *et al.*, 2012] A. Roussos, C. Russell, R. Garg, and L. de Agapito. Dense multibody motion estimation and reconstruction from a handheld camera. In *Proc. of IEEE ISMAR*, 2012.

[Schindler and Suter, 2006] 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.

[Silberman *et al.*, 2012] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2012.

[Stückler and Behnke, 2012] J. Stückler and S. Behnke. Model learning and real-time tracking using multi-resolution surfel maps. In *Proc. of the AAAI Conf. on Artificial Intelligence (AAAI)*, 2012.

[Sturm *et al.*, 2011] J. Sturm, C. Stachniss, and W. Burgard. A probabilistic framework for learning kinematic models of articulated objects. *J. on AI Research (JAIR)*, 2011.

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

[Unger *et al.*, 2012] 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)*, 2012.

[van de Ven *et al.*, 2010] 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. Conference on Robotics and Automation (ICRA)*, 2010.

[Wang *et al.*, 2004] 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.

[Wang *et al.*, 2012] S. Wang, H. Yu, and R. Hu. 3D video based segmentation and motion estimation with active surface evolution. *J. of Signal Processing Systems*, 2012.

[Zelnik-Manor *et al.*, 2006] L. Zelnik-Manor, M. Machline, and M. Irani. Multi-body factorization with uncertainty: Revisiting motion consistency. *Int. J. of Computer Vision*, 68(1):27–41, 2006.

[Zhang *et al.*, 2011] 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.