International Journal of Computer Vision (IJCV), to appear 2015.

Efficient Dense Rigid-Body Motion Segmentation and Estimation in RGB-D Video

Jörg Stückler · Sven Behnke

Received: 12 May 2014 / Accepted: 16 December 2014

Abstract Motion is a fundamental grouping cue in video. Many current approaches to motion segmentation in monocular or stereo image sequences rely on sparse interest points or are 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 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.

Keywords motion segmentation · rigid multi-body registration · multibody structure-from-motion

1 Introduction

Common motion is a fundamental grouping cue in video sequences. Bottom-up appearance-based segmentation approaches such as superpixels are frequently observed to not yield a segmentation into meaningful objects. E.g., such methods may oversegment objects with varying texture. Common motion, in contrast, can be used as an unsupervised, bottom-up cue that provides a segmentation into objects. In dynamic scenes, object boundaries can be inferred from the observed motion of objects without pre-trained a-priori knowledge on the visual appearance of an object.

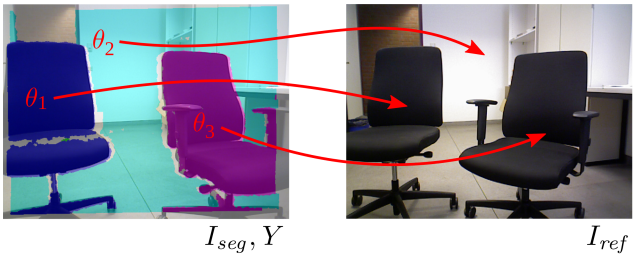

Fig. 1 We estimate a segmentation Y of an RGB-D image I_{seg} into coherent pixel regions that undergo rigid body motions θ_k towards a reference image I_{ref}.

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.

Dense motion segmentation is necessary for reconstructing scenes that contain moving objects. Many state-of-the-art scene reconstruction methods assume static scenes during mapping and may break down when objects move. Dense motion segmentation can also be used to segment and track moving objects from a moving platform such as a driving car. Moreover, if objects can be actively moved, e.g., by a robot, hypotheses on object boundaries can be explored and verified visually by dense motion segmentation.

Many motion segmentation approaches match images only sparsely at interest points and infer the groups of points with common 3D rigid-body motion [Gruber and Weiss, 2004, Schindler and Suter, 2006, Rothganger et al., 2007, Ross et al., 2010, Agrawal et al., 2005]. Most recent methods for dense 3D motion segmentation are still far from real-time performance [Sekkati and Mitiche, 2006, Zhang et al., 2011, Wang et al., 2012, Roussos et al., 2012].

In this article, we propose an efficient approach to dense 3D motion segmentation of rigid objects in RGB-D video. We formulate an expectation-maximization framework (see Fig. 2) that recovers motion segments, estimates their 3D rigid-body motion, and also finds the number of segments in the scene. Our formulation to rigid multi-body registration treats background and foreground objects equally 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

All authors
Computer Science Institute VI, University of Bonn
Friedrich-Ebert-Allee 144, 53113 Bonn, Germany
Tel.: +49-228-7354160, Fax: +49-228-734425
E-mail: stueckler@ais.uni-bonn.de, E-mail: behnke@cs.uni-bonn.de

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, 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. This article extends the work in [Stückler and Behnke, 2013] with a detailed derivation of our method and further comparative evaluation.

2 Related Work

Several approaches to 3D motion segmentation have been proposed that represent images sparsely through interest points. Multi-body factorization methods [Zelnik-Manor et al., 2006] 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 [Gruber and Weiss, 2004, Schindler and Suter, 2006, Rothganger et al., 2007]. Exploiting depth measurements for interest points from a calibrated stereo camera, Agrawal et al. [2005] propose a real-time capable framework for 3D motion segmentation based on RANSAC and SfM. These approaches, however, do not provide dense segmentations.

Some approaches segment 2D image motion densely based on optical flow. Cremers and Soatto [2005] propose motion competition, a variational framework for dense motion segmentation of monocular image sequences. They estimate the 2D parametric motion of multiple motion segments. Brox et al. [2006] extend this approach towards non-parametric motions. Occlusions and multiple data associations are explicitly modeled in the variational framework of Unger et al. [2012], but the method is far from real-time performance. In our approach, we also handle multiple data associations as additional pairwise labeling constraints during graph cut optimization of the motion segmentation. Ochs et al. [2014] estimate large-displacement optical-flow between subsequent RGB images. The approach tracks the optical flow of a subset of the image pixels throughout a sequence and groups pixels with common motion. The sparse set of motion tracks is turned into a dense labeling using a variational segmentation approach. Kumar et al. [2005] segment scenes into 2D motion layers using a conditional random field (CRF) model that incorporates occlusions and lighting conditions. The work by Ayvaci and Soatto [2009] 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. Weber and Malik [1997] proposed dense 3D motion segmentation between monocular images from optical flow assuming an affine camera model. Sekkati and Mitiche [2006] tackle dense 3D multibody structure-from-motion (SfM) from monocular video in a variational framework and demonstrate qualitative results. Recently, a variational framework has been proposed that integrates rigid-body motion segmentation with dense 3D reconstruction [Roussos et al., 2012] 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.

Dense 3D scene flow aims at the concurrent 3D reconstruction and motion estimation in dynamic scenes [Huguet and Devernay, 2007, Wedel and Cremers, 2011]. The dense depth available with RGB-D sensors can simplify scene flow estimation. Hadfield and Bowden [2014] propose a particle-based framework to scene flow. Quiroga et al. [2013] and Herbst et al. [2013] pose scene flow estimation for RGB-D images in a variational framework. In [Hornacek et al., 2014], local point sets within sphere neighborhoods at each pixel are aligned in a randomized multi-step process. The approach recovers smooth 6-DoF motion estimates at each pixel. Herbst et al. [2013] apply their estimated scene flow for 3D rigid-body motion segmentation in a subsequent RANSAC step. Our method performs dense rigid-body motion segmentation and estimation simultaneously. Scene flow can be determined from its outcome. In [Herbst et al., 2014], changes are detected against a background map to isolate moving objects. The approach assumes the underlying registration approach is sufficiently robust to changes in the scene in order to maintain the online mapping of the background. Our formulation to motion segmentation inherently dissects the moving parts in a scene through simultaneous multi-body segmentation and registration. Wang et al. [2012] transfer the approach of Cremers and Soatto [2005] 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.

With a stereo camera, Zhang et al. [2011] propose dense 3D multibody SfM using an energy minimiza-

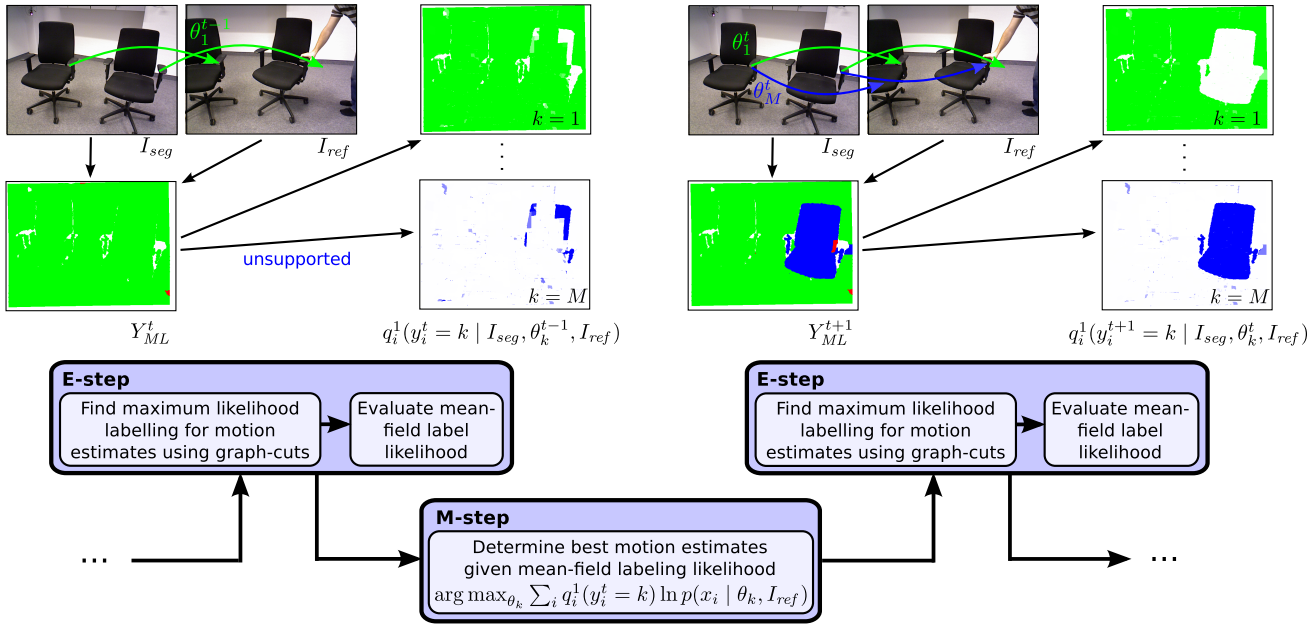

Fig. 2 We segment motion in an RGB-D image I_{seg} towards a reference image I_{ref} in an efficient expectation-maximization framework. In the E-step, we evaluate the likelihood of image site labels y_i under the latest motion estimates θ_k. Efficient graph cuts yield a maximum likelihood labeling y_{ML} 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.

tion framework. The approach relies on plane fitting to make the segmentation robust and is reported to require ca. 10 min per frame. Superpixel segmentation can also be formulated based on color, stereo depth, and stereo 3D flow simultaneously [Van den Bergh and van Gool, 2012]. 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.

Interactive vision is a line of research in robotics that frequently uses motion cues to identify novel objects [Fitzpatrick, 2003, Kenney et al., 2009]. Fitzpatrick [2003] proposed a background subtraction method in color images which segments the image into robot and object parts while the robot manipulates objects. He finds the point of first contact in an image sequence and determines the moving parts beforehand (robot) and afterwards (object). Kenney et al. [2009] also perform background subtraction and find coherent object segments using graph cuts. For segmentation, these approaches assume a static camera pose, whereas our approach recovers camera and object motion concurrently. Furthermore, our segmentation method is suitable for mobile manipulation scenarios, where keeping the moved object within the field of view would involve camera motion.

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,
- an efficient implementation based on a compact image representation and fast probabilistic registration techniques, and
- a novel benchmark dataset to compare our results with other approaches.

3 Efficient Rigid Multi-Body Registration of RGB-D Images

Our approach to rigid multi-body registration segments moving rigid parts between two RGB-D images, 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 $I = (x_1, \ldots, x_N)$ is partitioned into discrete sites with observations x_i such as pixels or map elements in a 3D representation. We index the sites with variable i. Let $Y = Y_1 \times \cdots \times Y_N$ be the labeling domain of the image. The site labeling $y_i \in Y_i$ denotes the membership of a site in one of M distinct motion segments or in the set of outliers O. Hence, the domains Y_i of the site labelings each are the label set $\mathcal{L} := \{O, 1, \ldots, M\}$. With $y = (y_1, \ldots, y_N)$ we denote a concrete labeling of the whole image.

All sites within a segment move with a common six degree-of-freedom (6-DoF) rigid-body motion $\theta_k \in$

symbol	meaning
I_{seg}, I_{ref}	segmented and reference image
i	site (e.g. pixel) in an image
x_i	observation at image site i
O	outlier segment
M	number of motion segments
$\mathcal{L} := \{O, 1, \dots, M\}$	label set
Y_i	label domain at a pixel, equals $\mathcal{L}$
$Y = Y_1 \times \cdots \times Y_N$	image labeling domain
$y_i \in Y_i$	labeling of image site i
$y = (y_1, \dots, y_N)$	image labeling
$\theta_k \in \mathbb{SE}(3)$	rigid-body motion of segment k
$\Theta = \{\theta_k\}_{k=1}^M$	set of rigid-body motions
$\overline{\Theta}$	motion estimates from last M-step
$\Omega = \{I_{seg}, \Theta, I_{ref}\}$	abbreviation
$\overline{\Omega} = \{I_{seg}, \overline{\Theta}, I_{ref}\}$	abbreviation
$\varphi(y_i, \Omega)$	unary potentials
$\varphi_{\mathcal{S}}(y_i, y_j, \Omega)$	pairwise smoothness potentials
$\varphi_{\mathcal{A}}(y_i, y_j, \Omega)$	pairwise disambiguation potentials
$\varphi(y_i, y_j, \Omega)$	both pairwise potentials
$\mathcal{N}_{\mathcal{S}}(i)$	neighborhood for smoothness
$\mathcal{N}_{\mathcal{A}}(y_i)$	neighborhood for disambiguation
$\mathcal{N}(y_i)$	neighborhood for both potentials
$\gamma(x_i, x_j)$	data-driven smoothness strength
α	disambiguation strength
$q_i(y_i \mid \overline{\Omega})$	mean-field label likelihood at site i
λ	label cost
s	surfel
μ, Σ, n	surfel mean, covariance, normal
$\mathcal{A}_k$	surfel associations of segment k

Table 1 List of symbols.

$\mathbb{SE}(3)$ between the segmented image I_{seg} and a reference image I_{ref}. Table 1 gives an overview on our notation.

3.1 An Expectation-Maximization Framework for Dense 3D Motion Segmentation of Rigid Parts

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:

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

The labeling of the image sites is a latent variable that we estimate jointly with the rigid-body motions of the segments using EM (e.g., [Bishop, 2006]). With the shorthands $\Omega = \{I_{seg}, \Theta, I_{ref}\}$ and $\overline{\Omega} = \{I_{seg}, \overline{\Theta}, I_{ref}\}$, the EM objective is

$$\arg\max_{\Theta} \sum_{y \in Y} p(y \mid \overline{\Omega}) \ \ln p(I_{seg}, y \mid \Theta, I_{ref}). \tag{2}$$

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

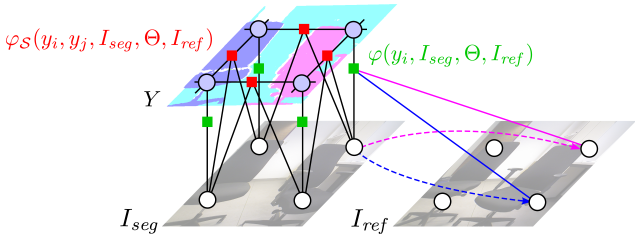

Fig. 3 We model the likelihood of an image labeling in a CRF with unary and pairwise potentials. The unary potentials measure the likelihood of observation between segmented and reference image under the motion estimate of a label. The pairwise potentials penalize differing labelings between image sites with low contrast and curvature.

and $p(y \mid \overline{\Omega})$ is the posterior distribution of the image labeling. Note the summation over all possible labelings $y \in Y$ of the image. Our EM approach is illustrated in Fig. 2.

We further factorize

$$p(I_{seg}, y \mid \Theta, I_{ref}) = p(I_{seg} \mid y, \Theta, I_{ref})\, p(y \mid \Theta, I_{ref}). \tag{3}$$

If we assume a uniform prior over labelings without knowing the image content, we can formulate our EM-objective as

$$\arg\max_{\Theta} \sum_{y \in Y} p(y \mid \overline{\Omega}) \ \ln p(I_{seg} \mid y, \Theta, I_{ref}). \tag{4}$$

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 labeling given the latest motion estimates $\overline{\Theta}$ to form the conditional expectation in (2).

M-step: Find new motion estimates Θ by maximizing the conditional expectation (2), given the posterior distribution of the image labeling.

3.2 Image Labeling Posterior

We model the likelihood of an image labeling y in a CRF

$$p(y \mid \Omega) = \prod_{i=1}^{N} \varphi(y_i, \Omega) \prod_{j \in \mathcal{N}_{\mathcal{S}}(i)} \varphi_{\mathcal{S}}(y_i, y_j, \Omega) \cdot \prod_{j \in \mathcal{N}_{\mathcal{A}}(y_i)} \varphi_{\mathcal{A}}(y_i, y_j, \Omega), \tag{5}$$

where $\varphi(y_i, \Omega)$ are unary potentials on the image sites, and the pairwise potentials model interactions between image sites i and j (see Fig. 3). We will introduce

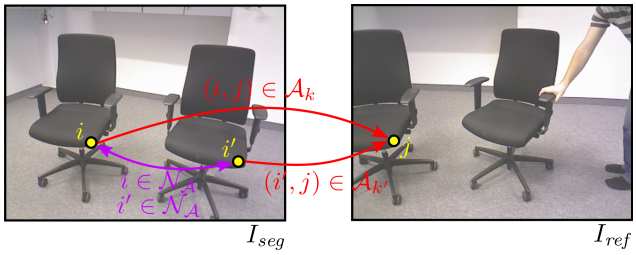

Fig. 4 Ambiguity resolution. If sites i and i' associate with the same site j in the reference image for motion segments k and k' (i.e. $(i,j) \in \mathcal{A}_k$ and $(i',j) \in \mathcal{A}_{k'}$), we include additional pairwise CRF terms between them. The likelihood of the assignment of both sites to labels k and k' is set to a small value (large negative log-likelihood α).

two kinds of pairwise potentials: smoothness potentials $\varphi_{\mathcal{S}}$ enforce spatial coherence of the segments, while ambiguity-resolving potentials $\varphi_{\mathcal{A}}$ prevent from reusing image sites in multiple data associations. The neighborhood $\mathcal{N}_{\mathcal{S}}(i)$ links direct neighbors in the image. The neighborhood $\mathcal{N}_{\mathcal{A}}(y_i)$ of an image site is label-dependent and consists of other image sites that, for the different labels, would transform to the same counterpart in the refence image. We abbreviate the pairwise potentials used with the combined neighborhood $\mathcal{N}(y_i)$ and the potential $\varphi(y_i, y_j, \Omega)$.

Unary Potentials: The unary potentials are given by the observation likelihood

$$\varphi(y_i, \Omega) := p(x_i \mid y_i, \Theta, I_{ref}) = p(x_i \mid \theta_{y_i}, I_{ref}), \tag{6}$$

which quantifies the likelihood to observe $x_i \in I_{seg}$ in I_{ref} under the motion estimate θ_{y_i} for label y_i. For the outlier label $l_i = O$, we set the observation likelihood to a constant p_O.

Pairwise Smoothness Potentials: Between direct neighbors i and j in the image representation, we use a contrast-sensitive Potts model [Boykov and Jolly, 2001]

$$\ln \varphi_{\mathcal{S}}(y_i, y_j, I_{seg}) = -\gamma(x_i, x_j)\ \delta(y_i, y_j), \tag{7}$$

where we define

$$\delta(y_i, y_j) := \begin{cases} 0 & \text{, if } y_i = y_j, \\ 1 & \text{, if } y_i \neq y_j, \end{cases} \tag{8}$$

and $\gamma(x_i, x_j) > 0$ controls the strength of the coupling in dependence on the difference between the observations at the image sites. It depends on the choice of the underlying image representation (see Sec. 3.6.2). We denote the set of direct neighbors of site i by $\mathcal{N}_{\mathcal{S}}(i)$.

Pairwise Disambiguation Potentials: We also need to avoid multiple associations of image sites in the segmented image with the same image site in the reference image (see Fig. 4). Otherwise, our 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.

The image site labelings decide on an association of sites between both images. In order to prevent the graph cut optimization from establishing labelings that would associate multiple times to a site in the reference image, we introduce additional pairwise couplings. We consider 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', respectively. We define the pairwise potential

$$\ln \varphi_{\mathcal{A}}(y_i, y_j) := \begin{cases} -\alpha & \text{if } y_i = k \wedge y_j = k' \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where α sets the strength of the couplings. We refer to the set of sites with the same association like site i under label y_i by $\mathcal{N}_{\mathcal{A}}(y_i)$. These sites are additionally coupled with i through pairwise potentials in the CRF.

3.3 Efficient Approximate Solution of the Expectation-Maximization Formulation

We propose an efficient approximate solution to the EM formulation. Firstly, we see that the observation likelihood of the segmented image in the reference image given motion estimates and labeling,

$$p(I_{seg} \mid y, \Theta, I_{ref}), \tag{10}$$

factorizes into the likelihood of the individual observations

$$p(I_{seg} \mid y, \Theta, I_{ref}) = \prod_{i=1}^{N} p(x_i \mid \theta_{y_i}, I_{ref}) \tag{11}$$

since we assume stochastic independence between the observations and each site is associated to exactly one segment given a specific labeling y. By this, eq. (2) becomes

$$\arg\max_{\Theta} \sum_{y \in Y} p(y \mid \overline{\Omega}) \sum_{i=1}^{N} \ln p(x_i \mid \theta_{y_i}, I_{ref}). \tag{12}$$

Note that each term of the inner sum only depends on one of the image labels.

Since exact inference of the joint label likelihood $p(y \mid \overline{\Omega})$ in a CRF is not tractable even for a single labeling y, we need to resort to approximations.

We apply a variational mean-field approximation [Saito et al., 2012] to the joint label likelihood

$$p(y \mid \overline{\Omega}) \approx \prod_{i=1}^{N} q_i(y_i \mid \overline{\Omega}) \tag{13}$$

to write

$$\arg\max_{\Theta} \sum_{y_1 \in Y_1} \dots \sum_{y_N \in Y_N} \left(\prod_{i=1}^{N} q_i(y_i \mid \overline{\Omega}) \right) \cdot \left(\sum_{i=1}^{N} \ln p(x_i \mid \theta_{y_i}, I_{ref}) \right) \tag{14}$$

in a principled way. Rearranging terms yields

$$\arg\max_{\Theta} \sum_{i=1}^{N} \left(\sum_{y_i \in Y_i} q_i \ \ln p(x_i \mid \theta_{y_i}, I_{ref}) \cdot \left(\sum_{y_1 \in Y_1} q_1 \dots \left(\sum_{y_{i-1} \in Y_{i-1}} q_{i-1} \cdot \left(\sum_{y_{i+1} \in Y_{i+1}} q_{i+1} \dots \left(\sum_{y_N \in Y_N} q_N \right) \right) \right) \right) \right), \tag{15}$$

where we use the shorthand $q_i := q_i(y_i \mid \overline{\Omega})$. Since the factors are normalized such that $\sum_{y_i \in Y_i} q_i(y_i \mid \overline{\Omega}) = 1$, we arrive at

$$\arg\max_{\Theta} \sum_{i=1}^{N} \sum_{y_i \in Y_i} q_i(y_i \mid \overline{\Omega}) \ln p(x_i \mid \theta_{y_i}, I_{ref}), \tag{16}$$

which is equivalent to

$$\arg\max_{\Theta} \sum_{k=0}^{M} \sum_{i=1}^{N} q_i(y_i = k \mid \overline{\Omega}) \ln p(x_i \mid \theta_k, I_{ref}). \tag{17}$$

Intuitively, each image site is assigned a weight $q_i(y_i = k \mid \overline{\Omega})$ for the reestimation of the rigid-body motion θ_k.

In the E-step, the factors $q_i(y_i \mid \overline{\Omega})$ are estimated in an iterative process $q_i^{\tau-1} \rightsquigarrow q_i^{\tau}$ using mean-field updates:

$$\ln q_i^{\tau}(y_i \mid \Omega) = const. + \ln \varphi(y_i, \Omega) + \sum_{j \in \mathcal{N}(y_i)} \sum_{y_j \in Y_j} q_j^{\tau-1}(y_j \mid \Omega) \ \ln \varphi(y_i, y_j, \Omega). \tag{18}$$

Since this process only performs local updates, the quality of the found local optimum strongly depends on the initial estimate $q_i^0(y_i)$. We therefore initialize the mean-field iterations with a ML-solution found by graph cuts [Boykov et al., 2001]

$$y_{ML} = \arg\max_{y \in Y} p(y \mid \overline{\Omega}) \tag{19}$$

such that

$$q_i^0(y_i \mid \overline{\Omega}) = \begin{cases} 1 & \text{if } y_i = y_{i,ML} \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

Due to the pairwise ambiguity-resolving potentials, the pairwise potentials define a semi-metric, since transitivity is not satisfied. While α-expansions require the pairwise potentials to be a metric, $\alpha\beta$-swaps are applicable for semi-metrics.

For an efficient algorithm, we are not required to run the mean-field iterations until convergence. A single iteration suffices to improve the estimate for $p(y \mid \overline{\Omega})$, which also improves the lower bound of the EM-algorithm. As we use graph cuts to seed the iterations, we typically obtain good solutions within a few cycles of EM by reducing the Kullback-Leibler divergence between $p(y \mid \overline{\Omega})$ and our approximation. We observe that according to eq. (18), after a single iteration the factors are

$$q_i^1(y_i \mid \Omega) = \eta_i \exp\left(\ln p(x_i \mid y_i, \Theta, I_{ref}) + \sum_{j \in \mathcal{N}(y_i)} \sum_{y_j \in Y_j} q_j^0(y_j \mid \Omega) \ \ln \varphi(y_i, y_j, I_{seg}) \right), \tag{21}$$

where η_i is a normalization factor such that $\sum_{y_i \in Y_i} q_i^1(y_i \mid \overline{\Omega}) = 1$. Plugging our ML-seed (eq. (20)) into eq. (21) yields

$$q_i^1(y_i \mid \Omega) = \eta_i \ p(x_i \mid y_i, \Theta, I_{ref}) \prod_{j \in \mathcal{N}(y_i)} \varphi(y_i, y_{j,ML}, I_{seg}). \tag{22}$$

Interestingly, the factors $q_i^1(y_i \mid \overline{\Omega})$ are local conditional probabilities

$$q_i^1(y_i \mid \Omega) = p(y_i \mid y_{ML} \setminus \{y_i\}, \overline{\Omega}) \tag{23}$$

in the CRF conditioned on the ML-solution. The weight intuitively is the likelihood that site i belongs to the segment with respect to the ML-labeling. Note that if the graph-cuts avoid ambiguous associations, the corresponding pairwise terms vanish from eq. (22).

3.4 Model Complexity

The pairwise 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

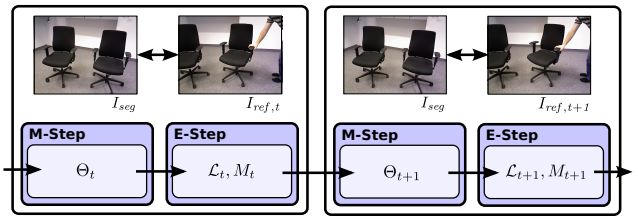

Fig. 5 Online EM. The EM framework is used to segment RGB-D images in video online by performing a few M- and E-steps per subsequent image. Typically, one iteration per image suffices.

redundant motion segments for the image segments. To control model complexity, we enhance the graph cut optimization in Sec. 3.3 with label costs [Delong et al., 2012], i.e., we use graph cuts to optimize the augmented CRF energy function (eq. (5))

$$E(y) = -\sum_{i=1}^{N} \ln \varphi(y_i, \Omega) - \sum_{j \in \mathcal{N}(y_i)} \ln \varphi(y_i, y_j, \Omega) - \sum_{l \in \mathcal{L}} \ln \varphi(l, y), \quad (24)$$

with per-label-costs

$$\ln \varphi(l, y) := \begin{cases} -\lambda & \text{if } l \neq 0 \wedge \exists y_i \in y : y_i = l \\ 0 & \text{otherwise.} \end{cases} \quad (25)$$

Each label is assigned the same cost λ except the outlier label for which we impose no cost. Label costs have a natural interpretation of implementing information criteria such as the Bayesian information criterion (BIC).

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. 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 by 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 discarded after the E-step.

3.5 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 y_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). The segmentation of temporally distant images also improves accuracy for motions with small velocities, while registration becomes more challenging for large motions. Our approach counteracts the latter by tracking motion over time.

3.6 Image Representation

The performance of our EM approach depends on the underlying image representation. Any representation is suitable that defines observation likelihood $p(x_i \mid \theta_{y_i}, I_{ref})$, image site neighborhood $\mathcal{N}_\mathcal{S}$, and dissimilarity $\gamma(x_i, x_j)$ for the pairwise interaction terms. To solve for the motion estimates of the segments in eq. (17), 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, Stückler and Behnke [2014]). 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 p is limited with its squared distance d^2,

$$\rho(p) = \max\left\{\rho_{\max}, 1/\left(\rho_d d^2\right)\right\}, \quad (26)$$

in order to capture the 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/ .

3.6.1 Observation Likelihood

We interpret voxels x in the MRSMap as image sites. Each voxel in a MRSMap contains a surfel s which is defined by mean $\mu \in \mathbb{R}^6$ and covariance $\Sigma \in \mathbb{R}^{6\times6}$ of the colored points falling into the voxel. The first three coordinates describe the Cartesian coordinates

of the points, while the latter are used for their color. In MRSMaps, the RGB values are represented with a Cartesian variant of the HSL color space. It consists of luminance L and two chrominances α and β.

Given the labeling y_i, the surfel $s_{seg,i}$ in voxel $x_{seg,i}$ is observed at a corresponding surfel $s_{ref,j}$ in voxel $x_{ref,j}$ under the rigid-body motion estimate θ_{y_i}, i.e., we model the observation likelihood

$$p(s_{seg,i}|\theta_{y_i}, s_{ref,j}) = \mathcal{N}\left(d^*(s_{seg,i}, s_{ref,j}, \theta_{y_i}); 0, \Sigma^*(s_{seg,i}, s_{ref,j}, \theta_{y_i})\right), \tag{27}$$

where we define

$$\begin{aligned} d^*(s_{seg,i}, s_{ref,j}, \theta_{y_i}) &:= \mu_{ref,j} - \left(R^*_{y_i}\mu_{seg,i} + t^*_{y_i}\right), \\ \Sigma^*(s_{seg,i}, s_{ref,j}, \theta_{y_i}) &:= \Sigma'_{ref,j} + R^*_{y_i}\Sigma'_{seg,i}\left(R^*_{y_i}\right)^T, \end{aligned} \tag{28}$$

If multiple surfels are contained within the voxels i and j for several view directions, we assign the best observation likelihood among all pairs of view directions. Here, we take spatial as well as color information into account such that

$$R^*_{y_i} = \begin{pmatrix} R_{y_i} & 0 \\ 0 & I_3 \end{pmatrix} \in \mathbb{R}^{6\times 6}, t^*_{y_i} = \begin{pmatrix} t_{y_i} \\ 0 \end{pmatrix} \in \mathbb{R}^6 \tag{29}$$

rotates the surfel coordinates according to the motion estimate and t_{y_i} is the translational part of θ_{y_i}. Correlations between the point and color distributions cannot be considered since the color distribution is not comparable for large spatial misalignments at which surface has not been measured. We hence remove these correlations by setting the corresponding entries in the surfel covariances $\Sigma'_{ref,j}$ and $\Sigma'_{seg,i}$ to zero. Furthermore, in order to improve robustness for illumination changes, we neglect small luminance and chrominance differences by setting differences below specific values to zero in each dimension.

For the unary potentials, we additionally examine the consistency of the surfel normals in the combined likelihood

$$\begin{aligned} \varphi(y_i, \Omega) = &\mathcal{N}\left(d^*(s_{seg,i}, s_{ref,j}, \theta_{y_i}); 0, \Sigma^*(s_{seg,i}, s_{ref,j}, \theta_{y_i})\right) \\ &\cdot \mathcal{N}\left(\arccos\left(n_{ref,j}, R(\theta_{y_i})\, n_{seg,i}\right), \sigma_n^2\right) \end{aligned} \tag{30}$$

with standard deviation σ_n. Since the rotation around the surface normal is not observable, we do not use the term for pose optimization.

The evaluation of the observation likelihood involves the association of the surfel $s_{seg,i}$ with a surfel $s_{ref,j} = A_k(s_{seg,i})$ from the reference image. The mean position of the surfel $s_{seg,i}$ is transformed to the reference image according to the motion estimate θ_{y_i}. We then search for a matching surfel in the reference image from coarse to fine resolutions. We scale the search radius with inverse resolution and find the association on the finest resolution possible. Each motion segment requires its own set of associations

$$\mathcal{A}_k := \{(s_{seg}, s_{ref}) \in I_{seg} \times I_{ref} \mid s_{ref,j} = A_k(s_{seg,i})\}. \tag{31}$$

Care has to be taken at image borders, background at depth discontinuities, and occlusions, since no association can be made. Assigning a low likelihood would be pessimistic and bias these parts to be explained as outliers, which would also affect connected image regions through the spatial smoothness potentials in the CRF. Instead, we assign the last observed data likelihood to such surfels.

3.6.2 Smoothness Cost Terms

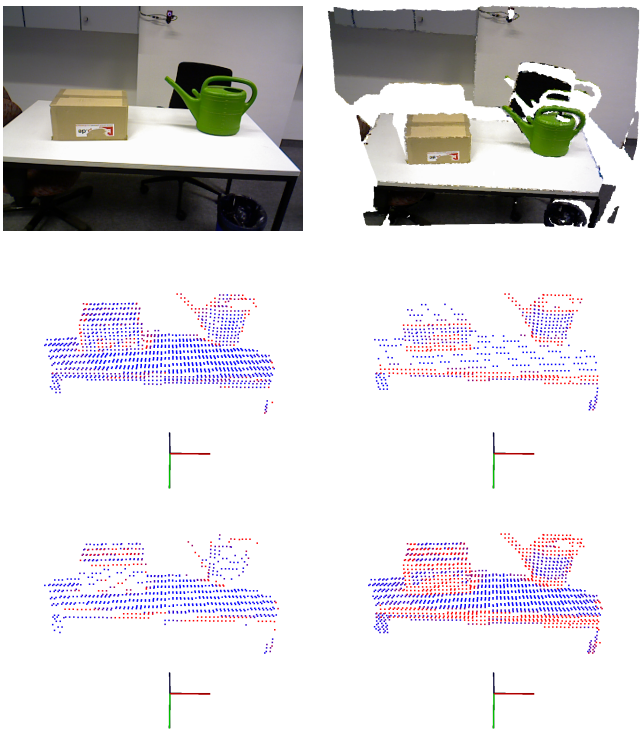

Fig. 6 Pairwise interactions in MRSMaps. We visualize the smoothness cost terms for direct voxel neighbors to the right (middle left), down (middle right), and forward (bottom left) directions. Directions are according to the shown camera frame (right: red, down: green, forward: blue axis). Bottom right: maximum cost over all neighbors. Costs are color-coded from blue (low) to red (high). Missing voxels either do not exist on the displayed resolution (0.025 m) or they have no valid neighbor in the specific direction.

We establish pairwise 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 3D representation and across the discrete changes of the depth-dependent resolution limit. We weaken pairwise couplings by the dissimilarity of surfels,

$$\gamma(x_i, x_j) := g_s \ \min\left\{1, \max\left\{0, \max\left\{g_n(1 - n_i^T n_j), \right.\right.\right. \\ \left.\left.\left. g_L d_L(s_i, s_j), \ g_\alpha d_\alpha(s_i, s_j), g_\beta d_\beta(s_i, s_j)\right\} - g_0\right\}\right\}, \quad (32)$$

where g_s is a scale parameter,

$$d_L(s_i, s_j) = |\mu_{L,i} - \mu_{L,j}|, \quad (33)$$
$$d_\alpha(s_i, s_j) = |\mu_{\alpha,i} - \mu_{\alpha,j}|, \quad (34)$$
$$d_\beta(s_i, s_j) = |\mu_{\beta,i} - \mu_{\beta,j}|, \quad (35)$$

and g_0 handles illumination differences and noise. Further parameters g_n, g_L, g_α and g_β are used to adjust the strengths of the individual cues. Fig. 6 illustrates our smoothness terms in an example.

3.6.3 Motion Estimation

The motion of the segments is estimated in the M-step. We apply our efficient rigid registration method for MRSMaps to the optimization of the EM-objective (eq. (17)). We augment the algorithm to incorporate the weighting by the mean-field factors

$$\arg\max_{\theta_{y_i}} \sum_{(s_i,s_j)\in\mathcal{A}_{\theta_{y_i}}} q_i^1(y_i \mid \overline{\Omega}) \\ \ln p(s_{seg,i} \mid \theta_{y_i}, s_{ref,j}), \quad (36)$$

with the set of surfel associations $\mathcal{A}_{\theta_{y_i}}$ of motion segment θ_{y_i} that has at most one association per surfel in the segmented image. This weighted least-squares objective is optimized using the Levenberg-Marquardt method.

Since this registration procedure performs local optimization, a good initialization is important. During incremental EM, parts of the scene may start to move at any time and split an existing segment. We initialize the motion estimate for yet unsupported segments m_k with an estimate of a supported segment $m_{\widehat{k}}$. We first identify which of the supported segments or the outlier set best explains m_k through

$$\widehat{k} = \arg\max_{k'\in\{0,1,\ldots,M\}} \sum_{y_i\in y_{ML}: y_{ML,i}=k'} q_i^1(y_i = k \mid \overline{\Omega}). \quad (37)$$

If this segment is not the outlier label, i.e., $\widehat{k} \neq 0$, we set $\theta_k = \theta_{\widehat{k}}$. Otherwise, we use the largest segment.

Fig. 7 Example segmentations (top, outliers dark red) towards a reference image (bottom) from the test sequences (left: small, middle: medium, right: large).

parameter	setting
α	100.0
λ	chairs seq.: 2000 , other seqs.: 50000
σ_n	$\pi/8$
g_s	0.4 (same resolution), 0.2 (different res.)
g_n	8
g_L	10
g_α	10
g_β	10
g_0	0.2
$\rho_{\max}$	$(0.0125\,m)^{-1}$
ρ_d	0.014

Table 2 Parameter settings.

sequence	small	medium	large
run-time in ms	200.2±42.3	213.1±54.7	138.7±37.5
median trans. error in m	0.012	0.018	0.034
median rot. error in rad	0.047	0.029	0.049

Table 3 Mean ± standard deviation of run-time and motion estimate accuracy of our method over all frames of the test sequences.

sequence	small	medium	large
median trans. error in m	0.013	0.020	0.030
median rot. error in rad	0.045	0.030	0.048

Table 4 Motion estimate accuracy of our method under real-time constraints.

4 Experiments

We evaluate segmentation and motion estimation accuracy of our approach on three RGB-D video sequences with ground-truth information[1]. 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. 7). 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 recorded with an Asus Xtion

[1] available from http://www.ais.uni-bonn.de/download/rigidmultibody

approach	sequence		
	small	medium	large
ours (all frames)	**0.95±0.11**	**0.94±0.12**	0.63±0.42
ours (real-time)	0.91±0.23	0.91±0.20	**0.65±0.41**
Ochs et al. [2014]	0.58±0.31	0.40±0.29	0.38±0.29

Table 5 Average segmentation accuracy (see Sec. 4.1) ± standard deviation of our method and the approach by Ochs et al. [2014]. The segmentation accuracy is averaged over objects and processed pairs of images in the sequences.

approach	sequence		
	small	medium	large
ours (all frames)	**0.05±0.29**	0.11±0.43	-0.58±1.01
ours (real-time)	-0.09±0.35	**0.04±0.45**	-0.43±0.92
Ochs et al. [2014]	2.14±1.19	1.44±0.96	**-0.05±1.22**

Table 6 Mean ± standard deviation of the error in the number of segments M of our method and the approach by Ochs et al. [2014].

sequence	number of segments M			
	1	2	3	4
small	139.5±15.9	181.5±27.9	232.6±36.9	–
medium	142.7±19.4	166.2±30.9	224.2±46.8	298.9±50.8
large	102.4±17.3	125.6±24.2	158.5±30.4	192.3±37.0

Table 7 Mean ± standard deviation of run-time (ms) for different number of segments.

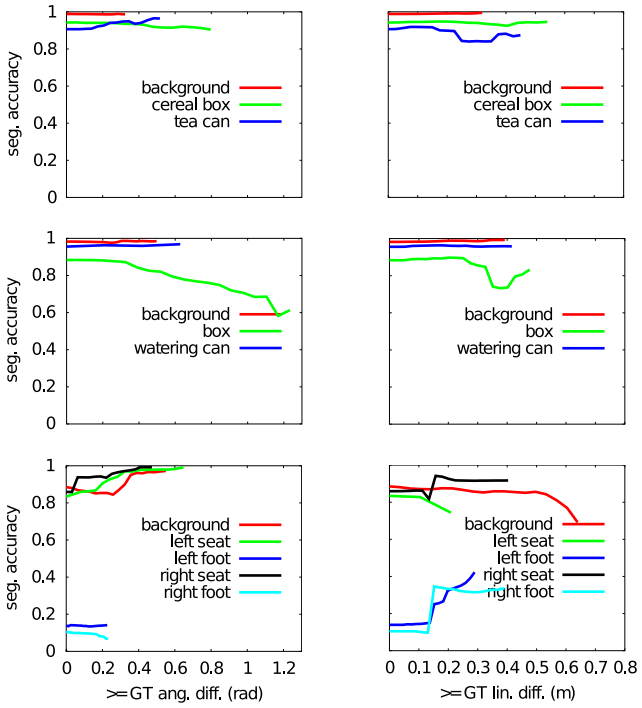

Fig. 8 Average segmentation accuracy vs. increasing rotational (left) and translational (right) ground-truth object motion (top: small, middle: medium, bottom: large objects). The mean is determined for segment motion greater or equal the value on the x-axis.

Pro Live camera. Ground truth of the 3D rigid-body motion has been obtained with an OptiTrack 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 and legs of persons are annotated with don't-care labels. Additionally, we set pixels to don't 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 automatically determine groups of objects that move jointly between the frames (0.12 rad rotational and 0.05 m translational motion) and merge their segments.

The sequences are processed sequentially, starting from each ground-truth labeled image as the image to be segmented. If not stated otherwise, the sequences are processed frame-by-frame, i.e. all frames of the 30 Hz recording are used. In real-time mode, we drop frames if they would arrive during the processing of a frame. The experiments have been run on an Intel Core i7-4770K CPU at a maximum clock speed of 3.50 GHz. We determined the parameters (Table 2) of our approach empirically.

We also provide comparisons with the state-of-the-art motion segmentation approach by Ochs et al. [2014] as a baseline. This algorithm processes RGB only and does not utilize the dense depth available in the RGB-D images, which is obtained through additional active sensing. The visible background in our benchmark also contains planar textureless surfaces such as walls and tables which renders motion segmentation purely based on RGB information difficult. Furthermore, the method does not constrain object motion to rigid-body motions. Hence, an advantage for our method is expectable on the datasets. We found a parameter setting for this baseline approach of $\nu = 0.1$, $\alpha = 100$, and a tracking subsampling factor of 8 for all three sequences by evaluating parameters on a grid. Run-time evaluation has been performed with an Intel Core i7-3610QM CPU and an NVidia GeForce GT 630M GPU.

4.1 Evaluation Measures

We quantify the average segmentation accuracy of the ground-truth segments with the measure proposed

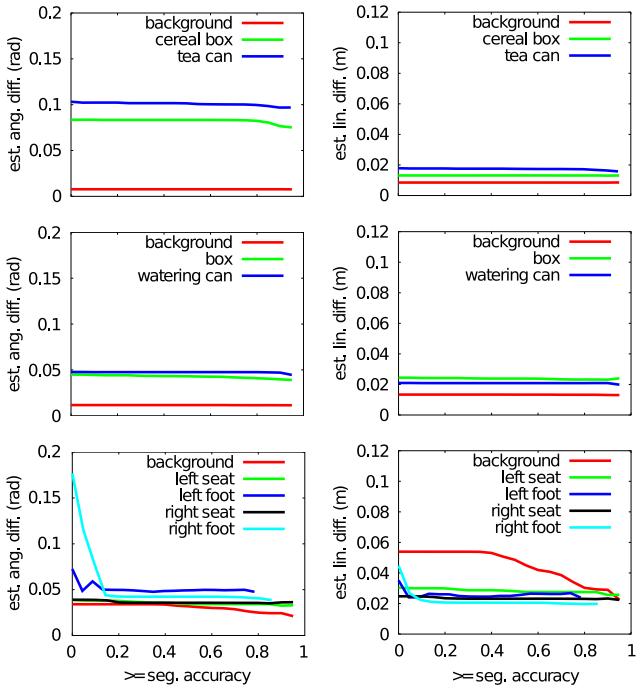

Fig. 9 Median rotational (left) and translational (right) error of the camera motion estimate vs. increasing object segmentation accuracy (top: small, middle: medium, bottom: large objects). The median is determined for segmentation accuracies greater or equal the value on the x-axis.

by Everingham et al. [2010],

$$seg.\ acc. = \frac{true\ positives}{true\ pos. + false\ pos. + false\ negatives}, \tag{38}$$

for which we back-project the resulting motion segmentation from the MRSMaps into the segmented RGB-D images and account for the labeling of each pixel. The segmentation accuracy for an individual object is determined as the best accuracy for any association of the ground truth segment to the estimated segments. We define average object segmentation accuracy, as given in Table 5, as the mean over the individual segmentation accuracies of the moving objects between images. We also measure translational and rotational errors between ground-truth and estimated motion.

4.2 Run-Time

The run-time of our approach is given in Tables 3 and 7. It segments images fast at a frame rate of about 2 to 10 Hz. As can be seen from Table 7 the run-time depends on the number of segments. It also depends on the distance of the camera to the measured surfaces which explains the qualitative difference in run-time between the large objects sequence to the other two sequences.

The approach by Ochs et al. demands significantly more computation time than our approach. Computing optical flow between a pair of images requires several seconds (approx. 10 to 30 s). Tracking and grouping motion uses run-time that depends on the length of the image sequence. We measured run-times in the minutes for sequence lengths of several hundred images (e.g., approx. 9 min for a sequence length of 512 frames on the small objects sequence). Finally, segmentation densification runs within a few seconds (approx. 4 s per label) on the GPU for one image.

4.3 Segmentation Accuracy

Fig. 8 shows average segmentation accuracy in dependency on the actual translational and rotational motion of the objects. To visualize the effect of different degrees of object motion on the segment accuracy, we vary a threshold for the translational and rotational motion and determine the average segmentation accuracy for those results for which the motion is above the threshold in Fig. 8.

Most objects and the background in the sequences can be very well segmented. The box-shaped objects show a drop in segmentation accuracy with rotation since sides of the boxes become occluded. For the chairs (bottom row) it can be seen that moderate object motion facilitates high segmentation accuracy. This is explained by the distant hence noisy, structure-less, and untextured background which allows only coarse misalignments to be detected. Note that in some subsequences, over- or undersegmentations could occur which can lower the average segmentation accuracy in specific intermediate intervals of object motion. 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 in the sequences, and achieves good overall segmentation accuracy (see Tables 5 and 6). Notably, if frames are dropped to operate in real-time, we obtain similar performance to processing all frames.

In comparison, the approach by Ochs et al. yields less accurate segmentation results than our method on the test sequences[2]. We observed, that the approach tends to undersegment the objects and to oversegment the background. In contrast to our method, the approach does not exploit the rigidity of objects and back-

[2] Due to the high run-time requirements of the method, we evaluated the approach at full frame-rate for sequence lengths that are multiples of 30 frames.

ground, and does not use depth information. The objects and the background are rather textureless, which renders motion estimation solely from photometric cues difficult.

4.4 Motion Estimate Accuracy

The results in Fig. 9 demonstrate that our approach recovers camera motion relative to the objects accurately. In Fig. 9, we determine the median pose error for all results above the varied segmentation accuracy threshold. While for many objects motion accuracy increases with segmentation accuracy, the motion also seems well estimated for low segmentation accuracies. Low segmentation accuracy often coincides with small displacements of the objects. For the small objects, or for the background at low segmentation accuracy, the pose estimates are less accurate. 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.

5 Conclusions

We presented an efficient dense motion segmentation approach for RGB-D image sequences. We employ EM to infer image labeling and motion estimates, and propose efficient approximations based on variational mean-field inference and graph cuts. Our approach recovers the number of motion segments and is suited for online operation in real-time. Our efficient probabilistic image representation in MRSMap and rapid registration method facilitate fast performance. In experiments, we demonstrated high accuracy of our method with regards to segmentation and motion estimates. Our approach also recovers the number of motion segments well.

We gain efficiency in our approach in several ways. For efficient CRF inference, we use a mean-field approximation which we initialize with graph cuts. We observed that a single step of mean-field updates suffices to obtain a soft labeling in the E-step. By using MRSMaps, RGB-D images are compressed from up to 307,200 pixels to only a few thousand surfels. MRSMaps support efficient data association and registration for the M-step. On-line processing is sped up by initializing the EM algorithm with the result of the previous frame.

The accuracy of our motion segmentation approach clearly depends on measurement accuracy as well as the underlying image representation. In order to improve the segmentation of fine-detailed structure and to increase the accuracy of motion estimation for small or repetitive objects, we could integrate interest points 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 consider degrading image overlap, segmentation evidence from multiple views could be beneficial to increase overlap. Also including physical priors into the segmentation could further support our approach in resolving ambiguous observations of motions, for instance, if an object moves along a planar, textureless surface of another object.

Future research could investigate the application of our EM framework to different image representations and registration methods. As the rigid registration method used for the M-step has local convergence properties, also motion estimation converges locally. By using a global alignment method for registration as for instance in [Drost et al., 2010], global convergence could be achieved. Finally, our dense motion segmentation could be suitable as a building block for the parsing and reconstruction of dynamic scenes. To this end, motion needs to be segmented between many pairs of images in video and fused into a map of the individual moving objects. This could be formulated as a joint motion labeling problem in the image sequence.

References

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

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

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

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.

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.

T. Brox, A. Bruhn, and J. Weickert. Variational motion segmentation with level sets. In *Proc. of the Europ.*

Conf. on Computer Vision (ECCV), Lecture Notes in Computer Science, pages 471–483. 2006.

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

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.

Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. Model globally, match locally: Efficient and robust 3D object recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

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.

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.

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.

S. Hadfield and R. Bowden. Scene particles: Unregularized particle based scene flow estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(3):564 – 576, 2014.

Evan Herbst, Xiaofeng Ren, and Dieter Fox. RGB-D flow: Dense 3-D motion estimation using color and depth. In *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2276–2282, 2013.

Evan Herbst, Peter Henry, and Dieter Fox. Toward online 3-D object segmentation and mapping. In *International Conference on Robotics and Automation (ICRA)*, 2014.

M. Hornacek, A. Fitzgibbon, and C. Rother. SphereFlow: 6 DoF scene flow from RGB-D pairs. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

F. Huguet and F. Devernay. A variational method for scene flow estimation from stereo sequences. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2007.

J. Kenney, T. Buckley, and O. Brock. Interactive segmentation for manipulation in unstructured environments. In *Proc. of the IEEE ICRA*, 2009.

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.

P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1187 – 1200, Jun 2014. Preprint.

J. Quiroga, F. Devernay, and J. L. Crowley. Local/global scene flow estimation. In *Proc. of the IEEE International Conference on Image Processing (ICIP)*, 2013.

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

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.

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.

M. Saito, T. Okatani, and K. Deguchi. Application of the mean field methods to mrf optimization in computer vision. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1680–1687, 2012.

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.

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.

J. Stückler and S. Behnke. Efficient dense 3D rigid-body motion segmentation in RGB-D video. In *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, 2013.

J. Stückler and S. Behnke. Multi-resolution surfel maps for efficient dense 3D modeling and tracking. *J. of Visual Communication and Image Representation*, 2014.

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.

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.

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.

J. Weber and J. Malik. Rigid body segmentation and shape description from dense optical flow under weak perspective. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19:139–143, 1997.

A. Wedel and D. Cremers. *Stereoscopic Scene Flow for 3D Motion Analysis*. 2011.

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

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.