2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Combining Motion and Appearance for Robust Probabilistic Object Segmentation in Real Time

Vito Mengers[1,2] Aravind Battaje[1,2] Manuel Baum[1,2] Oliver Brock[1,2]

***Abstract*— We present a robust method to visually segment scenes into objects based on motion and appearance. Both these cues provide complementary information that we fuse using two interconnected recursive estimators: One estimates object segmentation from motion as a probabilistic clustering of tracked 3D points, and the other estimates object segmentation from appearance as a probabilistic image segmentation. The interconnected estimators provide a probabilistic and consistent object segmentation in real time, which makes them well suited for many downstream robotic tasks. We evaluate our method on one such task, kinematic structure estimation, on a dataset of interactions with articulated objects and show that our fusion improves object segmentation by 70% and in turn estimated kinematic joints by 26% over a purely motion-based approach. Furthermore, we show the necessity of probabilistic modeling for downstream robotic tasks, achieving 339% of the performance of a recent multimodal but deterministic RNN for object segmentation on the estimation of kinematic structure.**

I. Introduction

Robots act upon objects. But visually segmenting a scene into objects is challenging due to ambiguities, as can be seen in Fig. 1. Approaches for object segmentation usually focus on accuracy [1], [2], but robotic behavior requires perception to be more than just accurate. It needs to (A) be robust, maintaining accuracy even under adverse conditions; (B) enable reasoning about uncertainty; and (C) be real-time capable. We present a method for object segmentation that fulfills these requirements.

To provide robust object segmentation (A), we resolve ambiguities in visual input by fusing information from two complementary cues. Parts of objects often move together, and parts of objects often look alike. These cues are complementary in the sense that their ambiguities usually differ, as depicted in Fig. 1. Motion cannot disambiguate similarly moving objects nor non-moving ones. Appearance cannot disambiguate similarly colored objects and oversegments textured ones. But we can fuse both cues to make object segmentation more robust to either kind of ambiguity.

To capture cues in motion and appearance, the perceptual system we propose is split into two estimators (Fig. 2). One estimates object segmentation from motion, and the other from appearance. Both estimators are interconnected, which allows them to jointly resolve ambiguities.

But even such complementary cues cannot alleviate all uncertainty. Robots may further act on uncertainty (B) by exploration or by acting cautiously, which benefits from an explicit, *probabilistic* representation. Because both cues suffer from different ambiguities and types of noise, each estimator benefits from a tailored belief distribution. Subsequently, performing Bayesian inference to fuse these beliefs yields a robust, disambiguated *object segmentation belief*.

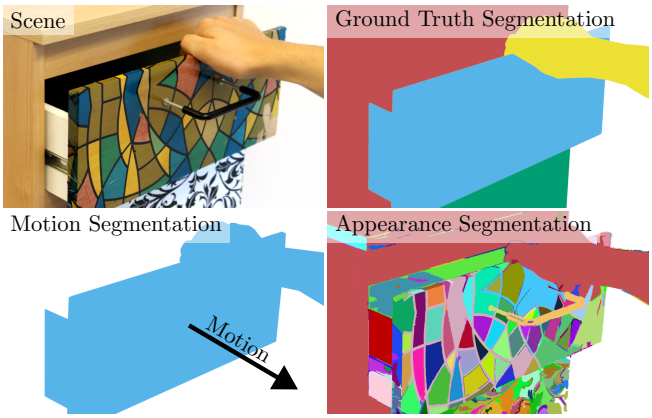

Fig. 1. Motion and appearance individually provide ambiguous information about objects in the scene: motion does not provide any information about non-moving parts, while appearance segments similarly colored objects and oversegments the textured areas. As a side effect, the hand is grouped wrongly. Because of these ambiguities, it is a challenge to segment the scene into objects such that the segmentation supports *robust* robot behavior.

We evaluate our approach in the context of kinematic structure estimation on a dataset [3] of interactions with articulated mechanisms. We compare against two other object segmentation methods and show that our fusion makes object segmentation more robust (A), scoring 170% of a purely motion-based approach. Our probabilistic modeling also enables downstream reasoning about segmentation uncertainty (B), leading to 339% of the performance of a recent *deterministic* RNN [1] on the estimation of kinematic structure. Additionally, our system is real-time capable (C), averaging 14 FPS on a mid-range desktop computer.

II. Related Work

Our main goal is to robustly segment a scene into objects for robotic manipulation. Hence, we will describe in Section II-A how recent approaches use different cues for object segmentation and their limitations for robotic tasks. In Section II-B, we examine the importance of object segmentation for kinematic structure estimation, a downstream perception task used in robotic manipulation.

[1] Robotics and Biology Laboratory, Technische Universität Berlin
[2] Science of Intelligence (SCIoI), Cluster of Excellence, Berlin, Germany
We gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135.

A. Object Segmentation

To determine the object segmentation of a scene, we need to group its parts into distinct objects. We can represent such grouping as a dense image segmentation [4], [5] or as sparse clusters of points [6], [7]. To estimate this grouping, many approaches rely only on a single cue. Motion-based approaches leverage the common motion of object-parts [6], [8], appearance-based approaches leverage the visual homogeneity of objects [4], [5], [9], and shape-based approaches leverage the closed surfaces of objects [10], [11]. However, such approaches group objects inadequately when the employed cue is ambiguous. Textured objects are often over-segmented, shapes can be irregular, and objects may be static. To address these ambiguities, recent approaches fuse several [1], [2], [12], [13], [14] or all [15] of these cues. We follow a similar strategy by fusing appearance and motion.

Although fusion approaches improve object segmentation beyond single cue methods, they often do not fit the requirements of robotic tasks. Some lack real-time capability [1], [12] or compute the segmentation newly each frame [13], [14], forgoing consistency. Others consider only predefined classes [15], focus only on the most salient object [12], or do not distinguish between different moving objects [2]. Lastly, they usually do not estimate uncertainty [1], [2], [12], [13], [14], [15], preventing its consideration in downstream tasks. In contrast, our method provides *consistent* estimates of object segmentation *and* uncertainty in *real time*, based on a probabilistic model factorized into two estimators.

B. Kinematic Structure Estimation

We apply our method to the task of kinematic structure estimation (KSE), where the goal is to estimate kinematic joints between objects. Recent approaches estimate these joints by analyzing objects' trajectories [6], [16], their 3D-segments over time [17], [18], [19], or by using learned semantic knowledge about the kinematics of visually identified objects [20], [21]. Since all these approaches center around objects, they require an object segmentation of the scene. Some rely directly on ground truth data [20], [21], while others either use only motion-based objects [6], [18], [19] or semantic knowledge about certain classes [21]. We use our method to provide KSE with probabilistic object segmentation and extend the KSE approach presented in [6].

III. Leveraging Motion and Appearance for Probabilistic Object Segmentation

We present a method to robustly segment visual scenes into objects and estimate the segmentation uncertainty. We use objects' motion and appearance, processing each of these cues with a specialized recursive estimator, and interconnect them to fuse their information as shown in Fig. 2. This leverages the cues' complementarity to make the object segmentation more robust in both representations. We first explain the belief representations of both estimators in Section III-A and then how we update each based on observations and the other's current belief in Sections III-B and III-C.

A. Probabilistic Representations of Object Segmentation

We can segment a scene into objects based on motion and appearance, because parts of an object often move and appear similarly. But to measure and interpret these similarities efficiently, we need different methods and representations. We measure motion by tracking visual 3D feature points as in [6, sec. IV-A] and appearance using traditional color image segmentation [4]. Based on these measurements, we can represent an object segmentation both as a clustering of tracked points c_t and as a dense image segmentation s_t.

However, both motion and appearance can be ambiguous and noisy, leading to uncertain object segmentation. But we can estimate and reduce this perceptual uncertainty by recursively estimating a probabilistic belief over each representation, i.e., the clustering *and* the image segmentation.

To represent object segmentation as a probabilistic clustering of points $\text{bel}(c_t)$, we associate with each point i an assignment belief over possible object clusters $o_1, \dots, o_M$. This belief over the assigned object $o_\text{a}^{[i]}$ is a Dirichlet distribution[1]:

$$\text{bel}(c_t) = \left\{\text{bel}(o_\text{a}^{[i]}) := \text{Dir}(o_1, \dots, o_M)\right\}_{i \in \{1,\dots,I\}} \tag{1}$$

To represent object segmentation as a probabilistic image segmentation we use a particle representation. This is necessary as the space of possible segmentations is high dimensional and a belief within it can have multiple modes. Thus, our belief over the current segmentation s_t is represented by a set of particles as shown in eq. (2)[2], each consisting of one possible dense image segmentation $\hat{s}_t^{[n]}$, the pixel-level velocities of region boundaries $v_t^{[n]}$, and a weight $w_t^{[n]}$.

$$\text{bel}(s_t) = \left\{\hat{s}_t^{[n]} \in \mathbb{N}^{H \times W}, v_t^{[n]} \in \mathbb{R}^{H \times W \times 2}, w_t^{[n]} \in [0,1]\right\}_{n \in \{1,\dots,N\}} \tag{2}$$

In the following, we will describe how we jointly filter the belief representations and account for new observations.

B. Recursively Tracking the Probabilistic Clustering

We track the probabilistic clustering of points $\text{bel}(c_t)$ based on their motion. To bolster this tracking, we leverage the parallely estimated image segmentation belief $\text{bel}(s_t)$, as it provides complementary appearance-based information (Section III-C).

To initialize clusters prior to motion in the scene, we rely on the appearance-based segmentation belief $\text{bel}(s_t)$. We determine which points should be grouped according to each particle and then account for uncertainty by marginalizing over the particles. This gives us pairwise neighborhood likelihoods, which we use to find an initial clustering using [22].

We now recursively filter this clustering by incorporating the motion of tracked points: when a point moves similar to an object cluster's motion, we increase their assignment likelihood. That is, we compare the motion of each point i as $x_t^{[i]} \in \mathbb{R}^3$ and the motion of each cluster tracked as rigid

[1] This distribution is suitable as it is the conjugate prior to the categorically distributed motion similarity as we will see in the following section.

[2] where H and W are the height and width of the image

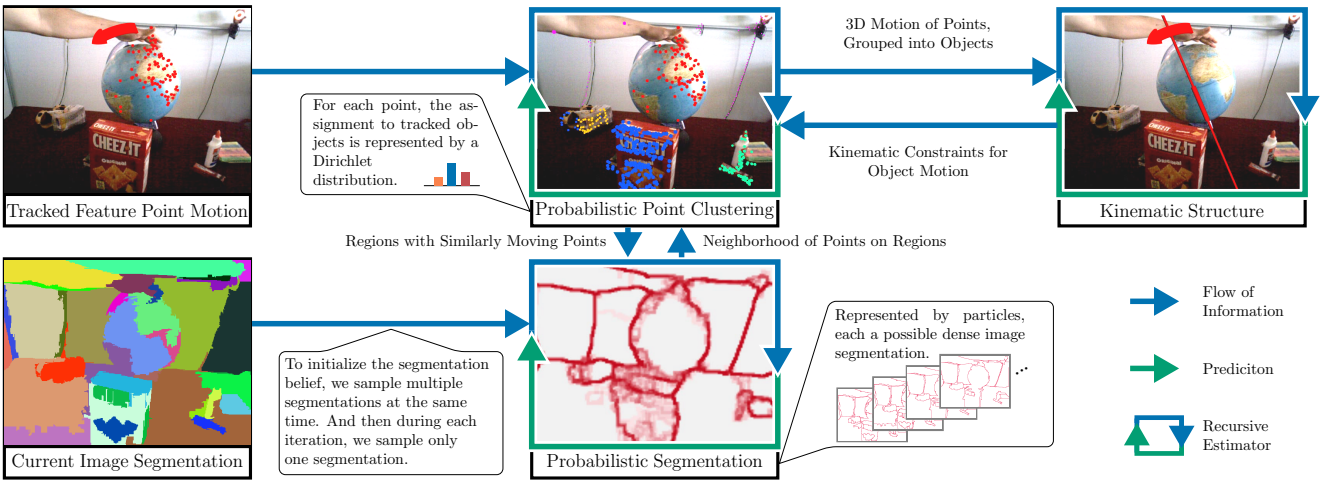

Fig. 2. Two interconnected recursive estimators (middle column) allow holistic fusion of motion and appearance cues. Each recursive estimator specializes in extracting object information from one cue, motion or appearance, probabilistically over time. By interconnecting these estimators, we allow them to benefit from each other: Information extracted from motion aids the interpretation of appearance, and vice versa. This scheme extends further as it provides consistent probabilistic estimates for downstream tasks such as kinematic structure estimation (right) in real time and can moreover leverage the higher-level information to again improve segmentation into objects.

body motion $H_t \in \mathrm{SE}(3)$ using the approach in [6, sec. IV-B] extended to take assignment uncertainty into account. By assuming an isotropic Gaussian distribution around each point, we determine how likely that point moves together with each object cluster. We normalize these likelihoods over objects, which results in categorical assignment likelihoods $p(o_{\mathrm{a}}^{[i]} | x_t^{[i]}, H_t^{[o_1]}, \ldots, H_t^{[o_M]})$ for each point i that we use to directly update its Dirichlet distributed assignment belief:

$$\mathrm{bel}(o_{\mathrm{a},t}^{[i]}) = \mathrm{bel}(o_{\mathrm{a},t-1}^{[i]}) + p(o_{\mathrm{a},t}^{[i]} | x_t^{[i]}, H_t^{[o_1]}, \ldots, H_t^{[o_M]}) \tag{3}$$

But motion does not provide reliable information for all points, because some do not move and others are tracked with high noise. However, the appearance-based image segmentation belief $\mathrm{bel}(s_t)$ is not affected by these problems. We can use it to identify neighboring points and use their assignment beliefs as additional measurements to reduce uncertainty in motion-based clustering. To account for the uncertainty of $\mathrm{bel}(s_t)$, we again marginalize over the particle set to obtain neighborhood likelihoods.

By incorporating both the tracked motion as well as the appearance-based segmentation belief we can track the clustering into objects robustly. However, initial clusters created solely based on appearance may not suffice when motion reveals new objects. In such cases, we use RANSAC to identify new clusters of points that obey a rigid body motion but cannot be assigned to any existing cluster.

C. Recursively Tracking the Probabilistic Segmentation

We track the probabilistic image segmentation $\mathrm{bel}(s_t)$ by recursively incorporating appearance-based image segmentations [4] as measurements s_t^*. To furthermore overcome ambiguities and reduce uncertainty, we leverage the current clustering belief $\mathrm{bel}(c_t)$, as it provides complementary motion-based information. For $\mathrm{bel}(s_t)$ we use a particle representation, as described in Section III-A. Thus, we design a particle filter over dense image segmentation.

In the forward model of the filter, we account for motion in image space. For this purpose, we estimate the velocities $v_t^{[n]}$ of boundary pixels between objects and use them to predict how the segmentation boundaries move in each particle. We estimate the velocities based on the closest measured region boundary and add sampled noise to disperse the particles.

To then weight the particles, we compare them for similarity to both the measured image segmentation s_t^* and the current clustering belief $\mathrm{bel}(c_t)$. The comparison to s_t^* requires a distance metric between segmentations $\mathrm{d}(s_1, s_2)$. As such metric, we use the sum of distances of each boundary pixel in one segmentation to the closest in the other. To now also compare the particles to $\mathrm{bel}(c_t)$, we cluster the points based on each segmentation particle $\hat{s}_t^{[n]}$. For each of these clusterings, we then determine its likelihood $p(\hat{s}_t^{[n]} | \mathrm{bel}(c_t))$ according to the clustering belief and weight the particle set according to eq. (4), with η as normalization factor.

$$w_t^{[n]} = \frac{1}{\eta} \frac{p(\hat{s}_t^{[n]} | \mathrm{bel}(c_t))}{\mathrm{d}(s_t^*, \hat{s}_t^{[n]})} \tag{4}$$

High dimensionality in the space of segmentations would require impractically many particles. To cope with this, we adapt each particle to increase its likelihood given the clustering belief $\mathrm{bel}(c_t)$ and the measured segmentation s_t^*. We adapt particles to be closer to $\mathrm{bel}(c_t)$ by merging regions when they contain points that are likely in the same cluster. We adapt particles to be closer to s_t^* by merging regions if they are jointly covered by one region in s_t^*. We also directly insert regions from s_t^* into particles. Each iteration, we apply these operations only to a few randomly chosen particles, but they sufficiently regularize the belief so that we only need $N = 50$ particles, enabling real-time estimation.

IV. Leveraging Probabilistic Object Segmentation in Kinematic Structure Estimation

The probabilistic object segmentation we derived in Section III is useful for downstream perception tasks. One such task, is Kinematic Structure Estimation (KSE) which finds kinematic relationships between objects in the scene.

We apply our method to robustly segment the scene into objects in KSE by extending the OMIP system [6]. As KSE requires a segmentation of the scene into objects (Section II-B), OMIP already has a clustering of tracked points into objects, but only motion-based and deterministic. We replace this clustering with our two estimators and hence make it more robust. This in turn also makes the overall KSE system more robust, as we will see in our experiments.

However, we can formulate this connection to KSE again as an *interconnection*, as shown in Fig. 2. The estimated kinematic structure provides constraints to the estimation of object motion, e.g. restricting it along an estimated prismatic axis. These constraints make object tracking more robust, as shown in [6]. Thereby, applying our object segmentation to KSE in turn also improves our object segmentation.

V. Experiments

We assess the performance of our estimators for object segmentation and their impact on kinematic structure estimation using a dataset of articulated objects. In this section, we outline our experimental setup and evaluation metrics (Section V-A), demonstrate how fusing motion and appearance information leads to better object segmentation (Section V-B), and show how our recursive estimators meet the needs of real-world robotic applications on the task of kinematic structure estimation (Section V-C).

A. Experimental Setup and Metrics

We restrict our evaluation to the RBO dataset of articulated objects and interactions [3] as it is the only dataset that provides ground truth for both object segmentation and kinematic structure in real-world RGB-D sequences. In the following, we explain how we derive ground truth from this dataset, introduce evaluation metrics, and discuss the baseline methods we use for comparison.

Deriving Ground Truth: The dataset [3] includes 3D shape models and corresponding 6D poses for each object. Using this information, we project the objects onto the image surface to create a "ground truth" object segmentation that we can directly compare to the segmentation obtained from motion and/or appearance. We use the kinematic joint information provided by the dataset as is. However, we exclude two objects from the evaluation—*pliers* and *foldingrule*—because they have inaccurate shape models, which prevents us from generating appropriate ground truth data.

Metric for Object Segmentation: To compare the different object segmentation approaches on segmentation accuracy, we define an objectness score O for the clustering of each ground truth object $\mathbb{O}_{\text{gt}}$ in eq. (5), which uses the Jaccard index as the similarity measure S_{o} between two objects represented as sets of points $\mathbb{O}_1$ and $\mathbb{O}_2$.

$$O(\mathbb{O}_{\text{gt}}, \text{bel}) = \max_{\mathbb{O}_{\text{bel}} \in \text{bel}} S_{\text{o}}(\mathbb{O}_{\text{bel}}, \mathbb{O}_{\text{gt}}) \tag{5}$$

$$K(j_{\text{gt}}, \text{bel}) = D(j_{\text{gt}}, \text{bel}) \cdot Q(j_{\text{gt}}, \text{bel}) \tag{6}$$

$$D(j_{\text{gt}}, \text{bel}) = \sum_{j_{\text{bel}} \in \text{bel}} p(\text{type}(j_{\text{bel}}) = \text{type}(j_{\text{gt}})) p_p(j_{\text{bel}}, j_{\text{gt}}) \tag{7}$$

$$Q(j_{\text{gt}}, \text{bel}) = \sum_{j_{\text{bel}} \in \text{bel}} \frac{S_{\text{a}}(j_{\text{bel}}, j_{\text{gt}}) + S_p(j_{\text{bel}}, j_{\text{gt}}) + S_{\text{s}}(j_{\text{bel}}, j_{\text{gt}})}{3} \cdot p_{\text{p}}(j_{\text{bel}}, j_{\text{gt}}) \tag{8}$$

$$S_{\text{o}}(\mathbb{O}_1, \mathbb{O}_2) = \frac{|\mathbb{O}_1 \cap \mathbb{O}_2|}{|\mathbb{O}_1 \cup \mathbb{O}_2|} \tag{9}$$

$$S_{\text{a}}(j_1, j_2) = 1 - \frac{\alpha(j_1, j_2)}{90^\circ} \tag{10}$$

$$S_{\text{p}}(j_1, j_2) = \frac{1}{1 + \text{dist}(j_1, j_2)} \tag{11}$$

$$S_{\text{s}}(j_1, j_2) = \frac{1}{1 + \left| |\text{state}(j_1)| - |\text{state}(j_2)| \right|} \tag{12}$$

$$p_p(j_1, j_2) = \frac{1}{\eta} \max\{S_{\text{o}}(\mathbb{O}_{j_1,1}, \mathbb{O}_{j_2,1}) \cdot S_{\text{o}}(\mathbb{O}_{j_1,2}, \mathbb{O}_{j_2,2}), S_{\text{o}}(\mathbb{O}_{j_1,2}, \mathbb{O}_{j_2,1}) \cdot S_{\text{o}}(\mathbb{O}_{j_1,1}, \mathbb{O}_{j_2,2})\} \tag{13}$$

SET OF EQUATIONS I
Evaluation Metrics used in our Experiments

Metrics for Kinematic Joint Evaluation: To compare the impact of different object segmentation approaches on the estimated kinematic structure, we compute kinematic joints for the different point clusterings using the joint tracker of OMIP [6, sec. IV-C]. We then compare the estimated joints to the ground truth joints of the dataset regarding the detection and parameter reconstruction of each joint with kinematic joint score $K(j_{\text{gt}}, \text{bel})$, defined in eq. (6).

The kinematic joint score should only score high if the joint is both *detected* and *reconstructed* well. Thus, it is the product of a score for detection D, defined in eq. (7) and a score for reconstruction quality Q, defined in eq. (8). They measure how well each ground truth joint j_{gt} is detected based on the estimated likelihood $p(\text{type}(j) = T)$ for a joint of the correct type T, and the similarity of the estimated joint parameters to the ground truth, respectively.

To compute this similarity between joint parameters, we consider all parameters of prismatic and revolute joints: the axis orientation depending on the angle α between their axes, the axis position depending on the shortest distance $\text{dist}(j_1, j_2)$ between their axes[3], and the articulation state depending on the difference in observed motion $|\text{state}(j)|$.

Finally, we combine above metrics to evaluate the kinematic joints. For this, we obtain the normalized[4] likelihood $p_{\text{p}}(j_{\text{bel}}, j_{\text{gt}})$ according to eq. (13).

Baseline Methods: We compare our object segmentation and kinematic structure estimation against two other methods. One is the purely motion-based technique of OMIP [6], to show how leveraging both motion and appearance makes estimation more robust. The other is a

[3] All parallel prismatic joint axes are equally valid. Hence, we always assume correct axis position for them ($S_p = 1$).

[4] The normalization over all the estimated object pairs via the factor η makes the joint evaluation indifferent to the object segmentation quality, only focusing on the estimated joints.

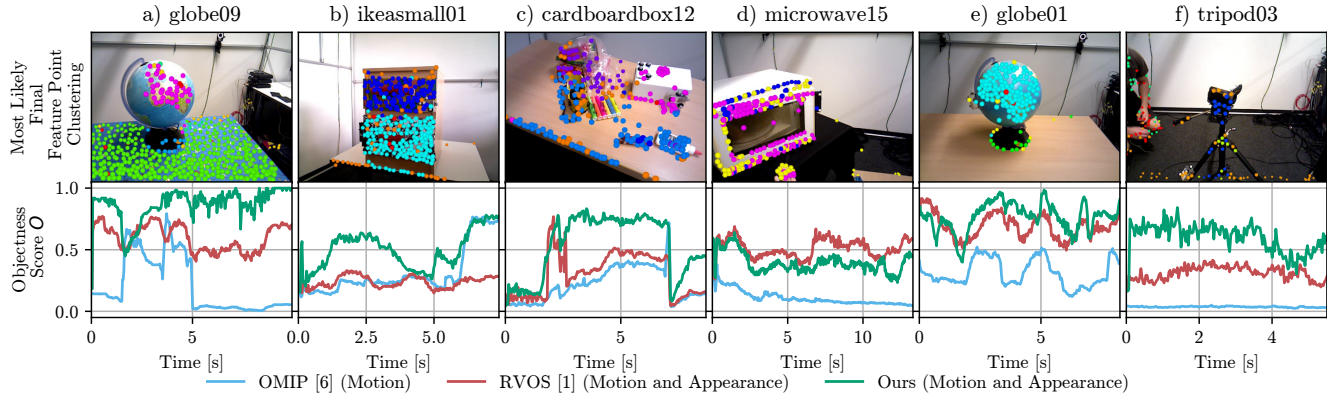

Fig. 3. Leveraging motion and appearance leads to better and robust clustering of objects. We show this by comparing motion-only OMIP [6], our fusion of motion & appearance, and RVOS [1] on six exemplary sequences. The clusters, indicated by overlaid points in the top row, stay consistent even when the tracked feature points have noisy motion or are partly occluded. Compared to purely motion-based estimation, this leads to continued tracking, as in a) after $t = 5\,\text{sec}$, and better fitting clusters, as in b)-e). Furthermore, our initial clustering from appearance allows us to track objects, that would otherwise be missed, as in f).

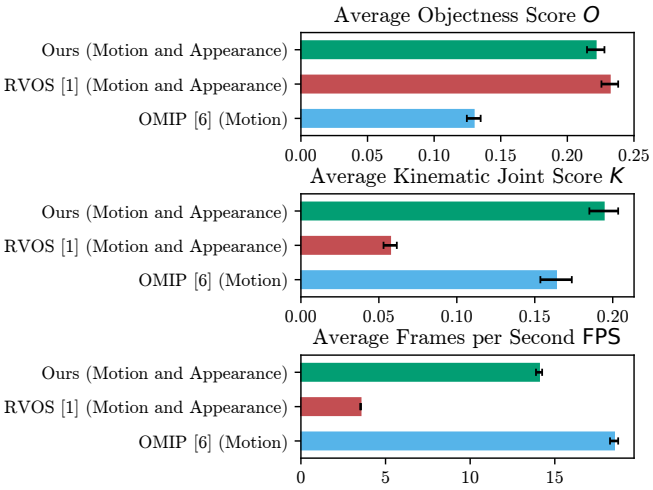

Fig. 4. Our method outperforms others on the RBO dataset of articulated objects and interactions [3]: Fusion of information from motion and appearance leads to more robust segmentation into objects over the motion-based estimation of OMIP [6]. The object segmentation of our interconnected recursive estimators and RVOS [1] are comparable, but our estimators fit the needs of real-world robotic tasks better by providing more consistent and probabilistic estimates. This leads to much better performance on the task of kinematic structure estimation, while also being real-time capable.

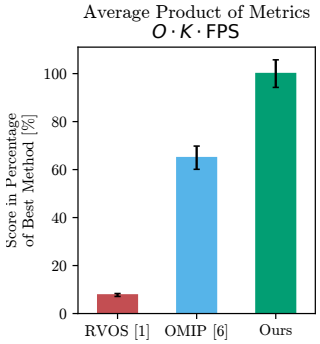

Fig. 5. We deem jointly high performance in all three measures as required: segmentation quality, downstream performance, and computation speed. Thus we compare the product of the metrics in Fig. 4 averaged over the dataset. Our *probabilistic* method achieves superior performance to OMIP [6], the *unimodal* approach that we extend, and both clearly outperform the *deterministic* RNN RVOS [1].

state-of-the-art multimodal RNN, called RVOS [1], which also leverages motion and appearance, but does not provide explicitly probabilistic estimates.

B. Fusion of Motion and Appearance Leads to Better Object Segmentation

Our method fuses information from motion and appearance to segment objects. In order to analyze the effects of this fusion, we first compare against the motion-based object segmentation of OMIP on some exemplary sequences (Fig. 3). Our method improves over the motion-based segmentation in two ways: Fusion leads to more consistent estimates, and appearance allows for better interpretation of the motion in the scene. We will now further elaborate on both.

Fusing motion and appearance information improves point clustering consistency, as appearance allows us to correctly assign newly detected points and points with ambiguous motion. This can be seen in Fig. 3a, where our method maintains the object throughout the sequence while motion-based OMIP loses it after $t = 5\,\text{sec}$. Similarly, the correct assignment of newly detected points allows us to outperform the motion-based method on most sequences (Fig. 3b-e).

The appearance based information allows for better interpretation of the motion in the scene, because we can immediately cluster and thus track objects from the beginning. We can see this effect in Fig. 3a, d, e, but especially in Fig. 3f with the black tripod. Here, the appearance-based initial clustering plays a crucial role as point tracking for the uniformly black parts of the tripod would be too uncertain. On the flip-side, this initial clustering using appearance may be over-segmented as in Fig. 3b and c. However, when the respective object moves the over-segmentations collapse.

These benefits of fusion also generalize over the dataset, where our method scores on average 70% higher than motion-based OMIP (Fig. 4). As this improvement stems from the fusion of motion and appearance, the multimodal RVOS performs similar to our method on object segmentation accuracy. However, our probabilistic estimates are very

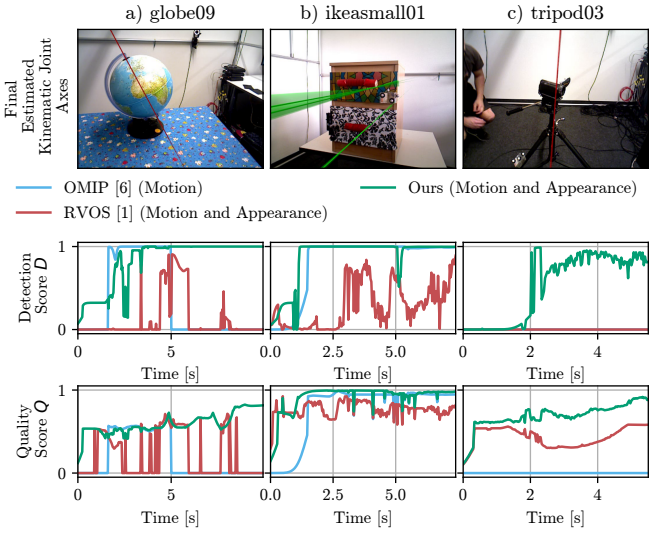

Fig. 6. The probabilistic object segmentation of our method in turn makes kinematic joint estimation more robust. We show this by comparing our approach to RVOS [1] and motion-only OMIP [6] on three exemplary sequences. The kinematic joints, indicated by their joint axes in the top row, are tracked more consistently than only based on motion, as in a). Moreover, the *probabilistic* estimate of our method allows for better joints than *deterministic* RVOS, as in a) and b). Lastly, our fusion of motion and appearance enable good estimation of previously undetected joints, as in c).

useful for other robotic tasks such as kinematic structure estimation, as we describe below.

C. Our Interconnected Recursive Estimators Fit the Needs of Real-World Robotic Tasks

We have shown in the previous subsection that our interconnected recursive estimators provide a segmentation into objects on par to state-of-the-art multimodal approaches like RVOS [1]. In the following, we show that our method is much better suited to real-world robotic tasks, because the provided estimates are more consistent and include a measure of uncertainty. We first illustrate this on exemplary sequences (Fig. 6) and then report the aggregated results.

Our probabilistic, fused object segmentation leads to better estimates of kinematic joints, because it is more consistent. This consistency lets us not only maintain objects longer (Fig. 3a), but thereby also the associated joints (Fig. 6a). Furthermore, the fusion of motion and appearance allows for the detection and reconstruction of previously undiscovered joints, as shown in Fig. 6c. The non-probabilistic object segmentation of RVOS is often unable to consistently track objects over the *entire* trajectory. This in turn leads to inconsistent kinematic joint estimates. See plots in Fig. 6a-c.

Thus, the better consistency and robustness of our probabilistic estimates also lead on average to an improvement in kinematic joint score of 19% over OMIP and 239% over RVOS (Fig. 4). This demonstrates that probabilistic modeling is key, when using object segmentation in downstream tasks. But our interconnected estimators have another advantage compared to RNN-based RVOS: They meet the requirement of real-time capability, running on average at 14 FPS on a mid-range desktop computer, where RVOS runs at 3.5 FPS.

Finally, if we assume the following characteristics are equally important for a given task, our method outperforms both OMIP and RVOS: object segmentation quality, downstream performance, and computation speed. We demonstrate this by comparing the product of our three performance metrics—objectness score, kinematic joint score, and frames per second—in Fig. 5.

VI. Limitations

In our experiments, we observe that our approach is limited by its measurement sources and modeling assumptions for object motion. We elaborate on this below:

Reliance on Image Segmentation: The appearance based estimation relies on image segmentations as measurements. If these measurements are *consistently* wrong, this can lead to overconfidence in an over- or undersegmentation. As our appearance-based estimator is agnostic to the chosen segmentation algorithm, this could be improved upon with a better segmentation technique, e.g. CNN-based [9], or a different input image, e.g. the current depth image.

Reliance on Feature Point Tracking: The basic unit for all motion-based estimation of our approach—clustering, object motion, kinematic joints—are the tracked feature points. If their tracking fails, our algorithm can thus not extract sufficient information from motion. Even when appearance-based estimation of the object segmentation is precise for textureless situations, we can still not track the object and fail to reconstruct its kinematic joints. We could improve on this reliance on feature points by introducing more basic visual features to measure motion in the scene, e.g. regions [23], contours [24], or dense optical flow [25].

Limitations from Modeling Assumptions: Our approach assumes rigid body motion and prismatic/revolute kinematic structure, which limits its effectiveness for deformable objects or other joint types. Although our probabilistic model can handle minor deviations, the motion-based estimation may not perform well in such cases.

VII. Conclusion

We presented an approach to segment scenes into objects. For such a perception method, robotic behavior poses three requirements: (A) It needs to be *robust* against adverse conditions, (B) enable reasoning about *uncertainty*, and (C) be *real-time* capable. Our approach is robust due to fusion of complementary information from motion and appearance (A), explicitly represents uncertainty over the object segmentation (B), and runs at 14 FPS on mid-range hardware (C).

In our experiments, we confirmed the necessity of robustness (A) and reasoning about uncertainty (B) on the downstream task of kinematic structure estimation, where we clearly outperform a recent *deterministic* RNN. Similarly, our approach could support other object-centric tasks by guiding robotic interactions with the environment and facilitating exploration or cautious behavior in uncertain conditions.

REFERENCES

[1] C. Ventura, M. Bellver, A. Girbau, A. Salvador, F. Marques, and X. Giro-i-Nieto, "RVOS: End-to-end recurrent network for video object segmentation," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5277–5286, 2019.

[2] T. Zhou, S. Wang, Y. Zhou, Y. Yao, J. Li, and L. Shao, "Motion-Attentive Transition for Zero-Shot Video Object Segmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 13 066–13 073, 2020.

[3] R. Martín-Martín, C. Eppner, and O. Brock, "The RBO Dataset of Articulated Objects and Interactions," *The International Journal of Robotics Research*, vol. 38, no. 9, pp. 1013–1019, 2019.

[4] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient Graph-Based Image Segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[5] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour Detection and Hierarchical Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 898–916, 2011.

[6] R. Martin Martin and O. Brock, "Online interactive perception of articulated objects with multi-level recursive estimation based on task-specific priors," *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2494–2501, 2014.

[7] E. Elhamifar and R. Vidal, "Sparse subspace clustering," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2790–2797, 2009.

[8] P. Bideau and E. Learned-Miller, "It's Moving! A Probabilistic Model for Causal Motion Segmentation in Moving Camera Videos," *European Conference on Computer Vision*, pp. 433–449, 2016.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[10] T. Rabbani, F. van den Heuvel, and G. Vosselman, "Segmentation of point clouds using smoothness constraints," *ISPRS commission V symposium: image engineering and vision metrology*, pp. 248–253, 2006.

[11] M. Danielczuk, M. Matl, S. Gupta, A. Li, A. Lee, J. Mahler, and K. Goldberg, "Segmenting Unknown 3D Objects from Real Depth Images using Mask R-CNN Trained on Synthetic Data," *2019 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7283–7290, 2019.

[12] S. Mahadevan, A. Athar, A. Ošep, S. Hennen, L. Leal-Taixé, and B. Leibe, "Making a Case for 3D Convolutions for Object Segmentation in Videos," *31st British Machine Vision Conference 2020 (BMVC)*, pp. 1–14, 2020.

[13] A. Valada, R. Mohan, and W. Burgard, "Self-Supervised Model Adaptation for Multimodal Semantic Segmentation," *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1239–1285, 2020.

[14] W. Wang and U. Neumann, "Depth-Aware CNN for RGB-D Segmentation," *European Conference on Computer Vision (ECCV) 2018*, vol. 11215, pp. 144–161, 2018.

[15] H. Rashed, A. El Sallab, S. Yogamani, and M. ElHelw, "Motion and Depth Augmented Semantic Segmentation for Autonomous Navigation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 364–370, 2019.

[16] K. Hausman, S. Niekum, S. Osentoski, and G. S. Sukhatme, "Active articulation model estimation through interactive perception," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 3305–3312.

[17] A. Jain, R. Lioutikov, C. Chuck, and S. Niekum, "ScrewNet: Category-Independent Articulation Model Estimation From Depth Images Using Screw Theory," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13 670–13 677, 2021.

[18] R. Staszak, M. Molska, K. Mlodzikowski, J. Ataman, and D. Belter, "Kinematic Structures Estimation on the RGB-D Images," *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, pp. 675–681, 2020.

[19] X. Wang, B. Zhou, Y. Shi, X. Chen, Q. Zhao, and K. Xu, "Shape2motion: Joint analysis of motion parts and attributes from 3d shapes," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8876–8884, 2019.

[20] V. Zeng, T. E. Lee, J. Liang, and O. Kroemer, "Visual identification of articulated object parts," *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2443–2450, 2021.

[21] X. Li, H. Wang, L. Yi, L. J. Guibas, A. L. Abbott, and S. Song, "Category-level articulated object pose estimation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3706–3715, 2020.

[22] M. Ester, H.-P. Kriegel, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.

[23] H. Tjaden, U. Schwanecke, E. Schömer, and D. Cremers, "A region-based gauss-newton approach to real-time monocular multiple object tracking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1797–1812, 2018.

[24] A. Yilmaz, X. Li, and M. Shah, "Contour-based object tracking with occlusion handling in video acquired using mobile cameras," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 26, no. 11, pp. 1531–1536, 2004.

[25] K. Kale, S. Pawar, and P. Dhulekar, "Moving object tracking using optical flow and motion vector estimation," *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, pp. 1–6, 2015.