RigidFusion: Robot Localisation and Mapping in Environments with Large Dynamic Rigid Objects

Citation for published version:

Digital Object Identifier (DOI):
10.1109/LRA.2021.3066375

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
IEEE Robotics and Automation Letters

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
RigidFusion: Robot Localisation and Mapping in Environments with Large Dynamic Rigid Objects

Ran Long1,∗, Christian Rauch1, Tianwei Zhang2, Vladimir Ivan1, Sethu Vijayakumar1,3,∗

Abstract—This work presents a novel RGB-D SLAM approach to simultaneously segment, track and reconstruct the static background and large dynamic rigid objects that can occlude major portions of the camera view. Previous approaches treat dynamic parts of a scene as outliers and are thus limited to a small amount of changes in the scene, or rely on prior information for all objects in the scene to enable robust camera tracking. Here, we propose to treat all dynamic parts as one rigid body and simultaneously segment and track both static and dynamic components. We, therefore, enable simultaneous localisation and reconstruction of both the static background and rigid dynamic components in environments where dynamic objects cause large occlusion.

We evaluate our approach on multiple challenging scenes with large dynamic occlusion. The evaluation demonstrates that our approach achieves better motion segmentation, localisation and mapping without requiring prior knowledge of the dynamic object’s shape and appearance.

I. INTRODUCTION

Mobile manipulation tasks, such as handling and transporting objects in an unmanned warehouse or collaborative manipulation [1], require a robot to localise against the static environment in which it moves while being robust to distractions from dynamic objects; as well as track the object which they manipulate. While these two problems have been previously addressed separately, only a few strands of work [2], [3] have attempted to solve these two problems together and track the camera and multiple objects at once.

In this work, we argue that localisation against the environment and tracking of objects are fundamentally the same problem, and that solving them simultaneously reduces ambiguity about the scene and improves localisation in cases of large dynamic occlusions.

The core problem – separating the scene into segments of transformations induced by ego-motion and/or moving objects – is challenging due to several factors:

1) **Unknown environments:** Robots may not have prior information about the semantic meaning, 3D model or appearance of the dynamic objects and static background.

2) **Large occlusion:** Dynamic parts of images are often discarded for robust visual odometry; but in many scenarios, they can occlude the majority of a camera view, such as for large moving objects or when manipulated objects are close to the camera. This ambiguity leads to tracking failures where a dynamic object is classified as part of the static environment. This is in contrast to driving/flying, where ego-motion effects dominate.

3) **Mutual static and dynamic transition:** Manipulated objects can transition between static and dynamic with respect to the world at any time during manipulation. Maintaining these state transitions purely with visual odometry can be difficult.

To address all three aspects concurrently, we treat localisation and object tracking as an integrated problem and formalise both as modelling and tracking any rigid movement. Consequently, we achieve improved motion segmentation, localisation and mapping in dynamic environments with large occlusion (Figure 1). For this, we assume that the motions of both static and dynamic components are rigid transforma-
tions. These motions can be identified using tightly coupled motion priors from odometry and kinematics.

In summary, this work contributes:

1) a new pipeline to simultaneously segment, track and reconstruct the static background and one dynamic rigid body from RGB-D sequences, using motion priors with potential drift,
2) a dense SLAM method that is robust to large occlusions in the visual input (over 65%) without relying on an initialisation of the static and dynamic models.
3) a new RGB-D SLAM dataset\(^1\) with dynamic objects that cause large occlusion in the scenes and ground truth trajectories.

II. RELATED WORK

Dynamic visual SLAM or visual odometry methods can be categorised into direct, indirect or multi-motion odometry methods. Robot proprioception can also be used to support localisation in dynamic environments.

Direct methods rely on prior information of static background or dynamic objects to distinguish between them. For certain dynamic objects, such as humans, PoseFusion [5] used OpenPose [6] to segment them against the environment. In addition, multi-object segmentation methods, such as Mask-RCNN [7], can provide accurate semantic segmentation, therefore supporting robot localisation when dynamic objects are included in the training set [8], [9]. Given object segmentation, different objects can be further assigned with different scores to indicate their probability to be dynamic [10]. However, a trained network can only recognise objects from the training set, and even if an object is recognised as static, the object can become dynamic if it is manipulated. Another strand of research distinguishes the static background through geometric properties, such as assuming all planes are static [11]. This would fail when objects that consist of planes, like boxes, are manipulated.

Indirect methods track the main rigid component of the visual input and discard the remaining components as outliers. Sun et al. [12] applied a RANSAC approach to sparse feature points of two consecutive images, and dynamic objects are removed as outliers. This work was later extended to scenarios with multi-cluster dynamics [11] and served as a pre-processing step for the input of SLAM algorithms. Li et al. [13] proposed a static pixel/point weighting method to represent the probability of a point being static, instead of classifying each point as either absolutely static or dynamic. Both StaticFusion (SF) [4] and Joint-VO-SF (JF) [14] applied a K-Means clustering method to separate the visual input into clusters and similarly assigned static pixel/point weights to each cluster. The dynamic clusters are detected as outliers and SF requires that dynamic components are less than 20-30% at the initial frame [4]. Rather than removing all outliers, Co-Fusion (CF) [2] treated outliers as an additional object and maintained the model of this object if the outliers are connected and occupy more than 3% of an image. While it maintains multiple objects, it is prone to over-segment the image, thus treating different parts of an object with the same transformation as different entities.

Multi-motion odometry methods, such as multi-body structure from motion (MBSfM) [15] and multimotion visual odometry (MVO) [3], directly separate and track multiple rigid bodies with distinct motions in the visual input. MBSfM requires all images in advance and cannot be processed online. MVO can also estimate the number of multiple moving objects online based on sparse feature points. However, it cannot provide dense mapping, and the rigid body with the largest number of feature points is treated as static. This means that a dynamic object could be recognised as static if it has a richer texture or occupies a larger part than static background in the visual input.

Robot proprioception can be combined with visual odometry to support localisation. Visual inertial odometry (VIO) methods [16], [17] fused IMUs and visual sensors. Wheel [18] or leg [19] odometry can be further combined with VIO to increase the accuracy of localisation. However, they are limited to static environments. Kim et al. [20] used an IMU to estimate and, thus, compensate camera ego-motion, therefore removing dynamic objects before camera tracking. However, this method relies on accurate robot proprioception.

In summary, state-of-the-art visual SLAM methods either 1) require full knowledge of objects in the scene (direct methods) and fail if the dynamic objects are not detected, or 2) cannot handle when dynamic objects become the predominant part of an image (indirect methods).

III. OVERVIEW

We propose a pipeline that treats the dynamic component as a single rigid body and uses motion priors to segment the static and dynamic components. The segmentation is used to track the camera, and to reconstruct the background and object models.

The overview of our pipeline is illustrated in Figure 2. Our approach takes two consecutive RGB-D frames A and B, static and dynamic motion priors, $\hat{\xi}_s, \hat{\xi}_d \in \mathbb{se}(3)$, and the previous segmentation of frame A, $\tilde{\Gamma}_A \in \mathbb{R}^{W \times H}$. Similar to [4], each new intensity and depth image pair $(I, D) \in \mathbb{R}^{W \times H}$ is over-segmented into K geometric clusters $V = \{ V_i | i = 1, \cdots, K \}$ by applying K-Means clustering [14]. We hypothesise that each cluster is as rigid as possible, and each rigid body can be approximated by the combination of clusters. We also assign each cluster a score $\gamma \in [0, 1]$ which represents the probability that a cluster belongs to the static rigid body: $\gamma = 0$ stands for dynamic clusters while $\gamma = 1$ means static clusters. For an RGB-D frame A, we denote the overall scores as $\gamma_A \in \mathbb{R}^K$.

If the difference between two motion priors $||\hat{\xi}_s - \hat{\xi}_d||_2$ is less than a threshold $\bar{d}$, we treat all clusters in the image as static and skip motion segmentation. Otherwise, we jointly optimise the scores $\gamma_B$ of the current frame and relative motions $\xi_s$ and $\xi_d$ of the static and dynamic rigid bodies (Section IV).

The pixel-wise segmentation $\Gamma_B \in \mathbb{R}^{W \times H}$ is then computed

---

\(^1\)http://conferences.inf.ed.ac.uk/rigidfusion/
Rigid Motion (b) clusters ... ∈ se(3). The projection function π : R3 → R2 projects 3D points onto the image plane using the camera intrinsic matrix.

The projection function π : R3 → R2 projects 3D points onto the image plane using the camera intrinsic matrix.

Considering two image frames A and B, the relation between T the camera and object poses are

\[ T = T_A T_B \]

using frame-to-model alignment (Section V).

and dynamic object and to refine the estimated camera pose using frame-to-model alignment (Section V).

We denote world-, camera-, and object-frames as F_W, F_C, F_O respectively (Figure 3). Similar to [18], we use \( T_{XY} \in SE(3) \) to transform homogeneous coordinates of a point in coordinate frame F_Y to F_X. In an image frame A, the camera and object poses are \( T_{WC_A} \) and \( T_{WO_A} \) respectively. Considering two image frames A and B, the relation between \( \xi_s \) and camera poses is: \( T(\xi_s) = T_{WC_A}^{-1} T_{WCB} = T_{CA} C_B \), and the relation between \( \xi_d \), camera and object poses is: \( T(\xi_d) = T_{WC_A} T_{WO_A} T_{WO_B} T_{WCB} = T_{CA} C_B C_D \). The motion priors \( \tilde{\xi}_s \) and \( \tilde{\xi}_d \) can be acquired from proprioceptive sensors, such as wheel odometry and arm forward kinematics.

In this paper, the static motion prior \( \tilde{\xi}_s \) is computed either from wheel odometry or by adding simulated drift on camera ground truth trajectories. We generate \( \tilde{\xi}_d \) by simulating drift on object ground truth trajectories.

IV. RIGID MOTION SEGMENTATION AND ESTIMATION

At the arrival of each RGB-D pair, we jointly segment and track both static and dynamic rigid bodies by minimising a combined cost that consists of four energy terms:

\[
\begin{align*}
\min_{\xi_s, \xi_d, \gamma} & \quad R(\xi_s, \gamma) + R(\xi_d, 1 - \gamma) + S(\xi_d, \gamma) + P(\xi_s, \xi_d) \\
\text{s.t.} & \quad \gamma \in [0, 1] \quad \forall i,
\end{align*}
\]

where \( \gamma \) represents the scores of all clusters. Specifically, the first two terms align the static and dynamic rigid bodies respectively. The third term \( S(\xi_d, \gamma) \) adds regularisation on both the spatial and time distribution of scores \( \gamma \) to maintain the smoothness of segmentation. The last term \( P(\xi_s, \xi_d) \) applies constraints on transformations \( \xi_s, \xi_d \) using motion priors \( \tilde{\xi}_s, \tilde{\xi}_d \).

from clusters and scores. Similar to StaticFusion, we compute the weighted RGB-D images of both static and dynamic rigid bodies from the segmentation \( \Gamma_B \). These weighted images are used to reconstruct models of the background and dynamic object and to refine the estimated camera pose using frame-to-model alignment (Section V).

A. Rigid Body Motion Estimation

Following previous RGB-D SLAM methods [21], [4], in static environments, the relative camera pose between two image frames A and B is estimated by minimising the intensity and depth residuals between the RGB-D image pairs \((I_A, D_A)\) and \((I_B, D_B)\). At a pixel \( p \) in frame A, the intensity residuals \( r^I_p \) and depth residuals \( r^D_p \) with respect to frame B are defined as:

\[
\begin{align*}
r^I_p & = I_B(W(x^I_p, T(\xi), D_A)) - I_A(x^I_p) \quad (2) \\
r^D_p & = D_B(W(x^D_p, T(\xi), D_A)) - |T(\xi)\pi^{-1}(x^D_p)| \quad (3)
\end{align*}
\]

where the image warping function \( W \) is given by:

\[
W(x^D, T, D) = \pi(T \pi^{-1}(x^D), D(x^D)) \quad (4)
\]

\( x^D \) represents the coordinate of pixel \( p \) in the 2D image, \(|\cdot|\) indicates the z-coordinate of a 3D point and \( D(x^D) \) is the depth of pixel \( p \). The homogeneous transformation matrix \( T(\xi) \in SE(3) \) is computed from its Lie algebra \( \xi \in se(3) \). The projection function \( \pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2 \) projects 3D points onto the image plane using the camera intrinsic matrix.
According to StaticFusion, given the scores \( \gamma \) of a rigid body, we can estimate the relative motion of this rigid body by applying the scores to weight residuals. Consequently, only pixels that belong to the rigid body have a high contribution:

\[
R(\xi, \gamma) = \sum_{p=1}^{N} \gamma_{i(p)} [C(\alpha_{p}w_{p}r_{p}^{2}(\xi)) + C(w_{p}r_{p}^{2}(\xi))] ,
\]

where \( N \) is the number of images pixels with valid depth reading in one image. \( i(p) \) indicates the index of the cluster that contains the pixel \( p \), and \( \gamma_{i(p)} \) represents the probability that this cluster belongs to the rigid body. \( \alpha_{p} \) is a scale parameter to weight photometric residuals so that they are comparable to depth residuals. The parameters \( w_{p} \) and \( w_{p} \) are computed according to the photometric and depth measurement noise. As in [4], we use the Cauchy robust penalty

\[
C(r) = \frac{c^{2}}{2} \log(1 + (\frac{c^{2}}{r}))
\]

to robustly control the minimisation of residuals, where \( c \) is the inflection point of \( C(r) \).

The novelty of our approach is that in equation 1, we treat the dynamic component as another rigid body with a different motion, where \( \gamma \) and \( 1 - \gamma \) represents the scores of the static and dynamic rigid body respectively. To simultaneously segment and track the two rigid bodies, we further encourage segmentation smoothness and use tightly coupled motion priors.

### B. Segmentation Smoothness

First, to maintain spatial smoothness, we use the regularisation term used in StaticFusion to penalise the score difference between adjacent clusters:

\[
S_{R}(\gamma) = \sum_{i=1}^{K} \sum_{j=i+1}^{K} E_{ij}(\gamma_{i} - \gamma_{j})^{2} ,
\]

where \( E_{ij} \) is the adjacent map for the cluster set \( V \). \( E_{ij} = 1 \) if clusters \( i \) and \( j \) are adjacent in space, otherwise \( E_{ij} = 0 \).

Furthermore, we consider the physical constraint that pixels that belong to the dynamic rigid body at the previous frame are likely to be dynamic at the current frame. Therefore, we use the segmentation result from the previous frame as segmentation prior to encourage segmentation smoothness over time:

\[
S_{T}(\xi_{d}, \gamma) = \sum_{i=1}^{K} (\gamma_{i} - \tilde{\gamma}(\xi_{d}))^{2} ,
\]

where \( \tilde{\gamma}(\xi_{d}) \) denotes the projection of \( \tilde{\gamma} \) from the previous frame \( B \) to the current frame \( A \) via \( \xi_{d} \):

\[
\tilde{\gamma}(\xi_{d}) = \sum_{\tilde{\gamma}_{i} \in V_{i}} \tilde{\Gamma}_{A}(W(\mathbf{x}_{\tilde{\gamma}_{i}}^{B}, T(\xi_{d})^{-1}, D_{B})) / |V_{i}|
\]

Here, \( V_{i} \) is the \( i \)-th cluster of the current frame \( B \) and \( \tilde{\Gamma}_{A} \) is the per-pixel segmentation from the previous frame \( A \). The warping function \( W \) (equation 4) transforms a pixel \( p \in V_{i} \) according to its coordinate in the current image \( \mathbf{x}_{\tilde{\gamma}_{i}}^{B} \) and the estimated motion of rigid body \( \xi_{d} \). \( |V_{i}| \) denotes the number of pixels in \( V_{i} \).

The spatial and time smoothness (equation 7 and 8) are combined and weighted by \( \lambda_{s} \) :

\[
S(\xi_{d}, \gamma) = \lambda_{s}(S_{R}(\gamma) + S_{T}(\xi_{d}, \gamma)) ,
\]

to represent the smoothness term \( S(\xi_{d}, \gamma) \) in equation 1.

### C. Tightly Coupled Motion Prior

Given the motion priors of both static and dynamic rigid bodies \( \tilde{\xi}_{s} \) and \( \tilde{\xi}_{d} \), we add a regularisation term on the motion of each rigid body:

\[
P(\tilde{\xi}_{s}, \tilde{\xi}_{d}) = \lambda_{s}||\tilde{\xi}_{s} - \tilde{\xi}_{d}||^{2}_{2} + \lambda_{d}||\tilde{\xi}_{d} - \hat{\xi}_{d}||^{2}_{2} ,
\]

where parameters \( \lambda_{s} \) and \( \lambda_{d} \) weight the regularisation terms. \( ||.||^{2}_{2} \) represents the square of the \( L_{2} \) norm. Because potential drift and noise in the motion prior could bias the solution, the prior information is neglected if the current estimated state is closer to the prior than a noise-related threshold. To achieve this, \( \lambda_{s} \) and \( \lambda_{d} \) are independently adapted online. Specifically, \( \lambda_{d} = 1 \) if \( ||\tilde{\xi}_{d} - \hat{\xi}_{d}|| > \hat{n} \), otherwise, \( \lambda_{d} = 0 \). \( \hat{n} \) is a threshold we choose and is related to the noise level of motion priors.

### D. Solver

The solver is based on StaticFusion. Since we directly align images in equation 1, the minimisation problem is solved via a coarse-to-fine scheme. We create an image pyramid for each image frame by iteratively down-sampling each image, which reduces the impact of depth noise. The optimisation starts from the coarsest level. The results of intermediate levels are used to initialise the following level.

For each level of the image pyramid, we decouple motions \( \tilde{\xi}_{s} \) and \( \tilde{\xi}_{d} \) from segmentation \( \gamma \). Specifically, at each iteration, we first fix \( \gamma \) and optimise \( R(\tilde{\xi}_{s}, \gamma) + R(\tilde{\xi}_{d}, 1 - \gamma) + P(\tilde{\xi}_{s}, \tilde{\xi}_{d}) \) over \( \tilde{\xi}_{s} \) and \( \tilde{\xi}_{d} \). Then \( \tilde{\xi}_{s} \) and \( \tilde{\xi}_{d} \) are fixed, and we optimise \( R(\tilde{\xi}_{s}, \gamma) + R(\tilde{\xi}_{d}, 1 - \gamma) + S(\tilde{\xi}_{d}, \gamma) \) over \( \gamma \).

### V. Mapping and Frame-to-Model Alignment

After the minimisation of equation 1, we use the optimal scores \( \gamma \) and \( 1 - \gamma \) to compute the weighted images for static and dynamic rigid bodies respectively. The weighted images are fused to the model of rigid bodies, and the estimated motions \( \tilde{\xi}_{s} \) and \( \tilde{\xi}_{d} \) are used to initialise the frame-to-model alignment. We use ElasticFusion without loop closure [21] to build the model and conduct frame-to-model alignment.

### VI. Evaluation

### A. Setup

The proposed method is evaluated on RGB-D sequences that are collected with an Asus Xtion PRO Live in plane-parallel movement (2 DoF translation, 1 DoF rotation) showing different characteristic object movements. The camera is either hand-held or mounted on an omnidirectional robot base (Figure 4a). The object is a remote controlled KUKA youBot with stacked boxes (Figure 4b). The camera and the object are equipped with Vicon markers for ground truth comparisons and to simulate motion prior drift for
camera-only sequences. The motion estimation performance is quantitatively evaluated via the absolute trajectory error (ATE) and the relative pose error (RPE) [22] against the Vicon ground truth for the optical frame. The visualised trajectories are aligned by the initial camera pose.

In the implementation of RF, we set $\lambda_r = 2$, and the thresholds $\hat{d}$ and $\hat{h}$ are both chosen as 0.01. We extend StaticFusion to use motion priors by appending the regularisation term $\lambda_s||\hat{\xi}_s - \hat{\xi}_s||^2$ to the loss function. The method that StaticFusion with ground truth camera motion prior is denoted as SF true. We control the impact of adding camera motion prior by choosing the same $\hat{h} = 0.01$ for SF true.

For camera-only sequences, the average simulated drift on camera trajectories is 6 cm/s (trans.) and 0.4 rad/s (rot.), while the average drift on object trajectories is 1.5 cm/s (trans.) and 0.1 rad/s (rot.). The camera and object speed is less than 60 cm/s. In robot experiments, we use wheel odometry as camera motion priors and keep the object motion prior by choosing the same $\hat{d} = 0.01$.

**TABLE I: Camera sequence description.**

<table>
<thead>
<tr>
<th>sequence</th>
<th>camera</th>
<th>frame motion</th>
<th>object</th>
<th>difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>straight</td>
<td>orbit</td>
<td>orthogonal crossing</td>
<td>orbit</td>
<td>straight</td>
</tr>
<tr>
<td>orbit</td>
<td>orbit</td>
<td>rotation to camera</td>
<td>orbit</td>
<td>orbit</td>
</tr>
<tr>
<td>takeover</td>
<td>orbit</td>
<td>lateral orbital crossing</td>
<td>orbital</td>
<td>orbital</td>
</tr>
</tbody>
</table>

**Fig. 4:** Omnidirectional platforms for moving (a) camera and (b) stacked boxes ($0.4 \times 0.6 \times 1$ m) with Vicon markers.

**B. Synthetic Experiments**

We hypothesise that the proposed objective with motion priors improves the estimation for dynamic objects that occupy more than 50% of valid image pixels. To study this effect in a controlled environment, we synthesised a simple scene with an object of varying size moving across the image from left to right. The relation of trajectory error to drift magnitude (Figure 5) supports this hypothesis.

**C. Camera Experiments**

We collected four sequences involving plane-parallel movement of the camera and the object within the camera frame. These sequences have different characteristics of camera and object motion (Table I). Figure 6 (top) shows the 2D plane projection of the true trajectories.

Our approach RigidFusion (RF) is compared against Joint-VO-SF (JF, [14]), StaticFusion (SF, [4]), StaticFusion with true motion priors (SF true) and Co-Fusion (CF, [2]). The quantitative evaluation in Table II shows that our method outperforms the state-of-the-art on difficult sequences. Although Co-Fusion achieves best results on the easier straight sequence, it tends to over-segment dynamic objects and treats parts of the static background as dynamic. This effect is more dominant in the more difficult sequences, leading to worsen performance of CF.

The visualisation of the estimated trajectories in Figure 6 (bottom) confirms that our method outperforms the state-of-the-art in dynamic scenes. The improved localisation performance stems from a better segmentation of dynamic parts in the image (Figure 7). In our frame-to-frame odometry setting, the improved motion segmentation performance directly affects the estimation performance and additionally leads to a better reconstruction of the static environment.

**D. Robot Experiments**

In four additional robotic experiments, we use the camera on the floating base of an omnidirectional robot and replace the simulated drift with wheel odometry. The true trajectories

**TABLE II: ATE and RPE for camera-only sequences.**

<table>
<thead>
<tr>
<th>RGB-D sequence</th>
<th>Motion prior (drift)</th>
<th>JF</th>
<th>SF</th>
<th>SF true</th>
<th>CF</th>
<th>RF (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>straight orbit</td>
<td>17.6</td>
<td>48.4</td>
<td>34.8</td>
<td>14.5</td>
<td>3.84</td>
<td>7.57</td>
</tr>
<tr>
<td>orbit</td>
<td>44.2</td>
<td>74.4</td>
<td>74.7</td>
<td>20.1</td>
<td>10.2</td>
<td>10.4</td>
</tr>
<tr>
<td>takeover orbit</td>
<td>50.1</td>
<td>55.3</td>
<td>47.3</td>
<td>48.1</td>
<td>48.2</td>
<td>13.1</td>
</tr>
</tbody>
</table>

**Fig. 5:** ATE of estimated camera trajectories on a synthetic sequence with different object sizes relative to the amount of valid image pixels. Co-Fusion and StaticFusion break around a dynamic ratio of 0.5 or less. Using the true motion priors in StaticFusion allows larger objects up to a ratio of 0.6, while our method with drift on the motion priors can track up to 0.75.
For sequence swing, which contains four dynamic objects, the
4 translational, which contains four dynamic objects, the
camera trajectory, static object for sequenceswing-
ing
4 translational, which contains four dynamic objects, the

(fig. 6: true and estimated trajectories (units in meter). Top: Top-down view of true camera and object trajectories in
evaluation sequences. The green trajectory represents the true object position in the Vicon reference frame. The red/blue
trajectory segments represent the camera trajectory and if the object is static (blue) or dynamic (red) within the image. Black
arrows point in the camera view direction. Bottom: True and estimated trajectories for our RigidFusion (with and without
drift on motion priors), the baselines StaticFusion [4] (with and without true motion priors) and Co-Fusion [2]. Trajectories
start at the origin (black solid dot) and end at the circle-cross marker. Our proposed method is closer to the ground truth
trajectory even with drift on the motion priors (red, dashed), while StaticFusion fails even with true prior (blue, solid).

TABLE III: ATE and RPE for sequences collected with Ada

<table>
<thead>
<tr>
<th>RGB-D sequence</th>
<th>Wheel odometry</th>
<th>Method</th>
<th>JF</th>
<th>SF</th>
<th>SF true</th>
<th>CF</th>
<th>RF (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sideway1</td>
<td>3.27</td>
<td>37.1</td>
<td>62.8</td>
<td>36.8</td>
<td>32.1</td>
<td>7.58</td>
<td></td>
</tr>
<tr>
<td>overtake</td>
<td>3.16</td>
<td>23.8</td>
<td>79.1</td>
<td>24.7</td>
<td>16.5</td>
<td>14.0</td>
<td></td>
</tr>
<tr>
<td>sideway2</td>
<td>3.64</td>
<td>51.9</td>
<td>86.3</td>
<td>21.9</td>
<td>12.3</td>
<td>7.98</td>
<td></td>
</tr>
<tr>
<td>camera ground truth</td>
<td>camera prior with drift</td>
<td>camera prior with true camera motion prior</td>
<td>Co-Fusion baseline</td>
<td>RigidFusion (ours) with true motion prior</td>
<td>StaticFusion baseline</td>
<td>StaticFusion with true camera motion prior</td>
<td></td>
</tr>
</tbody>
</table>

(a) Trans. Absolute Trajectory Error RMSE (cm)

(b) Trans. Relative Pose Error RMSE (cm/s)

The quantitative results in Table III show that using real
wheel odometry as motion priors, RF outperforms all other
four methods in terms of both ATE and RPE on all four
sequences. The estimated trajectories for sequences sideway1
and overtake are shown in Figure 8 (bottom).

E. Object Reconstruction

We compare the reconstructed dynamic object for CF and
RF in Figure 9. Since CF tends to over-segment objects, we
only show the first detected model. Results show that RF
 generates a more complete dynamic model than CF. This
suggests that the segmentation estimated by RF is consistent
over time and more accurate than CF.

F. Impact of Odometry Drift on Trajectory Estimation

We amplify the wheel odometry drift to test RF’s robustness
against different levels of camera motion prior drift. We
also test RF’s performance without the object motion prior
(fix \( \lambda_d = 0 \)). The relation between the RPE of the estimated
trajectories and the drift over all robot sequences is shown
in Figure 10. Even without the object motion prior, RF still
achieves better performance than CF for up to 17 cm/s drift
in terms of average RPE. Using both motion priors, RF
performs even better and is more robust to large odometry
drift. This demonstrates that our method can handle large
odometry drift and the absence of an object motion prior.

G. Impact of Multiple Dynamic Objects

RF assumes that the dynamic motion can be explained
by a single rigid transformation. To test RF’s performance
when this assumption is violated, we conduct qualitative
experiments on two OMD sequences [23] where multiple
dynamic objects are present (Figure 11).

For sequence swing_4_translational, which contains
one large and one small dynamic object, the motion prior
for the larger object is provided. For sequence swing_4_translational, which contains four dynamic objects, the
Fig. 7: Segmentation and 3D reconstructed background for our proposed algorithm RigidFusion (RF), StaticFusion (SF) [4] (with and without true motion priors), Joint-VO-SF (JF) [14] and Co-Fusion (CF) [2] on camera-only sequence sideways. Our proposed method is the only one that can consistently segment the large rigid dynamic object (compare first row with highlighted boxes against red dynamic segmentation) and reconstruct the background even the motion priors have a significant drift.

motion prior for the top-left object is provided. Despite this under-representation of the dynamic motion, RF outperforms SF and is able to correctly segment the static environment against all the dynamic objects. However, similar to SF, RF cannot independently track multiple dynamic objects with different motions.

VII. CONCLUSION

We have presented a robot localisation and mapping approach in environments where dynamic components can occupy the major portions of the visual input. To address this problem, we assume that the dynamic component is rigid, and jointly segment and track the static and dynamic rigid bodies. Detailed evaluation shows that our method RigidFusion outperforms state-of-the-art when a dynamic rigid object occludes more than 65% of the camera view. We demonstrate its robustness to odometry drift up to 17 cm/s and the absence of object motion priors.

Our method treats the whole dynamic component as a single rigid body and is thus unable to track multiple dynamic objects independently in the scene. Our future research direction involves extending the current pipeline to enable multiple rigid objects segmentation, tracking and reconstruction in dynamic environments. To handle dynamic objects that are not in contact with the manipulator, and thus have no kinematic prior, we intent to estimate motion priors using visual correspondences.

ACKNOWLEDGEMENT

This research is supported by the EPSRC UK Robotics and AI ISCF ORCA Hub (EP/R026173/1), the Alan Turing Institute and the Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS). We thank Raluca Scona for answering questions about StaticFusion and Tin Lun Lam from AIRS for critical comments.

REFERENCES

Fig. 8: True (top) and estimated (bottom) trajectories (units in meter). Our method (RF) outperforms all state-of-the-art methods. Although CF has closer end-position in x-y plane, it has a larger drift in the z position than RF.

---
camera/orbit camera/overtake robot/sideway2
---
CF
RF (ours)

Fig. 9: Reconstructed dynamic object. CF can only reconstruct parts of the dynamic object, while RF reconstructs a more complete model with inaccurate wheel odometry.

---
Fig. 10: RPE of the estimated trajectories impacted by the drift magnitude of wheel odometry. Our method can handle up to 17 cm/s drift without the object motion prior (solid red) before breaking down to comparable results with CF. Using both motion priors (solid blue), RF has a even better performance and stronger robustness.

---
Fig. 11: Segmentation results of two OMD [23] sequences with multiple dynamic objects. Although multiple objects can only be represented by a single transformation, RigidFusion (RF) is able to segment the static environment (blue) against multiple dynamic objects (red), while StaticFusion (SF) maps dynamic objects into the static environment.

---

---