Joint Ego-motion Estimation Using a Laser Scanner and a Monocular Camera Through Relative Orientation Estimation and 1-DoF ICP

Kaihong Huang Cyrill Stachniss

Abstract— **Pose estimation and mapping are key capabilities of most autonomous vehicles and thus a number of localization and SLAM algorithms have been developed in the past. Autonomous robots and cars are typically equipped with multiple sensors. Often, the sensor suite includes a camera and a laser range finder. In this paper, we consider the problem of incremental ego-motion estimation, using both, a monocular camera and a laser range finder jointly. We propose a new algorithm, that exploits the advantages of both sensors—the ability of cameras to determine orientations well and the ability of laser range finders to estimate the scale and to directly obtain 3D point clouds. Our approach estimates the five degree of freedom relative orientation from image pairs through feature point correspondences and formulates the remaining scale estimation as a new variant of the iterative closet point problem with only one degree of freedom. We furthermore exploit the camera information in a new way to constrain the data association between laser point clouds. The experiments presented in this paper suggest that our approach is able to accurately estimate the ego-motion of a vehicle and that we obtain more accurate frame-to-frame alignments than with one sensor modality alone.**

I. Introduction

The ability to estimate the ego-motion of a vehicle is a vital part of most autonomous navigation systems. Mobile robots and autonomous cars typically use different techniques for pose estimation, such as scan matching, visual odometry, or integrated GPS/IMU systems. Often, such systems are equipped with multiple sensors. Laser scanners are important for obstacle detection and tracking, while cameras are frequently used to interpret the scene using semantic segmentation or visual object detection systems.

Estimating the ego-motion using a 2D or 3D laser range finder through point cloud alignment is often referred to as scan-matching. Scans are either matched pair-wise or with respect to a local or global map in order to compute the relative transformation between the robot's poses at the different points in time. Popular approaches for that are the iterative closest point (ICP) algorithm [2], [22] and its variants, such as [19], [20], or correlative scan matching [15].

The ego-motion can also be estimated using visual odometry with stereo or monocular cameras [4], [5], [9], [11]. In case of a calibrated monocular camera, only five out of the six degrees of freedom can be estimated, since the scale cannot be determined. The most popular algorithm for that is Nistér's 5-point algorithm [14]. When comparing systems using cameras to those using lasers, we often see that cameras are slightly better in estimating the angular components, i.e., the rotation of the movement, whereas laser scanners are superior for estimating the translation and for obtaining 3D points. Furthermore, cameras provide dense color information, which can simplify the data association using feature correspondences. Thus, coupling laser scanners and camera can yield advantages.

In this paper, we address the problem of combining the laser range and camera information to jointly estimate the ego-motion of a mobile platform. We propose a coupled laser-visual scan matching method for frame-to-frame pose estimation. In this way, we combine the advantages of both sensing modalities, i.e., the abilities to accurately estimate orientations (camera) and the scale (laser range finder). Our method works with general sensor configurations and does not require an overlap of the fields of view of the camera and the laser scanner. Besides a calibrated monocular camera, we only assume that the laser and camera are extrinsically calibrated, which means that the relative pose between the two devices on the robot are known.

The main contribution of this paper is a novel approach to joint laser-camera pose estimation. It estimates the 5-DoF relative orientation from image pairs through feature point correspondences and formulates the remaining scale estimation problem as a variant of the ICP problem with only one degree of freedom. This can be solved effectively through a 1D grid search followed by a refinement for computing the minimum of the error function for solving the scale problem. Furthermore, our approach exploits the camera information to effectively constrain the data association between laser point clouds, even if the fields of view of both sensors do not overlap. In sum, we make two key claims: our approach (i) allows for accurate frame-to-frame alignments from monocular vision and laser range data and (ii) is able to exploit the advantages of both modalities.

II. Related Work

A problem similar to the visual-laser scan matching appears in RGB-D image registration [3], [8], [13]. Devices such as the Microsoft Kinect camera or stereo cameras can provide dense 3D information in the form of depth images along with regular RGB color information. Numerous methods are tailored to such dense 3D measurements with large overlapping fields of view. Recent examples are KinectFusion [13], DVO [8], and MPR [3]. Projective data association approaches are often used in such cases to speed up the registration process and to jointly exploit the depth and color cues. These approaches are in nearly all aspects

Both authors are with the University of Bonn, Germany. This work has partly been supported by the DFG under the grant number FOR 1505: Mapping on Demand and by the Chinese Scholarship Committee CSC.

different from our proposed one, even though they solve a related problem.

The ICP algorithm by Besl and McKay [2] for aligning point clouds is a popular approach for laser-based scan matching. For registering two point clouds recorded at unknown relative positions, the ICP algorithm iteratively performs two steps: data association and transformation estimation given the data association. The point-to-point distance and the point-to-plane distance are two commonly used metrics. We refer to the work of Pomerleau *et al.* [17] and Rusinkiewicz *et al.* [18] for a detailed review and comparison for different ICP variants.

A straight forward and common practice to combine laser and cameras for pose estimation is to use the visual odometry result as an initial guess for the ICP pipeline, see [16], [21]. The work of Zhang *et al.* [21] starts with visual odometry to roughly estimate the ego-motion, and the result is then refined by ICP. Another way of combining camera and laser information is to use image/color information to guide and accelerate the data association process [1], [7], [10], [12]. Andreasson *et al.* [1], Joung [7], and Men [10] represent methods that treat the color information as the fourth channel input to the ICP. This strategy reveals a faster convergence rate than normal ICP according to Men [10]. The color information is not used in the error minimization process in Men's approach, which is in contrast to the work of Joung [7], whose error function incorporates the color consistency of matched points. They both use color data to filter out unlikely point candidates before ICP.

The work of Naikal *et al.* [12] achieves the same goal but employs a different strategy. The data association is established through image patch matching instead of using a k-d tree-based closest point assignment. They project scan points onto the respective images so that the 3D points can be associated to image patches around the projected location. A patch matching process is then carried out across images by minimizing a bidirectional sum of absolute differences. The resulting patch correspondences eventually determine the scan point correspondences. The visual odometry result is furthermore used to provide a search window for the patch matching process.

All of the aforementioned methods have the limitation that they are only applicable to laser points that are visible in the camera image and not to all scan points. Furthermore, only the laser points with the improved correspondences are used to estimate the relative transformation. Additionally, these approaches typically cannot handle the fact that the rotation information obtained from visual matching is often more accurate than the ones obtained by a laser scanner. Our approach, in contrast, naturally respects this and works even if there is no overlap between the fields of view of laser and camera.

III. Background: ICP and Relative Orientation

A. Iterative Closest Point for Aligning Point Clouds

Consider two point clouds, the *previous* point cloud $\{\boldsymbol{a}_i \in \mathbb{R}^3\}_{i=1}^{M}$ and the *current* point cloud $\{\boldsymbol{b}_j \in \mathbb{R}^3\}_{j=1}^{N}$ that are generated from two, often consecutive laser scans. We would like to estimate the relative rotation $\boldsymbol{R} \in \mathrm{SO3}$ and translation $\boldsymbol{t} \in \mathbb{R}^3$ between the two scanning locations by registering the two point clouds. If a point pair $(\boldsymbol{a}_i, \boldsymbol{b}_j)$ of two measured points belongs to the same scene point, we have the relation

$$\boldsymbol{a}_i = \boldsymbol{R}\boldsymbol{b}_j + \boldsymbol{t}. \tag{1}$$

The point correspondences are usually unknown and need to be estimated. If we have an initial guess $\hat{\boldsymbol{R}}$, $\hat{\boldsymbol{t}}$ for the transformation, we can match a current points $\boldsymbol{b}_j$ to its closest points in the previous scan $\{\boldsymbol{a}\}$, i.e,

$$\boldsymbol{m}(\boldsymbol{b}_j) \stackrel{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{p}\in\{\boldsymbol{a}\}} \left\|\boldsymbol{p} - \boldsymbol{b}'_j\right\|^2 \tag{2}$$

with $\boldsymbol{b}'_j \stackrel{\text{def}}{=} \hat{\boldsymbol{R}}\boldsymbol{b}_j + \hat{\boldsymbol{t}}$. Point-to-point ICP solves

$$\boldsymbol{R}, \boldsymbol{t} = \operatorname*{argmin}_{\boldsymbol{R},\boldsymbol{t}} \sum_j \left\|\boldsymbol{t} + \boldsymbol{R}\boldsymbol{b}_j - \boldsymbol{m}(\boldsymbol{b}_j)\right\|^2 \tag{3}$$

and point-to-plane solves

$$\boldsymbol{R}, \boldsymbol{t} = \operatorname*{argmin}_{\boldsymbol{R},\boldsymbol{t}} \sum_j \left|[\boldsymbol{t} + \boldsymbol{R}\boldsymbol{b}_j - \boldsymbol{m}(\boldsymbol{b}_j)] \cdot \boldsymbol{\eta}_j\right|^2, \tag{4}$$

where $\boldsymbol{\eta}_j$ is the (rotated) normal information of the surface around the point $\boldsymbol{b}_j$.

ICP is an intuitive and straightforward to implement algorithm, but the unknown point associations can be limiting in practice and thus a good initial guess for the correspondences is of great value.

B. Relative Orientation of the Image Pair

The task of matching distinct features across camera images works quite reliably. In a monocular camera setup, we can estimate five out of the six degrees of freedom of the transformation between camera viewpoints purely based on image point correspondences. These are three parameters for rotation $\boldsymbol{R}$ and two parameters for translation direction $\boldsymbol{t}_{\text{dir}} \in \mathbb{R}^3$. The scale, i.e., the length of $\boldsymbol{t}_{\text{dir}}$, cannot be determined and thus one uses $\|\boldsymbol{t}_{\text{dir}}\| = 1$. This set of five parameters is often referred to as the relative orientation of the image pair and it can be efficiently estimated by exploiting the coplanarity constraint. This constraint is formulated by $\mathbf{x}_i^{\mathsf{T}} \boldsymbol{E} \mathbf{x}'_i = 0$, where $\mathbf{x}_i, \mathbf{x}'_i$ are the 2D image coordinates of a corresponding point pairs and $\boldsymbol{E}$ is the so-called essential matrix, from which the orientation parameters can be extracted. Various direct solutions for computing the essential matrix exist. We use Nistér's five-point algorithm [14] and SIFT features together with a standard RANSAC procedure. The relative orientation parameters are extracted from the essential matrix $\boldsymbol{E}$. We verify the parameters by standard checks such as the fact that triangulated points must lie in front of the camera and we handle typical corner cases, such as zero translations.

IV. Our Approach to Incremental Pose Estimation using Camera and Laser Information

Our approach starts with estimating the 5-DoF relative orientation of the image pair and then uses the laser information for scale estimation.

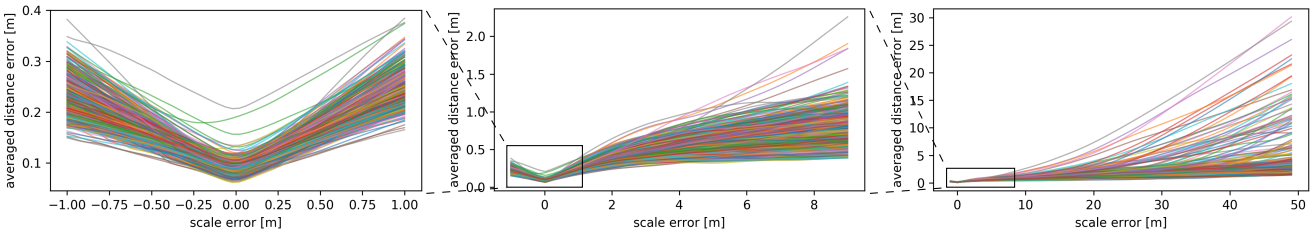

Fig. 1: 1-DoF ICP point-to-point cost function evaluated on the KITTI dataset. The x axis shows the scale deviation from the ground truth (0) and the y axis the averaged point matching error distance. The function reveals a smooth surface and appears to be mostly convex.

A. 1-DoF ICP for Scale Estimate

Given the relative orientation $\boldsymbol{R}^0, \boldsymbol{t}_{\text{dir}}$ computed from the image pair, the metric scale of the translation $\|\boldsymbol{t}_{\text{true}}\|$ is unknown. We denote the unknown scale parameter as s and express the scale through the translation vector between the two poses as

$$\boldsymbol{t}_{\text{true}} = s\,\boldsymbol{t}_{\text{dir}}, \quad s \in [0, \infty). \tag{5}$$

To estimate s, we propose to solve a novel variant of the ICP problem with only one degree of freedom, which can be expressed through

$$s = \underset{s \geq 0}{\operatorname{argmin}} \sum_i \left\| s\boldsymbol{t}_{\text{dir}} + \boldsymbol{R}^0\boldsymbol{b}_j - \boldsymbol{m}(\boldsymbol{b}_j) \right\|^2 \tag{6}$$

$$\text{or} \quad s = \underset{s \geq 0}{\operatorname{argmin}} \sum_i \left| [s\boldsymbol{t}_{\text{dir}} + \boldsymbol{R}^0\boldsymbol{b}_j - \boldsymbol{m}(\boldsymbol{b}_j)] \cdot \boldsymbol{\eta}_j \right|^2, \tag{7}$$

for the point-to-point and point-to-plane cost function respectively. Efficient closed form solution can be derived for both equations. To solve Eq. (6), we define $\boldsymbol{e}_j \overset{\text{def}}{=} \boldsymbol{m}(\boldsymbol{b}_j) - \boldsymbol{R}\boldsymbol{b}_j$ and obtain:

$$\Phi(s) \overset{\text{def}}{=} \sum_j \left\| s\boldsymbol{t}_{\text{dir}} + \boldsymbol{R}^0\boldsymbol{b}_j - \boldsymbol{m}(\boldsymbol{b}_j) \right\|^2 \tag{8}$$

$$= \sum_j \| s\boldsymbol{t}_{\text{dir}} - \boldsymbol{e}_j \|^2 \tag{9}$$

$$= \sum_j s^2 - 2s\boldsymbol{e}_j^\mathsf{T}\boldsymbol{t}_{\text{dir}} + \boldsymbol{e}_j^\mathsf{T}\boldsymbol{e}_j. \tag{10}$$

By setting $\frac{\partial \Phi}{\partial s} = 0$, we obtain $\sum_j s - \boldsymbol{e}_j^\mathsf{T}\boldsymbol{t}_{\text{dir}} = 0$ and thus

$$s_{\text{new}} = \frac{1}{N} \sum_j \boldsymbol{e}_j^\mathsf{T}\boldsymbol{t}_{\text{dir}}, \tag{11}$$

where N is total number of matched point pairs.

Similarly, for point-to-plane distances according to Eq. (7), we define $w_j \overset{\text{def}}{=} \boldsymbol{\eta}_j^\mathsf{T}\boldsymbol{t}_{\text{dir}}$ and obtain

$$\Phi(s) \overset{\text{def}}{=} \sum_j \left| [s\boldsymbol{t}_{\text{dir}} + \boldsymbol{R}^0\boldsymbol{b}_j - \boldsymbol{m}(\boldsymbol{b}_j)] \cdot \boldsymbol{\eta}_j \right|^2 \tag{12}$$

$$= \sum_j \left| s\boldsymbol{\eta}_j^\mathsf{T}\boldsymbol{t}_{\text{dir}} + \boldsymbol{\eta}_j^\mathsf{T}[\boldsymbol{R}^0\boldsymbol{b}_j - \boldsymbol{m}(\boldsymbol{b}_j)] \right|^2 \tag{13}$$

$$= \sum_j \left| sw_j - \boldsymbol{\eta}_j^\mathsf{T}\boldsymbol{e}_j \right|^2. \tag{14}$$

Setting $\frac{\partial \Phi}{\partial s} = 0$ leads to $\sum_j sw_j^2 - w_j\boldsymbol{\eta}_j^\mathsf{T}\boldsymbol{e}_j = 0$ and thus

$$s_{\text{new}} = \frac{\sum_j w_j \boldsymbol{\eta}_j^\mathsf{T}\boldsymbol{e}_j}{\sum_j w_j^2}. \tag{15}$$

This 1-DoF ICP problem possess several attractive properties. First and foremost, in all our analyzed cases, the cost function has a well distinguishable global minimum, especially in feature-rich environment. Consider the KITTI [6] dataset "odometry sequence 00" as an example. Fig. 1 shows a plot of the point-to-point cost function evaluated over keyframes covering the whole scene. Although the cost function depends on the scene structure, we found that the curve is smooth, appears to be mostly convex, and that the global minimum represents the true scale well. Beside from that, the solution space of the scale parameter can also be easily bounded if additional information is available, as the scale parameter represents the physical straight line distance that the vehicle has traveled.

Based on the above observation, we propose a two-step approach to solve the 1-DoF ICP instead of a standard ICP implementation: 1D grid search followed by an iterative refinement. The grid search operates in a branch and bounce fashion to locate the basin of the global minimum. Given a search boundaries $[s_{\min}, s_{\max}]$, we generate a small number of linearly spaced hypotheses and evaluate the cost (e.g., total point distance) for each hypothesis. We then select two of the hypotheses with minimum cost as the new search boundaries and repeat the process. In this way, the solution space can be drastically reduced in just a few iterations and one has a higher chance to avoid shallow local minima that may impair the ICP performance. If a prior s^0 exists, for example from wheel encoders, we can also incorporate it into the the grid search as an extra hypothesis in the first iteration.

After the grid search has been performed, we iteratively refine s using the closed form solution in Eq. (11) or Eq. (15) according to the use cost function, starting with the data association that results from the grid search solution and updating it in every iteration.

The algorithmic view of our method is given in Alg. 1. It is worth pointing out that in the point-to-plane version, we can safely discard any points of the current scan that has a normal vector perpendicular to $\boldsymbol{t}_{\text{dir}}$, without compromising the solution. Because $\boldsymbol{\eta}_j^\mathsf{T}\boldsymbol{t}_{\text{dir}} = w_j = 0$ implies that these points have no influence on the solution. Such a filtering can greatly reduce the computational effort and is a big advantage for real-time processing. For example, in the KITTI dataset

sequence 00, we can remove up to 40% of the scan points because they lie on the ground plane and do not contribute to the scale estimate.

After convergence, we execute one final ICP step that corrects all three translational parameters (but not the rotation parameters nor changes the data association). We observed that this steps leads to slightly better results for the estimated trajectory in the end.

Algorithm 1 1-DoF ICP for scale estimate

Input:
- Previous point cloud $\boldsymbol{a}$, current point cloud $\boldsymbol{b}$;
- Relative orientation $R^0, \boldsymbol{t}_{\text{dir}}$;
- Initial search boundary $s_{\min}, s_{\max}$;
- Initial guess s^0;

Parameter:
- Number of hypotheses per iteration $n \in [3, \infty)$;
- Outlier distance threshold d_{out};

Output: Estimated scale s.

▷ Step 1: Grid Search
repeat
  Hypothesis $\{s_1, \cdots, s_n\} \leftarrow \text{linspace}(s_{\min}, s_{\max}, n)$;
  if first iteration **then** $s_{n+1} \leftarrow s^0$;
  for $s \in \{s_1, \cdots, s_{n+1}\}$ **do**
    Transform current cloud $\boldsymbol{b}' \leftarrow R^0\boldsymbol{b} + s\boldsymbol{t}_{\text{dir}}$;
    Match previous cloud $\boldsymbol{m} \leftarrow \text{argmin}_{\boldsymbol{a}} \left\|\boldsymbol{a} - \boldsymbol{b}'\right\|$;
    Calculate cost $C(s) \leftarrow \sum_{\boldsymbol{b}'} \left\|\boldsymbol{m} - \boldsymbol{b}'\right\|$;
  Update $s_{\min}, s_{\max} \leftarrow$ the two s with lowest cost.
until converge **or** maximum iterations reached
$s \leftarrow \text{argmin}_{\{s_{\min}, s_{\max}\}} C(s)$
▷ Step 2: Refinement
repeat
  Transform current cloud $\boldsymbol{b}' \leftarrow R^0\boldsymbol{b} + s\boldsymbol{t}_{\text{dir}}$;
  Match previous cloud $\boldsymbol{m} \leftarrow \text{argmin}_{\boldsymbol{a}} \left\|\boldsymbol{a} - \boldsymbol{b}'\right\|$;
  Remove points pair whose distance exceeded d_{out};
  if using point-to-point **then**
    Update $s \leftarrow \frac{1}{N}\sum_j \boldsymbol{e}_j^{\mathsf{T}}\boldsymbol{t}_{\text{dir}}$ ▷ from Eq. (11)
  else // using point-to-plane
    Update $s \leftarrow \frac{\sum_j w_j \boldsymbol{\eta}_j^{\mathsf{T}}\boldsymbol{e}_j}{\sum_j w_j^2}$ ▷ from Eq. (15)
until converge **or** maximum iterations reached
return s

B. Relative Orientation Constrained Data Association

Beside reducing the normal ICP problem to a one degree of freedom one, the relative orientation can also be used to guide the data association that has to take place in all ICP iterations. Considering the fact that the current frame must be located in the direction $\boldsymbol{t}_{\text{dir}}$ with respect to the previous one, the same must hold for the corresponding points, see Fig. 2 for an illustration. Therefore, for a current point $\boldsymbol{b}_j$, we can restrict its matching candidates $\boldsymbol{a}_i$ to be located near to the ray $\boldsymbol{r}_j = \boldsymbol{b}'_j - \lambda\boldsymbol{t}_{\text{dir}}$, instead of arbitrary points in the whole previous point cloud. Ideally, the point $\boldsymbol{a}_i$ should lie

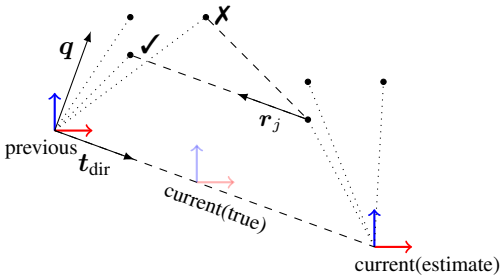

Fig. 2: The relative orientation supports point cloud data association. Given $R^0, \boldsymbol{t}_{\text{dir}}$, the point set $\boldsymbol{b}'_j$ should be matched to a point in $\boldsymbol{a}_i$ that is lying on the ray $\boldsymbol{r}_j = \boldsymbol{b}'_j - \lambda\boldsymbol{t}_{\text{dir}}$ even if it is not the closest point. Thus, several wrong associations can be excluded.

exactly on the ray, but due to noise, we relax the constraint and allow the candidate point to slightly deviate from the ray.

To achieve this, we propose a modified closest point association procedure as listed in Alg. 2. The main idea is to use a temporary coordinate system with $\boldsymbol{t}_{\text{dir}}$ being the x-axis and the previous frame's origin being the origin of that frame. Any point correspondences that are inconsistent with the direction $\boldsymbol{t}_{\text{dir}}$ will have nonzero Y and Z components in its error vector in this temporary frame. Thus, we can define a weighted Euclidean distance metric, which punishes the Y and Z components in this frame, i.e.,

$$d^2(\boldsymbol{a}_i, \boldsymbol{b}'_j) \overset{\text{def}}{=} (Q\boldsymbol{a}_i - Q\boldsymbol{b}'_j)^{\mathsf{T}} \begin{bmatrix} 1 & & \\ & \alpha & \\ & & \alpha \end{bmatrix} (Q\boldsymbol{a}_i - Q\boldsymbol{b}'_j) \tag{16}$$

$$= (\boldsymbol{a}_i - \boldsymbol{b}'_j)^{\mathsf{T}} Q^{\mathsf{T}} \begin{bmatrix} 1 & & \\ & \alpha & \\ & & \alpha \end{bmatrix} Q(\boldsymbol{a}_i - \boldsymbol{b}'_j) \tag{17}$$

$$\overset{\text{def}}{=} (\boldsymbol{a}_i - \boldsymbol{b}'_j)^{\mathsf{T}} W(\boldsymbol{a}_i - \boldsymbol{b}'_j), \tag{18}$$

where $\alpha \gg 1$ is the penalty weight for the Y and Z components and Q is a rotation matrix, which is used to transform the points $\boldsymbol{a}_i, \boldsymbol{b}'_j$ from the previous frame into the new temporary frame.

The rotation matrix Q depends on vector $\boldsymbol{t}_{\text{dir}}$ and can be generated by applying QR decomposition to $\boldsymbol{t}_{\text{dir}}$. The orthonormal matrix of the decomposition result is used as Q after transpose. The rows of Q consist of $\boldsymbol{t}_{\text{dir}}$ and two orthogonal complements of $\boldsymbol{t}_{\text{dir}}$ in $\mathbb{R}^3$, i.e., $Q = \begin{bmatrix}\boldsymbol{t}_{\text{dir}} & \boldsymbol{q}_1 & \boldsymbol{q}_2\end{bmatrix}^{\mathsf{T}}$ and $\boldsymbol{q}_1 \perp \boldsymbol{q}_2 \perp \boldsymbol{t}_{\text{dir}}$.

The proposed distance metric can be used in a standard k-d tree algorithm with minor modifications, while the other parts of ICP algorithm remain the same.

With this distance metric, we can efficiently transfer the knowledge gained from image feature correspondences into the process of laser point association without requiring an overlap in the field of views of both sensors.

Algorithm 2 Constrained Data Association

1: **Input:**
- Previous point cloud $\boldsymbol{a}$, current point cloud $\boldsymbol{b}$;
- Relative orientation and scale $R^0, \boldsymbol{t}_{\text{dir}}, s$;

2: **Parameter:**
- Penalty weight $\alpha \gg 1$;
- Outlier distance threshold d_{out};

3: **Output:** Matched point pairs $(\boldsymbol{m}, \boldsymbol{b})$.

4: Calculate QR decomposition: $QR = \boldsymbol{t}_{\text{dir}}$;

5: Weight matrix $W \leftarrow Q \begin{bmatrix} 1 & & \\ & \alpha & \\ & & \alpha \end{bmatrix} Q^{\mathsf{T}}$;

6: Transform current points $\boldsymbol{b}' \leftarrow R^0 \boldsymbol{b} + s\boldsymbol{t}_{\text{dir}}$;

7: Match previous points $\boldsymbol{m} \leftarrow \text{argmin}_{\boldsymbol{a}} (\boldsymbol{a} - \boldsymbol{b}')^{\mathsf{T}} W (\boldsymbol{a} - \boldsymbol{b}')$;

8: Remove point pairs with $\left\| \boldsymbol{m} - \boldsymbol{b}' \right\| > d_{\text{out}}$;

9: **return** point correspondences $(\boldsymbol{m}, \boldsymbol{b})$

V. Experimental Evaluation

The main focus of this work is a novel approach to joint laser-camera ego-motion estimation. We make the claims that our approach (i) allows for accurate frame-to-frame alignment from monocular vision and laser range data and that (ii) it is able to exploit the advantages of both modalities. Our experiments are designed to support these two claims.

We perform our evaluations on the KITTI dataset as it is a standard dataset for these type of problems and we have ground truth available. Note that, compared to several other methods, our approach achieves it performance without any loop-closing.

Within this evaluation, we use the geodesic distance on a unit sphere to parameterize the rotational error. Given a rotation matrix ΔR, we compute

$$e_{\text{rot}} = \arccos\left(\frac{\text{trace}(\Delta R) - 1}{2}\right) \tag{19}$$

as the error angle.

A. Error Evaluation

The first set of experiments is designed to support both claims, i.e., that our approach can accurate align frames pairwise and that it is able to exploit the advantages of both modalities. Fig. 3 shows the cumulative error plots for the rotation (left plot) and the translation (right plot) of our approach in comparison to an optimized, laser-only ICP. These are the cumulative plots over all sequences of the KITTI dataset but the plots for the individual datasets looks similar and show the same characteristics.

Based on the left plot of Fig. 3, it is clearly visible that the relative orientation information from the camera provides a better estimate of rotational component of the ego-motion (blue line) than laser-based ICP (dashed orange line). The blue line shows the performance of Nistér's 5-point algorithm and our approach (as we use the 5-point algorithm for the rotation estimation). The fact that this approach is better than laser-based ICP can be seen because blue curve is always above the dashed, orange one.

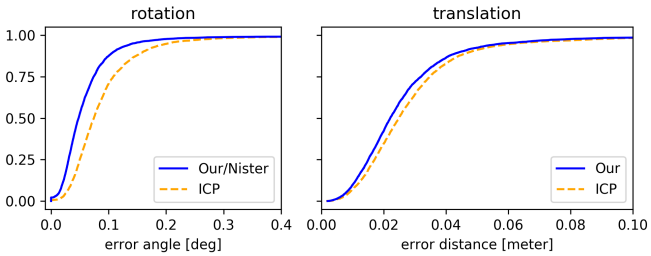

Fig. 3: Cumulative error distribution: Percentage (y-axis) of cumulative errors (x-axis) in rotation and translation for our approach and laser-only ICP.

We can furthermore show that our approach outperforms laser-only ICP when estimating the translational part, see right plot of Fig. 3. The blue line represents our approach and is always above the dashed, orange one, which corresponds to laser-only ICP. This is the case for two reasons: First, our point-to-point data association described in Sec. IV-B is better than the regular ICP data association as we drastically reduce the number of potential matches since we only need to consider points that are in line with the rotation. This avoids several wrong data associations. Second, the orientation estimates of our approach are better than those of laser-only ICP and they also impact the translation estimation. Note that no translational error can be provided for the camera-only case as the scale, i.e., the length if the motion vector, cannot determined using a monocular camera.

Thus, we can conclude that our approach outperforms visual odometry from the monocular camera (because we obtain an accurate scale estimate) as well as laser-based ICP (more accurate orientation and translation).

B. Trajectory Estimation

This second part of the evaluation also supports the first claim and furthermore provides a better visual impression about the quality of the estimated trajectories. We plot the ground truth trajectories, our estimates, and the ones of laser-only ICP for several KITTI sequences in Fig. 4 (only a subset is shown due to space reasons). In all sequences except the top left one, it is rather clear from the shown X/Y plots that our trajectory estimate is always closer to the ground truth than the ones obtained by laser-only ICP. For the top left trajectory (sequence 00), this is more difficult to see. When inspecting the error in the Z component, however, we can see in Fig. 5 that our approach clearly outperforms laser-only ICP. For nearly all keyframes, the error in the height estimate (Z axis) is larger for the laser-only ICP estimate.

VI. Conclusion

In this paper, we presented novel approach to ego-motion estimation using a monocular camera and a laser range finder jointly. Our approach estimates the 5-DoF relative orientation from the camera images and uses a novel variant of ICP

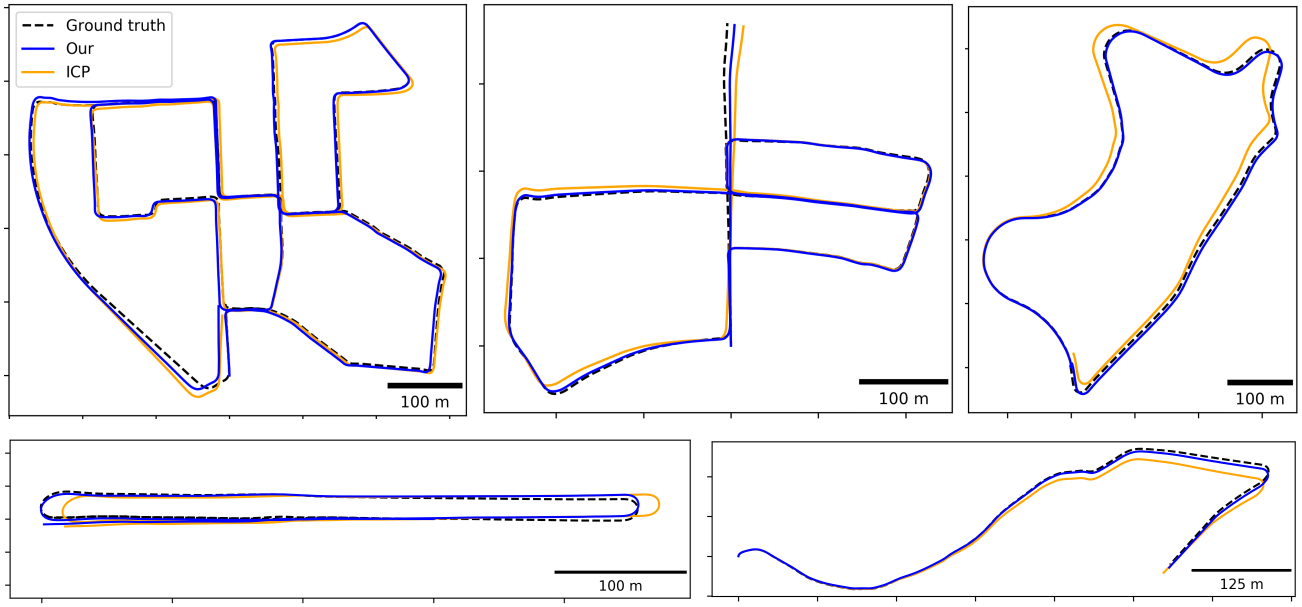

Fig. 4: Resulting trajectories from a subset of the KITTI sequence through our frame-to-frame registration without any loop-closing.

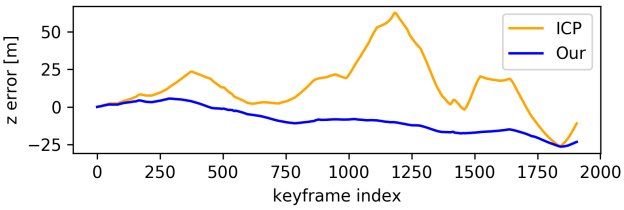

Fig. 5: Error in Z direction of sequence 00 for the keyframes.

with 1-DoF to estimate the scale. We can furthermore constrain the possible data associations among the point clouds given constraints derived from the relative orientation. We implemented our approach and evaluated it using KITTI data. In sum, our approach provides accurate trajectory estimates, which are better than those of each sensing modality alone.

References

[1] H. Andreasson, R. Triebel, and W. Burgard. Improving plane extraction from 3d data by fusing laser data and vision. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 2656–2661, 2005.

[2] P.J. Besl and N.D. McKay. A method for registration of 3-d shapes. *IEEE Trans. on Pattern Analalysis and Machine Intelligence (TPAMI)*, pages 14–2, 1992.

[3] B. Della Corte, I. Bogoslavskyi, C. Stachniss, and G. Grisetti. A General Framework for Flexible Multi-Cue Photometric Point Cloud Registration. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2018.

[4] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Trans. on Pattern Analalysis and Machine Intelligence (TPAMI)*, 2017.

[5] J. Engel, T. Schöps, and D. Cremers. Lsd-slam: Large-scale direct monocular slam. In *European Conference on Computer Vision*, pages 834–849. Springer, 2014.

[6] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[7] J.H. Joung, K.H. An, J.W. Kang, M.J. Chung, and W. Yu. 3d environment reconstruction using modified color icp algorithm by fusion of a camera and a 3d laser range finder. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2009.

[8] C. Kerl, J. Sturm, and D. Cremers. Robust odometry estimation for rgb-d cameras. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, pages 3748–3754. IEEE, 2013.

[9] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *Proc. of the IEEE and ACM Intl. Symp. on Mixed and Augmented Reality (ISMAR)*, 2007.

[10] H. Men, B. Gebre, and K. Pochiraju. Color point cloud registration with 4d icp algorithm. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2011.

[11] R. Mur-Artal, J.M.M. Montiel, and J. D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Trans. on Robotics (TRO)*, 31(5):1147–1163, 2015.

[12] N. Naikal, J. Kua, G. Chen, and A. Zakhor. Image augmented laser scan matching for indoor dead reckoning. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2009.

[13] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *Proc. of the Intl. Symposium on Mixed and Augmented Reality (ISMAR)*, pages 127–136, 2011.

[14] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Trans. on Pattern Analalysis and Machine Intelligence (TPAMI)*, 26(6):756–770, 2004.

[15] E.B. Olson. Real-time correlative scan matching. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, pages 4387–4393, 2009.

[16] G. Pandey, J. McBride, S. Savarese, and R. Eustice. Visually bootstrapped generalized icp. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2011.

[17] F. Pomerleau, F. Colas, and R. Siegwart. A review of point cloud registration algorithms for mobile robotics. 4:1–104, 05 2015.

[18] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on*, pages 145–152, 2001.

[19] A Segal, D. Haehnel, and S. Thrun. Generalized-icp. In *Proc. of Robotics: Science and Systems (RSS)*, volume 2, page 435, 2009.

[20] J. Serafin and G. Grisetti. Nicp: Dense normal based point cloud registration. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 742–749, 2015.

[21] J. Zhang and S. Singh. Visual-Lidar Odometry and Mapping: Low-Drift, Robust, and Fast. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2015.

[22] Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *Intl. Journal of Computer Vision (IJCV)*, 13(2):119–152, 1994.