Joint 45th International Symposium on Robotics (ISR) and 8th German Conference on Robotics (ROBOTIK), Munich, June 2014.

Combining the Strengths of Sparse Interest Point and Dense Image Registration for RGB-D Odometry

Jörg Stückler, stueckler@ais.uni-bonn.de
Arno Gutt, s6argutt@uni-bonn.de
Sven Behnke, behnke@cs.uni-bonn.de

Autonomous Intelligent Systems, Computer Science Institute, University of Bonn, Germany

Abstract

Visual odometry, i.e., estimating ego-motion from camera images, is frequently used as a building block in robot navigation systems. In this paper, we propose an efficient approach that provides robust and accurate visual odometry from RGB-D cameras in a wide range of settings. We seamlessly combine dense RGB-D image registration with the alignment of sparse interest points. While the former approach is robust and accurate when perceiving the depth towards structures well in less textured parts of an environment, the latter often performs better, if well textured but less structured parts are visible. Our formulation also integrates interest points with strongly uncertain or no depth to make best use of the available images. In experiments, we demonstrate advantages of our approach over methods that either are based on dense image or sparse interest point matching.

1 Introduction

Robot navigation systems such as those for rough terrain rovers or flying robots, frequently employ visual odometry as one important cue to estimate the six degree-of-freedom motion of the vehicle. In this paper, we propose robust and accurate visual odometry from RGB-D cameras that potentially covers a wide range of settings by the combination of shape and texture cues.

We combine dense RGB-D image registration with sparse interest point matching in a coherent framework. Our dense image registration method uses all available depth to align observed surfaces. This process is supported by the texture information contained in the RGB images. We complement dense shape registration with sparse interest point matching to also incorporate detailed information from the RGB images in regions that only have far and noisy depth readings or no depth measurements at all. Our approach seamlessly integrates both objectives and performs dense RGB-D image registration and sparse bundle adjustment concurrently.

Figure 1: Dense RGB-D registration is particularly well suited in close-by scenes with strong shape variations (top left). It also performs well if strong shape but weak texture cues are available (top right). The registration of sparse interest points is advantagous if shape does barely constrain the registration, but texture variations provide landmarks in the scenes. This may be the case with mostly far and noisy depth measurements (bottom left), or if a flat structure is observed (bottom right).

2 Related Work

Using multi-camera setups such as stereo cameras, the 3D geometry of interest points and the camera motion can be directly estimated between two images. In his seminal work, Nister [8] proposes real-time visual odometry for monocular as well as stereo cameras. On sequences of stereo images, camera motion is estimated first between pairs of frames using a 3-point RANSAC algorithm which is further refined using bundle adjustment. The approach of Howard et al. [4] also formulates visual odometry as sliding window bundle adjustment. It enforces rigidness in the arrangement of the interest point matches which further improves robustness. Fovis [5] applies concepts from stereo visual odometry to RGB-D cameras. The approach initializes interest point matching and bundle adjustment with a coarse rotation estimate that is obtained through image correlation. Droeschel et al. [3] match interest points in the intensity images of time-of-flight depth cameras and register the 3D coordinates of the points.

With dense depth measurements available, images can be

aligned as a whole to estimate camera motion. Typical approaches in robotics and computer graphics are variants of the ICP [1] algorithm which register 3D point clouds. On current CPUs, ICP methods require subsampling to achieve high frame rate. In robotics, GICP [10] has been proposed which unifies the ICP formulation for various error metrics such as point-to-point, point-to-plane, and plane-to-plane. Magnusson et al. [7] propose the 3D-NDT which transforms point clouds into 3D grids and represents the points by their Gaussian statistics within the voxels. Stoyanov et al. [12] register a 3D-NDT of a scene point cloud to a model 3D-NDT. To the best of our knowledge, these registration methods are not reported to support real-time capable scan-matching of VGA RGB-D images.

Recently, Steinbrücker et al. [11] proposed a dense photogrammetric matching method for estimating visual odometry from RGB-D images. They model how images transform through view-pose change and optimize for the pose in order to align gray-scale images taking into account the measured depth. Kerl et al. [6] extended this approach with a student t-distribution noise model and iteratively reweighted least squares optimization.

In previous own work [15], we proposed multi-resolution surfel maps (MRSMaps), which enables high frame-rate registration on CPUs. Similar to the 3D-NDT, we overlay a 3D grid at multiple resolutions onto the point cloud measurements. Within the voxels, we store the Gaussian statistics on the shape and color distribution of the points. To support data association during registration, we describe local context of each voxel in shape-texture descriptors. The data association is performed in an efficient multi-resolution strategy, and the alignment pose between two maps is found through optimization of the matching of the Gaussian statistics. Our highly efficient implementation registers 640×480 RGB-D images at a frame rate of about 23 Hz on a CPU.

3 Dense Image and Sparse Interest Point Registration

We combine approaches to dense image and sparse interest point registration in a common optimization framework. For dense image registration, we employ multi-resolution surfel maps (MRSMaps, [15]). Sparse interest point matching is performed through bundle adjustment. Both objectives are integrated in a single one to combine their complementary strengths.

3.1 Dense Image Registration with Multi-Resolution Surfel Maps

Multi-resolution surfel maps (MRSMaps [15]) represent the RGB-D image content as Gaussian color and shape statistics within voxels of an octree. We denote the content of a voxel as surface element (surfel). The maximum resolution at each point in the map is adapted to the noise properties of the individual measurements. It is limited in proportion to the squared distance from the sensor. A map not only stores surfels in the leaves of the octree, but also on all coarser resolutions. This allows for aligning maps taken from different view points efficiently on the finest common resolution. Registration alternates iteratively between surfel association and pose optimization.

3.1.1 Surfel Association

Associations of surfels $\mathcal{A}_S = \{(s_s, s_t)\}$ are established from the current estimate x of the view pose difference between the maps. For each scene surfel s_s, we search for an association with a model surfel s_t. Surfels that have not been associated in the previous iteration are transformed with the current pose estimate, and a corresponding surfel is searched in a local volume around the transformed position. The size of this local volume scales with the resolution of the surfel. We determine associations for surfels that could be associated in the previous iteration in a more efficient way by either selecting the associated surfel from the previous association or one of its direct neighbors in the voxel grid. We only establish associations on the finest resolution common between both maps to save redundant calculations on coarser resolutions.

3.1.2 Pose Optimization

To estimate pose, its likelihood given the observations of scene MRSMap M_s and target map M_t is optimized,

$$p(x \mid M_s, M_t) = \eta\, p(M_s \mid x, M_t)\, p(x), \qquad (1)$$

where η is a normalization constant independent of x. Through the probability $p(x)$, we can include prior knowledge on the pose. Without such knowledge, the prior is given by the uniform distribution. The observation likelihood considers the matching of associated surfels given the pose estimate

$$p(M_s \mid x, M_t) = \prod_{(s_s, s_t) \in \mathcal{A}_S} p(s_s \mid x, s_t). \qquad (2)$$

The probability of a surfel match is the likelihood of the matching of the normal distributions of both surfels,

$$p(s_s \mid x, s_t) = \mathcal{N}\left(0; \mu_t - T(x)\mu_s, \Sigma_t + R(x)\Sigma_s R(x)^T\right), \qquad (3)$$

where μ is the sample mean of a surfel and Σ its sample covariance.

We efficiently optimize the logarithm of our probabilistic objective function using the Levenberg-Marquardt (LM) method. The negative logarithm of the objective is

$$L_M(x) := const. + \frac{1}{2} \sum_{(s_s, s_t) \in \mathcal{A}_S} \log |\Sigma(x; s_s, s_t)| + \frac{1}{2} \sum_{(s_s, s_t) \in \mathcal{A}_S} (\mu_t - g(x, s_s))^T \Sigma(x; s_s, s_t)^{-1} (\mu_t - g(x, s_s)), \qquad (4)$$

where we defined $g(x, s_s) := T(x)\mu_s$ and $\Sigma(x; s_s, s_t) := \Sigma_t + R(x)\Sigma_s R(x)^T$. We neglect the

effect of the pose variable on the matching covariances to write the objective as

$$L_M(x) \approx const. + \frac{1}{2} e_M(x)^T W_M e_M(x). \quad (5)$$

Here, $e_M(x)$ is a vector of residuals stacked from the residuals $\mu_t - g(x, s_s)$ for each surfel association. The weight matrix W_M contains the inverse matching covariances $\Sigma(x; s_s, s_t)^{-1}$ of the associations on its main diagonal. Using this approximation, LM optimization performs damped Gauss-Newton steps

$$\Delta x = \left[J_M(x)^T W_M J_M(x) + \lambda I \right]^{-1} J_M(x)^T W_M e_M(x), \quad (6)$$

where $J_M(x)$ is the Jacobian of $e_M(x)$.
For further details on RGB-D image aggregation and registration with MRSMaps, we kindly refer the reader to [15].

3.2 Sparse Interest Point Registration

We extract multi-scale ORB [9] interest points F from the RGB image and use the ORB descriptor for matching them.

3.2.1 Interest Point Detection

The ORB interest point detector finds corners in the image and describes the texture pattern in the local vicinity of the corner with binarized pixel comparisons. For pose estimation, a uniform distribution of interest points across the image is beneficial. We overlay a 8×6 grid over the image and select the 25 strongest interest points in a cell. We parametrize the position of interest points f by 2D pixel location f_x, f_y and inverse depth $f_\rho = 1/d$. Its covariance $\Sigma_f = \mathrm{diag}(\sigma^2_{f,x}, \sigma^2_{f,y}, \sigma^2_{f,\rho})$ quantifies uncertainties in pixel position and inverse depth. One reason for using inverse depth is the disparity measurement principle of textured-light projecting RGB-D sensors. A good approximation for such sensors is that the standard deviation of a depth measurement scales quadratically with depth. Inverse depth also allows for modeling large depth measurement noise [2]. Points without depth can be assigned a very large constant.
Additionally, to consider imprecisions of the pixel position of the interest point, we examine the depth in the local image neighborhood of the interest point. We use the empirical mean of the depth in a local radius and add its variance to the modeled depth uncertainty. As a positive side-effect, if the interest point is at a depth discontinuity, depth variance covers the range between foreground and background.

3.2.2 Interest Point Matching

For matching interest points F between images, we first reject self-similar features in each image according to the Hamming distance of their binary descriptors. From the remaining interest points, we build a LSH index [9] to efficiently match the binary descriptors between the images. The set of associations $\mathcal{A}_F = \{(f_s, f_t)\}$ consists of the best three matches per interest point of the source image. A match is required to be mutual, i.e., it needs to be found in both directions between the images.

3.2.3 Pose Optimization

As in bundle adjustment, we concurrently optimize for the view pose difference between the images as well as the *landmark* positions L of the interest points in a common reference frame. We parametrize landmark position as pixel position and inverse depth, while the common reference frame is naturally chosen as the camera frame of the target image. Using a bundle-adjustment approach potentially improves the estimate over pure 3D registration, when only coarse depth or barely depth is available for the interest points.
We formulate this optimization objective as the likelihood of the view pose estimate x_s of the scene image and the landmark positions L_s and L_t of the interest points in both images,

$$p(x_s, L_s, L_t \mid x_t = 0, F_s, F_t) = \eta\, p(F_s, F_t \mid x_s, x_t = 0, L_s, L_t)\, p(x_s)\, p(L_s)\, p(L_t), \quad (7)$$

given that the pose of the target image is fixed and coincides with the common reference frame, i.e., $T(x_t) = I$. The factorization includes prior probabilities on poses and landmark positions, for which we assume uninformed uniform probabilities.
The observation likelihood of the interest points further factorizes into

$$p(F_s, F_t \mid x_s, x_t = 0, L_s, L_t) = \prod_{a=(f_s,f_t)\in\mathcal{A}_F} p(f_s \mid x_s, l_a)\, p(f_t \mid x_t = 0, l_a). \quad (8)$$

We model the individual observation likelihood of an interest point as being normal distributed, i.e.

$$p(f \mid x, l) = \mathcal{N}\left(f; h(x, l), \Sigma_f\right). \quad (9)$$

The observation model

$$h(x, l) = \pi^{-1}\left[T(x)\, \pi\left(l\right)\right] \quad (10)$$

projects the landmark position l into the camera frame in which the interest point has been observed. The projective mapping π transforms positions parametrized in pixel location and inverse depth into Cartesian 3D coordinates, i.e.

$$\pi\left[\begin{pmatrix} f_x \\ f_y \\ f_\rho \end{pmatrix}\right] = K^{-1} \begin{pmatrix} f_\rho^{-1} & 0 & 0 \\ 0 & f_\rho^{-1} & 0 \\ 0 & 0 & f_\rho^{-2} \end{pmatrix} \begin{pmatrix} f_x \\ f_y \\ f_\rho \end{pmatrix}, \quad (11)$$

where

$$K = \begin{pmatrix} fl_x & 0 & c_x \\ 0 & fl_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \tag{12}$$

is the instrinsic camera calibration matrix parametrized by focal lengths fl_x, fl_y and optical center c_x, c_y. The inverse projective mapping is

$$\pi^{-1}\left[\begin{pmatrix} l_x \\ l_y \\ l_z \end{pmatrix}\right] = \begin{pmatrix} l_z^{-1} & 0 & 0 \\ 0 & l_z^{-1} & 0 \\ 0 & 0 & l_z^{-2} \end{pmatrix} K \begin{pmatrix} l_x \\ l_y \\ l_z \end{pmatrix}, \tag{13}$$

The negative log likelihood of the objective in Eq. (8) is

$$\begin{aligned} L_F(y) &= const. \\ &+ \frac{1}{2} \sum_{a=(f_s,f_t)\in\mathcal{A}_F} \Big[(f_s - h(x,l_a))^T \Sigma_{f,s}^{-1} (f_s - h(x,l_a)) \\ &\quad + (f_t - h(0,l_a))^T \Sigma_{f,t}^{-1} (f_t - h(0,l_a)) \Big]. \end{aligned} \tag{14}$$

In y, we stack the view pose x of the source image and the landmark positions l_i of the $N = |\mathcal{A}_F|$ associated interest points. This non-linear least squares problem can be written in the form

$$L_F(y) = const. + \frac{1}{2} e_F(y)^T W_F e_F(y), \tag{15}$$

where $e_F(y)$ stacks the individual residuals $f - h(x,l)$ of the interest points in both images, and W_F is a block-diagonal matrix with corresponding inverse covariances Σ_f^{-1} on the diagonal.
We use the LM method to optimize for the view pose of the source image and the landmark positions concurrently. The Jacobian $J_F(y)$ of $e_F(y)$ has a special structure,

$$J_F(y) := -\left(\begin{array}{c|ccc} \frac{dh}{dx}(x,l_1) & \frac{dh}{dl}(x,l_1) & 0 & \cdots & 0 \\ & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \frac{dh}{dx}(x,l_N) & 0 & \cdots & 0 & \frac{dh}{dl}(x,l_N) \\ \hline & \frac{dh}{dl}(0,l_1) & 0 & \cdots & 0 \\ & 0 & \ddots & \ddots & \vdots \\ 0 & \vdots & \ddots & \ddots & 0 \\ & 0 & \cdots & 0 & \frac{dh}{dl}(0,l_N) \end{array}\right). \tag{16}$$

The LM update is

$$\Delta y = (H_F(y) + \lambda I)^{-1} b_F(y) \tag{17}$$

with $H_F(y) = J_F(y)^T W_F J_F(y)$ and $b_F(y) = J_F(y)^T W_F e_F(y)$. Also $H_F(y)$ has a special form, i.e.

$$H_F(y) = \left(\begin{array}{c|ccc} H_{F,xx} & H_{F,xl_1} & \cdots & H_{F,xl_N} \\ \hline H_{F,l_1x} & H_{F,l_1l_1} & 0 & \cdots & 0 \\ & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ H_{F,l_Nx} & 0 & \cdots & 0 & H_{F,l_Nl_N} \end{array}\right), \tag{18}$$

where we define

$$\begin{aligned} H_{F,xx} &:= \sum_{i=1}^{N} \left(\frac{dh}{dx}(x,l_i)\right)^T \Sigma_{f_{s,i}}^{-1} \left(\frac{dh}{dx}(x,l_i)\right), \\ H_{F,xl_i} &:= \left(\frac{dh}{dx}(x,l_i)\right)^T \Sigma_{f_{s,i}}^{-1} \left(\frac{dh}{dl}(x,l_i)\right), \\ H_{F,l_ix} &:= \left(\frac{dh}{dl}(x,l_i)\right)^T \Sigma_{f_{s,i}}^{-1} \left(\frac{dh}{dx}(x,l_i)\right), \\ H_{F,l_il_i} &:= \left(\frac{dh}{dl}(x,l_i)\right)^T \Sigma_{f_{s,i}}^{-1} \left(\frac{dh}{dl}(x,l_i)\right) \\ &\quad + \left(\frac{dh}{dl}(0,l_i)\right)^T \Sigma_{f_{t,i}}^{-1} \left(\frac{dh}{dl}(0,l_i)\right). \end{aligned} \tag{19}$$

Hence, the LM update step can be subdivided into two steps. The first step updates the pose

$$\Delta x = S_F^{-1} \left(b_F(x) - \sum_{i=1}^{N} H_{F,xl_i} \left(H_{F,l_il_i} + \lambda I\right)^{-1} b_F(l_i) \right), \tag{20}$$

using the Schur complement

$$S_F := H_{F,xx} + \lambda I - \sum_{i=1}^{N} H_{F,xl_i} \left(H_{F,l_il_i} + \lambda I\right)^{-1} H_{F,l_ix} \tag{21}$$

of $H_F(x) + \lambda I$. The update on the landmark positions L is individual in each landmark l_i,

$$\Delta l_i = \left(H_{F,l_il_i} + \lambda I\right)^{-1} \left(b_F(l_i) - H_{F,xl_i} \Delta x\right). \tag{22}$$

This reduces run-time complexity from quadratic to linear in the number of interest point matchings. We neglect matches with a likelihood below a threshold in each iteration of the LM optimization to improve accuracy and robustness for outliers.

3.3 Combined Registration

Considering cues of dense image and sparse interest point registration concurrently, the likelihood of source image view pose and landmark positions is

$$\begin{aligned} p(x_s, L_s, L_t \mid x_t = 0, M_s, M_t, F_s, F_t) = \\ \eta\, p(M_s \mid x_s, M_t)\, p(F_s, F_t \mid x_s, x_t = 0, L_s, L_t) \\ p(x_s)\, p(L_s)\, p(L_t). \end{aligned} \tag{23}$$

The negative logarithm of this likelihood sums the objectives in Eqs. (4) and (14),

$$L(y) = L_M(x) + L_F(y) = const. + e(y)^T W e(y), \tag{24}$$

where $e(y)$ stacks the residuals $e_M(x)$ and $e_F(y)$, and W is a block-diagonal matrix composed of W_M and W_F. The LM update is

$$\Delta y = (H(y) + \lambda I)^{-1} b(y). \tag{25}$$

The matrix $H(y)$ is still sparse and decomposes into the form of Eq. (18). Compared with $H_F(y)$, the entries involving landmark positions remain the same, i.e., $H_{xl_i} = H_{F,xl_i}$, $H_{l_ix} = H_{F,l_ix}$, and $H_{l_il_i} = H_{F,l_il_i}$. Component H_{xx} is

$$H_{xx} = H_{F,xx} + \sum_{(s_s,s_t)\in\mathcal{A}_S} \left(\frac{dg}{dx}(x,s_s)\right)^T \Sigma(x;s_s,s_t)^{-1} \left(\frac{dg}{dx}(x,s_s)\right). \tag{26}$$

Hence, we can also apply the Schur decomposition of $H(y) + \lambda I$ to efficiently compute individual updates on the view pose and the landmark positions as in Eqs. (20) and (22).

4 Experiments

We evaluate our approach with an Intel Core i7-4770K QuadCore CPU with a maximum clock rate of 3.50 GHz. We used sequences of the RGB-D benchmark [16] which includes evaluation measures and pose ground truth captured with an optical motion capture system. The selected sequences provide diversity in scenes such as close or far average measurements, well or less textured scenes, or scenes with little or strong shape variations. We compare our combined approach (MRS+IP) with dense RGB-D registration using MRSMaps, our sparse interest point matching method alone (IP), warp [11], fovis [5], and GICP [10]. If not stated otherwise, an open-source implementation of warp contained in the OpenCV library has been used.

Table 1 summarizes average run-time of several approaches. Pure sparse interest point registration methods such as IP or fovis, achieve frame rates beyond 30 Hz. The frame rate of our method is about 15.6 Hz in average which is slightly lower than warp or MRSMap registration without interest points.

Tables 2 and 3 report median and minimum accuracy, respectively. The combination of MRSMap and interest point registration often improves the performance of the individual approaches, while it retains their individual strengths. This can also be seen from the distribution of errors in Fig. 3. While the dense method is very accurate and robust in close-by scenes with less noisy depth (e.g., fr2_desk), sparse interest point matching also succeeds if distant scenery is observed with noisy depth (e.g., fr2_large_with_loop). While not every frame can be processed in real-time, Fig. 2 demonstrates that our method is suitable for large frame skips, far beyond real-time requirements (2 frames skipped).

Table 1: Comparison of average run-time.

method	MRSMap	IP	MRS+IP	warp	fovis
run-time in ms	49.9	13.3	64.2	54.5	**8.3**
frame rate in Hz	20.0	75.2	15.6	18.3	**120.5**

Table 2: Median relative pose error (RPE) in mm.

sequence	MRSMap	IP	MRS+IP	warp	fovis
fr1 360	5.0	6.3	**4.8**	5.9	7.1
fr1 desk	**4.6**	7.2	5.1	5.8	6.3
fr1 room	**3.5**	4.9	3.7	4.6	5.4
fr1 rpy	3.0	3.6	**2.8**	5.1	5.4
fr1 xyz	**2.4**	4.3	2.7	4.1	4.6
fr2 360 hemisphere	27.3	**9.1**	9.2	40.6	10.4
fr2 desk	2.2	2.4	**1.9**	2.1	2.5
fr2 large with loop	25.8	10.2	**9.3**	94.7	12.1
fr2 pioneer slam 2	11.4	**4.8**	6.1	6.4	7.7
fr2 rpy	1.7	1.4	**1.0**	1.7	1.7
fr2 xyz	1.6	1.5	**1.0**	2.0	1.9
fr3 nostruct notext far	**9.3**	21.1	12.2	40.4	11.3
fr3 nostruct notext near	15.3	11.8	13.6	28.2	**11.2**
fr3 nostruct text far	18.1	19.0	**16.3**	19.2	20.8
fr3 nostruct text near	11.6	**6.3**	6.8	7.0	7.3
fr3 struct notxt far	**2.2**	10.5	2.5	8.6	9.1
fr3 struct notxt near	2.1	12.7	**2.0**	8.6	9.3
fr3 struct txt far	**5.5**	6.9	6.7	8.1	8.8
fr3 struct txt near	**3.2**	4.4	4.1	5.9	6.5

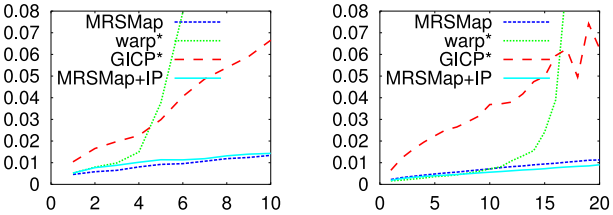

Figure 2: Median trans. error (m) for frame skips on fr1_desk (left) and fr2_desk (right) (*from [11]).

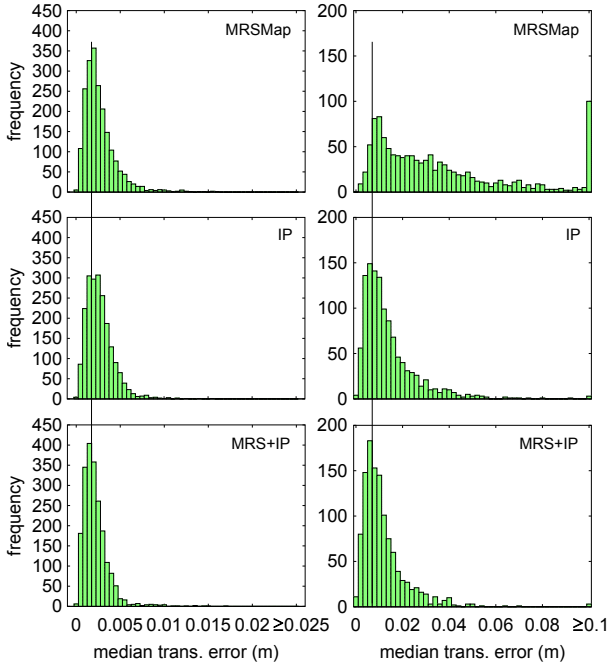

Figure 3: Histograms of median translational error on fr2_desk (left) and fr2_large_with_loop (right) sequence.

Table 3: Maximum relative pose error (RPE) in mm.

sequence	MRSMap	IP	MRS+IP	warp	fovis
fr1 360	**41.8**	169.4	95.1	75.4	43.1
fr1 desk	**25.9**	45.4	39.8	131.8	34.2
fr1 room	**45.0**	103.5	78.7	167.8	55.1
fr1 rpy	**28.9**	38.3	50.6	41.8	38.7
fr1 xyz	**9.6**	18.7	13.6	18.1	25.8
fr2 360 hemisphere	6296	2396	1823	5.6e5	**537.7**
fr2 desk	17.0	15.3	16.9	**14.1**	15.5
fr2 large with loop	2177	202.4	**144.6**	5.1e5	220.6
fr2 pioneer slam 2	908.6	5441	904.8	1.6e5	**902.7**
fr2 rpy	30.0	**8.0**	15.9	189.5	11.0
fr2 xyz	27.3	6.0	**5.5**	8.8	9.9
fr3 nostruct notext far	**49.4**	353.8	**49.4**	6.0e4	108.4
fr3 nostruct notext near	**57.8**	1198	65.4	3.2e4	79.3
fr3 nostruct text far	**57.7**	2375	67.1	1230	101.5
fr3 nostruct text near	60.8	**39.2**	62.1	100.5	41.6
fr3 struct notxt far	17.1	239.4	**13.2**	2579	62.4
fr3 struct notxt near	13.2	8493	**11.4**	1108	86.9
fr3 struct txt far	**23.0**	23.7	35.6	39.0	45.2
fr3 struct txt near	**14.5**	32.5	26.7	34.8	38.2

5 Conclusions

In this paper, we propose an approach to visual odometry with RGB-D cameras that combines the complementary strengths of dense image and sparse interest point registration. Dense image registration is performed by transforming the full RGB-D image content into multi-resolution surfel maps, and by registering these representations of successive images. Sparse ORB interest points are detected in texture corners in the RGB image and matched between the images using local texture descriptors and reprojection distance. Bundle adjustment then optimizes for the pose difference between the images and the positions of the interest points.

Both objectives are formulated in a probabilistic way and have been integrated in a single optimization framework. We evaluate our approach on a RGB-D benchmark dataset with state-of-the-art approaches and demonstrate superior results in accuracy over both sparse or dense approaches alone. Our approach is fast enough to perform real-time visual odometry for 640×480 RGB-D video on a CPU.

In future work, we plan to integrate our approach for simultaneous localization and mapping (SLAM) with RGB-D cameras. Also the use for motion segmentation of multiple moving rigid objects [13] and deformable registration [14] is a potential direction for future research.

References

[1] P. J. Besl and N. D. McKay. A method for registration of 3-D shapes. *IEEE Trans. on Pattern Anal. and Mach. Intell. (PAMI)*, 14(2):239–256, 1992.

[2] J. Civera, A.J. Davison, and J. Montiel. Inverse depth parametrization for monocular SLAM. *IEEE Trans. on Rob.*, 24(5):932–945, 2008.

[3] D. Droeschel, S. May, D. Holz, P. Ploeger, and S. Behnke. Robust ego-motion estimation with ToF cameras. In *Proc. of Europ. Conf. on Mobile Robots (ECMR)*, 2009.

[4] A. Howard. Real-time stereo visual odometry for autonomous ground vehicles. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 3946–3952, 2008.

[5] A. S. Huang, A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy. Visual odometry and mapping for autonomous flight using an RGB-D camera. In *ISRR*, 2011.

[6] C. Kerl, J. Sturm, and D. Cremers. Robust odometry estimation for RGB-D cameras. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2013.

[7] M. Magnusson, T. Duckett, and A. J. Lilienthal. Scan registration for autonomous mining vehicles using 3D-NDT. *Journal of Field Robotics*, 24(10):803–827, 2007.

[8] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 652–659, 2004.

[9] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An efficient alternative to SIFT or SURF. In *ICCV*, 2011.

[10] A. Segal, D. Haehnel, and S. Thrun. Generalized-ICP. In *Proc. of Robotics: Science and Systems (RSS)*, 2009.

[11] F. Steinbruecker, J. Sturm, and D. Cremers. Real-time visual odometry from dense RGB-D images. In *ICCV Workshops*, 2011.

[12] T. Stoyanov, M. Magnusson, H. Andreasson, and A. J Lilienthal. Fast and accurate scan registration through minimization of the distance between compact 3D NDT representations. *Int. J. of Robotics Research*, 31(12):1377–1393, 2012.

[13] J. Stückler and S. Behnke. Efficient dense 3D rigid-body motion segmentation in RGB-D video. In *Proc. of British Mach. Vision Conf. (BMVC)*, 2013.

[14] J. Stückler and S. Behnke. Efficient deformable registration of multi-resolution surfel maps for object manipulation skill transfer. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2014.

[15] J. Stückler and S. Behnke. Multi-resolution surfel maps for efficient dense 3D modeling and tracking. *Journal of Visual Communication and Image Representation*, 25(1):137–147, 2014.

[16] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Proc. of IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, 2012.