Information-Driven Direct RGB-D Odometry

Alejandro Fontán[1,2] alejandro.fontanvillacampa@dlr.de
Javier Civera[1] jcivera@unizar.es
Rudolph Triebel[2,3] rudolph.triebel@dlr.de

[1] University of Zaragoza [2] German Aerospace Center (DLR)
[3] Technical University of Munich

Abstract

This paper presents an information-theoretic approach to point selection for direct RGB-D odometry. The aim is to select only the most informative measurements, in order to reduce the optimization problem with a minimal impact in the accuracy. It is usual practice in visual odometry/SLAM to track several hundreds of points, achieving real-time performance in high-end desktop PCs. Reducing their computational footprint will facilitate the implementation of odometry and SLAM in low-end platforms such as small robots and AR/VR glasses. Our experimental results show that our novel information-based selection criteria allows us to reduce the number of tracked points an order of magnitude (down to only 24 of them), achieving an accuracy similar to the state of the art (sometimes outperforming it) while reducing 10× the computational demand.

1. Introduction

In the last years, we have witnessed an impressive progress in the accuracy and robustness of visual odometry and Simultaneous Localization and Mapping (SLAM) [17, 19, 18, 9, 20]. This boost in the performance has enabled the transfer of visual odometry and SLAM to several commercial products related to augmented reality (AR), virtual reality (VR) and robotics.

In spite of their respective successes, visual odometry and SLAM are still facing significant challenges. The high computational demand of the state of the art is among the most critical ones for a widespread use in real applications. The embodiment of localization and mapping algorithms into small robotic/AR/VR platforms will impose constraints on their computational and memory footprints [8]. Most algorithms currently require a hardware that exceeds the capabilities of many existing and foreseeable platforms.

Find more information in our project website rmc.dlr.de/rm/en/staff/alejandro.fontanvillacampa/IDNav

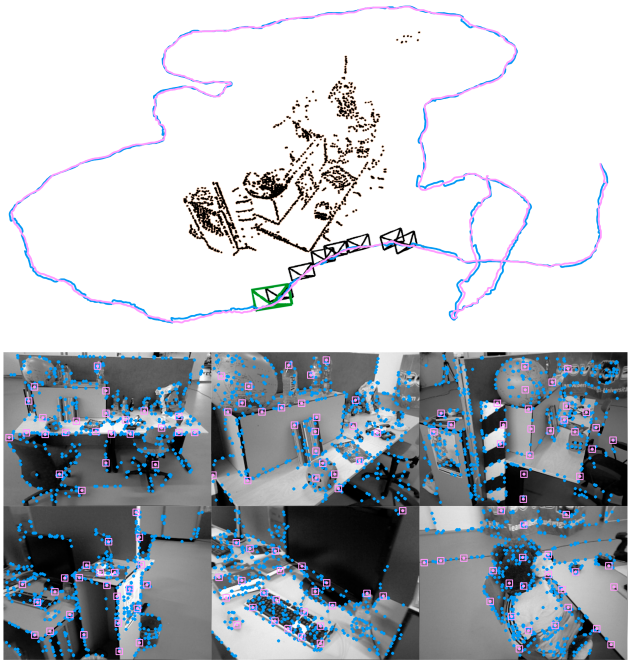

Figure 1: **Top:** Trajectories and maps estimated by our RGB-D odometry (ID-RGBDO) in two cases: tracking 500 image points (blue), and tracking only the 24 most informative points (magenta) with considerable computational savings. The difference between the two is almost unnoticeable. **Bottom:** Sample frames and tracks for the 500 points case (blue dots) and the 24 most informative ones (magenta squares).

In this paper we aim to drastically reduce the computational load of direct RGB-D odometry with a negligible loss in accuracy. For that, we propose a novel and efficient information-based criterion to keep only the most informative point in the local Bundle Adjustment and pose tracking optimizations. We implemented a RGB-D odometry (that we denote ID-RGBDO) and evaluated our approach in the TUM dataset, demonstrating that we can achieve substantial

reductions in the number of tracked features without noticeably degrading the accuracy. We outperform the naive selection approaches used in the literature, that mainly select points on a grid to maximize coverage.

Observe the two estimated trajectories in Figure 1, one tracking the 24 most informative points and the second one 500 points –a reasonable number in the state of the art. Notice that they have almost the same accuracy, but the one using 24 points requires roughly $10\times$ less computation. Our proposed information criteria are able to select the small set of highly informative points that makes this possible.

2. Related work

Graph reduction is a relevant topic in the SLAM community, with a considerably large literature [12, 1, 11]. We focus here on the main approaches using information theory and in particular those developed for visual SLAM.

Information was first used in EKF-based monocular SLAM in [7] in order to guide sequential search. Based on it, [2, 3, 10] introduced a multi-hypothesis formulation able to address ambiguous cases robustly. An information analysis of filtering and Bundle Adjustment was used in [23] to prove the advantages of the latter. Up to our knowledge, ours is the first work that addresses information in a direct odometry framework.

Information-based approaches have been also used in laser-based SLAM. [13] proposes a method to add only non-redundant and informative links to a pose graph, and [16] uses mutual information to remove low informative laser scans from the graph. The approach in [4] is able to reduce not only poses, but also landmarks, based on information theoretic criteria. [26, 28] use the Kullback-Leibler divergence to sparsify a SLAM graph.

3. Notation and fundamentals

Our direct information-driven odometry minimizes the photometric reprojection error in a sliding window of frames. Our formulation, based on direct Bundle Adjustment and Tracking, is related to recent approaches to direct visual odometry and SLAM, namely [9, 6, 15, 14]. However, we implemented ID-RGBDO in order to have a higher degree of control in the evaluation. Notice, in any case, that our contribution can be applied to any RGB-D odometry system and should give similar improvements.

This section will cover the necessary background and notation, and the specifics of our RGB-D odometry and contributions will be detailed in Section 4 (camera pose tracking) and Section 5 (sliding-window Bundle Adjustment).

3.1. Photometric model

Point representation. For a point p, its image coordinates are denoted as $\mathbf{p} = \begin{bmatrix} p_u & p_v \end{bmatrix}^\top \in \mathbb{R}^2$ and its inverse depth in the camera frame as $d \in \mathbb{R}$. For its photometric appearance, we use a set of intensity values spread in a patch centered in $\mathbf{p}$ [9].

Keyframe representation. A keyframe j is defined by its RGB-D channels, its 6DOF camera pose as a transformation matrix $\mathbf{T} \in \mathbf{SE}(3)$, two brightness parameters $\{a_j, b_j\}$ and a set of reference points to track. The Lie-algebras pose-increments $\widehat{\mathbf{x}}_{\mathfrak{se}(3)} \in \mathfrak{se}(3)$, with $\widehat{\cdot}_{\mathfrak{se}(3)}$ being the mapping operator from the vector to the matrix representation of the tangent space [22], are expressed as a vector $\mathbf{x} \in \mathbb{R}^6$. During the optimization, we update the transformations at step (k) using left matrix multiplication and the exponential map operator $\exp(\cdot)$, i.e.,

$$\mathbf{T}^{(k+1)} = \exp(\widehat{\mathbf{x}}_{\mathfrak{se}(3)}) \cdot \mathbf{T}^{(k)}. \tag{1}$$

Residual function. The photometric residual r_i of an image point p_i in a frame i is the intensity difference with the corresponding point in a reference keyframe j, combined with an affine brightness transformation and a robust norm [9]

$$r_i = \left\| e^{-a_j}(I_j(\mathbf{p}_j) - b_j) - e^{-a_i}(I_i(\mathbf{p}_i) - b_i) \right\|_\gamma. \tag{2}$$

Although some works use the t-distribution [14, 15], we observed a higher accuracy using the Huber norm (as in [9]) and saturating large values (as in [5]).

The image points $\mathbf{p}_i$ and $\mathbf{p}_j$ are related by

$$\mathbf{p}_i = \Pi(\mathbf{R}\Pi^{-1}(\mathbf{p}_j, d_j) + \mathbf{t}), \tag{3}$$

where $\Pi(\mathbf{P})$ projects in the image plane the point $\mathbf{P}$ in the camera frame; and $\Pi^{-1}(\mathbf{p}, d)$ back-projects the image point with coordinates $\mathbf{p}$ at inverse depth d. $\mathbf{R} \in \mathbf{SO}(3)$ and $\mathbf{t} \in \mathbb{R}^3$ are the relative rotation and translation between keyframe j and frame i.

Optimization. We do Gauss-Newton optimization, that can be written as

$$(\mathbf{J}^T \mathbf{\Sigma}_r^{-1} \mathbf{J})\mathbf{y} = -\mathbf{J}^T \mathbf{\Sigma}_r^{-1} \mathbf{r}, \tag{4}$$

where the rows of the matrix $\mathbf{J} = \begin{bmatrix} \mathbf{J}_x & \mathbf{J}_d & \mathbf{J}_{a,b} \end{bmatrix} \in \mathbb{R}^{n \times m}$ contains the derivatives of the residual function (equation (2)) with respect to the Lie-algebra increments $\mathbf{J}_x$, the point inverse depths $\mathbf{J}_d$ and the photometric parameters $\mathbf{J}_{a,b}$. The diagonal matrix $\mathbf{\Sigma}_r \in \mathbb{R}^{n \times n}$ contains the covariances σ_r^2 of the photometric residuals. The residual vector $\mathbf{r} \in \mathbb{R}^n$ stacks the n individual residuals to minimize. $\mathbf{y} \in \mathbb{R}^m$ stands for the state correction containing the increments for poses, inverse depths and photometric parameters.

Residual covariance. Our residual covariance σ_r^2 includes the impact from geometry and appearance. We propose to model it by multiplying a photometric term σ_Φ^2 with a geometric one $h(\delta A)$ that comes from projecting a differential area surrounding the 3D point:

$$\sigma_r^2 = h(\delta A) \cdot \sigma_\Phi^2. \tag{5}$$

Figure 2 illustrates how the differential area around a point changes with the viewpoint. This change δA can be modeled as the determinant of the derivative of the image point $\mathbf{p}_i$ in frame i with respect to the coordinates $\mathbf{p}_j$ of the corresponding point in a reference keyframe j:

$$\delta A = \left| \frac{\partial \mathbf{p}_i}{\partial \mathbf{p}_j} \right|. \tag{6}$$

With this, we define the geometric weight $h(\delta A)$ as the following function, that penalizes the residual covariance for large perspective distortions

$$h(\delta A) = e^{c_h (\delta A - 1)^2}, \tag{7}$$

where c_h is a constant to ponder the influence of the model.

The photometric term σ_Φ^2 is computed from a first order propagation of the inverse depth covariance σ_d^2

$$\sigma_\Phi^2 \approx \left[\left(g_u \frac{\partial p_u}{\partial d} \right)^2 + \left(g_v \frac{\partial p_v}{\partial d} \right)^2 \right] \sigma_d^2, \tag{8}$$

where the intensity gradients $\begin{bmatrix} g_u & g_v \end{bmatrix}$ come from a first-order Taylor expansion of the intensity in the vicinity of $\mathbf{p}$

$$I(\mathbf{p} + \delta\mathbf{p}) \approx I(\mathbf{p}) + \begin{bmatrix} g_u & g_v \end{bmatrix} \begin{bmatrix} \delta p_u \\ \delta p_v \end{bmatrix}. \tag{9}$$

Using the stereo model for RGB-D cameras based on structured light patterns, and assuming a focal length f and a baseline b, the inverse depth error covariance σ_d is [6]

$$\sigma_d = \frac{1}{fb} \sigma_{px}, \tag{10}$$

where σ_{px} is the disparity error.

3.2. Information metrics

Information theory provides a mean to quantify and formalize all processes related with information. In the context of SLAM the special case of multivariate Gaussians is comprehensively well founded [7, 2]. The information-driven formulation proposed in this paper is based on the following classical information metrics.

Differential entropy of a k-dimensional Gaussian distribution $\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$. It can be seen as the expected information content of a future event, given the set of possible results and their probability distribution [2]

$$H(\mathbf{X}) = \frac{1}{2} \log((2\pi e)^k |\boldsymbol{\Sigma}_X|). \tag{11}$$

Entropy reduction, which is the relative difference between two Gaussian distributions

$$\Delta H(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) - H(\mathbf{Y}) = \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_X|}{|\boldsymbol{\Sigma}_Y|}, \tag{12}$$

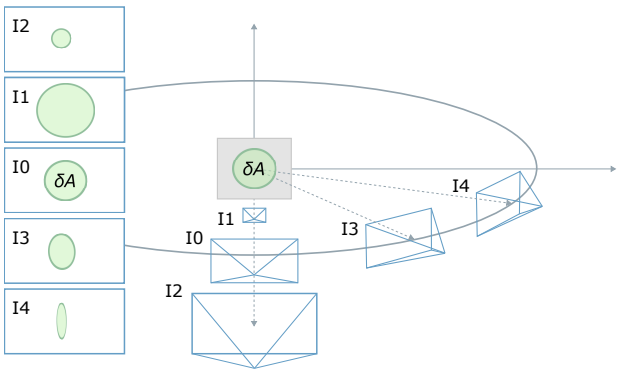

Figure 2: Illustration of the projective distortion of a differential 2D patch δA.

that is, how much more accuracy is obtained by measuring $\mathbf{Y}$ instead of $\mathbf{X}$ [23].

Conditional covariance. Assuming $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^M$ are combined in a joint Gaussian $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z)$,

$$\boldsymbol{\Sigma}_Z = \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}, \tag{13}$$

the conditional covariance $\boldsymbol{\Sigma}_{x \backslash y}$ of $\mathbf{x}$ given $\mathbf{y}$, is the Schur complement of $\boldsymbol{\Sigma}_{yy}$ in $\boldsymbol{\Sigma}_Z$:

$$\boldsymbol{\Sigma}_{x \backslash y} = \boldsymbol{\Sigma}_x^* = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{yx}. \tag{14}$$

Mutual information between two random variables. It measures how much knowing one of the variables reduces the uncertainty about the other [21]:

$$MI(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_{xx}|}{|\boldsymbol{\Sigma}_x^*|}. \tag{15}$$

Throughout the paper the entropy is measured in absolute numbers of bits (*i.e.*, *log* stands for base-2 logarithm).

4. ID-RGBDO - tracking

We now apply to direct RGB-D pose tracking the ideas above, presented in this section theoretically and evaluated experimentally in Section 6.1.

4.1. Informative point selection

Most direct methods are either dense or semi-dense approaches, aiming to use as many pixels as possible. In order to achieve real-time performance, they rely on a high-end computational platform or make use of sub-optimal approximations.

Sparse direct methods, on the contrary, reduce the number of points by extracting those with a significant photometric gradient (Step 1) and widely spread across the image

(Step 2). These heuristics work reasonably well in a wide array of scenarios, although several aspects are left unexplored: Are we reaching the lowest possible error given our data? Are we using redundant information and hence wasting computation? Is there enough visual information for the problem to be well conditioned at all times? Our proposal is to add an algorithm (Step 3) that selects points in a manner that, together with the previous two conditions, maximizes the entropy of the camera pose.

The camera pose entropy depends on the determinant of its covariance matrix $\mathbf{\Sigma}_x$, as shown in equation (11). Each point p contributes with $\Delta_p\mathbf{\Lambda}_x$ to the information matrix $\mathbf{\Lambda}_x$, that can be obtained as the sum of the Jacobian autoproduct for the whole set of points $\mathcal{P}$

$$\mathbf{\Lambda}_x = \mathbf{\Sigma}_x^{-1} = \sum_{p\in\mathcal{P}} \Delta_p\mathbf{\Lambda}_x = \sum_{p\in\mathcal{P}} \mathbf{j}_{x,p}^T \sigma_r^{-2} \mathbf{j}_{x,p}, \tag{16}$$

where $\mathbf{j}_{x,p}$ is the row of the Jacobian $\mathbf{J}_x$ that corresponds to the photometric residual of point p.

The addition of point p also yields to a variation of the information matrix determinant $\Delta_p|\mathbf{\Lambda}_x|$, that has the very satisfying property[1] that can be expressed individually per point, depending on the p^{th} row of the Jacobian $\mathbf{j}_{x,p}$ and the current adjoint information matrix $\mathbf{\Lambda}_x^{adj}$:

$$\begin{aligned}\Delta_p|\mathbf{\Lambda}_x| &= |\mathbf{\Lambda}_x + \mathbf{j}_{x,p}^T\sigma_r^{-2}\mathbf{j}_{x,p}| - |\mathbf{\Lambda}_x| \\ &= |\mathbf{\Lambda}_x||\mathbf{I} + \mathbf{\Lambda}_x^{-1}\mathbf{j}_{x,p}^T\sigma_r^{-2}\mathbf{j}_{x,p}| - |\mathbf{\Lambda}_x| \\ &= |\mathbf{\Lambda}_x|(1 + \sigma_r^{-2}\mathbf{j}_{x,p}\mathbf{\Lambda}_x^{-1}\mathbf{j}_{x,p}^T) - |\mathbf{\Lambda}_x| \\ &= \sigma_r^{-2}\mathbf{j}_{x,p}\mathbf{\Lambda}_x^{adj}\mathbf{j}_{x,p}^T.\end{aligned} \tag{17}$$

Based on this, our algorithm works as follows. We start from a pre-filtered set of high-gradient pixels by using a grid with a region-adaptative gradient threshold (as in [9]). We prioritize points that belong to Canny edges (as in [6]) but also keep some points in areas with weaker gradient (Step 2). From here we follow Algorithm 1. We choose for each degree of freedom (each of the six columns of $\mathbf{J}_x$) the image point p with maximum derivative, and build with them an initial information matrix. We then iteratively select the point that maximizes the following function (Step 3)

$$f(p\in\mathcal{P}, z, \mathbf{\Lambda}_x) = \Delta_p|\mathbf{\Lambda}_x| + \frac{1}{c_z(z_p - z)^2 + 1}. \tag{18}$$

The first addend in the function takes into account the increment of information described above. The second one contributes to spread the points in the image, in order to compensate for effects that are not modeled in the projection function. This last expression increases its value when the radial coordinate z_p of a point p approaches z. z is initialized at the image border and its value is reduced by Δz for each selected point until reaching the principal point. c_z models the importance of this second term with respect to the information increment of each point.

[1]For simplicity we applied a consequence of the Sylvester's determinant theorem $|(I_m + cr)| = 1 + rc$.

Algorithm 1 Informative point selection.

function SELECT INF. POINTS $(m, \mathcal{P}, \boldsymbol{J}_x)$
  ▷ m = number of points to be selected
  ▷ $\mathcal{P}$ = set of available points
  $\mathcal{Q} \leftarrow \emptyset$ ▷ $\mathcal{Q}$ = set of selected points
  $\mathbf{\Lambda}_x \leftarrow \mathbf{0}$ ▷ Init. Information matrix
  for $k \leftarrow 1$ to DOF **do** ▷ DOF = 6
    i $\leftarrow$ arg max $(\mathbf{j}_{x,p}[k]))$
    $\mathbf{\Lambda}_x \leftarrow \mathbf{\Lambda}_x + \Delta_p\mathbf{\Lambda}_x(\mathcal{P}[i])$
    $\mathcal{Q} \leftarrow \mathcal{Q} \cup \mathcal{P}[i]$ ▷ Add selected point
    $\mathcal{P} \leftarrow \mathcal{P} - \mathcal{P}[i]$
  end for
  ▷ Informative selection
  $z \leftarrow$ image border
  while $(\mathcal{P} \neq \emptyset \ \ \& \ \ \dim(\mathcal{Q}) < m)$ **do**
    i $\leftarrow$ arg max $(f(\mathcal{P}, z, \mathbf{\Lambda}_x))$ ▷ Most inf. point
    $\mathbf{\Lambda}_x \leftarrow \mathbf{\Lambda}_x + \Delta_p\mathbf{\Lambda}_x(\mathcal{P}[i])$
    $\mathcal{Q} \leftarrow \mathcal{Q} \cup \mathcal{P}[i]$
    $\mathcal{P} \leftarrow \mathcal{P} - \mathcal{P}[i]$
    $z \leftarrow z - \Delta z$
  end while
  return $\mathcal{Q}$
end function

4.2. Pose estimation

With our selected set of points, we aim to find the motion $\Delta\mathbf{x}$ between the closest keyframe and the current frame, that minimises the photometric residual vector $\mathbf{r}$ (see equation (2)). This optimization is initialized with a constant velocity model and a multi-scale pyramid image to aid convergence.

The addition of a kinematic model has been extensively used in odometry and SLAM. [15] showed that adding a motion prior in direct odometry helps in cases such as lack of texture, motion blur or dynamic content. The motion estimation with such a prior can be written as

$$(\mathbf{J}_x^T\mathbf{\Sigma}_r^{-1}\mathbf{J}_x + \mathbf{\Sigma}_m^{-1})\Delta\mathbf{x} = -\mathbf{J}_x^T\mathbf{\Sigma}_r^{-1}\mathbf{r} + \mathbf{\Sigma}_m^{-1}(\mathbf{x}_{t-1} - \mathbf{x}_t^{(k)}), \tag{19}$$

where $\mathbf{x}_{t-1}$ and $\mathbf{x}_t^{(k)}$ are the camera speeds for the previous frame and the last iteration of the current frame respectively. The diagonal covariance matrix $\mathbf{\Sigma}_m \in \mathbb{R}^{6\times6}$ models the strenght of the motion prior. As explained in [15], assigning high values to this covariance matrix decreases the influence of the motion prior with respect to the image residuals and vice versa. Tuning the values of the matrix is left to the

knowledge of the agent motion or the availability of another type of sensor (such as an IMU).

As in [9], we consider outliers and discard those points whose photometric error exceeds three times the standard deviation of the distribution. This reduces the effect that occlusions and false matches have on the accuracy and robustness of the odometry.

5. ID-RGBDO - windowed optimization

5.1. Keyframe creation

There are several different strategies to select keyframes from an image sequence, with the aim of estimating a local map. Conservative strategies privilege the use of already existing keyframes before constituting a new one. Only if there is no previous candidate with enough overlap, the system assumes that a new area is being explored and creates a new keyframe [6]. An alternative approach is to first initialize a large number of keyframes and later, in the local mapping step, cull down and marginalize the redundant ones [9, 17]. We use this latest method, as it makes the tracking more robust to rapid motions and allows to maintain a sliding window optimization with close keyframes.

Keyframe creation is mainly associated with visual change, related to rotation and/or translation or due to lighting changes. This task is commonly addressed by setting thresholds to the following criteria: 1) a maximum rotation and translation distance, 2) a minimum number of inlier points, 3) after a fixed number of tracked frames or, 4) due to a strong change in brightness parameters.

Similar to [14], we propose the keyframe creation to be associated with the entropy reduction ΔH of the camera pose. Differently from [14], we obtain the entropy reduction independently for each degree of freedom $x \in \mathbf{x}$ using the Schur complement on the covariance matrix. We set the entropy $H^*(x_0)$ of the first frame immediately after the keyframe as the reference. This means that, in essence, our system creates a new keyframe when a certain entropy decrement is observed in at least one of the degrees of freedom of the camera.

$$H^*_\Delta(x, x_0) = 1 - \frac{H^*(x)}{H^*(x_0)}. \tag{20}$$

It may seem paradoxical, within this information framework, to establish a threshold for the process of keyframe creation. However, in contrast to other systems that define multiple and ambiguous thresholds, it is worth noting the entropy decrement allows us to use a single value that is related to tracking information. Disaggregating the information for each particular degree of freedom adds robustness and accuracy, as the aggregated information might compensate low information values in some degrees of freedom with higher ones in some others.

5.2. Keyframe marginalization

Keyframe marginalization is essential to keep the optimization size-bounded, enabling real-time operation [9, 18]. The marginalization criteria depend on whether we optimize a local map or a sliding window of keyframes. For the first case, the aim should be detecting and removing redundant keyframes, allowing lifelong operation in the same environment without unlimited growth of the number of keyframes unless the visual content of the scene changes [18]. The second technique, adopted by odometries, maintains a sliding window around the last keyframe, sufficiently spaced for an accurate optimization of the point depths.

Our marginalization belongs to the second group. However, instead of using a heuristically designed function to keep the keyframes spatially distributed, we use the mutual information measurement in order to delete the redundant ones.

Partial marginalization using the Schur complement. Instead of simply dropping out keyframes and points from the optimization, and in order to preserve most of the information, we substitute the non-linear terms with a linearized expression of the photometric error (as in [9, 25, 27]).

The state vector update in equation (4) is first written in the following form

$$\begin{bmatrix} \mathbf{H}_{\alpha\alpha} & \mathbf{H}_{\alpha\beta} \\ \mathbf{H}_{\beta\alpha} & \mathbf{H}_{\beta\beta} \end{bmatrix} \begin{bmatrix} \mathbf{y}_\alpha \\ \mathbf{y}_\beta \end{bmatrix} = \begin{bmatrix} \mathbf{b}_\alpha \\ \mathbf{b}_\beta \end{bmatrix}, \tag{21}$$

where α and β are the blocks of variables we would like to keep and marginalize respectively. Applying the Schur complement we obtain

$$\mathbf{H}^*_\alpha = \mathbf{H}_{\alpha\alpha} - \mathbf{H}_{\alpha\beta}\mathbf{H}^{-1}_{\beta\beta}\mathbf{H}_{\beta\alpha} \tag{22}$$

$$\mathbf{b}^*_\alpha = \mathbf{b}_\alpha - \mathbf{H}_{\alpha\beta}\mathbf{H}^{-1}_{\beta\beta}\mathbf{b}_\beta, \tag{23}$$

which represents again a linear system for the state vector update, but in this case with variables β marginalized out. We can hence write a quadratic function on $\mathbf{y}$ that can be added to the photometric error during all subsequent optimization and marginalization operations, replacing the corresponding non-linear terms:

$$r(\delta\mathbf{y}_\alpha)\big|_{\mathbf{y}_\alpha} = \frac{1}{2}\delta\mathbf{y}^T_\alpha\mathbf{H}^*_\alpha\delta\mathbf{y}_\alpha - \delta\mathbf{y}^T_\alpha\mathbf{b}^*_\alpha. \tag{24}$$

Note that partial marginalization fixes the linearization point of the variables involved, and then this would require the tangent space to remain the same over all subsequent optimization and marginalization steps. To reduce this problem we perform a relinearization of $r(\delta\mathbf{y}_\alpha)\big|_{\mathbf{y}_\alpha}$, as in [25], every time the state is updated, i.e.,

$$\begin{aligned} &r(\delta\mathbf{y}_\alpha)\big|_{\mathbf{y}_\alpha+\Delta\mathbf{y}_\alpha} \\ &= r(\Delta\mathbf{y}_\alpha)\big|_{\mathbf{y}_\alpha} + \frac{1}{2}\delta\mathbf{y}^T_\alpha\mathbf{H}^*_\alpha\delta\mathbf{y}_\alpha - \delta\mathbf{y}^T_\alpha(\mathbf{b}^*_\alpha - \mathbf{H}^*_\alpha\Delta\mathbf{y}_\alpha). \end{aligned} \tag{25}$$

Similar to [9], when dropping a keyframe we first marginalize all points referred to it and then the keyframe itself.

Redundancy detection using Mutual Information. As in [21], the redundancy $\psi(\mathcal{K}_j)$ of a keyframe with respect to the others can be expressed by

$$\psi(\mathcal{K}_j) = \sum_{i \in \mathcal{K}} MI(i, j, \mathbf{\Sigma}_{(i,j)\setminus\mathcal{K}-\{i,j\}}), \qquad (26)$$

where the Mutual Information between every pair of keyframes (i, j) is computed from their conditional covariance matrix $\mathbf{\Sigma}_{(i,j)\setminus\mathcal{K}-\{i,j\}}$ with respect to the rest. This metric is used to remove, when necessary, the less informative keyframe within the window.

6. Experimental Results

For our evaluation we use the public TUM RGB-D benchmark [24]. This dataset contains several indoor sequences, captured with an RGB-D camera and annotated with ground truth camera poses. Specifically, we use all static sequences except those beyond the range of the sensor (see Table 1 for the sequence list).

This section is divided into four sets of experiments. The first set evaluates the informative point selection procedure introduced in section 4.1. The next set analyses the keyframe creation criterion that we propose in section 5.1. The third set shows an analysis of computational performance. Finally, we compare our system against several state-of-the-art RGB-D odometry and SLAM systems.

The error metrics chosen for the following figures and tables are the translational keyframe-to-frame error (K2FE), used for evaluating our informative point selection, and the root-mean-square errors of traslational drift in m/s (RPE) and Absolute Trajectory Error (ATEs) for comparing against state-of-the-art baselines.

6.1. Informative point selection

We evaluate the performance of our system both quantitatively and qualitatively in terms of trajectory estimation and computational performance.

Figure 3 shows the translational keyframe-to-frame error (K2FE) using a number of points between 24 and 256 for all sequences we evaluated (more than $20,000$ frames). The four configurations shown refer to different alternatives for point selection: completely random (**rand**), distributed in a grid and above an intensity gradient threshold (**grid**), based on our criterion to maximize the entropy of the pose (see equation (18)) (**inf**) and with a mixed approach between the last two (**inf+grid**). The figure shows that our information-based criterion, both combined with the grid approach and not, leads to the highest accuracy. The difference between the four alternatives is smaller as the number of points increases, but the information-based selection always results in a higher accuracy. The negligible difference in accuracy between **inf** and **inf+grid** is relevant for real-time performance, as grid-based point pre-selection is significantly faster than choosing them only based on information criteria. This is why in our RGB-D odometry we adopt this mixed approach.

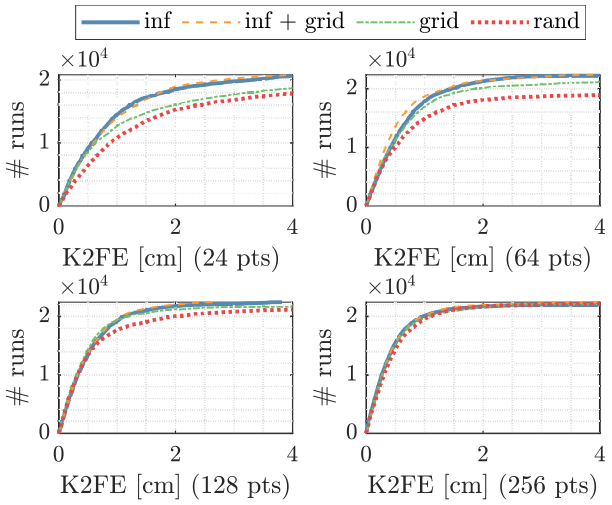

Figure 3: **Point Informative Selection**. Accumulated translational keyframe-to-frame error (K2FE) in all sequences. Different lines correspond to point selection modes.

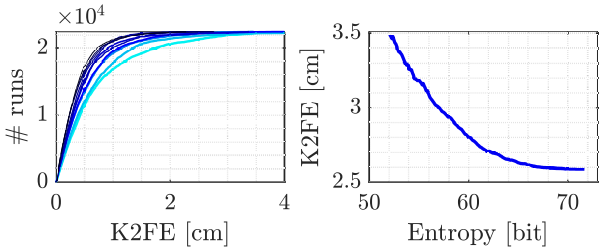

Figure 4: **Accuracy vs Entropy**. Left: Accumulated K2FE. Color degradation from black to blue indicates a higher entropy reduction. Right: Accumulated K2FE vs absolute entropy values.

The relation between entropy reduction and accuracy is shown in Figure 4. In short, the cost (the number of points needed) of improving pose accuracy increases with the absolute value of the entropy. A limitation of our current research is that, for different sequences, the specific shape of the entropy-accuracy curve is slightly different. As shown in Figure 5, two sequences with similar entropy values have

		RPE (m/s)				ATE (m)		
		[14]	[18]†	[29]	Ours	[18]†	[29]	Ours
1	fr1 desk‡	**0.024**	0.051	0.031	0.029	0.065	**0.044**	0.051
2	fr1 floor‡	0.232	0.038	**0.010**	0.011	0.061	0.021	**0.020**
3	fr1 plant‡	0.025	0.044	0.036	**0.024**	0.067	0.059	**0.039**
4	fr1 rpy‡	0.032	0.037	0.034	**0.026**	0.066	0.047	**0.045**
5	fr1 xyz‡	0.018	**0.014**	0.019	0.019	**0.009**	0.043	0.043
6	fr2 desk	-	0.030	**0.008**	0.011	0.213	0.037	**0.030**
7	fr2 dishes	-	0.035	**0.012**	0.015	0.104	**0.033**	0.041
8	fr2 rpy	-	0.004	0.004	**0.003**	**0.004**	0.007	0.007
9	fr2 xyz	-	0.005	0.004	**0.003**	0.008	0.008	**0.007**
10	fr3 cabinet	-	0.071	**0.036**	0.058	0.312	**0.057**	0.063
11	fr3 large cabinet	-	0.100	0.167	**0.049**	0.154	0.317	**0.096**
12	fr3 long office household	-	0.019	**0.010**	**0.010**	0.276	0.085	**0.038**
13	fr3 nostr. text. far	0.073	0.121	**0.035**	0.037	0.147	**0.026**	0.049
14	fr3 nostr. text. near	0.028	0.050	0.043	**0.015**	0.111	0.090	**0.062**
15	fr3 str. notext. far	0.039	**0.013**	0.027	0.016	**0.008**	0.031	0.018
16	fr3 str. notext. near	**0.021**	0.060	-	-	**0.091**	-	-
17	fr3 str. text. far	0.039	0.018	0.013	**0.012**	0.030	0.013	**0.010**
18	fr3 str. text. near	0.041	0.017	**0.010**	0.011	0.045	0.025	**0.013**

Table 1: RMSE of translational drift RPE (m/s) and ATE (m) for state-of-the-art baselines and ID-RGBDO (Ours). Remarkably, ID-RGBDO (Ours) tracks only 24 points per keyframe. † stands for ORB-SLAM2-based odometry, where loop closure was deactivated from the original implementation of [18]. ‡ stands for special initialization for tracking convergence.

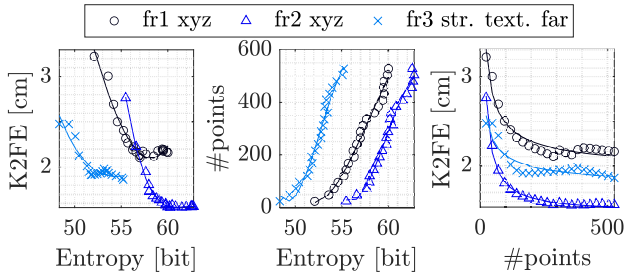

Figure 5: **Entropy reduction**. Left: translational keyframe-to-frame error (K2FE) vs. entropy reduction. Center: number of points vs. entropy reduction. Right: K2FE vs. number of points. The three different colors stand for three different sequences.

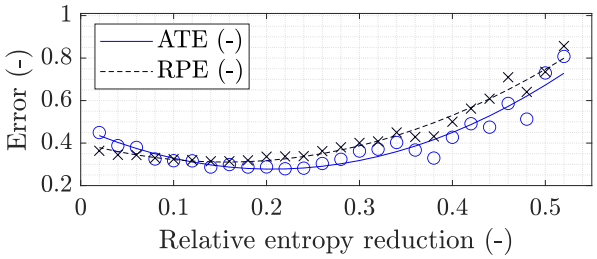

Figure 6: **Keyframe Creation**. Both RPE and ATE trajectory errors are influenced by the keyframe creation strategy. The figure shows a value for the relative entropy decrement H^*_Δ where both errors are minimal.

different translational errors. These discrepancies may be due to the need of a better photometric model, as for example scenes with strong presence of motion blur give poor performance. This is not relevant for our current selection criterion, that uses relative entropy. However, future work to understand this effect could lead to further improvements.

6.2. Informative Keyframe Creation

Here we demonstrate the adequacy of our entropy-based criterion for keyframe creation. Figure 6 shows the variation of the normalized trajectory errors (RPE and ATE), aggregated over all sequences, versus the threshold on the relative entropy reduction H^*_Δ to create a new keyframe. Low values lead to a high number of keyframes, which might increase the drift. Increasing the threshold on the relative entropy reduction decelerates the keyframe creation, reducing the overlap and increasing the error and eventually leading to tracking failure. Notice how this effect is modeled in the curves of Figure 6, and that they can be used to choose a reasonable threshold.

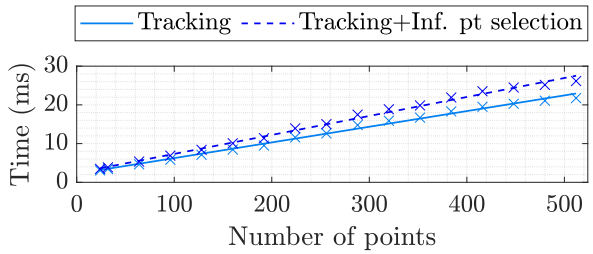

Figure 7: **Tracking cost**. Observe its linear growth with the number of points, and hence the convenience of using a small number of them. Observe also the small overhead introduced by our informative point selection.

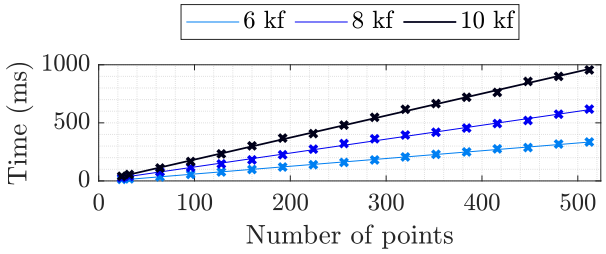

Figure 8: **Bundle Adjustment cost**. Notice the steep growth with the number of points, that our selection algorithm reduces with a minimal impact in the accuracy.

6.3. Computational Performance

We run all experiments on a laptop with an Intel Core i7-7500U CPU at 2.70 GHz and 8 GB of RAM. Figure 7 shows the linear dependence of the tracking cost (with and without informative point selection) with the number of image points. Time is reduced between $5\times$ and $10\times$ from the usual practice of tracking hundreds of points to our minimal setup of 24 points. Notice also that the overhead introduced by our informative point selection algorithm is small compared to the total tracking cost, in particular for a low number of points.

Figure 8 shows the cost of our Direct Bundle Adjustment depending on the number of points and cameras. For our minimal configuration of 24 points per keyframe, the cost is reduced approximately $10\times$ with respect to more usual setups that optimize hundreds of points.

6.4. Evaluation against SotA baselines

We compare our system against three different baselines. Firstly, against Canny-VO [29], a recent RGB-D odometry based on geometric edge alignment. Secondly, against an ORB-SLAM2-based odometry, for which the original ORB-SLAM2 [18] was used with its loop closure deactivated. And, thirdly, against DVO_SLAM [14], a dense direct RGB-D SLAM. The results for ORB-SLAM2-based odometry were taken from [29]. Table 1 shows the trajectory errors for these three baselines and ID-RGBDO. In our ID-RGBDO we use 24 points per keyframe and 8 keyframes in the sliding-window Bundle Adjustment. In the case of fr1, as these sequences have high motion blur due to quick rotations, we use initially a higher amount of points to aid tracking convergence but then within the Bundle Adjustment we stick to the 24 most informative points per keyframe and 8 keyframes configuration.

Notice how, for a large part of the fr2 and fr3 camera sequences, that are rich in texture and/or structure, our algorithm outperforms the three baselines. Our tracking fails in sequence 16, as all direct odometries do, while the feature-based ORB-SLAM2 succeeds. We have detected that this is due to the fact that the problem is not well conditioned with a photometric cost function and however it is conditioned enough if features are used. This result tells us how beneficial a mixed direct-features system managed with information measurements could be.

7. Conclusions and Future Work

In this paper we have proposed a novel criterion to select the most informative points to be tracked in a RGB-D odometry framework. We have shown experimentally that using a small number of very informative points and keyframes can have a significant impact in the computational cost of RGB-D odometry, while keeping an accuracy similar to the state of the art. Specifically, our experimental results show that tracking the 24 most informative points is enough to match the performance of the state of the art while reducing the computational cost up to a factor $10\times$.

Up to our knowledge, this is the first time that information theory is applied to direct odometry and SLAM methods. We believe that our results will facilitate the use of visual odometry and SLAM in small robotic platforms and AR/VR glasses, that are limited in computation and power.

There are several lines of research that build on and could improve the results of this work. Firstly, the development of a probabilistic photometric model could improve the accuracy of the information metrics. And secondly, we also think that further analysis on the information of the windowed keyframe optimization could offer even better results. We plan to investigate both topics in the near future.

Acknowledgments This project has received funding from the Spanish Government (PGC2018-096367-B-I00) and the Aragón Government (DGA T45 17R/FSE).

References

[1] Luca Carlone, Zsolt Kira, Chris Beall, Vadim Indelman, and Frank Dellaert. Eliminating conditionally independent sets in factor graphs: A unifying perspective based on smart factors. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4290–4297. IEEE, 2014. 2

[2] Margarita Chli and Andrew J Davison. Active matching. In *European conference on computer vision*, pages 72–85. Springer, 2008. 2, 3

[3] Margarita Chli and Andrew J Davison. Active matching for visual tracking. *Robotics and Autonomous Systems*, 57(12):1173–1187, 2009. 2

[4] Siddharth Choudhary, Vadim Indelman, Henrik I Christensen, and Frank Dellaert. Information-based reduced landmark SLAM. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4620–4627, 2015. 2

[5] Alejo Concha and Javier Civera. An evaluation of robust cost functions for RGB direct mapping. In *2015 European Conference on Mobile Robots (ECMR)*, pages 1–8. IEEE, 2015. 2

[6] Alejo Concha and Javier Civera. RGBDTAM: A cost-effective and accurate RGB-D tracking and mapping system. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6756–6763. IEEE, 2017. 2, 3, 4, 5

[7] Andrew J Davison. Active search for real-time vision. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 66–73. IEEE, 2005. 2, 3

[8] Andrew J Davison. FutureMapping: The computational structure of spatial AI systems. *arXiv preprint arXiv:1803.11288*, 2018. 1

[9] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 1, 2, 4, 5, 6

[10] Ankur Handa, Margarita Chli, Hauke Strasdat, and Andrew J Davison. Scalable active matching. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1546–1553. IEEE, 2010. 2

[11] Jerry Hsiung, Ming Hsiao, Eric Westman, Rafael Valencia, and Michael Kaess. Information sparsification in visual-inertial odometry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1146–1153. IEEE, 2018. 2

[12] Guoquan Huang, Michael Kaess, and John J Leonard. Consistent sparsification for graph optimization. In *2013 European Conference on Mobile Robots*, pages 150–157, 2013. 2

[13] Viorela Ila, Josep M Porta, and Juan Andrade-Cetto. Information-based compact pose SLAM. *IEEE Transactions on Robotics*, 26(1):78–93, 2009. 2

[14] Christian Kerl, Jürgen Sturm, and Daniel Cremers. Dense visual SLAM for RGB-D cameras. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2100–2106. IEEE, 2013. 2, 5, 7, 8

[15] Christian Kerl, Jürgen Sturm, and Daniel Cremers. Robust odometry estimation for RGB-D cameras. In *2013 IEEE International Conference on Robotics and Automation*, pages 3748–3754. IEEE, 2013. 2, 4

[16] Henrik Kretzschmar and Cyrill Stachniss. Information-theoretic compression of pose graphs for laser-based slam. *The International Journal of Robotics Research*, 31(11):1219–1230, 2012. 2

[17] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 1, 5

[18] Raul Mur-Artal and Juan D Tardós. ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 1, 5, 7, 8

[19] Taihú Pire, Thomas Fischer, Javier Civera, Pablo De Cristóforis, and Julio Jacobo Berlles. Stereo parallel tracking and mapping for robot localization. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1373–1378. IEEE, 2015. 1

[20] Tong Qin, Peiliang Li, and Shaojie Shen. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. 1

[21] Patrik Schmuck and Margarita Chli. On the redundancy detection in keyframe-based slam. In *2019 International Conference on 3D Vision (3DV)*, pages 594–603, 2019. 3, 6

[22] Hauke Strasdat. *Local accuracy and global consistency for efficient visual SLAM*. PhD thesis, Department of Computing, Imperial College London, 2012. 2

[23] Hauke Strasdat, José MM Montiel, and Andrew J Davison. Visual slam: why filter? *Image and Vision Computing*, 30(2):65–77, 2012. 2, 3

[24] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A Benchmark for the Evaluation of RGB-D SLAM Systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012. 6

[25] Vladyslav Usenko, Jakob Engel, Jörg Stückler, and Daniel Cremers. Direct visual-inertial odometry with stereo cameras. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1885–1892. IEEE, 2016. 5

[26] John Vial, Hugh Durrant-Whyte, and Tim Bailey. Conservative sparsification for efficient and consistent approximate estimation. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 886–893, 2011. 2

[27] Lukas Von Stumberg, Vladyslav Usenko, and Daniel Cremers. Direct sparse visual-inertial odometry using dynamic marginalization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2510–2517. IEEE, 2018. 5

[28] Yue Wang, Rong Xiong, Qianshan Li, and Shoudong Huang. Kullback-leibler divergence based graph pruning in robotic feature mapping. In *2013 European Conference on Mobile Robots*, pages 32–37. IEEE, 2013. 2

[29] Yi Zhou, Hongdong Li, and Laurent Kneip. Canny-VO: Visual Odometry With RGB-D Cameras Based on Geometric 3-D–2-D Edge Alignment. *IEEE Transactions on Robotics*, 35(1):184–199, 2018. 7, 8