SDICP: Semi-Dense Tracking based on Iterative Closest Points

Laurent Kneip[1,2]
laurent.kneip@anu.edu.au
Zhou Yi[1,2]
yi.zhou@anu.edu.au
Hongdong Li[1,2,3]
hongdong.li@anu.edu.au

[1)] Research School of Engineering
Australian National University
[2)] ARC Centre of Excellence
for Robotic Vision
[3)] NICTA Canberra Labs

Abstract

This paper introduces a novel strategy for real-time monocular camera tracking over the recently introduced, efficient semi-dense depth maps. We employ a geometric iterative closest point technique instead of a photometric error criterion, which has the conceptual advantage of requiring neither isotropic enlargement of the employed semi-dense regions, nor pyramidal subsampling. We outline the detailed concepts leading to robustness and efficiency even for large frame-to-frame disparities. We demonstrate successful real-time processing over very large view-point changes and significantly corrupted semi-dense depth-maps, thus underlining the validity of our geometric approach.

1 Introduction

Camera tracking denotes the continuous image-based computation of a camera's position and orientation with respect to a reference frame. This task lies at the heart of the visual odometry and visual SLAM problems, making it the key to accuracy, efficiency, and reliability in visual motion and structure estimation. Although tracking frequently exploits the temporal order and regularity of an image sequence—for instance by employing a dynamic motion model that imposes smoothness in the estimated camera trajectory—, we focus here on the most general case that does not rely on such assumptions: a single image plus a reference frame (e.g. another image) in which depth information is available. All camera tracking instances can be reduced to this basic scenario. The present paper discusses tracking of regular cameras. This significantly complicates the motion estimation process, as 3D-3D registration methods applicable to RGB-D cameras are no longer an option. A *tracker* in our context describes an algorithm that solves the *absolute pose* or *2D-3D registration* problem.

The particular form of a tracker largely depends on the form of the depth information. Traditional approaches for instance extract sparse sets of local invariant keypoints in the images, thus leading to depth information in form of a sparse point cloud. The classical way of dealing with such data consists of establishing 2D-3D correspondences, and then solving the perspective *n*-point problem. Fischler and Bolles [7] present a prominent variant able to handle outlier-affected data by solving the perspective 3-point problem within a robust hypothesize-and-test architecture, followed by nonlinear refinement over the inlier subset.

More recently, Newcombe et al. [17] have exploited dense depth maps for entire images, in which case camera tracking can rely on photometric error criteria. That is, the tracker attempts to find a camera pose for which the warping function leads to minimal photometric difference[1]. Klein and Murray [12] already presented a related concept by using depth information of image patches along with the optimized camera pose in order to warp the patches into neighboring frames via individual affine transformations. Engel et al. [4] recently presented yet another simplified version of [17] by reducing the depth map estimation to image regions with sufficient gradient. It leads to a set of regions in the image that correspond to boundaries in the texture or along occlusions, called *semi-dense region*. [12] and [4] are conceptually similar to [17], however gain computational efficiency by reducing the computation from dense to sparse or semi-dense regions.

Experience has shown that photometric error minimization is in general a superior paradigm compared to classical pose estimation based on sparse correspondences. Photometric methods have the more general advantage of compensating for appearance variations caused by perspective view-point changes, whereas classical sparse methods often rely on static feature descriptors only (providing at most rotation and scale invariant properties [14, 15]). Furthermore, the amount of data in dense or semi-dense methods leads to good signal-to-noise ratio, which is why dense or semi-dense photometric error minimization has become the state-of-the-art in camera tracking. However, photometric registration techniques inherently suffer from the disability to overcome large disparities, where *large* sometimes means even just a couple of pixels [16]. Many photometric registration techniques therefore depend on pyramidal subsampling schemes in order to alleviate this problem.

The goal of the present paper is a novel 2D-3D registration paradigm for semi-dense depth maps that relies on the Iterative Closest Point (ICP) technique, and thus a reintroduction of geometric error minimization as a valid alternative for semi-dense visual SLAM. ICP has been used almost exhaustively in 3D-3D registration problems (e.g. with Laser point clouds [19]). Typical issues with applying ICP are missing data, noise, outliers, and local minima. Bouaziz et al. [1] argue that it is unreliable and difficult to address these issues by introducing heuristics to prune or reweight individual points. Instead, they propose a new formulation of the ICP algorithm using sparsity-induced norms. A further relevant work is given by Fitzgibbon [8], who proposes to enlarge the basin of convergence by smoothing the objective function. Yang et al. [24] investigate a new globally optimal strategy for Euclidean registration in 3D under the L2-norm error metric. Though the above works show an improvement in the registration result, they mostly aim at surface registration in 3D. The works of Feldmar et al. [6], Tomono [23], and **?**] are more related to ours, as they attempt curve or edge registration in 2D using ICP. Based on a hypothesized relative pose, they *warp* a reference curve into the tracked image based on a prior 3D model (e.g. in virtual visual servoing) or depth inside a reference frame. The quality of the 3D reference structure in those applications is rather good, thus reducing the challenge in terms of noise, outliers, missing data, and occlusions. Comport et al. [3] present a non-ICP based method for minimizing point-to-curve distances, however again relying on clean models in form of shape primitives.

This paper introduces a novel ICP-based camera tracking procedure adapted to the case of noisy, outlier-affected semi-dense depth maps. By employing a geometric error criterion, we are naturally able to overcome large disparities, and avoid the need for subsampling of

[1] The warping function is well explained in [11], and it consists of generating an artificial image from pixel-wise depth information in a reference frame, plus a relative pose hypothesis.

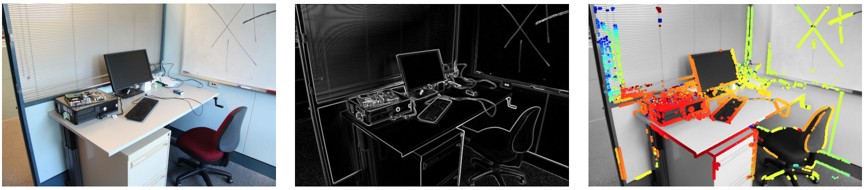

Figure 1: Example of a semi-dense depth map. The left figure shows the original image. The center image shows the approximate absolute image gradient derived from simple Sobel filters. The semi-dense region in the image plane is defined by thresholding this value. Every pixel within the region is finally tracked in a neighbor frame based on the epipolar constraint, thus leading to the inverse depth map in the right image (hot colors=close, cold colors=far). As can be observed, the initial depth map is typically affected by outliers, especially in new regions of the image (left part) and on self-similar background.

images and depth maps or isotropic enlargement of semi-dense regions. Large convergence basins and robustness with respect to outliers are furthermore supported by prior evaluation of sparse hypotheses (cf. Section 2). We speed up the computation by employing a distance transformation and the bin sort technique (cf. Section 3). Distance transformations have already proven useful for 3D-3D registration based on RGB-D cameras [2], and 2D-3D registration within the context of virtual visual servoing [9]. We conclude our work with a successful application to real data, demonstrating real-time tracking over large disparities, significant view-point changes, and severly corrupted depth maps (cf. Section 4).

2 Robust ICP-based camera tracking

The present section introduces our ICP-based camera tracking concept. We start with a clear problem definition and a summary of ICP-based 2D-3D registration. The section concludes with a sparse extension that increases the convergence basin as well as the resilience with respect to outliers.

2.1 ICP-based tracking

Let $\mathcal{F}_k$ be a reference camera frame for which depth information is available. Without loss of generality, we assume that the position of $\mathcal{F}_k$ coincides with the origin, and that its orientation equals to identity. Tracking of a single moving camera consists of retrieving the pose of a subsequent frame $\mathcal{F}_{k+1}$ given by position $\mathbf{t}$ and orientation $\mathbf{R}$, such that $\mathbf{s}^{\mathcal{F}_k} = \mathbf{R} \cdot \mathbf{s}^{\mathcal{F}_{k+1}} + \mathbf{t}$, where $\mathbf{s}^{\mathcal{F}_k}$ represents a point in frame $\mathcal{F}_k$.

Depth information in our case originates from semi-dense regions in the image, following the approach presented in [4]. Let $\mathcal{P} = \{\mathbf{p}_i\}$ be the set of pixel locations defining the semi-dense region. As illustrated in Figure 1, those regions are obtained by thresholding the norm of the image gradient, which is derived from a simple convolution with Sobel kernels[2]. For each pixel where the norm of the gradient is high enough, depth is estimated by tracking along the epipolar line in a previous frame. An example result is indicated in Figure 1,

[2] Note that the Sobel filter may not return the most stable results. It is however acceptable in our application, as a certain degree of variation in the edge-map is easily tolerated.

and given by inverse depth information for each $\mathbf{p}_i$—denoted d_i—as well as its variance—denoted σ_i. We assume the initialization to be given (for further details please refer to [4]), and focus on the tracking part. We furthermore assume that the camera is calibrated, and that we have accurate knowledge about a camera-to-world transformation function $\pi(\mathbf{p}_i) = \mathbf{f}_i$ transforming points in the image plane into unit direction vectors located on the unit sphere around the camera center. The inverse transformation $\pi^{-1}(\lambda \mathbf{f}_i) = \mathbf{p}_i$ projecting any point along the ray defined by $\mathbf{f}_i$ onto image location $\mathbf{p}_i$ is also known.

We propose a geometric approach for semi-dense tracking. The idea consists of finding a camera pose for which the semi-dense depth map in $\mathcal{F}_k$ reprojects near the semi-dense region extracted in 2D in frame $\mathcal{F}_{k+1}$. We may intuitively consider the semi-dense depth map as a curve in 3D, and its projection into $\mathcal{F}_{k+1}$ as a curve in 2D. The goal is to minimize the 2D distances between the reprojected curve and the curve corresponding to the semi-dense region extracted from image gradients. The registration is not simply solved in 2D (for instance using a 2D orthogonal procrustes technique[20]), which would ignore the projective distortion of our semi-dense region. Instead, the incremental updates of the camera pose lead to new, *warped* locations in the image plane for the reprojected semi-dense depth map. The technique is motivated by the observation that the gradient image—and thus the extracted semi-dense region—is typically a stable feature throughout an image sequence [3].

As in [6], [23], and [3], we propose ICP to solve this problem. Let

$$\mathcal{S}^{\mathcal{F}_k} = \{\mathbf{s}_i^{\mathcal{F}_k}\} = \{(d_i^{\mathcal{F}_k})^{-1}\pi(\mathbf{p}_i^{\mathcal{F}_k})\} \tag{1}$$

denote the semi-dense depth map in $\mathcal{F}_k$. Reprojection into $\mathcal{F}^{k+1}$ gives the semi-dense region

$$\mathcal{O}^{\mathcal{F}_{k+1}} = \{\mathbf{o}_i^{\mathcal{F}_{k+1}}\} = \{\pi^{-1}\left(\mathbf{R}^T\left(\mathbf{s}_i^{\mathcal{F}_k} - \mathbf{t}\right)\right)\}. \tag{2}$$

Now let $\mathcal{P}^{\mathcal{F}_{k+1}} = \{\mathbf{p}_i^{\mathcal{F}_{k+1}}\}$ be the set of pixels belonging to the semi-dense region extracted by thresholding the norm of the image gradient in $\mathcal{F}_{k+1}$. We define

$$n(\mathbf{o}_i^{\mathcal{F}_{k+1}}) = \underset{\mathbf{p}_j^{\mathcal{F}_{k+1}} \in \mathcal{P}^{\mathcal{F}_{k+1}}}{\operatorname{argmin}} \|\mathbf{p}_j^{\mathcal{F}_{k+1}} - \mathbf{o}_i^{\mathcal{F}_{k+1}}\| \tag{3}$$

to be a function that returns the pixel from $\mathcal{P}^{\mathcal{F}_{k+1}}$ that is closest to $\mathbf{o}_i^{\mathcal{F}_{k+1}}$ (in the image plane) under the Euclidean distance metric. ICP-based 2D-3D registration finally consists of minimizing the sum of the distances to the closest points over the pose of $\mathcal{F}_{k+1}$, namely $\mathbf{t}$ and $\mathbf{R}$. Our objective hence results in

$$\{\hat{\mathbf{t}}, \hat{\mathbf{R}}\} = \underset{\mathbf{t},\mathbf{R}}{\operatorname{argmin}} \sum_{i=1}^{N} \left\{ \sigma_i^{-1} \|n(\mathbf{o}_i^{\mathcal{F}_{k+1}}) - \mathbf{o}_i^{\mathcal{F}_{k+1}}\| \right\} \tag{4}$$

Note that this error criterion employs a robust L1-norm metric. Directly minimizing the above energy therefore already leads to a successful frame-to-frame tracking mechanism for short baselines. We employ gradient descent over the pose parameters using numerical Jacobian computation and a line search along the gradient direction, and furthermore weight the residuals according to the variance of the original depth estimate[4].

[3] The contours of curved surfaces in 3D could violate the assumption of purely projective distortion, which is however a general problem that affects all approaches (including photometric ones).

[4] Note that (4) is not continuously differentiable around 0. However, rather than relying on an iteratively reweighted optimization scheme, we found that in practice, over a large number of pixels (e.g. 30000), a direct minimization of the distances based on gradient descent still turns out to converge successfully.

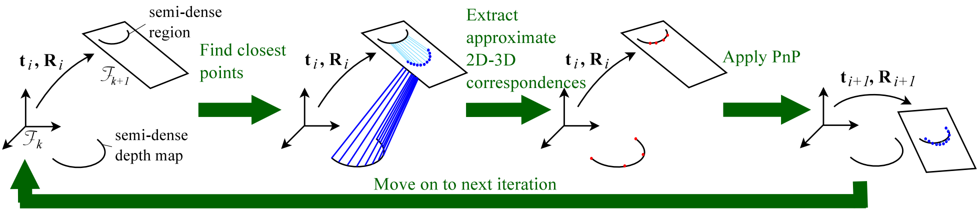
semi-dense
region
$\mathbf{t}_i, \mathbf{R}_i$
$\mathcal{F}_{k+1}$
$\mathcal{F}_k$
semi-dense
depth map
Find closest
points
$\mathbf{t}_i, \mathbf{R}_i$
Extract
approximate
2D-3D
correspondences
$\mathbf{t}_i, \mathbf{R}_i$
Apply PnP
$\mathbf{t}_{i+1}, \mathbf{R}_{i+1}$
Move on to next iteration

Figure 2: Illustration of our iterative sparse initialization procedure. In each iteration, the algorithm randomly picks a small number of depth points along with the closest points to the depth points' reprojections into the tracked frame. They serve as 2D-3D correspondences, which are then fed to a perspective n-point solver to update the current pose estimate.

2.2 Robust sparse initialization

The above mechanism works well in short baseline situations, in which the distance between reprojected depth points and their closest points in the extracted semi-dense region typically remains below a few pixels only. However, this is not always the case, and depending on the framerate of the camera and the dynamics of the camera motion, we may have to bridge larger distances, potentially leaving the convergence basin for a direct, iterative error reduction strategy based on all distances.

Our remedy consists of including a sparse update mechanism prior to the direct iterative error minimization in (4), which works as follows:

- We randomly pick j depth points $\{\mathbf{s}_{i_1}^{\mathcal{F}_k},\ldots,\mathbf{s}_{i_j}^{\mathcal{F}_k}\}$ from our semi-dense depth map.
- We project them into frame $\mathcal{F}_{k+1}$ using (2) and based on our current transformation parameters $\mathbf{t}$ and $\mathbf{R}$. The result is the set $\{\mathbf{o}_{i_1}^{\mathcal{F}_{k+1}},\ldots,\mathbf{o}_{i_j}^{\mathcal{F}_{k+1}}\}$.
- We derive the closest points $\{n(\mathbf{o}_{i_1}^{\mathcal{F}_{k+1}}),\ldots,n(\mathbf{o}_{i_j}^{\mathcal{F}_{k+1}})\}$ for every such point using (3).
- We obtain approximate 2D-3D correspondences $\{\left(n(\mathbf{o}_{i_1}^{\mathcal{F}_{k+1}}),\mathbf{s}_{i_1}^{\mathcal{F}_k}\right),\ldots,\left(n(\mathbf{o}_{i_j}^{\mathcal{F}_{k+1}}),\mathbf{s}_{i_j}^{\mathcal{F}_k}\right)\}$.
- We feed those points to a perspective n-point solver in order to update our pose.
- We evaluate the remaining distances, and go back to step 1 if they remain too high, or move on to gradient descent as a final optimization step.

The strategy is illustrated in Figure 2. We chose $j = 4$, which is the minimum number of 2D-3D correspondences required to compute a unique hypothesis for the absolute pose, and we use three of those correspondences to execute the P3P algorithm presented in [13], and the remaining one to disambiguate the at most four solutions returned by this algorithm. Note that our sparse initialization is not to be confused with RANSAC [7]. RANSAC typically fits a model to fixed 2D-3D correspondences, whereas in our case, approximate correspondences are searched within each iteration by finding closest points. Our sparse initialization therefore is still to be understood as a robust iterative error minimization strategy (similar to robust lmeds). We thereby avoid the need to extract and match local invariant keypoints, and consistently use the concept of "closest points".

3 Towards robust real-time performance

This section introduces the important extensions to SDICP increasing both computational efficiency and robustness against outliers. Note that it does not introduce any conceptual changes to the pipeline, but only robust and efficient ways of evaluating the errors in the image plane.

3.1 Preemptive scoring

We extend the sparse update strategy presented in 2.2 towards a preemptive, multi-hypothesis scoring mechanism. The approach is inspired by [18], and consists of establishing multiple hypotheses during each iteration, for which we then evaluate an incrementing part of the distances to the closest points, each time pruning the worst half of the remaining hypotheses. In our implementation, we evalute 10% of the distances for 8 hypotheses, then 5% more for 4 (i.e. 15%), then 10% more for 2 (i.e. 25%), and the remaining ones into the remaining hypothesis (i.e. 100%). Furthermore, the remaining hypothesis is applied if and only if it leads to a reduction of the summed distances with respect to the previous iteration.

3.2 The Chebychev distance transform

Computing the exact location of the closest points of each reprojected depth point (i.e. $n(\mathbf{o}_i^{\mathcal{F}_{k+1}})\ \forall i \in \{1,\ldots,N\}$) is a time-consuming task, especially as this procedure is embedded into an iterative optimization scheme. However, it is important to realize that for most of the time, we are not interested in the exact location of the closest pixel within the semi-dense region, but only in the distance to them. In fact, the exact location of the closest points is only needed for our random samples during the sparse initialization process, which means $4 \times \#\text{iterations} \times \#\frac{\text{hypotheses}}{\text{iteration}}$ times. This number typically remains below a few hundred for the entire tracking of a single frame. As a result, we propose to compute the distances directly based on a distance field. The distance field is an image generated for $\mathcal{F}_{k+1}$ and that—for each pixel in the image—indicates the distance to the closest pixel in the original semi-dense region $\mathcal{P}^{\mathcal{F}_{k+1}}$ obtained by thresholding the norm of the gradients[5]. We propose the use of a distance metric for which the distance field can be extracted very fast, namely the *Chebychev distance*. This metric is also called the *chessboard distance*, as it corresponds to the minimum number of moves a king would need to reach the semi-dense region[6].

3.3 Bin sort

As mentioned earlier, our criterion for evaluating poses is given by the sum of distances to the closest points in the semi-dense region. While this corresponds to an L1-norm metric—and thus already provides robustness with respect to outliers—, we include an additional, helpful robustness measure during the sparse initialization step. Instead of simply suming up all the distances, we sort the distances and consider the value of the 95th percentile. Assuming that the semi-dense region contains no more than 5% outliers, this eliminates all outlier distances from the evaluation. The evaluation of the 95th percentile however requires another precaution in the implementation, as the explicit sorting of all distances would be

[5] Distance transforms have already been used in various contexts, such as for instance in 3D-3D registration for RGB-D cameras [2].

[6] Note that we perform bilinear interpolation when retrieving distances in the distance field.

too computationally intensive. Instead, we solve this problem by filling a histogram with constant number and size of bins, which can be done in linear time and with reduced memory requirements. It has two consequences: a) The complexity of finding the value of the 95th percentile depends linearly on the number of bins, and b) the accuracy of the 95th percentile is limited by the bin size. In our implementation, we chose a bin-size of 0.1 pixel and 200 bins, which keeps the retrieval of the 95th percentile very efficient[7], and leads to largely sufficient accuracy for the sparse initialization step.

3.4 Finding closest points

Explicit locations of closest points are still required during our sparse initialization process. The most straightforward solution for finding closest points in the semi-dense region consists of simply iterating through a window around a location in the image plane starting from the top left corner. This is however very inefficient as we potentially have to loop through half the window even as we converge to the correct result. Our solution consists of grouping the local neighborhood pixels into classes of increasing distance from the center of a window, and then looping through those classes starting with the smallest distance. We furthermore limit the search to radii below the current value of the 95th percentile, as we want to perform updates with (approximate) inlier correspondences.

4 Experimental validation

The focus of our experimental evaluation lies on a comparison to a well established, classical sparse alternative. Our aim is to demonstrate an improvement in robustness over traditional methods, similar to what we have seen from recent dense or semi-dense photometric approaches. A brief direct opposition between SDICP and a semi-dense photometric error minimizer is also provided, as well as an analysis of computational efficiency.

4.1 Robustness and accuracy

In order to demonstrate SDICP's robustness advantage with respect to classical sparse approaches, we applied both methods to a publically available dataset for which ground truth information is available. We use the sequence *freiburg2_xyz* from the TUM RGB-D SLAM benchmark suite [22], as this one is well suited for a detailed evaluation of local tracking performance. The dataset is captured with a Kinect sensor running at 30Hz. We use RGB information only, converted to monochrome images within our algorithm. The VGA images in the dataset are already undistorted, and the intrinsic camera parameters are given by $f_x = f_y = 525$, $c_x = 319.5$, and $c_y = 239.5$.

Our experiment consists of initializing a sparse point cloud as well as a semi-dense depth map in one of the first frames of the sequence, and then performing both sparse and semi-dense tracking with respect to this reference frame throughout the entire sequence. We use a sparse method to initialize the relative pose between the very first frame in the sequence and our chosen reference frame for the tracking. It relies on homogeneous Harris corner extraction [10], patch descriptors, and a brute-force matcher to establish 2D-2D correspondences. We then apply the five-point algorithm [21] embedded into a RANSAC scheme [7] in order

[7] Note that the value of the 95th percentile is very efficiently computed by starting on the right side and subtracting bin counts from the total number of elements in the histogram, rather than counting from the left side.

to identify the initial relative pose as well as the inlier subset. We finally run two-frame bundle adjustment to complete the initialization of the sequence[8].

In order to obtain the reference point cloud for our sparse tracking scheme, we use the optimized inlier correspondences obtained from the initialization step. We then use the same type of features and descriptors to continuously establish sparse 2D-3D correspondences towards our reference point cloud, and run the perspective 3-point algorithm [13] again embedded into RANSAC [7] in order to track the pose of the camera. Each tracking step is finalized by nonlinear refinement. The sparse reference point cloud is free of outliers and contains moderate noise due to the robust two-frame initialization procedure.

The reference depth map for SDICP is obtained as explained in Section 2.1. In contrast to keypoint matching, the simplified epipolar tracking of [5] is prone to errors, thus leading to a significant amount of noise in the semi-dense depth map as well as a large number of outliers being either too close or too distant with respect to the camera. However, we intentionally use this pre-mature initialization, as this poses a more difficult challenge to our algorithm in terms of accuracy and robustness with respect to outliers.

The results are presented in Figure 3. The first two columns show a qualitative evaluation of the tracking performance by reprojecting the semi-dense depth map into the current frame once using the pose retrieved by our novel SDICP algorithm (first column), and once using the sparse baseline implementation (second column). The sparse tracker quickly loses track of the reference map as we move away from the initialization spot. On the other hand, SDICP successfully tracks throughout the entire sequence, and copes with impressive variations of the view-point pushing the reprojected semi-dense depth map into all corners of the image. The result may also be observed in the supplemental video file. The last two columns in Figure 3 finally show the detailed performance of both trackers in comparison to ground truth. The plots clearly show the frequent tracking losses of the sparse implementation, as well as increased smoothness and thus accuracy in the results produced by SDICP. Both methods suffer from the same bias in the translation estimation, which we track back to the inaccuracy of the commonly used, initial relative pose computation. However, it is impressive to see that SDICP still maintains an accurate orientation estimate throughout the entire sequence, despite the presence of many outliers in the depth map. We interpret this result as an impressive state-of-the-art improvement over traditional sparse approaches, able to cope with large variations of the view-point as well as significantly corrupted data.

It is important to understand that our experiment aims at evaluating the accuracy and robustness of the tracker alone. We therefore track continuously with respect to one and the same, outlier-affected initialization. It is clear that propagating and updating the semi-dense depth map would drastically improve its accuracy, which in turn would lead to improved tracking results as well. It is notably for this reason that a plain comparison to the publicly accessible framework presented in [5] would be unfair, as this framework performs continuous propagation and updates.

4.2 Convergence in case of large disparities

We carried out a direct comparison to semi-dense photometric error minimization. Although our implementation of the latter works on regular sequences, we quickly noticed that it fails on instances of elevated frame-to-frame disparity. For example, if the actual disparity goes

[8] Note that there may be other possibilities for bootstrapping the sequence, such as directly using the semi-dense regions. While this is certainly an interesting research topic, it is not the subject of this paper. We evaluate here only the performance of the tracker.

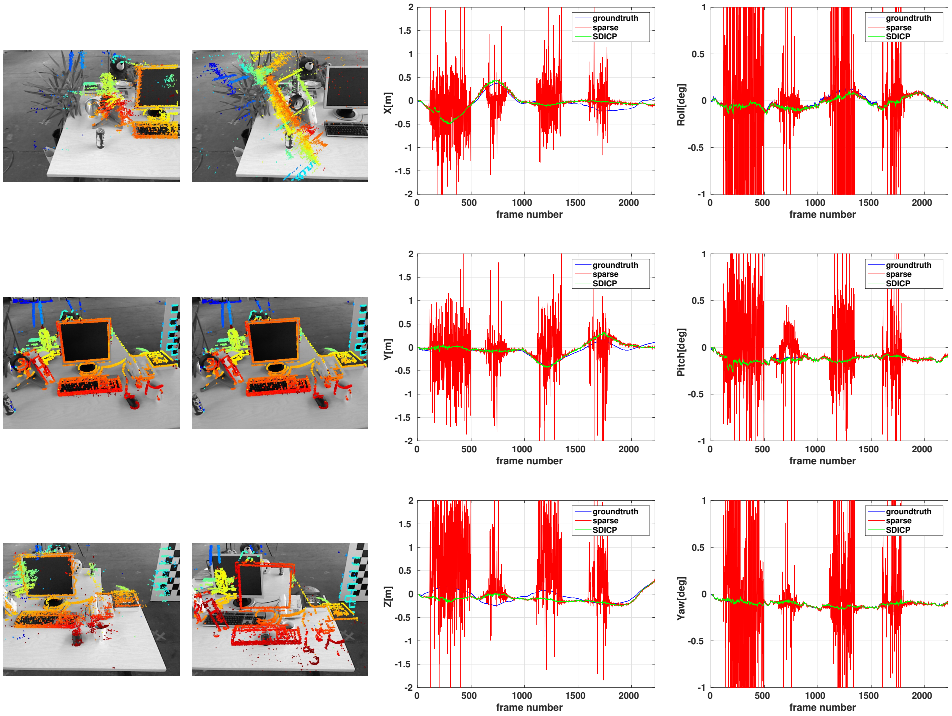

Figure 3: Columns 1 and 2 show a qualitative comparison between the proposed SDICP algorithm and a sparse baseline implementation. The tracking performance is visualized by reprojecting the semi-dense depth map into the current frame, once using the pose obtained by SDICP (first column), and once using the pose obtained by a sparse method (second column). It can be observed that SDICP maintains tracking throughout large view-point changes (top and bottom rows), whereas the sparse reference implementation maintains successful tracking only in the vicinity of the initialization spot (central row). The images also visualize the significant amount of noise and outliers in the tracked semi-dense depth map. The figure furthermore shows a comparison of the 6DoF trajectory estimation of SDICP and a sparse baseline implementation with respect to ground truth (columns 3 and 4). Besides the frequent tracking failures of the sparse alternative, the plot shows increased smoothness in the results of SDICP, thus suggesting superior accuracy.

beyond 20 pixels for major parts of the image, we would require a width of more than ±10 pixels in the semi-dense region such that we would have a potential overlap and thus even a chance to converge. Furthermore, it is commonly known that photometric error minimization schemes have difficulties to overcome disparities going beyond a couple of pixels, which is why successful convergence often requires pyramidal subsampling techniques. This in turn poses further requirements on the minimum width of the semi-dense region. Engel et al. [4] solve this problem by an isotropic enlargement of the semi-dense region, stating that it helps the convergence of the tracker.

If choosing four pyramidal layers as proposed in [4] and limiting the width of the semi-dense region to pixels with sufficient image gradient, our implementation of semi-dense photometric error minimization fails on instances with more than 20 pixels frame-to-frame disparity. In contrast, SDICP naturally bridges this gap, even if completely disabling the widening of the semi-dense region. The maximum disparity for SDICP is of course depending on the structure. Simpler structures may lead to larger convergence basins, but at the same time deliver fewer data points, and thus reduced accuracy. Intuitively spoken, the convergence basin is limited to half the distance between neighbouring edges in the image plane. Abundant texture may also cause failure of convergence, as this can easily lead to wrong local minima.

4.3 Computational efficiency

Our algorithm is implemented in C++ and runs on 8 cores in parallel. All our results were produced on a 2.5GHz Core i7 machine. The total time consumption for tracking a frame with large disparity (> 20 pixels) is 0.12s. This however represents an extreme case with 13 iterations during the sparse robust initialization. This step usually takes only a single iteration during regular tracking situations. Furthermore, the efficiency of the entire algorithm can be scaled by evaluating only a part of the depth map, thus easily ensuring real-time capability.

5 Discussion

The present work is conceived as a more practial contribution in form of a novel concept for semi-dense monocular camera tracking. In contrast to several recent works in the literature which all rely on photometric error minimization techniques, we achieve state-of-the-art results by relying on a geometric error criterion. Our innovation lies in using ICP—a technique which has proven successful in countless rigid registration contexts—and extending it to the case of noisy, outlier-affected semi-dense depth maps.

While we are able to demonstrate outstanding performance with respect to large disparities, large view-point changes, and significant data corruption, the present paper at the same time reveals a number of important, fundamental questions. The most important one is given by the question which error minimization strategy is most appropriate for semi-dense features: geometric or photometric? Photometric error minimization requires an isotropic enlargement of the semi-dense region, which seems unnatural given that the semi-dense region represents a border that—from a theoretical standing—is infinitely thin. Our implementation does not require this enlargement, which is why we consider it a more natural solution for image gradient based semi-dense registration. Another question is whether gradients are a possibly more robust feature than intensity-based appearance. As an example, consider a change in the background intensity of an occlusion. A photometric error measure will increase, whereas our method based on thresholded gradients may see no change at all.

ACKNOWLEDGMENT

The research leading to these results is supported by the ARC Centre of Excellence for Robotic Vision. The work is furthermore supported by ARC grants DP120103896 and DE150101365.

References

[1] S. Bouaziz, A. Tagliasacchi, and M. Pauly. Sparse iterative closest point. *Computer Graphics Forum (Symposium on Geometry Processing)*, 32(5):1–11, 2013.

[2] E. Bylow, J. Sturm, C. Kerl, F. Kahl, and D. Cremers. Direct camera pose tracking and mapping with signed distance functions. In *Demo Track of the RGB-D Workshop on Advanced Reasoning with Depth Cameras at Robotics: Science and Systems (RSS)*, Berlin, Germany, 2013.

[3] A.I. Comport, E. Marchand, M. Pressigout, and F. Chaumette. Real-time markerless tracking: the virtual visual servoing framework. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):615–628, 2006.

[4] J. Engel, J. Sturm, and D. Cremers. Semi-dense visual odometry for a monocular camera. In *Proceedings of the International Conference on Computer Vision (ICCV)*, Sydney, Australia, 2013.

[5] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Zurich, Switzerland, 2014.

[6] J. Feldmar, N. Ayache, and F. Betting. 3D-2D projective registration of free-form curves and surfaces. In *Proceedings of the International Conference on Computer Vision (ICCV)*, Cambridge, MA, 1995.

[7] M.A. Fischler and R.C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[8] A. W. Fitzgibbon. Robust registration of 2D and 3D point sets. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 411–420, Manchester, UK, 2001.

[9] X. Gratal, J. Romero, and D. Kragic. Virtual visual servoing for real-time robot pose estiamtion. In *Proceedings of the 18th IFAC world congress*, Milano, Italy, 2011.

[10] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of Alvey Vision Conference*, pages 147–151, Manchester, UK, 1988.

[11] C. Kerl, J. Sturm, and D. Cremers. Robust odometry estimation for RGB-D cameras. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Karlsruhe, Germany, 2013.

[12] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, Nara, Japan, 2007.

[13] L. Kneip, D. Scaramuzza, and R. Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, USA, 2011.

[14] Stephan Leutenegger, Margarita Chli, and Roland Siegwart. BRISK: Binary Robust Invariant Scalable Keypoints. In *Proceedings of the International Conference on Computer Vision (ICCV)*, Barcelona, Spain, 2011.

[15] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[16] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Vancouver, Canada, 1981.

[17] R. Newcombe, S. Lovegrove, and A. Davison. DTAM: Dense Tracking and Mapping in Real-Time. In *Proceedings of the International Conference on Computer Vision (ICCV)*, Barcelona, Spain, 2011.

[18] D. Nistér. Preemptive RANSAC for live structure and motion estimation. *Machine Vision & Applications*, 16(5):321–329, 2005.

[19] F. Pomerleau. *Applied Registration for Robotics: Methodology and Tools for ICP-like registration*. PhD thesis, Eidgenössische Technische Hochschule ETH Zürich, 2013.

[20] P.H. Schonemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31:1–10, 1966.

[21] H. Stewénius, C. Engels, and D. Nistér. Recent developments on direct relative orientation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60(4):284–294, 2006.

[22] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, Vilamoura-Algarve, Portugal, 2012.

[23] M. Tomono. Robust 3D SLAM based on an edge-point ICP algorithm. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Kobe, Japan, 2009.

[24] J. Yang, H. Li, and Y. Jia. Go-ICP: Solving 3D registration efficiently and globally optimally. In *Proceedings of the International Conference on Computer Vision (ICCV)*, Sydney, Australia, 2013.