Tracking Planes with Time of Flight Cameras and J-linkage

Loren Arthur Schwarz Diana Mateus Joé Lallemand Nassir Navab

Computer Aided Medical Procedures (CAMP)
Technische Universität München, 85748 Garching, Germany
{schwarz,mateus,lalleman,navab}@cs.tum.edu

Abstract

In this paper, we propose a method for detection and tracking of multiple planes in sequences of Time of Flight (ToF) depth images. Our approach extends the recent J-linkage algorithm for estimation of multiple model instances in noisy data to tracking. Instead of randomly selecting plane hypotheses in every image, we propagate plane hypotheses through the sequence of images, resulting in a significant reduction of computational load in every frame. We also introduce a multi-pass scheme that allows detecting and tracking planes of varying spatial extent along with their boundaries. Our qualitative and quantitative evaluation shows that the proposed method can robustly detect planes and consistently track the hypotheses through sequences of ToF images.

1. Introduction

Recent developments in Time of Flight (ToF) camera technology allow the direct and fast acquisition of depth images and avoid problems associated with stereo systems (such as lack of texture, repetitive structures or poor lighting). Despite the low resolution of current commercial ToF devices when compared to stereo or laser range systems and the presence of several sources of noise, the depth information provided by ToF cameras is very valuable for scenarios where fast modeling and interpretation of a scene is required. In this context, relevant applications include augmented reality [24], robotics [23, 19, 12] and human-machine interaction [22].

Given the camera calibration, a ToF depth image can be transformed into a set of 3D points. Such a point representation of a scene is general but inefficient to store and process [19]. In particular, difficulties arise when dealing with sequences of point sets (i.e. depth videos). Temporal processing and measurement of motion require establishing correspondences between the point sets in different frames. Therefore, scene modeling methods seek to build a compact representation of the environment that is consistent over time. This is a well studied problem, for instance in robotics [11, 19]. Commonly used methods for creating a global representation of the scene consist in incrementally matching the sets of 3D points over time, for instance, using standard Iterative Closest Point (ICP) approaches [28]. Simultaneous Localization and Mapping (SLAM) approaches focus on fusing the information over time.

Such global, high-quality representations can be expensive to compute and thus unsuitable for online applications, where only the currently observed portion of the scene is relevant and where local representations suffice. An intuitive approach for building local, temporally consistent representations is to partition the scene into elementary components that are subsequently tracked. A popular choice is to decompose the scene into planar primitives. The motivation for using planes comes from the compactness, simplicity and stability of this representation [3, 10]. Indeed, planes have proven to be important geometric features in several applications. Examples include feature matching and grouping [8], robot localization [19] and 3D reconstruction and modeling [1, 9, 17].

In this paper, we address the problem of extracting a local, temporally consistent representation of a scene from sequences of depth images acquired with a ToF camera. We rely on the assumption that important, dominant or stable regions of the scene are planar, which is reasonable for structured, man-made interiors, such as laboratories or industrial environments [3]. More specifically, our proposed method consists in detecting and tracking multiple instances of a plane model, given sequences of noisy 3D point data. We build upon the J-linkage clustering algorithm for model estimation, recently introduced by Toldo and Fusiello [26, 27]. What makes J-linkage particularly suitable in our setting is its ability to identify *multiple* instances of a given model in the presence of large amounts of noise. In order to extract a temporally consistent representation, we extend the J-linkage algorithm, mainly used

for detection, to *tracking*. This is achieved by linking the processing of individual frames together by means of plane hypotheses that are propagated over time. Additionally, as the planes observed with ToF cameras appear as point-sets with variable size, point-density and degree of noise, we adapt the plane detection method to handle these situations with an iterative multi-pass scheme.

Our method provides as an output the detected planes and their boundaries throughout the given sequence of ToF images. Figures 1 and 2 give a pictorial overview of the method. The experiments presented in section 4 provide a qualitative and quantitative evaluation on different sequences of a ToF camera moving in a static environment. Comparisons are performed against the full J-linkage detection method.

2. Related Work

Detecting and building piecewise representations of an observed scene based on 3D planes is a broadly studied problem in computer vision. The success of piecewise planar models arises from the strong planarity constraints which improve the robustness of 3D reconstructions and from the compactness of the representation, suitable for storage and rendering [3, 10]. Many of the approaches for planar scene representation focus on detecting 3D planes based on 2D point correspondences in multiple views, making use of projective geometry [16]. Detected planes are then used for various applications, such as filtering, generating hypotheses of new 2D point correspondences [14], or creating a 3D reconstruction of the environment [25, 1, 17]. We target a similar application, namely a plane piecewise representation, however, using a ToF camera and thus overcoming some of the problems associated to finding 2D correspondences in multiple views (e.g., the lack of texture or repetitive structures). Furthermore, we focus on the temporal consistency of the representation. Some attempts towards detecting and tracking 3D planar representations over time have been made, but from 2D image correspondences, as proposed in [24, 21].

In robotics, planar representations of the environment are commonly used to facilitate mapping and localization tasks [12]. Examples of efforts towards planar piecewise representations using range data acquired with laser scanners include [15, 4]. Recently, some attempts have also been done using ToF images [13, 20]. Holz *et al*. [13] target the application of grasp planning and object manipulation. Poppinga *et al*. [20] propose a sequential plane detection method from ToF videos. Like in our case, the temporal consistency in [20] is achieved by modifying the plane estimation and taking into account the estimates from previous frames. However, the basic plane detection method differs from ours, as it is based on the region growing algorithm in [11]. Indeed, different approaches to planar surface detection and estimation of the corresponding plane parameters exist. The authors of [6] categorize the approaches into iterative methods [2], voting-based methods [23, 9] or methods employing a growing procedure [11, 29]. As opposed to these approaches, we formulate the problem in terms of robust model estimation in the presence of multiple instances of a model, where the model is a plane.

Standard methods for robust model estimation exist, RANSAC [7] being perhaps one of the most popular approaches in computer vision. Although robust, the original RANSAC algorithm is not able to handle multiple model instances. Bartoli *et al*. [3] have proposed a method to detect planes in 3D sparse correspondences that adapts RANSAC to cope with multiple instances. Methods for model estimation under multiple instances include mean-shift [5] and the Hough transform. The former is based on a clustering approach that finds the modes of the distribution of models. The latter works in the transformed Hough space, where agreeing models can be easily detected. The Randomized Hough Transform (RHT) has been used for range image segmentation [18, 6] in planar regions. However, RHT-based methods have limited accuracy and low computational efficiency [27]. Recently, J-linkage clustering [27] has been proposed as a method for fitting multiple instances of a model to data corrupted by noise and outliers. The algorithm is based on random sampling and an agglomerative clustering method. First, the input data points are represented with feature vectors that indicate the set of random models consistent with every point. In this feature space, J-linkage clustering is used to group the points belonging to the same model. Fouhey *et al*. [8] have recently applied J-linkage to the detection of planes from 2D images.

The use of the J-linkage model estimation algorithm allows us to successfully handle the noise present in ToF images. However, the direct application of the method to this type of data is not sufficient for detecting all plane instances and, moreover, does not provide a solution for the temporal consistency problem. In the following, we describe our modifications to the original algorithm, extending it to tracking and allowing to overcome problems of different plane densities and sizes.

3. Plane Detection and Tracking Method

Given a sequence of T ToF depth images $\mathcal{I}_1, \ldots, \mathcal{I}_T$, each consisting of $N = N_x \times N_y$ pixels, our objective is to detect an arbitrary number of planes in the first frame and to track the planes throughout the sequence. Here, a plane $\mathbf{p} = (\mathbf{n}, d)$ is represented by its normal $\mathbf{n}$, its distance d to the origin and the coordinates of its four boundary points $\{\mathbf{b}_i\}_{i=1}^4$, with $\mathbf{b}_i \in \mathbb{R}^3$. For each image $\mathcal{I}_t$ with $1 \leq t \leq T$, we first compute the corresponding set of 3D points $\mathcal{X}_t = \{\mathbf{x}_{t,i}\}_{i=1}^N$, where $\mathbf{x}_{t,i} \in \mathbb{R}^3$, based on the intrinsic parameters of the ToF camera.

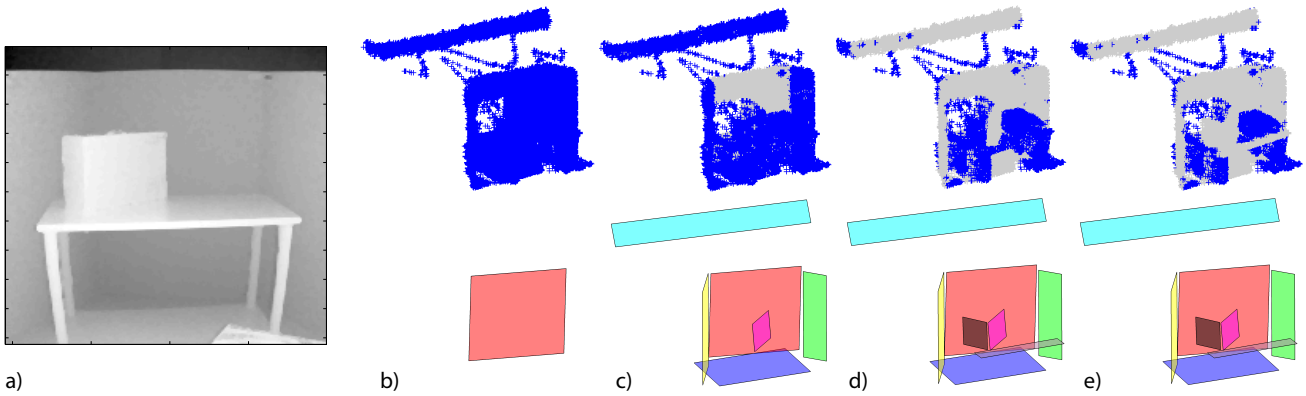


Figure 1. Multi-pass plane detection scheme on a single ToF image. a) Input depth image. b)-e) 3D point set before every pass of the algorithm (top), planes detected incrementally (bottom). Points belonging to detected planes are removed from the initial point set after each pass. In the last pass, no additional planes are detected and the multi-pass scheme terminates.

Our method consists of a plane detection algorithm, applied to the first frame of the sequence, and a plane tracking algorithm for all other frames. Both parts build upon a multi-pass strategy that allows dealing with noise in the ToF data and with planes of varying size. In each pass, plane hypotheses are extracted using the J-linkage algorithm [26], followed by refinement and global aggregation. Our method does not require any prior information for initialization and uses the full J-linkage algorithm in the first frame. For all subsequent frames, we propose a modification, where plane hypotheses from previous frames are used for initialization and are propagated in time. Figure 2 gives an overview of the proposed method. We will first describe our plane detection approach for individual images in sections 3.1, 3.2 and 3.3, before turning to tracking in section 3.4.

3.1. Extraction of Plane Hypotheses (J-linkage)

The J-linkage algorithm was recently introduced for robust detection of multiple model instances in data that contains significant amounts of noise and outliers [26]. In our setting, model instances correspond to planes that we wish to detect in ToF images. Let $\mathcal{X}$ be a set of input points extracted from a ToF image. Initially, the algorithm generates *plane hypotheses* $\hat{\mathcal{P}} = \{\hat{\mathbf{p}}_1, \ldots, \hat{\mathbf{p}}_{\hat{M}}\}$, by selecting $\hat{M}$ triplets of points from $\mathcal{X}$. The first point of each triplet is chosen at random, while the other two points are taken from the vicinity of the first point [26].

Next, the *preference set* is computed for every point $\mathbf{x}_i \in \mathcal{X}$, indicating which of the plane hypotheses the point supports. The preference set is represented by an $\hat{M}$-dimensional binary vector, where the j-th entry is 1 if $d_\perp(\mathbf{x}_i, \hat{\mathbf{p}}_j) < \tau$, and 0 otherwise. Here, $d_\perp(\mathbf{x}_i, \hat{\mathbf{p}}_j)$ denotes the orthogonal distance of a point $\mathbf{x}_i$ to a plane $\hat{\mathbf{p}}_j$ and τ is the maximum allowed point-to-plane distance. The preference set representation is used to cluster the points in $\mathcal{X}$ in order to extract plane hypotheses with the largest support. An agglomerative clustering approach is used together with the Jaccard distance metric. The Jaccard distance between two preference sets ranges from 0, for identical sets, to 1, for disjoint sets [26]. Starting with separate clusters for all points in $\mathcal{X}$, pairs of clusters with the smallest Jaccard distance are repeatedly merged by intersecting their preference sets. Clustering terminates when all preference sets are disjoint. The points in the remaining clusters are finally used to estimate the $\bar{M}$ plane hypotheses $\bar{\mathcal{P}} = \{\bar{\mathbf{p}}_1, \ldots, \bar{\mathbf{p}}_{\bar{M}}\}$, where $\bar{M} \ll \hat{M}$. We only retain clusters that consist of at least μ points.

3.2. Refinement of Plane Hypotheses

The plane hypotheses obtained so far are often supported by points that belong to spatially separated 3D objects. As this effect is undesirable, our aim is to refine the plane hypotheses such that each retained plane is supported by evenly distributed points. To this end, we proceed as follows for each of the $\bar{M}$ plane hypotheses. Let $\mathcal{X}^m = \{\mathbf{x} \in \mathcal{X} \mid d_\perp(\mathbf{x}, \bar{\mathbf{p}}_m) < \tau\}$ denote the set of points supporting the m-th plane hypothesis, $1 \leq m \leq \bar{M}$. We define the average distance of a point $\mathbf{x}$ to its closest neighbors as

$$d_{\kappa nn}(\mathbf{x}) = \frac{1}{|\kappa nn(\mathbf{x})|} \sum_{\mathbf{y} \in \kappa nn(\mathbf{x})} \|\mathbf{x} - \mathbf{y}\|^2 \qquad (1)$$

where $\kappa nn(\mathbf{x})$ represents the κ nearest neighbors of $\mathbf{x}$. We then remove outlier points that are consistent with the plane hypothesis but that are spatially separated from the majority of points on the given plane. All points $\mathbf{x}_i^m \in \mathcal{X}^m$ are discarded if

$$d_{\kappa nn}(\mathbf{x}_i^m) > \frac{1}{|\mathcal{X}^m|} \sum_{\mathbf{x} \in \mathcal{X}^m} d_{\kappa nn}(\mathbf{x}). \qquad (2)$$

After outlier elimination, we re-estimate each of the $\bar{M}$ plane hypotheses $\bar{\mathbf{p}}_m$ using only the remaining points in

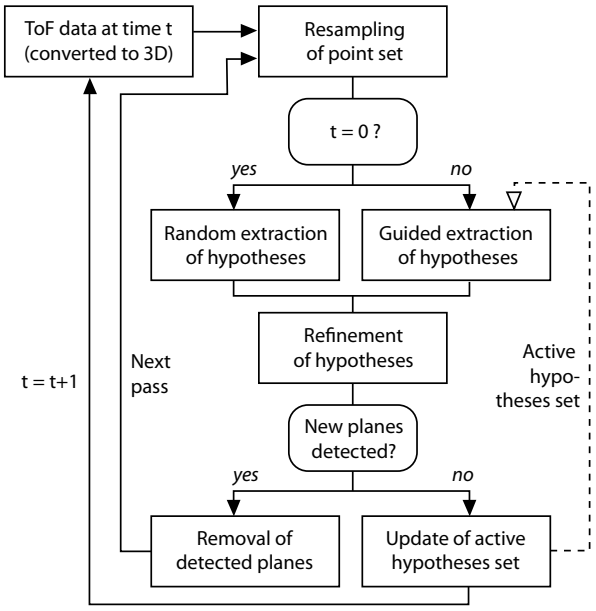


Figure 2. Schematic of the plane detection and tracking method. At each time t, a ToF image is converted to a 3D point set, followed by multiple passes of hypothesis extraction and refinement. After every pass, the points belonging to detected planes are removed from the original point set. When no additional planes are found, the active plane hypotheses are updated and used as an initialization for the following image.

$\mathcal{X}^m$. These points are also used to estimate the boundary points $\mathbf{B}^m = \{\mathbf{b}_{m,i}\}_{i=1}^4$ for each plane. We transform the point set $\mathcal{X}^m$ using PCA, such that the directions of largest variation coincide with the coordinate axes, allowing us to easily obtain the bounds of the point set. As a final refinement step, we compute the area a_m of each plane hypothesis and discard planes for which $a_m < \alpha$ or $|\mathcal{X}^m|/a_m < \beta$. Here, α is a minimum plane area and β is a minimum point density that we require for all planes. We denote the final plane hypotheses for the point set $\mathcal{X}$ as $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^M$ and the bounding points as $\mathcal{B} = \{\mathbf{B}^i\}_{i=1}^M$, where $M \leq \bar{M}$.

3.3. Multi-Pass Strategy

The hypothesis refinement step ensures that retained planes have an even point distribution and do not stretch across large spatial gaps. However, while J-linkage is designed for detection of *multiple* model instances, detection of planes at very different sizes remains challenging. By the discrete nature of ToF imaging, small structures will be represented with fewer points than larger structures. In addition, the number of points per unit surface area decreases with the distance of a structure from the camera. Increasing sensitivity by allowing for clusters with fewer points in the J-linkage algorithm unfortunately leads to numerous false plane hypotheses, and thus to computational overhead.

We therefore propose a multi-pass strategy, where the hypothesis extraction and refinement steps described above are repeated iteratively. More specifically, we restrict the point set before the k-th pass to $\mathcal{X}[k] = \mathcal{X} - \mathcal{Y}$, where $\mathcal{Y}$ contains all points belonging to planes detected in the previous passes:

$$\mathcal{Y} = \bigcup_{c=1}^{k-1} \bigcup_{m=1}^{M_c} \mathcal{X}[c]^m. \tag{3}$$

Here, $\mathcal{X}[c]^m$ are the points supporting the m-th plane hypothesis at the c-th pass and M_c is the number of refined plane hypotheses for that pass. Note that, in each pass, we remove points of previously detected planes from the *original* point set $\mathcal{X}$ to prevent duplicate plane detections. Moreover, we resample $\mathcal{X}[k]$ before each pass to contain $n \ll N$ points, where n is equal at each pass. This way, we increase the probability of detecting small planes, since a plane undetected in a previous pass will be represented by a larger number of points after resampling. The multi-pass scheme terminates if no additional plane hypotheses are found after the refinement step. An illustration of the procedure can be found in Figure 1, a summary is provided in Algorithm 1.

Algorithm 1 Multi-Pass Plane Detection

input $\mathcal{X}$
$\mathcal{P} = \emptyset, \mathcal{Y} = \emptyset, \mathcal{B} = \emptyset, k = 1$
repeat
  $\mathcal{X}[k] \leftarrow \mathcal{X} - \mathcal{Y}$
  $\bar{\mathcal{X}}[k] \leftarrow \text{resample}(\mathcal{X}[k], n)$
  $\bar{\mathcal{P}}[k] \leftarrow \text{extractHypotheses}(\bar{\mathcal{X}}[k])$
  $\mathcal{P}[k], \mathcal{B}[k] \leftarrow \text{refineHypotheses}(\bar{\mathcal{P}}[k], \mathcal{X}[k])$
  $\mathcal{Y}[k] \leftarrow \emptyset$
  for $m \leftarrow 1$ to $|\mathcal{P}[k]|$ **do**
    $\mathcal{X}[k]^m \leftarrow \{\mathbf{x} \in \mathcal{X}[k] \mid d_\perp(\mathbf{x}, \mathbf{p}_m) < \tau\}$
    $\mathcal{Y}[k] \leftarrow \mathcal{Y}[k] \cup \mathcal{X}[k]^m$
  end for
  $\mathcal{Y} \leftarrow \mathcal{Y} \cup \mathcal{Y}[k], \mathcal{P} \leftarrow \mathcal{P} \cup \mathcal{P}[k], \mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{B}[k]$
  $k \leftarrow k + 1$
until $\mathcal{P}[k] = \emptyset$
return $\mathcal{P}, \mathcal{B}$

3.4. Plane Tracking Over Time

In the previous sections, we have described our approach for extraction and refinement of plane hypotheses from individual ToF images using a multi-pass strategy. We will now turn to the problem of tracking detected planes through a sequence of ToF images $\mathcal{I}_1, \ldots, \mathcal{I}_T$. The proposed extension to plane tracking consists of a modification that significantly increases the efficiency of the J-linkage-based hypothesis extraction step and, at the same time, links plane detections from one frame to the next.

As described section 3.1, the J-linkage algorithm initially generates a set of plane hypotheses $\hat{\mathcal{P}}$ by randomly selecting triplets of points. Following [26], the number of hypotheses $\hat{M}$ needs to be chosen large to ensure robustness against noise and outliers in the point set. While this procedure favors that suitable plane hypotheses are generated, it also causes a significant computational load, in particular for the clustering step. For the case when planes have already been detected in a previous frame, we therefore propose to use the existing plane hypotheses as prior knowledge.

We introduce a set of active hypotheses $\mathcal{P}_{\text{act}}$ that are tracked and updated from frame to frame. After plane detection in the first frame ($t = 0$) using full J-linkage, we set $\mathcal{P}_{\text{act}} = \mathcal{P}_0$, where $\mathcal{P}_0$ refers to the planes detected in the first frame. Let $M_{\text{act}} = |\mathcal{P}_{\text{act}}|$. In all subsequent frames, we do not initialize the J-linkage algorithm with random plane hypotheses $\hat{\mathcal{P}}_t$, but instead create a set of *guided* hypotheses $\tilde{\mathcal{P}}_t$ from $\mathcal{P}_{\text{act}}$. The guided hypotheses are comprised of the active plane hypotheses and transformed duplicates thereof, to account for motion occurring between the frames.

The transformations we apply to generate the guided hypotheses correspond to incremental rotations of the ToF camera around its center with random translations. More formally, we define the set of guided hypotheses $\tilde{\mathcal{P}}_t$ before any frame $t > 0$ as

$$\tilde{\mathcal{P}}_t = \left\{ \tilde{\mathcal{P}}_t^1 \cup \ldots \cup \tilde{\mathcal{P}}_t^{M_{\text{act}}} \right\}, \tag{4}$$

where $\tilde{\mathcal{P}}_t^i$ is the set of plane hypotheses derived by transformation of the plane $\mathbf{p}_i^{\text{act}} \in \mathcal{P}_{\text{act}}$. Let $\mathbf{T}(\theta, \phi, \psi, \mathbf{t})$ denote a 4×4 rigid homogeneous transformation matrix that rotates by θ, ϕ and ψ around the three coordinate axes and translates by a vector $\mathbf{t}$. We can then write

$$\tilde{\mathcal{P}}_t^i = \{\mathbf{T}(\theta, \phi, \psi, \mathbf{t})^{-\top} \mathbf{p}_i^{\text{act}}\} \quad \forall \theta, \phi, \psi \in \Omega, \tag{5}$$

where $\mathbf{t} \in \mathbb{R}^3$ is a random vector, unique for every element of the set, and Ω contains ρ rotation angles from an interval $[-\gamma; \gamma]$. The maximum rotation angle parameter γ has a significant influence on tracking results, as demonstrated in section 4. In total, the number of guided hypotheses in $\tilde{\mathcal{P}}_t$ is $\tilde{M} = \rho^3 M_{\text{act}}$, and in practice $\tilde{M} \ll \hat{M}$.

Using the guided hypotheses as an initialization for J-linkage, we proceed as described above with hypothesis extraction, refinement and multi-pass processing. Once the final plane hypotheses $\mathcal{P}_t$ for frame t have been obtained, we update the set of active hypotheses. For each plane $\mathbf{p}_i^{\text{act}} \in \mathcal{P}_{\text{act}}$ we look for the closest plane $\mathbf{p}_*^t \in \mathcal{P}_t$, such that $\|\mathbf{n}_i^{\text{act}} \times \mathbf{n}_*^t\| < \mu$ and $|d_i^{\text{act}} - d_*^t| < \epsilon$. If such a plane exists, we set $\mathbf{p}_i^{\text{act}} = \mathbf{p}_*^t$, otherwise the plane $\mathbf{p}_i^{\text{act}}$ remains unchanged. The plane boundaries are updated accordingly, such that the boundaries corresponding to $\mathbf{p}_i^{\text{act}}$ are replaced by the boundaries corresponding to $\mathbf{p}_*^t$. This way, the plane hypotheses established in the first frame of the sequence are continuously tracked and updated.

4. Experiments and Results

We evaluated our plane detection and tracking method on ToF sequences acquired using a PMDVision CamCube camera with a resolution of 204×204 pixels. Our 3 testing sequences consist of 60 frames, each, and show a scene with a table, boxes and walls at various angles. The ToF camera was handheld and underwent arbitrary rotations and translations such that most of the planar structures in the scene were visible in all frames.

We used the following two error metrics to assess our plane detection and tracking approach. The residual error $e_{\text{res}}(t)$ measures the average distance of the points $\mathbf{x}_{i,t} \in \mathcal{X}_t$, $1 \leq i \leq N$, from the planes estimated for frame t, and is defined as

$$e_{\text{res}}(t) = \frac{1}{N} \sum_{i=1}^{N} \min_{\mathbf{p} \in \mathcal{P}_{\text{act}}} d_{\perp}(\mathbf{x}_{t,i}, \mathbf{p}). \tag{6}$$

While $e_{\text{res}}(t)$ quantifies the plane detection quality for a single frame, we assess the tracking precision across multiple frames using the angular error $e_{\text{ang}}(t)$. Since the scene is rigid, the relative orientation of the plane normals should remain constant over time, despite the movement of the ToF camera. We therefore measure the deviation of all pairwise plane orientations in frame t from those in the first frame,

$$e_{\text{ang}}(t) = \frac{1}{M_{\text{act}}^2} \sum_{i,j=1}^{M_{\text{act}}} |\mathbf{A}_{i,j}^t - \mathbf{A}_{i,j}^0|, \tag{7}$$

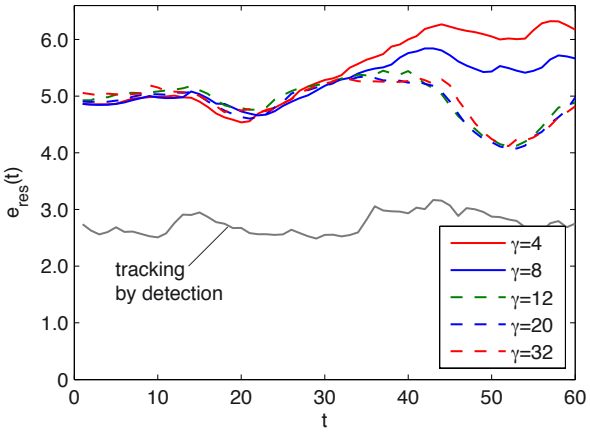
where $\mathbf{A}^t$ is a symmetric matrix containing the angles between all pairs of planes in the active hypotheses set $\mathcal{P}_{\text{act}}$ at time t. We expect $e_{\text{ang}}(t)$ to be small and nearly constant for all $1 \leq t \leq T$.

Figure 4(a) shows plots of $e_{\text{res}}(t)$ for one of the testing sequences, averaged over 10 repetitions per graph, in order to decrease the influence of random effects. We varied the parameter γ that represents the maximum rotation angle used in the creation of guided hypotheses (Section 3.4). For each rotation axis, we selected $\rho = 5$ values from the interval $[-\gamma, \gamma]$. For values of γ up to 8 degrees, the residual error grows significantly after frame 30, indicating that the tracked planes increasingly deviate from the point data. The rotation angles provided for the hypothesis extraction step are thus insufficient to capture the true rotation induced by the moving ToF camera. Values of $\gamma \geq 12$ allow keeping track of the planes throughout the sequence, without a deterioration of residual error even for very large γ. Obviously, the chosen value for γ should reflect the expected maximum camera movement for a sequence. The dashed graphs follow a common shape corresponding to the camera movement that, after rotating to the side, returns to the initial orientation. Between frames 25 and 40, the residual increases, caused by an appearing side of an object that

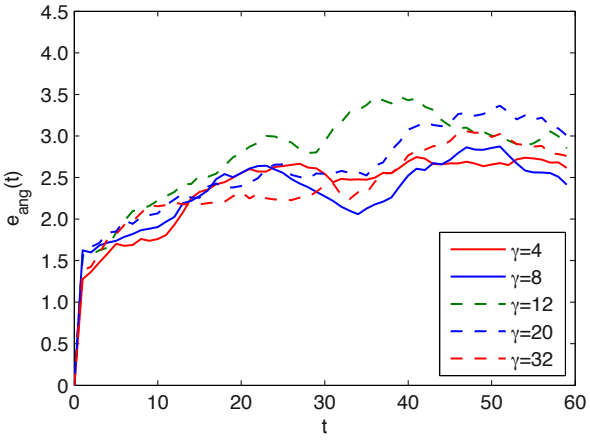
was hidden during the first frame, when plane hypotheses were initialized. For comparison, Figure 4(a) also shows the residual graph for a pure detection approach, where the full J-linkage algorithm is used in every frame without propagation of plane hypotheses. In this case, the residual error is lower and less dependent on changes in the camera orientation. The reason for this behavior is intuitive: the tracking approach relies on the planes detected in the first frame, while the pure detection variant is able to identify new planar structures as they appear during the sequence. However, our proposed tracking extension to J-linkage results in a significant performance gain. While a multi-pass detection run (consisting of 3-4 passes) using the full J-linkage algorithm takes 50 seconds for a single frame on average, a frame can be processed in 10 seconds with our tracking extension. Our Matlab implementation is based on the J-linkage algorithm code of Toldo and Fusiello, available online [26].

Figure 4(b) gives the angular deviations $e_{\text{ang}}(t)$ corresponding to the above experiments. For all settings of γ, low deviations around 2 degrees are measured, with a slight increase by about 1.5 degrees through the sequence. This indicates that the mutual orientation of the tracked planes remains stable. Incorrect plane estimations would lead to inconsistent plane orientations, and thus, to a significant increase of the deviation. Note that $e_{\text{ang}}(t)$ does not deteriorate for small values of γ, as is the case for $e_{\text{res}}(t)$. For these settings, tracking is lost for multiple planes at the same time, with the effect that their plane equations are not updated any more and their mutual orientation does not change.

A qualitative assessment of the proposed plane tracking method can be made by means of Figure 3. Selected frames from three testing sequences are shown with superimposed points indicating detected and tracked planes. Note that planes disappear in both sequences and are correctly tracked after reappearance. The reason is that plane hypotheses are kept in the set of active hypotheses and simply do not get updated if no corresponding plane is detected.



(a) Residual error between points and tracked planes.



(b) Deviation of plane orientations (degrees).

Figure 4. Average residual error and angular deviation for 10 experiments on a testing sequence of 60 frames. The maximum rotation angle parameter γ is varied. The residual error is also plotted for a pure tracking-by-detection variant using full J-linkage.

5. Conclusion

We have presented a method for detection and tracking of multiple planes in depth images from a ToF camera. Our approach builds upon the recent J-linkage algorithm for detection of multiple instances of a model in noisy data [26] and extends it to plane tracking. The proposed extension significantly reduces the computational complexity required on each image by propagating plane hypotheses from one image to the next. In order to cope with noise and sampling artifacts of ToF cameras, we also introduce a multi-pass strategy applied for every frame. Our evaluation indicates that planes can be consistently tracked in image sequences acquired with a moving ToF camera. The achievable accuracy is close to that of a detection approach using full J-linkage, at a significantly improved computational performance. Future work includes an extension that allows adding new planes during the tracking phase by dynamically extending the set of active plane hypotheses. An optimization with respect to processing speed seems promising, considering recent work on efficient real-time implementation of the J-linkage algorithm [27].

Acknowledgements This work was partially supported by the German Federal Ministry of Education and Research (AVILUS project, grant no. 01 IM 08 001 A).

References

[1] C. Baillard and A. Zisserman. A plane-sweep strategy for the 3d reconstruction of buildings from multiple images. In *ISPRS Journal of Photogrammetry and Remote Sensing*, pages 56–62, 2000. 1, 2

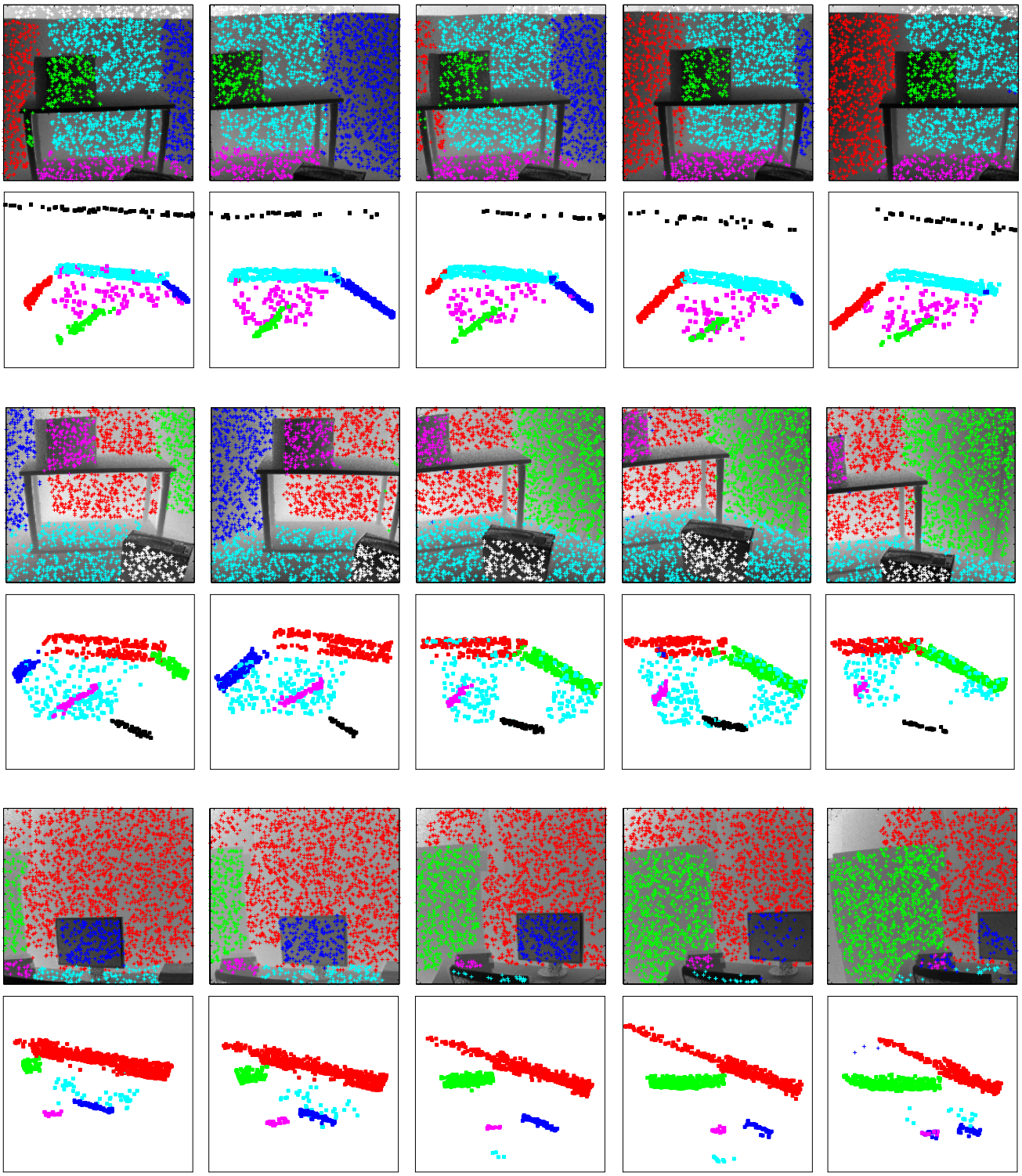


Figure 3. Qualitative illustration of plane tracking results for selected frames of three testing sequences. Top rows: Sampled points, colored according to their plane membership, overlaid ontop of original ToF images. Bottom rows: Corresponding views from above the 3D scene with sampled points colored accordingly. Notice that planes disappear and are tracked again after reappearance.

[2] H. Baltzakis and P. E. Trahanias. Iterative computation of 3d plane parameters. *Image and Vision Computing*, 18(14):1093 – 1100, 2000. 2

[3] A. Bartoli. A random sampling strategy for piecewise planar scene segmentation. *Computer Vision and Image Understanding (CVIU)*, 105:2007, 2007. 1, 2

[4] J. M. Biosca and J. L. Lerma. Unsupervised robust planar segmentation of terrestrial laser scanner point clouds based on fuzzy clustering methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, 63(1):84 – 98, 2008. 2

[5] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 24(5):603 –619, may 2002. 2

[6] Y. Ding, X. Ping, M. Hu, and D. Wang. Range image segmentation based on randomized hough transform. *Pattern Recogn. Lett.*, 26(13):2033–2041, 2005. 2

[7] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 2

[8] D. Fouhey, D. Scharstein, and A. Briggs. Multiple plane detection in image pairs using j-linkage. *Int. Conf. on Pattern Recognition (ICPR)*, 2010. 1, 2

[9] F. Fraundorfer, K. Schindler, and H. Bischof. Piecewise planar scene reconstruction from sparse correspondences. *Image Vision Comput.*, 24(4):395–406, 2006. 1, 2

[10] D. Gallup, J. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1418–1425, 2010. 1, 2

[11] D. Hahnel, W. Burgard, and S. Thrun. Learning compact 3d models of indoor and outdoor environments with a mobile robot. *Robotics and Autonomous Systems*, 44(1):15 – 27, 2003. 1, 2

[12] M. Heracles, B. Bolder, and C. Goerick. Fast detection of arbitrary planar surfaces from unreliable 3d data. In *Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 5717–5724, 2009. 1, 2

[13] D. Holz, R. Schnabel, D. Droeschel, J. Stückler, and S. Behnke. Towards semantic scene analysis with time-of-flight cameras. *RoboCup International Symposium*, 2010. 2

[14] A. Imiya and I. Fermin. Voting method for planarity and motion detection. *Image and Vision Computing*, 17(12):867 – 879, 1999. 2

[15] A. Leonardis, A. Gupta, and R. Bajcsy. Segmentation of range images as the search for geometric parametric models. *Int. J. of Computer Vision (IJCV)*, 14(3):253–277, 1995. 2

[16] M. Lourakis, A. Argyros, and S. Orphanoudakis. Detecting planes in an uncalibrated image pair. In *British Machine Vision Conference (BMVC)*, page Poster Session, 2002. 2

[17] B. Micusik and J. Kosecka. Piecewise planar city 3d modeling from street view panoramic sequences. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 2906–2912, 2009. 1, 2

[18] K. Okada, S. Kagami, M. Inaba, and H. Inoue. Plane segment finder: algorithm, implementation and applications. In *Int. Conf. on Robotics and Automation (ICRA)*, volume 2, pages 2120 – 2125 vol.2, 2001. 2

[19] K. Pathak, A. Birk, N. Vaskevicius, and J. Poppinga. Fast registration based on noisy planes with unknown correspondences for 3-d mapping. *Robotics, IEEE Transactions on*, 26(3):424 –441, jun. 2010. 1

[20] J. Poppinga, N. Vaskevicius, A. Birk, and K. Pathak. Fast plane detection and polygonalization in noisy 3d range images. *International Conference on Intelligent Robots and Systems (IROS)*, 2008. 2

[21] J. Prankl, M. Zillich, B. Leibe, and M. Vincze. Incremental model selection for detection and tracking of planar surfaces. In *British Machine Vision Conference (BMVC)*. BMVA Press, 2010. 2

[22] L. Schwarz, D. Mateus, V. Castaneda, and N. Navab. Manifold learning for ToF-based human body tracking and activity recognition. In *British Machine Vision Conference (BMVC)*. BMVA Press, 2010. 1

[23] G. Silveira, E. Malis, and P. Rives. Real-time robust detection of planar regions in a pair of images. In *Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 49 –54, 2006. 1, 2

[24] G. Simon, A. Fitzgibbon, and A. Zisserman. Markerless tracking using planar structures in the scene. In *Int. Symp. on Augmented Reality (ISAR)*, pages 120 –128, 2000. 1, 2

[25] S. Sinha, D. Steedly, and R. Szeliski. Piecewise planar stereo for image-based rendering. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 1881–1888, 2009. 2

[26] R. Toldo and A. Fusiello. Robust multiple structures estimation with j-linkage. *European Conference on Computer Vision (ECCV)*, pages 537–547, 2008. 1, 3, 5, 6

[27] R. Toldo and A. Fusiello. Real-time incremental j-linkage for robust multiple structures estimation. *International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, 2010. 1, 2, 6

[28] Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *Int. J. of Computer Vision (IJCV)*, 13(2), 1994. 1

[29] M. Zucchelli, J. Santos-Victor, and H. I. Christensen. Multiple plane segmentation using optical flow. In *British Machine Vision Conference (BMVC)*, pages 313–322, 2002. 2