UC Santa Cruz

UC Santa Cruz Previously Published Works

Title

Multi-Planar Fitting in an Indoor Manhattan World

Permalink

https://escholarship.org/uc/item/2238w2k7

Authors

Kim, Seongdo
Manduchi, Roberto

Publication Date

2017-03-01

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Multi-Planar Fitting in an Indoor Manhattan World

Seongdo Kim Roberto Manduchi
UC Santa Cruz
{seongdo,manduchi}@soe.ucsc.edu

Abstract

We present an algorithm that finds planar structures in a Manhattan world from two pictures taken from different viewpoints with unknown baseline. The Manhattan world assumption constrains the homographies induced by the visible planes on the image pair, thus enabling robust reconstruction. We extend the T-linkage algorithm for multi-structure discovery to account for constrained homographies, and introduce algorithms for sample point selection and orientation-preserving cluster merging. Results are presented on three indoor data set, showing the benefit of the proposed constraints and algorithms.

1. Introduction

The topic of geometry reconstruction of indoor environments has received substantial attention in recent years. Both standard and range cameras can be used for this task; while range cameras provide obvious advantages, standard cameras are far more ubiquitous. We are particularly interested in structure-from-motion geometric reconstruction on commodity smartphones as a potential assistive technology tool for blind persons. We envision a system that, from two or more pictures taken by a blind user from slightly different viewpoints, could infer the dominant planar geometry of the visible scene, and communicate it via a suitable non-visual interface (e.g., synthetic speech) to the user. This paper tackles the geometric reconstruction problem for standard cameras, specializing it for a Manhattan World (MW) scene [3], where all surfaces are assumed to be planar and mutually parallel or orthogonal. The MW assumption has been considered frequently in the literature; its use is justified for indoors environments as the geometry of the space inside most buildings can be expected to comply with a regular MW.

The MW assumption places strict constraints on the variety of surfaces that need to be considered. It also provides a convenient way to estimate the camera orientation at any viewpoint, provided that the vanishing points of the lines visible in the scene can be computed, and that the camera intrinsic calibration is known [17]. In fact, given that the vanishing points in a Manhattan World have mutually orthogonal directions, and assuming that the smartphone has an accelerometer that can be used to estimate the phone inclination with respect to the vertical (all modern smartphones do), only one vanishing point from a horizontal line bundle needs to be computed. Knowledge of the cameras' orientation is very helpful, as it constrains (reduces the degrees of freedom of) the homographies that map planes seen from the two different viewpoints. The main contribution of this paper is in the careful use of this constraint in the context of a RANSAC-based multi-structure estimation procedure for scenes with a relatively small number of available feature points. Note that, while most feature-based geometry reconstruction works in the literature assume that many hundreds or thousands of feature points are available in an image, indoor scenes are often characterized by the presence of large textureless extents (flat walls), as well as of substantial specularities, which limit the number of available feature matches.

In this work, we use the MW assumption to condition T-linkage [22], a clustering technique that builds on the original J-linkage [29] algorithm. Unlike traditional sequential RANSAC approaches that find one structure at a time (where each structure is associated with exactly one model), J- and T-linkage cluster points that are associated to possibly multiple similar models. This enables points belonging to the same structure to coalesce, even when no one model can perfectly represent the structure. We enforce MW-induced geometric constraints on the individual models found through random sampling at the beginning of J- or T-Linkage. We also propose two techniques for selecting samples that are likely to belong to the same planar structure. In addition, we introduce a simple technique for geometry-aware cluster merging, in order to reduce the oversegmentation effect typical of J- and T-linkage.

2. Previous Work

RANSAC [5] is arguably the most popular method for robust estimation of parametric models, and its extension to multiple model estimation has been studied extensively.

Examples include sequential RANSAC [32, 15], multi-RANSAC [35] , FLoSS [18], and CC-RANSAC [8]. An extensive survey of RANSAC-based methods can be found in [24].

While RANSAC looks to maximize the cardinality of the consensus (or inliers) set, different criteria can be used to identify dominant models. For example, J-Linkage [29] and T-Linkage [22] consider *preference sets* and *functions*, which measure how many models explain a given data point or a cluster of points. Fouhey et al. [6] used J-linkage to discover planes in the scene, and proposed a model reduction step to reduce oversegmentation. Other approaches used mode seeking over hypergraphs [34] or energy minimization criteria [30]. For example, SA-RCM [23] and PEaRL [14] defined cost functions that measure goodness of fit and degree of model complexity, along with spatial coherence.

Our approach to multi-planar estimation exploits Manhattan world constraints on the homography induced by a plane. In previous work, Saurer et al. [26] incorporated weak Manhattan constraints (where planes are only assumed to be all parallel to the vertical direction) in homography estimation, using an accelerometer to measure the gravity direction. Szpak et al. [19] estimated multiple homographies from two images, using two sets of explicit constraints derived from the epipolar geometry.

A number of researchers have proposed techniques for layout estimation of environments from single images using Manhattan or box world hypotheses, showing excellent results [20, 11, 12, 4]. Our work, which uses two images from different viewpoints and thus enables depth computation via triangulation, could certainly be combined with single-image layout estimation, for example to propagate depth values computed for feature points to whole planar surfaces. In fact, we use a very simple single-image estimation technique (orientation map [20]) for sample point selection in a variant of our algorithm.

3. Method

3.1. MW-Constrained Homographies

Saurer et al. [26] showed that the homography relating two views of a vertical plane (that is, parallel to the gravity direction) can be represented by a matrix $\mathbf{H}$ with only 6 unknown entries (assuming that a proper homography has been pre-applied to both images, effectively rotating the cameras so that their focal planes are also vertical). If the normal $\mathbf{n}$ of the plane is known, there is one additional linear constraint between two entries of $\mathbf{H}$. In addition, one non-linear constraint can be found due to the fact that one singular value of $\mathbf{H}$ must be equal to 1. Based on these observations, Saurer et al. [26] showed that a pair of feature matches across views determines 4 solutions for $\mathbf{H}$.

In the Manhattan World (MW) case, observation of at least one horizontal vanishing point, together with information from the accelerometer (which indicates the direction of gravity, assumed to be aligned with one of the three MW cardinal directions), enable estimation of the orientation of both cameras with respect to the cardinal MW reference system. It is thus possible to compute and apply an homography to both images, that is equivalent to rotating both cameras such that their axes are mutually parallel to the MW cardinal axes. Note the difference with the weak Manhattan world assumption [26]: knowledge of the cameras' orientation allows us to set the rotation matrix $\mathbf{R}$ between the two rotated cameras to the identity. In addition, the normal to any given plane in the MW scene, expressed with respect to the MW frame, is a one-hot vector (two entries equal to 0 and one equal to 1). We can enumerate the three plane normals by the index of their non-null entry index, e.g. $\mathbf{n}_2 = [0\ 1\ 0]^T$.

Assuming without loss of generality that the intrinsic calibration matrices are equal to the identity, the canonical homography decomposition [10]

$$\mathbf{H} = \mathbf{R} + \frac{1}{d}\mathbf{t}\mathbf{n}^T \tag{1}$$

that relates the two images of the same plane with normal $\mathbf{n}$ taken from viewpoints separated by $\mathbf{t}$ (where d is the distance of the first viewpoint from the plane) takes one of three possible forms depending on the plane orientation. For example, for a plane oriented as $\mathbf{n}_1$, the $\mathbf{n}_1$*–constrained homography* is:

$$\mathbf{H}^1 = \begin{bmatrix} 1 + t_x/d & 0 & 0 \\ t_y/d & 1 & 0 \\ t_z/d & 0 & 1 \end{bmatrix} \tag{2}$$

Hence, as already noted by Saurer et al. [26], the homography $\mathbf{H}$ has only 3 degrees of freedom (DOF), rather than 8. By constraining the space of possible homographies, more robust planar estimation can be expected (under the assumption that the orientation of the cameras with respect to the cardinal axes of the Manhattan World has been computed correctly). The direct linear transformation (DLT) method [10] can be used to estimate $\mathbf{t}/d$ an thus the (constrained) homography $\mathbf{H}$. We use two feature matches across the image pair to build a planar hypothesis, although it would be possible to use a single feature match using a method similar to [26].

3.2. MW-Constrained Multiplanar Clustering

3.2.1 J- and T-linkage Clustering

J-linkage [29] is a method for robust estimation of multiple parametric structures. Like RANSAC, it generates a

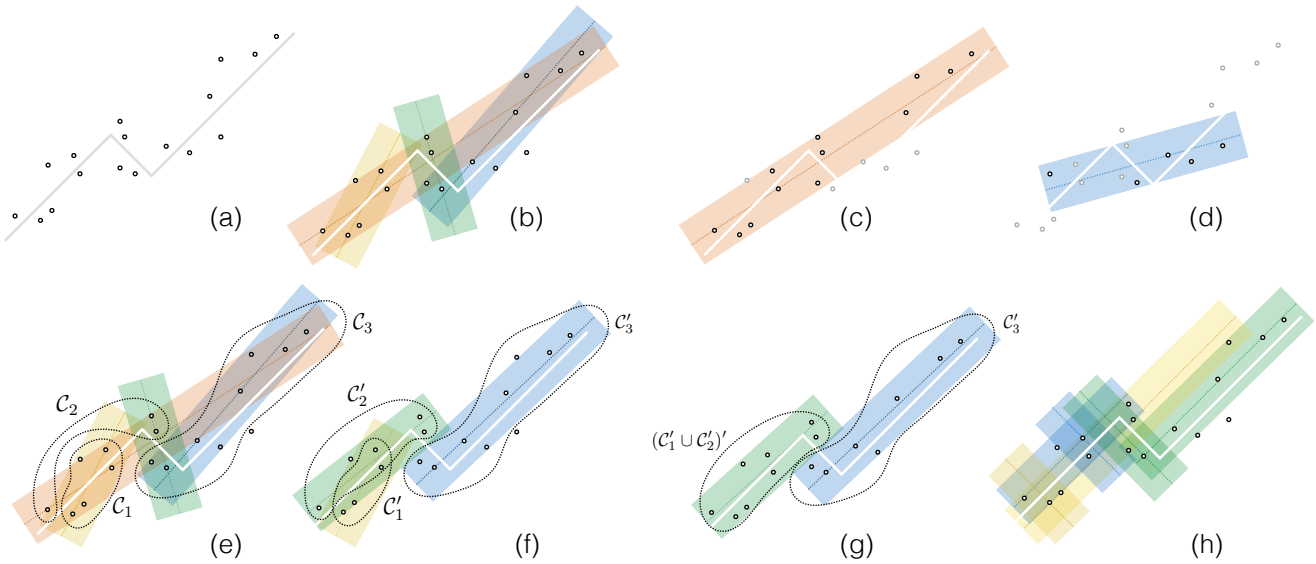


Figure 1. A toy example to illustrate some of the main concepts. (a): Original set of points. The goal is to cluster these points using a piecewise linear model. (b): A set of hypothesized models from randomly chosen point pairs, shown with their consensus sets. Sequential RANSAC determines the model with the largest consensus set (c) and removes it from the set of points, then proceeds iteratively (d). Note that this operation may remove some of the model points of other true models. (e): The clusters determined by J-linkage. These clusters can be refined (f) and merged (g) as explained in Sec. 3.4. (h): A set of hypothesized models constrained by the two possible orientations of the underlying model.

large number of hypotheses (models) using randomly sampled minimal sample sets (that is, sets of points[1] with the smallest cardinality required to build a model). The hope is that at least one of these randomly generated models is close (under a proper metric) to the"true" model that generated some of the data.

We define *model point* of a "true" model any point that is generated by this model. The *consensus set (CS)* of a candidate model is defined by the set of points that are within a certain distance ϵ to the model (see e.g. Fig. 1 (b)). RANSAC and similar methods analyze the consensus sets of models generated from randomly chosen minimal sample sets, in hopes to identify which of these models is closest to a true model. The underlying assumption is that the consensus set of a true model (1) has relatively large cardinality, and (2) is composed for the most part of model points. This assumption is at the basis of sequential RANSAC and its derivatives (e.g. MultiRANSAC [35]), which iteratively select the model with the largest consensus set, remove the points in its consensus set from the pool of data points, and recompute the models on the remaining points. The expectation is that, by removing the consensus set of a model, all traces of this model from the data will disappear, thus facilitating discovery of other true models. In fact, due to noise, model points of a given true model may appear in the consensus sets of several other models: removing the consensus set of a model may have the undesired effect of removing model points from other true models (see Fig. 1 (b) and (c)). In addition, models with large consensus sets often span across multiple true model, as in the case of Fig. 1 [27]. As a result, the models produced by sequential RANSAC are often unsatisfactory.

The J-linkage algorithm [29] takes a different approach. After computing the consensus sets of all hypotheses from randomly generated minimal sample sets, it clusters the data in a way that satisfies the following criteria: **Criterion 1**: Points in the same cluster must all belong to the consensus set of one model, or to the intersection of the consensus sets of two or more models. **Criterion 2**: Points in two different clusters cannot all belong to the same consensus set. Using the terminology in [29], the *preference set (PS)* of a point is the set of models that have this point in their consensus set, while the preference set of a cluster of points is the intersection of the preference sets of these points. Hence, a cluster produced by J-linkage must have non-void preference set (Criterion 1), and two different clusters must have disjoint preference sets (Criterion 2). Criterion 1 allows clusters to be proper subsets of consensus sets, and consensus sets to split across clusters. It thus mitigates a critical problem associated with sequential RANSAC, which greedily removes whole consensus sets. Criterion 2 prohibits two proper sub-

[1]We use the term "point" to indicate a generic datum used for model estimation. In the context of this application, a "point" is a feature match between two images, and a "model" is a (constrained) homography.

sets of the same consensus set from forming distinct clusters, thus reducing the risk of oversegmentation.

J-linkage computes a clustering satisfying these two constraints using an agglomerative approach based on the intuition that two clusters are likely to contain points of the same true model when they both belong to the consensus set of one or more models (that is, when they have overlapping preference sets). Formally, J-linkage starts by assigning a cluster to each point. It then iteratively selects and merges the two clusters with the smallest *Jaccard distance JD* (one minus intersection over union) of their preference sets, provided that this distance is less than 1, until no more merging is possible (Fig. 1 (e)). The preference set of the merged cluster is equal to the intersection of the preference sets of the two clusters. Note that this greedy strategy is guaranteed to produce a clustering that satisfies both Criterion 1 and 2. T-linkage [22] is a variation of J-linkage that defines the *preference function PF* of a cluster of points as a vector whose i-th component represents the minimum over the points in the cluster of a decreasing function of each point's distance to the i-th model. T-linkage greedily merges the two clusters with minimum *Tanimoto distance TD* of their preference functions, where the Tanimoto distance is a generalization of the Jaccard distance to real-valued vectors. T-linkage is simple to implement, has very few parameters to tune, and produces state of the art results [22].

3.2.2 MW-Constrained J- and T-linkage

When using J- or T-linkage to compute 8-DOF homographies induced by multiple planes in arbitrary orientations, a minimal sample set has 4 points and defines one homography. In the case of MW geometry, we modify this baseline algorithm to account for the reduced model DOF (see Algorithm 1). Specifically, we sample sets of 2 points (feature matches); each sample set defines 3 models (homographies), one per possible plane normal direction ($\mathbf{n}_1$, $\mathbf{n}_2$, or $\mathbf{n}_3$ – see Fig. 1 (h) for a 2-D case.) Then, agglomerative clustering, based on J- or T-distance of preference sets or functions is performed on the model sets, independently for the three normal directions. This results in a final set of clusters, each characterized by its normal direction. In addition, we enforce a simple visibility constraint [20] that states that the image of a plane with normal $\mathbf{n}_k$ cannot cross the line L_k defined by the two vanishing points other than $\mathbf{n}_k$. Specifically, two clusters are merged (lines 4 and 14 in Algorithm 1) for normal $\mathbf{n}_k$ only if their points are on the same side of L_k.

Algorithm 1 MW-constrained T-linkage

Input: Points $\{p_i\}$, normal directions $(\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3)$
Output: Co-planar point clusters $\{\mathcal{C}_j^k\}$ for each normal direction $\mathbf{n}_k$

1: **for** j from 1 to MAXITER **do**
2: sample 2 points (p_{j_1}, p_{j_2})
3: **for all** normal directions $\mathbf{n}_k$ **do**
4: **if** p_{j_1}, p_{j_2} on same side of L_k **then**
5: set $\mathcal{C}_j^k = \{p_{j_1}, p_{j_2}\}$
6: compute $\mathbf{n}_k$–constrained homography H_j^k
7: **end if**
8: **end for**
9: **end for**
10: **for** k from 1 to 3 **do**
11: **repeat**
12: find the two clusters $\mathcal{C}_i^k, \mathcal{C}_j^k$ on same side of L_k with minimal Tanimoto distance of their PFs
13: **if** TD(PF($\mathcal{C}_i^k$,)PF($\mathcal{C}_j^k$)) < 1 **then**
14: merge $\mathcal{C}_i^k, \mathcal{C}_j^k$ into one cluster $\mathcal{C}_n^k = \mathcal{C}_i^k \cup \mathcal{C}_j^k$
15: **end if**
16: **until** no more merges are possible
17: **end for**

3.3. Region-Based Sample Selection

The efficiency of random sampling can be increased if points pairs that are unlikely to be coplanar are removed from the sample set. We propose two algorithms that use single-image analysis to only select samples with good likelihood of being coplanar. The first algorithm computes the *orientation map* [20] in one of the images in the pair (see Fig. 3 (f)). The orientation map defines connected regions in the image for each normal orientation $\mathbf{n}_k$. These regions are obtained by sweeping edge segments around the vanishing point of their supporting line, until the sweep is "blocked" by another line. The intersection of two sweeps around two vanishing points forms a region of points that are assumed to belong to a plane with normal oriented along the remaining vanishing point. (See [20] for details.) In this algorithm (MW-OM), point pairs are only sampled from within the same region in the orientation map, and $\mathbf{n}_k$–constrained homography models are built only for the orientation represented by that region.

Our second algorithm (MW-RS; see Fig. 2) for sample selection defines a quadrilateral region around each feature point p as follows. For each vanishing point v_i, the closest edge segments S_{i_1}, S_{i_2} intersecting the line joining p and v_i are found on either side of p. Let d_i be the distance between p and the closest segment intersecting the line joining p and v_i. The vanishing points v_i, v_j with associated smallest values d_i, d_j in $\{d_1, d_2, d_3\}$ are found; the segments $\{S_{i_1}, S_{i_2}, S_{j_1}, S_{j_2}\}$ are used to define a quadrilateral region $\mathcal{R}(p)$ with orientation $\mathbf{n}_k$ equal to the direction to the remaining vanishing point.

We found that, in typical indoor scenarios, this simple algorithm finds local coplanar regions quite robustly. Rather

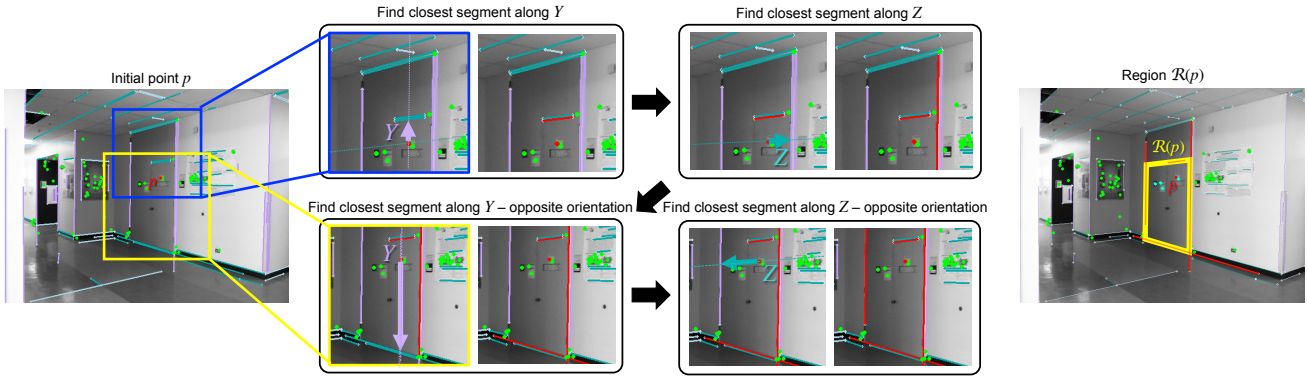


Figure 2. An example of construction of region $\mathcal{R}(p)$ as described in Sec. 3.3 (only search along 2 axes shown). The point p is shown in red; lines in the image oriented along the $n_1 = Y$ and $n_3 = Z$ Manhattan world axes are shown in purple and teal, respectively. The resulting region $\mathcal{R}(p)$ is shown in yellow. All feature points inside $\mathcal{R}(p)$, shown in pale blue along with p, are used to compute a n_1-constrained homography.

than sampling two points, we actually consider all feature points contained within $\mathcal{R}(p)$ to create a $\mathbf{n}_k$–constrained homography using DLT. (In our experiments, the regions $\mathcal{R}(p)$ contained 9 feature points on average.) Hence, the total number of models considered is equal to the number of feature matches in the image.

3.4. Cluster Merging

J- and T-linkage are often prone to over-segmentation. This is due to the rather restrictive requirement that all points in a cluster must belong to the consensus set of at least one model. If no model is sufficiently aligned with a true model, multiple models may need to be employed to explain the model points of the same true model. Fouhey et al. [6], who first pointed out this problem, proposed a post-processing cluster merging algorithm that aims to find large consistent models. A distance measure between two clusters is defined as the mean residual of the least square parametric fit to all points in the two clusters. At each iteration, the two clusters with minimal distance are merged, provided that this distance is smaller than the threshold ϵ used to define the consensus set of a model. This is a reasonable strategy, which however carries the risk of merging clusters that are close in space, even when they come from different true models. This is because a low-residual fit to nearby clusters of points can often be found, regardless of whether the clusters are generated by the same or different true models.

We propose an alternative strategy based on the following simple idea: two clusters $\mathcal{C}_i, \mathcal{C}_j$ should be merged if this results in a super-cluster that explains the data in substantially the same way as the individual clusters. This notion could be formalized by fitting an homography (as in [6]) to the points (matches) of the merged cluster $\mathcal{C}_i \cup \mathcal{C}_j$. If the consensus set of the resulting homography is similar (e.g. as measured by the Jaccard distance JD) to $\mathcal{C}_i \cup \mathcal{C}_j$, then $\mathcal{C}_i$ and $\mathcal{C}_j$ are good candidates for merging. We found it beneficial to first fit an homography to each individual cluster, and then consider the consensus set of this homography (called the *refined version* of the cluster), indicated as $\bar{\mathcal{C}}_i$. Note that the clusters of points produced by J- or T-linkage do not, in general, span the consensus set of any model, and thus one should expect $\bar{\mathcal{C}}_i \neq \mathcal{C}$ in general. In fact, the refined clusters $\bar{\mathcal{C}}_i$ do not, in general, satisfy Criterion 1 in Sec. 3.2. See Fig. 1 (f). Our algorithm proceeds iteratively by considering, for each normal direction $\mathbf{n}_k$, all pairs of clusters $\mathcal{C}_i^k, \mathcal{C}_j^k$ with Jaccard distance less than a threshold τ, and substituting two such clusters with the union of their refined versions $\bar{\mathcal{C}}_i^k \cup \bar{\mathcal{C}}_j^k$ when the Jaccard distance between this set and its refined version is less than a threshold τ (see Algorithm 2).

4. Experiments

4.1. Implementation Details

Feature points are found using the SIFT algorithm [21] and matched across images in a pair; features that don't have a correct match within a radius of 3 pixels are set as outliers. Then, feature points are hand-labeled according to the planar structure (if any) they belong to in the image (each visible plane is assigned an index). This represents the "ground truth" used to assess our algorithms. Consensus sets in the planar clustering algorithms are computed with maximum reprojection error threshold of 2 pixels. The threshold τ on the Jaccard distance in Algorithm 2 was set to 0.5. The number of iterations (MAXITER) in Algorithm 1 was set to 5000. However, note that in the MW-RS algorithm of Sec. 3.3, the number of iterations is equal to the number of feature points in the image (144 on average). On a Intel i7 3.5GHz machine with 8GB memory, Algorithm 1

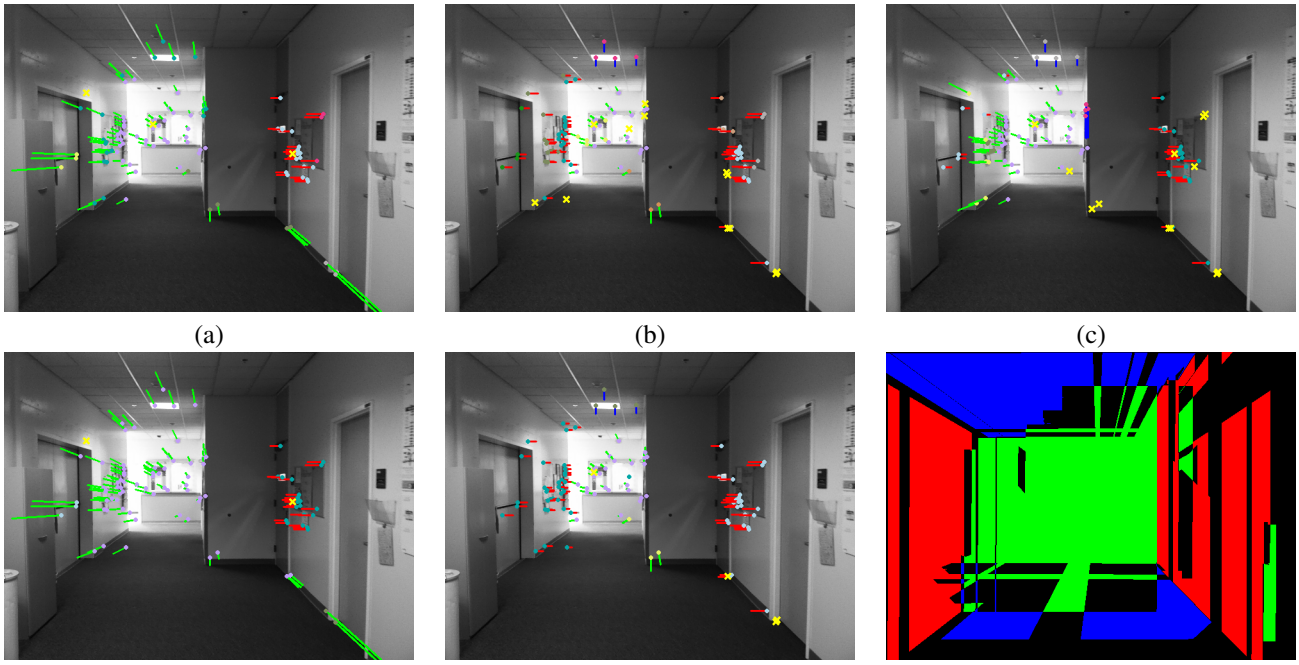


Figure 3. Examples of planar clustering for one frame pair in the sequence New:Corridor3 using different algorithms. (a) MW (ARI=0.488); (b) MW-RS (ARI=0.589); (c) MW-OM (ARI=0.448); (d) MW+mrg (ARI=0.416); (e) MW-RS+mrg (ARI=0.676). The orientation map used for sample selection in MW-OM is shown in (f). The colored segments represent the estimated normal for the cluster, with color indicating the normal direction. Colored dots represent feature points, with the color marking the cluster identity. Outliers are shown as yellow crosses.

Algorithm 2 Cluster merging

Input: Point clusters $\{\mathcal{C}_j^1\}, \{\mathcal{C}_j^2\}, \{\mathcal{C}_j^3\}$ from MW-constrained T-linkage

Output: Refined clusters $\{\mathcal{C}_j^1\}, \{\mathcal{C}_j^2\}, \{\mathcal{C}_j^3\}$

repeat
  for k from 1 to 3 **do**
    find the two clusters $\mathcal{C}_i^k, \mathcal{C}_j^k$ on same side of $\mathcal{L}_k$ with minimum Jaccard distance
    if JD$(\mathcal{C}_i^k, \mathcal{C}_j^k) < \tau$ **then**
      compute $\mathbf{n}_k$–constrained homographies from $\mathcal{C}_i^k, \mathcal{C}_j^k$ and associated consensus sets $\bar{\mathcal{C}}_i^k, \bar{\mathcal{C}}_j^k$
      compute $\mathbf{n}_k$–constrained homography from $\bar{\mathcal{C}}_i^k \cup \bar{\mathcal{C}}_j^k$ and associated consensus set $\overline{\bar{\mathcal{C}}_i^k \cup \bar{\mathcal{C}}_j^k}$
      if JD$(\bar{\mathcal{C}}_i^k \cup \bar{\mathcal{C}}_j^k, \overline{\bar{\mathcal{C}}_i^k \cup \bar{\mathcal{C}}_j^k}) < \tau$ **then**
        merge $\mathcal{C}_i^k, \mathcal{C}_j^k$ into one cluster $\mathcal{C}_n^k = \mathcal{C}_i^k \cup \mathcal{C}_j^k$
      end if
    end if
  end for
until no more merges are possible

took on average 1.56 second per frame, while MW-RS took on average 0.94 seconds per frame (both algorithms were implemented in C++).

To detect vanishing directions, we use an algorithm similar to [28]. We first detect line segments using the Line Segment Detection algorithm [9], and keep those with length larger than a threshold τ_l, which is equal to the diagonal length of the image divided by 30. We then cluster these segments using T-linkage (where each pair of lines hypothesizes a vanishing point, and consensus sets are found based on the consistency measure between a line and a vanishing point proposed in [28]). We consider a total number of 500 vanishing point hypotheses. For each output line cluster, we fit a vanishing point via least squares, then select the three approximately mutually orthogonal vanishing directions $\mathbf{v}_k$ with the largest consensus sets, and order them according to the size of their consensus sets. (Two vanishing directions are considered to be approximately orthogonal if their orientation differ by less than $10°$.) We then orthogonalize the triplet of vanishing directions thus found via QR decomposition. The vanishing direction that forms the smallest angle with the vector [0 1 0] in the camera reference frame is chosen to represent the gravity direction. We refine $\mathbf{v}_k$ using a

non-linear solver [2, 1] with the following cost function:

$$\underset{\mathbf{R}\in SO(3)}{\arg\min} \sum_{k=1}^{3} \sum_{L\in\mathcal{L}_k} \|\mathbf{u}_L \mathbf{R} \mathbf{v}_k\| \tag{3}$$

where $\mathcal{L}_k$ is the cluster of image lines associated with the estimated vanishing direction $\mathbf{v}_k$, and $\mathbf{u}_L$ is the lever vector associated with image line $L \in \mathcal{L}_k$. The *lever vector* [16] $\mathbf{u}_L$ is the normal vector of the plane through the camera's optical center and the image line L (note that $\mathbf{u}_L$, which is easily computable given the intrinsic camera parameters, is orthogonal to the k-th vanishing direction.) Finally, we recompute $\{\mathcal{L}_k\}$ based on the new vanishing direction, this time using a smaller threshold $\tau_s = \tau_l/2$.

4.2. Data Sets

We evaluated our algorithm on three data sets. (1) The Michigan-Milan data set [7] contains a variety of indoor environments, and provides the camera calibration. We only considered 3 sequences from this data set in our experiments (Entrance 1, Entrance 2, and Room 4). This is because the other sequences either contained too few point features (due to large flat walls or periodic tile patterns) for reliable planar fitting using point matches, or because of the presence of objects or clutter that would break the multiplanar Manhattan world assumption. Images in this data set contain an average number of 2.5 planes visible, with 36 feature points per plane on average. (2) The Michigan indoor data set [31], for which camera calibration was available (although substantial residual radial distortion had to be removed via manual calibration). 4.5 planar surfaces are visible in each image on average, and 33 feature points were detected per plane on average. (3) A new data set was collected with images taken inside two buildings in our campus, which we found to be particularly challenging due to the presence of large untextured areas and multiple specularities (Figs. 3, 4 (d)) (We will make this new data set openly available, along with the camera calibration.) Images were taken with an iPhone 6 camera. An average of 4.3 planes are visible per image, with 21 features detected per plane on average.

4.3. Metric

We evaluated the proposed Manhattan World-constrained T-linkage multiplanar fitting algorithm (MW) against regular T-linkage (T-L) and against T-linkage using the weak Manhattan world (WMW) constraint of Saurer et al. [26]. We also considered the point sampling strategies described in Sec. 3.3, and specifically the use of orientation maps [20] (MW-OM) and of regions grown around each feature point (MW-RS). Additionally, we evaluated the benefit of the proposed cluster merging algorithm (Sec. 3.4) against the technique of Fouhey et al. [6]. Note that our cluster merging technique requires knowledge of the normal direction for the planes represented by each cluster, and thus can only be used with variants of the MW algorithm.

All of these algorithms cluster the pool of feature matches into supposedly co-planar sets (plus an outlier cluster). Their results are benchmarked against the ground truth, created as described in the Sec. 4.1. Since we don't have ground truth quantitative information about the 3-D geometry of the scene and the motions of two cameras, we evaluate the results of our algorithms solely based on their ability to generate clusters that resemble the ground truth clusters. For this purpose, we use the adjusted Rand Index (ARI [13, 33]), a standard metric used to compare two partitions. Like the original Rand Index [25], ARI considers all pairs of points and defines the numbers N_{ij}, where i (j) is equal to 1 if the two points belong to the same cluster in the first (second) partition, to 0 otherwise. Then,

$$\text{ARI} = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00}+N_{01})(N_{01}+N_{11}) + (N_{00}+N_{10})(N_{10}+N_{11})} \tag{4}$$

ARI takes on a value of 1 when the two partitions are equivalent, and of 0 when the Rand Index equals its expected value for random clustering [33].

For each sequence in our data sets, several frame pairs were selected for processing, and the resulting ARI values were averaged over the selected frame pairs.

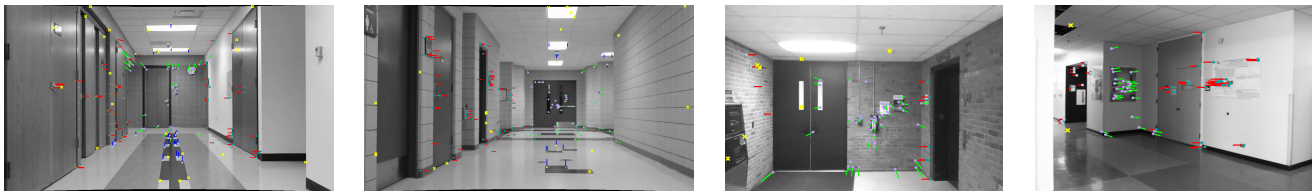
4.4. Results

The ARI results on the considered image data sets are shown in Tab. 1. It is seen that MW give better results on average than WMW, and that both are better than regular T-linkage. Cluster merging using the algorithm of Fouhey et al. [6] is shown to always be detrimental. Our merging technique (Sec. 3.4) applied to MW also decreases performance. When analyzing this phenomenon, we observed that in several cases, a cluster of points found by MW may contain points from the same planar surface, but with incorrectly estimated orientation $\mathbf{n}_k$. During the merging process, this cluster may be merged with another cluster on a plane oriented as $\mathbf{n}_k$, resulting in an incorrectly merged super-cluster. For example, the cluster of points shown in teal in Fig. 3 (a) was assigned an incorrect orientation; it was then incorrectly merged (b) with the set of points marked in purple, which itself spanned multiple planar surfaces.

Both point sampling strategies (MW-OM and MW-RS) gave similar or slightly inferior results than MW. However, after cluster merging, both ARIs increased, with MW-RS achieving the top score. This is because both proposed point sampling techniques ensure that points in each sample are chosen with high likelihood from the same planar surface, and that only one model per sample is built for the

	T-L	T-L+mrgF	WMW	WMW+mrgF	MW	MW+mrgF	MW+mrg	MW-RS	MW-RS+mrg	MW-OM	MW-OM+mrg
Mich:EECSBuilding	0.406	0.270	0.444	0.263	0.443	0.250	0.420	0.520	**0.564**	0.475	**0.596**
Mich:Library	0.637	0.517	0.719	0.450	**0.730**	0.168	0.559	0.682	**0.756**	0.665	0.719
Mich:Library2	0.464	0.220	0.578	0.226	**0.627**	0.234	0.609	0.588	**0.632**	0.576	0.557
Mich:LockerRoom	0.425	0.410	0.514	0.337	0.502	0.347	**0.553**	0.480	**0.601**	0.519	0.539
Mich:Object	0.545	0.356	0.541	0.339	**0.572**	0.368	0.541	0.554	0.552	**0.595**	0.477
MM:Entrance1	0.615	0.698	0.556	0.706	0.518	0.662	0.670	0.439	**0.713**	0.494	**0.753**
MM:Entrance2	0.355	**0.832**	0.531	0.820	0.392	0.692	0.647	0.439	**0.894**	0.319	0.697
MM:Room4	0.683	0.636	0.729	0.678	0.876	0.673	0.952	0.666	**0.970**	0.859	**0.975**
New:Corridor1	0.465	0.289	0.482	0.243	0.663	0.260	0.482	**0.692**	0.677	**0.745**	0.622
New:Corridor2	0.751	0.349	0.331	0.334	0.750	0.331	0.331	0.717	0.760	**0.766**	**0.797**
New:Corridor3	0.496	0.368	0.416	0.342	**0.507**	0.362	0.416	0.494	**0.648**	0.457	0.427
Median	0.496	0.368	0.531	0.339	0.572	0.347	0.553	0.554	**0.677**	0.576	**0.620**
Average	0.531	0.450	0.531	0.431	0.598	0.395	0.562	0.570	**0.706**	0.588	**0.603**

Table 1. ARI results (larger is better) on sequences from three different data sets: the Michigan indoor (Mich:) data set [31], the Michigan-Milan (MM:) data set [7], and a new data set collected in our campus (New:). The algorithms considered are: T-linkage (T-L), Manhattan World-constrained T-linkage (MW), Weak Manhattan World-constrained T-linkage (WMW) [26], MW with samples constrained by the orientation map (MW-OM), and MW with samples from regions grown around each point (MW-RS). The suffix *+mrg* indicates application of our cluster merging procedure (Sec. 3.4), while *+mrgF* uses the algorithm of Fouhey et al. [6]. The best result for each data set is shown in red, the second best in boldface.



(a) (b) (c) (d)

Figure 4. Examples of clustering using MW-RS+mrg, aling with the ARI values computed for these specific pairs. (a) Frame pair from sequence Mich:EECS-Building (ARI=0.595). (b) From Mich:Library2 (ARI=0.590). (c) From MM:Entrance1 (ARI=0.827). (d) From New:Corridor1 (ARI=0.814).

orientation estimated for that surface. Thus, the estimated orientation of the clusters is for the most part correct, mitigating the risk of merging clusters belonging to differently oriented surfaces. It should be noticed that the quality of the results with the proposed point sampling techniques depends on the accuracy of planar segmentation in individual images. For example, several regions in the orientation map shown in Fig. 3 (f) have incorrect orientation, which affects the resulting MW-OM clustering (c). The MW-RS appears to provide a more robust local clustering of coplanar points. Remarkably, the MW-RS algorithm, as noted earlier, uses a much smaller number (3%) of samples than the other methods, where each sample contains 9 points on average.

A sample of results on images from all three data sets using the MW-RS algorithm with final cluster merging is shown in Fig. 4.

5. Conclusions

We have introduced an algorithm for the computation of multiple planar structures in a Manhattan world from a stereo pair with unknown baseline. By constraining the homographies induced by the visible planes on the two images, the Manhattan world assumptions enables robust planar detection. We also introduced two algorithms for selecting sample points for hypothesis generation with high likelihood to belong to the same planar surface, as well as a cluster merging technique that reduces over-segmentation, still using the geometric constraints induced by the Manhattan world hypothesis. Our experiments with three different data sets have shown the benefit of using the proposed geometric constraints. Our cluster merging algorithm has also proven beneficial, but only with the proposed point sampling strategies.

References

[1] S. Agarwal, K. Mierle, and Others. Ceres solver. http://ceres-solver.org.

[2] A. Bjorck. *Numerical Methods for Least Squares Problems*.

[3] J. M. Coughlan and A. L. Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 941–947. IEEE, 1999.

[4] S. Dasgupta, K. Fang, K. Chen, and S. Savarese. Delay: Robust spatial layout estimation for cluttered indoor scenes.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 616–624, 2016.

[5] 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.

[6] D. F. Fouhey, D. Scharstein, and A. J. Briggs. Multiple plane detection in image pairs using j-linkage. In *20th International Conference on Pattern Recognition (ICPR)*, pages 336–339. IEEE, 2010.

[7] A. Furlan, S. D. Miller, D. G. Sorrenti, F.-F. Li, and S. Savarese. Free your camera: 3d indoor scene understanding from arbitrary camera motion. In *BMVC*, 2013.

[8] O. Gallo, R. Manduchi, and A. Rafii. CC-RANSAC: Fitting planes in the presence of multiple surfaces in range data. *Pattern Recognition Letters*, 32(3):403–410, 2011.

[9] R. Grompone, J. Jakubowicz, J. M. Morel, and G. Randall. LSD: A line segment detector. *Image Processing On Line*, 2:35–55, 2012.

[10] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[11] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *European Conference on Computer Vision (ECCV)*, pages 224–237, 2010.

[12] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007.

[13] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[14] H. Isack and Y. Boykov. Energy-based geometric multi-model fitting. *International Journal of Computer Vision*, 97(2):123–147, 2014.

[15] Y. Kanazawa and H. Kawakami. Detection of planar regions with uncalibrated stereo using distributions of feature points. In *Proc. British Machine Vision Conference*, 2004.

[16] C. Kim and R. Manduchi. Planar structures from line correspondences in a manhattan world. In *Computer Vision–ACCV 2014*, pages 509–524. Springer, 2015.

[17] J. Košecká and W. Zhang. Video compass. In *Computer Vision—ECCV 2002*, pages 476–490. Springer, 2002.

[18] N. Lazic, G. Inmar, F. Brendan, and A. Parham. Floss: Facility location for subspace segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 825–832, 2009.

[19] N. Lazic, G. Inmar, F. Brendan, and A. Parham. Robust multiple homography estimation: An ill-solved problem. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2132–2141, 2015.

[20] D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[22] L. Magri and A. Fusiello. T-linkage: a continuous relaxation of j-linkage for multi-model fitting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[23] T. T. Pham, T. J. Chin, J. Yu, and D. Suter. The random cluster model for robust geometric fitting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1658–1671, 2014.

[24] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J. Frahm. USAC: A universal framework for random sample consensus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):2022–2038, 2013.

[25] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[26] O. Saurer, F. Fraundorfer, and M. Pollefeys. Homography based visual odometry with known vertical direction and weak manhattan world assumption. In *Proc. IEEE/IROS Workshop on Visual Control of Mobile Robots*, 2012.

[27] C. V. Stewart. Bias in robust estimation caused by discontinuities and multiple structures. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(8):818–833, 1997.

[28] J. P. Tardif. Non-iterative approach for fast and accurate vanishing point detection. In *Proc. IEEE Conference on Computer Vision*, 2009.

[29] R. Toldo and A. Fusiello. Robust multiple structures estimation with J-linkage. In *Proc. European Conference on Computer Vision*, 2008.

[30] P. H. Torr. An assessment of information criteria for motion model selection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 47–52, 1997.

[31] G. Tsai, C. Xu, J. Liu, and B. Kuipers. Real-time indoor scene understanding using bayesian filtering with motion cues. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 121–128. IEEE, 2011.

[32] E. Vincent and R. Laganiére. Detecting planar homographies in an image pair. In *Proc. 2nd International Symposium on Image and Signal Processing and Analysis*, 2001.

[33] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010.

[34] H. Wang, G. Xiao, Y. Yan, and D. Suter. Mode-seeking on hypergraphs for robust geometric model fitting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2902–2910, 2015.

[35] M. Zuliani, C. S. Kenney, and B. Manjunath. The multi-ransac algorithm and its application to detect planar homographies. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 3, pages III–153. IEEE, 2005.