Extracting Vanishing Points across Multiple Views

Michael Hornáček[1,2] and Stefan Maierhofer[1]
[1]VRVis Research Center
[2]Institute of Photogrammetry and Remote Sensing
Vienna University of Technology

Abstract

The realization that we see lines known to be parallel in space as lines that appear to converge in a corresponding vanishing point has led to techniques employed by artists since at least the Renaissance to render a credible impression of perspective. More recently, it has also led to techniques for recovering information embedded in images pertaining to the geometry of their underlying scene.

In this paper, we explore the extraction of vanishing points in the aim of facilitating the reconstruction of Manhattan-world scenes. In departure from most vanishing point extraction methods, ours extracts a constellation of vanishing points corresponding, respectively, to the scene's two or three dominant pairwise-orthogonal orientations by integrating information across multiple views rather than from a single image alone. What makes a multiple-view approach attractive is that in addition to increasing robustness to segments that do not correspond to any of the three dominant orientations, robustness is also increased with respect to inaccuracies in the extracted segments themselves.

1. Introduction

In this paper, we explore the extraction of vanishing points in the aim of facilitating the reconstruction of Manhattan-world scenes (cf. Coughlan and Yuille [7]), in a manner most closely akin to that of Sinha *et al*. [21]. Owing to the geometry of image formation, a set of lines parallel in space project to lines in the image plane that converge in a corresponding vanishing point (cf. Figure 2). Under known camera geometry, that vanishing point back-projects to a ray through the camera center itself parallel with the projecting lines. Accordingly, if we are able to compute the vanishing points corresponding to the scene's dominant three pairwise-orthogonal line orientations, we have in our possession normal vectors corresponding to each of the scene's dominant three pairwise-orthogonal plane orientations.

In departure from most vanishing point extraction methods, ours extracts a constellation of vanishing points across *multiple views* rather than in a single image alone. Doing so makes the method more robust both to segments that do not correspond to any of the three dominant orientations and to inaccuracies in the extracted segments themselves. By making use of a strong orthogonality criterion per view, optimal segment intersection estimation and a novel tripod fitting technique, our approach thus allows for the extraction of results that correspond closely to the dominant three pairwise-orthogonal orientations of a Manhattan-world scene.

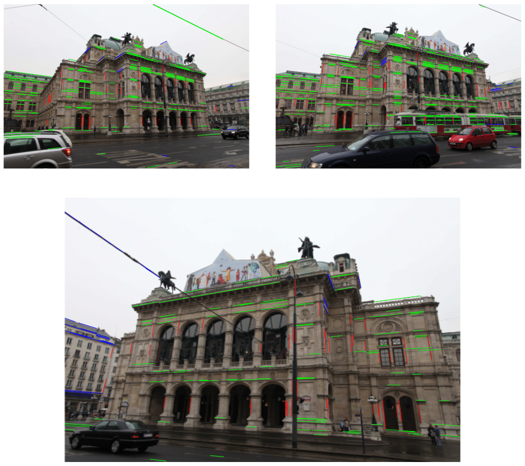


Figure 1. A Manhattan-world scene with extracted line segments respectively corresponding to its dominant three pairwise-orthogonal orientations, identified using our approach.

2. Related Work

The literature on the extraction of vanishing points dates back to the late 1970's and straddles the fields of photogrammetry, computer vision and robotics. Knowledge of vanishing points has been put to use in scene reconstruction, partial camera calibration and the navigation of robots and autonomous vehicles. Since our focus is on scene reconstruction, however, we direct our attention to extraction approaches accordingly.

2.1. Extraction Approaches

Extraction techniques tend to involve what amount to an *accumulation* (or *grouping*) step followed by an *estimation* (or *search*) step, perhaps repeated for some number of iterations. In the accumulation step, line segments are grouped according to the condition that they come close enough to sharing a common point of intersection, which is interpreted as a candidate vanishing point. In the estimation step, one or more optima are chosen from among the results of the accumulation step. Finally, a subsequent re-estimation of the corresponding candidate vanishing points is often performed vis-à-vis their respective inlier segments.

Tessellating the Gaussian Sphere. The Euclidean unit sphere S^2 centered on the camera center $\mathbf{C} \in \mathbb{R}^3$ is (locally) topologically equivalent to the corresponding camera's image plane π. One extraction strategy in the literature involves tessellating this *Gaussian* sphere and tallying the number of great circles that pass through each accumulation cell, with maxima assumed to represent the vanishing points corresponding to dominant scene orientations.

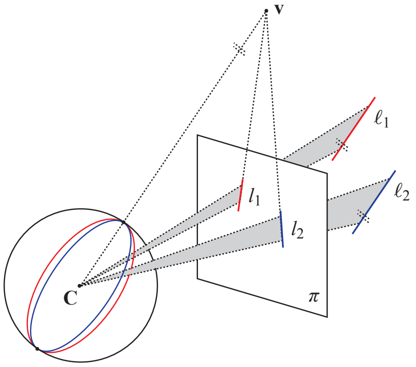


Figure 2. The projections $l_1, l_2 \subset \pi$ of two lines ℓ_1, ℓ_2 parallel in space converge in a corresponding vanishing point $\mathbf{v}$ in the image plane π. Note that the lines ℓ_1, ℓ_2 in space have the selfsame orientation as the ray extending from the camera center $\mathbf{C}$ through $\mathbf{v}$. We call that ray the back-projection of $\mathbf{v}$ with respect to the given camera. We call the plane through $\mathbf{C}$ and a line ℓ the line's interpretation plane. The unit sphere centered on the camera center $\mathbf{C}$ is called the Gaussian sphere.

Barnard [3] was the first to avail himself of the Gaussian sphere as an accumulation space for extracting vanishing points. Quan and Mohr [17] improve upon Barnard's approach by carrying out a hierarchical sampling and by making use of a better tessellation. Lutton *et al.* [14] first extract candidate vanishing points using a related sampling approach and subsequently use a second sampling approach to choose three vanishing points assumed to correspond closely to the scene's dominant three pairwise-orthogonal scene orientations. Shufelt [20] observes that spurious maxima on the Gaussian sphere can arise both on account of weak perspective effects, and on account of textural effects leading to segments that do not correspond to dominant scene orientations. Accordingly, he introduces one Gaussian sphere technique that incorporates *a priori* knowledge about the geometry of objects of interest, and another that incorporates edge precision estimates in the aim of compensating for the influence of textural effects.

The Intersection Constraint. For three lines in the image plane to be the projection of lines parallel in space, the normals of their interpretation planes must (ideally) be coplanar. This fact motivates van den Heuvel's [22] introduction of an *intersection constraint* for triplets of image segments. Given n image segments, van den Heuvel clusters over the subset of the total $\binom{n}{3}$ possible triplets of interpretation plane normals that satisfy his intersection constraint, with clusters themselves constrained such that each triplet of interpretation plane normals they respectively contain satisfy the intersection constraint. Roughly speaking, the largest cluster is then chosen to correspond to the first vanishing point; another two are subsequently extracted, constrained to be collectively close to pairwise-orthogonal with the orientation estimated from the first vanishing point.

The Image Plane as Accumulation Space. Magee and Aggarwal [15] compute the intersections of all $\binom{n}{2}$ pairs of lines through image segments and cluster them on the unit sphere. Rother [18] presents an approach that likewise operates over the set of all such intersections, but instead uses a voting scheme coupled with single-view constraints on camera geometry (cf. Liebowitz and Zisserman [12]). Part of Rother's contribution is a distance function $d(\mathbf{v}, s)$ for determining the extent to which an image line segment s corresponds to a given (candidate) vanishing point $\mathbf{v}$.

Expectation Maximization. Košecká and Zhang [10] cast the problem of extracting the vanishing points corresponding to the scene's dominant three pairwise-orthogonal orientations in terms of an expectation maximization (EM) framework. Pflugfelder [16] introduces his own EM framework, and integrates segment information over a video stream for a static camera. Advantages of making use of a video stream include greater robustness to single-frame sensor noise and the ability to incorporate additional dynamic information that may appear in the scene, due for instance to human activity or changes in lighting conditions.

Extraction across Multiple Views. Werner and Zisserman [23] present a multiple-view approach for extracting the dominant three pairwise-orthogonal orientations across k uncalibrated views of the scene. They begin by computing vanishing points per view assumed to correspond closely to the scene's dominant three pairwise-orthogonal orientations, which they proceed to match combinatorially across the k views. They estimate the corresponding orientations by minimizing the reprojection error with respect to each corresponding vanishing point's inlier segments.

Antone and Teller [1] combine a Hough approach with an EM framework and knowledge of intrinsic camera parameters to extract candidate vanishing points across multiple views and then match them across cameras. Candidate scene orientations are found by fitting a plane to line segment interpretation plane normals. Matching of vanishing points corresponding to common orientations is then used to carry out a refinement of relative camera rotations.

Most closely akin to ours is the method of Sinha *et al.* [21]. They begin by extracting up to n candidate vanishing points per view using a RANSAC-based approach (cf. Fischler and Bolles [8]) with support defined in terms of inlier count with respect to the distance measure $d(\mathbf{v}, s)$ of Rother; a segment s is an inlier of a candidate vanishing point $\mathbf{v}$ if $d(\mathbf{v}, s) < T_{\text{Roth}}$ for some threshold T_{Roth}. Once up to n candidate vanishing points have been extracted in each of the k available views, Sinha *et al.* back-project each candidate vanishing point to its corresponding normalized direction vector, which they place on a unit sphere. Next, they cluster the points on that unit sphere, extracting the cluster center best aligned with the up vector for most of the cameras. From among the remaining clusters, they obtain another two, collectively constrained to correspond closely to pairwise-orthogonal orientations.

2.2. Re-estimation

Given a set $\mathcal{S}_{\mathbf{v}}$ of image segments determined to be inliers of a candidate vanishing point $\mathbf{v} \in \mathbb{P}^2$, Caprile and Torre [4] re-estimate $\mathbf{v}$ by computing a weighted mean of the intersections of the lines $\mathbf{l} \in \mathbb{P}^2$ corresponding to the segments $s \in \mathcal{S}_{\mathbf{v}}$. A more accurate approach involves fitting a point $\mathbf{v} \in \mathbb{P}^2$ to the set of lines $\mathbf{l} \in \mathbb{P}^2$ corresponding to the segments in $\mathcal{S}_{\mathbf{v}}$ by minimizing with respect to point-line incidence (cf. Collins and Weiss [6], Cipolla and Boyer [5]), which can be solved using the SVD. It is this re-estimation approach that we use in our implementation (cf. Section 3). An approach that produces potentially even better intersection estimations is the maximum likelihood intersection estimation technique of Liebowitz [11], which can be solved using Levenberg-Marquardt (cf. Lourakis [13]).

3. Implementation

We begin with the recovery of camera geometry for each of the k available views (cf. Irschara *et al.* [9]). Individually in each of those views, we then extract a set of image line segments $\mathcal{S}$ and compute a constellation $\mathcal{C}$ of two or three candidate vanishing points, collectively constrained to satisfy an orthogonality criterion and individually refined by computing an optimal point of intersection vis-à-vis candidate vanishing point inlier segments. We then map the orientations corresponding to those candidate vanishing points to antipodal points on the unit sphere, given by corresponding unit direction vectors. We proceed to extract three pairwise-orthogonal orientations—which we expect to already correspond closely with the dominant three pairwise-orthogonal orientations of the underlying urban scene—by fitting a tripod centered at the sphere's origin to those said points (cf. Figure 3).

3.1. Extracting a Constellation from a Single View

Individually in each of the k available views, we attempt to extract a pair or triplet of vanishing points that already come close to corresponding to the dominant three pairwise-orthogonal orientations of the underlying scene. We do this in each given view by iterating over a number N_1 of constellations of candidate vanishing points chosen at random, and choosing the constellation with best support that satisfies an orthogonality criterion. We then compute an optimal re-estimation of each candidate vanishing point $\mathbf{v} \in \mathcal{C}$ of the winning constellation $\mathcal{C}$ with respect to the inlier segments $\mathcal{S}_{\mathbf{v}} \subseteq \mathcal{S}$ of $\mathbf{v}$.

Candidate Vanishing Points. Given a calibrated view of the scene and a set $\mathcal{S}$ of line segments s that we have extracted from that view, we compute candidate vanishing points from the intersections of the image lines $l \subset \mathbb{R}^2$ corresponding to the segments $s \in \mathcal{S}$. We obtain the homogeneous representation $\mathbf{l} \in \mathbb{P}^2$ of a line l in the image plane corresponding to an extracted image segment s from the homogeneous endpoints $\mathbf{p}_1, \mathbf{p}_2 \in \mathbb{P}^2$ of s,

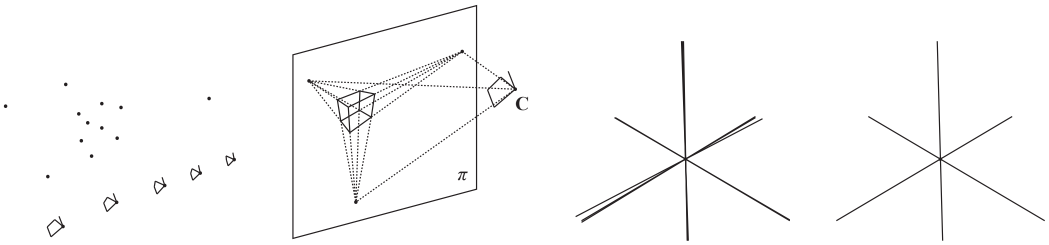
$$\mathbf{l} \sim \mathbf{p}_1 \times \mathbf{p}_2. \tag{1}$$

Given the homogenous vectors $\mathbf{l}, \mathbf{l}' \in \mathbb{P}^2$ that represent the two lines $l, l' \subset \mathbb{R}^2$, we compute the intersection of l, l' once again using the vector product,

$$\mathbf{v} \sim \mathbf{l} \times \mathbf{l}', \tag{2}$$

yielding the candidate vanishing point $\mathbf{v} \in \mathbb{P}^2$ corresponding to the segments s, s'.

Accumulation. It is the distance measure $\mathcal{F}^{(i)}_{\text{min}} = \mathcal{F}^{(i)}(\hat{\mathbf{l}}_i) = d^2_{\perp}(\mathbf{l}, \mathbf{x}^a_i) + d^2_{\perp}(\mathbf{l}, \mathbf{x}^b_i)$ of Liebowitz [11] (cf. Figure 4) that we use in our grouping of segments $s_i \in \mathcal{S}$ with respect to a candidate vanishing point $\mathbf{v} \in \mathbb{P}^2$; accordingly, given a candidate vanishing point $\mathbf{v}$, we consider each segment s_i for which $\mathcal{F}^{(i)}_{\text{min}} < T_{\text{Lieb}}$ to be an inlier of $\mathbf{v}$, and the set $\mathcal{S}_{\mathbf{v}} \subseteq \mathcal{S}$ to be the set of all such inliers. Pflugfelder [16] at length describes the merits of calling on this distance measure over other distance measures available in the literature, such as the distance function $d(\mathbf{v}, s)$ of Rother [18] used in Sinha *et al.*



(a) Recovery of camera geometry for the k available views.

(b) Extraction of a constellation of two or three candidate vanishing points in a single view, constrained to correspond closely to pairwise-orthogonal scene orientations.

(c) Antipodal unit direction vectors corresponding to the orientations computed from candidate vanishing points extracted across the k available views.

(d) The pairwise-orthogonal orientations corresponding to the best-fit tripod.

Figure 3. The processing pipeline of our approach.

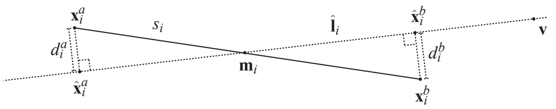


Figure 4. The line $\hat{\mathbf{l}}_i = \arg\min_{\mathbf{l}} \mathcal{F}^{(i)}(\mathbf{l})$ is the line through $\mathbf{v}$ that minimizes the quantity $d_\perp^2(\mathbf{l}, \mathbf{x}_i^a) + d_\perp^2(\mathbf{l}, \mathbf{x}_i^b) = d_i^a \cdot d_i^a + d_i^b \cdot d_i^b$ with respect to the segment s_i. Note that $\mathbf{m}_i$ is not necessarily the midpoint of s_i.

Optimal Intersection Estimation. Given a set of n lines $\mathbf{l}_i \in \mathbb{P}^2$, the least-squares point of intersection with respect to point-line incidence is given by the vector $\hat{\mathbf{v}}_{\text{SVD}} \in \mathbb{P}^2$ that minimizes the quantity

$$\left\| \begin{bmatrix} \mathbf{l}_1 & \cdots & \mathbf{l}_n \end{bmatrix}^\top \hat{\mathbf{v}}_{\text{SVD}} \right\|^2, \tag{3}$$

where each vector $\mathbf{l}_i$ is scaled to unit length (cf. Cipolla and Boyer [5]). This minimizing vector $\hat{\mathbf{v}}_{\text{SVD}}$ is the vector corresponding to the smallest singular value of the singular value decomposition of the $n \times 3$ matrix $\begin{bmatrix} \mathbf{l}_1 & \cdots & \mathbf{l}_n \end{bmatrix}^\top$.

Orthogonality Criterion. For a pair of candidate vanishing points $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{P}^2$, our orthogonality criterion requires that the unit direction vectors $\mathbf{d}_1, \mathbf{d}_2 \in \mathbb{R}^3$ corresponding to their back-projections be within a tight threshold of orthogonality (cf. Figure 3(b)); i.e., $|\mathbf{d}_1^\top \mathbf{d}_2| < T_{\text{ortho}}$. For a triplet $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in \mathbb{P}^2$, we check each pair $\mathbf{d}_i, \mathbf{d}_j, i \neq j$, of corresponding back-projections for orthogonality in the same manner.

A Constellation's Vote. Given a constellation $\mathcal{C}$ of two or three candidate vanishing points, its vote is given by

$$\text{vote}(\mathcal{C}) = \sum_{\mathbf{v} \in \mathcal{C}} \sum_{s_i \in \mathcal{S}_\mathbf{v}} 1 - \frac{\mathcal{F}_{\min}^{(i)}}{T_{\text{Lieb}}}, \tag{4}$$

where $\mathcal{F}_{\min}^{(i)}$ is, once again, the error of the optimal line $\hat{\mathbf{l}}_i$ through the candidate vanishing point $\mathbf{v}$ with respect to the segment s_i; the set $\mathcal{S}_\mathbf{v}$ contains all inlier segments s_i of $\mathbf{v}$, such that as before, each $\mathcal{F}_{\min}^{(i)}$ constrained to be smaller than the threshold T_{Lieb}.

Pseudocode. For each of the k views of the scene, we extract a constellation $\mathcal{C}$ of two or three vanishing points corresponding to the (ideally) dominant three pairwise-orthogonal orientations of the scene. We provide the pseudocode in Algorithm 1.

Algorithm 1 Extracting a Constellation of Vanishing Points from a Single View

```
for $N_1$ iterations do
    take 6 distinct image line segments at random from $\mathcal{S}$ and compute the candidate vanishing points $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$
    for all 4 constellations $\mathcal{C} \in \{\{\mathbf{v}_1, \mathbf{v}_2\}, \{\mathbf{v}_1, \mathbf{v}_3\}, \{\mathbf{v}_2, \mathbf{v}_3\}, \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}\}$ do
        $\text{vote}_\mathcal{C} \leftarrow \text{vote}(\mathcal{C})$
        if $|\mathcal{C}| = 3$ yet the constellation with the greatest vote thus far encountered contains only a pair of candidate vanishing points, and the constellation $\mathcal{C}$ satisfies the orthogonality criterion then
            store $\mathcal{C}$ as the constellation with best support
        else if $\text{vote}_\mathcal{C}$ is the greatest constellation vote thus far encountered and the constellation $\mathcal{C}$ satisfies the orthogonality criterion then
            store $\mathcal{C}$ as the constellation with best support
        end if
    end for
end for
return the re-estimation $\hat{\mathbf{v}}_{\text{SVD}}$ of each candidate vanishing point $\mathbf{v}$ in the constellation with best support
```

3.2. Optimizing across k Views

A vanishing point back-projects to a ray through the view's camera center $\mathbf{C}$ whose *direction* can be given as either of an antipodal pair of unit vectors. Let the set $\mathcal{T}$—which we call a *tripod*—contain three orthonormal vectors $\mathbf{t} \in \mathbb{R}^3$. Let the set $\mathcal{K}$ contain the k constellations $\mathcal{C}$ of two or three candidate vanishing points extracted across k available views. Let $\mathcal{X}$ be the set of antipodal pairs of unit vectors corresponding to the back-projection of each vanishing point from the union of the k constellations $\mathcal{C} \in \mathcal{K}$. We proceed by fitting a tripod $\mathcal{T}$ to the antipodal unit direction vectors in $\mathcal{X}$ by iteratively rotating the tripod $\mathcal{T}$ with respect to the vectors in $\mathcal{X}$ close to the tripod's axes.

An Iteration of Tripod Fitting. Given, without loss of generality, a vector $\mathbf{t}_1 \in \mathcal{T} = \{\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3\}$ and the set $\mathcal{X}_1 \subset \mathcal{X}$ of the unit vectors in $\mathcal{X}$ within an angle T_{axis} of $\mathbf{t}_1$, the mean unit vector $\boldsymbol{\mu}_1$ of the vectors in $\mathcal{X}_1$ is given by the normalized sum of all $\mathbf{x} \in \mathcal{X}_1$,

$$\boldsymbol{\mu}_1 = \sum_{\mathbf{x}\in\mathcal{X}_1} \mathbf{x} \Big/ \left\| \sum_{\mathbf{x}\in\mathcal{X}_1} \mathbf{x} \right\|. \tag{5}$$

Let the matrix $\mathtt{R}_1$ be the matrix that rotates the vector $\mathbf{t}_1$ into the vector $\boldsymbol{\mu}_1$. We treat the denominator of the right-hand side of (5) as a measure of confidence $\omega_1 = \|\sum_{\mathbf{x}\in\mathcal{X}_1} \mathbf{x}\|$ in the rotation given by $\mathtt{R}_1$, the magnitude of which depends on the cardinality of $\mathcal{X}_1$ and on the extent to which the vectors $\mathbf{x} \in \mathcal{X}_1$ are clustered together. Having also computed the rotation matrices $\mathtt{R}_2, \mathtt{R}_3$ and weights ω_2, ω_3 corresponding, respectively, to the axes $\mathbf{t}_2, \mathbf{t}_3 \in \mathcal{T}$, an axis $\mathbf{t} \in \mathcal{T}$ rotates to $\mathbf{t}'$ by our tripod fitting technique according to

$$\begin{aligned} \mathbf{t}' &= \frac{\omega_1\mathtt{R}_1\mathbf{t} + \omega_2\mathtt{R}_2\mathbf{t} + \omega_3\mathtt{R}_3\mathbf{t}}{\|\omega_1\mathtt{R}_1\mathbf{t} + \omega_2\mathtt{R}_2\mathbf{t} + \omega_3\mathtt{R}_3\mathbf{t}\|} \\ &= \frac{(\omega_1\mathtt{R}_1 + \omega_2\mathtt{R}_2 + \omega_3\mathtt{R}_3)\mathbf{t}}{\|(\omega_1\mathtt{R}_1 + \omega_2\mathtt{R}_2 + \omega_3\mathtt{R}_3)\mathbf{t}\|} \\ &= \frac{\mathtt{A}\mathbf{t}}{\|\mathtt{A}\mathbf{t}\|} = \mathtt{R}\mathbf{t}. \end{aligned} \tag{6}$$

In order to express the transformation in (6) in closed form, we seek the orthogonal matrix $\mathtt{R}$ for which $\mathtt{R}\mathbf{t}$ gives $\mathbf{t}'$. By the SVD, we can decompose the matrix $\mathtt{A}$ such that $\mathtt{A} = \mathtt{U}\Sigma\mathtt{V}^\top$, where $\mathtt{U}, \mathtt{V}^\top$ are orthogonal matrices and Σ is a diagonal matrix; the orthogonal matrix closest in a least-squares sense to the matrix $\mathtt{A}$ is $\hat{\mathtt{R}} = \mathtt{U}\mathtt{V}^\top$ (cf. Schönemann [19]). For a single iteration of our tripod fitting algorithm, the tripod $\mathcal{T}$ thus rotates to $\mathcal{T}'$ according to

$$\mathcal{T}' = \bigcup_{\mathbf{t}\in\mathcal{T}} \{\hat{\mathtt{R}}\mathbf{t}\} \tag{7}$$

Initialization. We run our fitting algorithm k times, once for a tripod corresponding to the back-projections of the candidate vanishing points in each of the k available constellations $\mathcal{C} \in \mathcal{K}$. If a constellation $\mathcal{C}$ contains only a pair of candidate vanishing points, we compute the third axis of the corresponding tripod $\mathcal{T}$ from the vector product of its first two. Since we demand that our final tripod have pairwise-orthogonal axes, we orthogonalize every tripod $\mathcal{T}$ that we use to initialize our tripod fitting algorithm. This reduces to orthogonalizing the matrix $\mathtt{T} = \begin{bmatrix}\mathbf{t}_1 & \mathbf{t}_2 & \mathbf{t}_3\end{bmatrix}$ in the same manner as presented above; i.e., $\mathtt{T} = \mathtt{U}\Sigma\mathtt{V}^\top$, and so $\hat{\mathtt{T}} = \mathtt{U}\mathtt{V}^\top = \begin{bmatrix}\hat{\mathbf{t}}_1 & \hat{\mathbf{t}}_2 & \hat{\mathbf{t}}_3\end{bmatrix}$.

Support. From among k runs of our tripod fitting algorithm, we choose our best-fit tripod from among the k outcomes based on cosine similarity (cf. Banerjee *et al.* [2]). For each of the k outcome tripods $\mathcal{T}$, we compute

$$\gamma_{\mathcal{T}} = \sum_{\mathbf{t}\in\mathcal{T}} \sum_{\mathbf{x}\in\mathcal{X}_\mathbf{t}} \cos(\mathbf{x}^\top\mathbf{t}), \tag{8}$$

which expresses the aggregate cosine similarity between each tripod axis $\mathbf{t} \in \mathcal{T}$ and every vector $\mathbf{x} \in \mathcal{X}_\mathbf{t}$, and is thus a measure of the tripod's support. We accordingly choose the tripod with best support as our best-fit tripod.

Pseudocode. We obtain a best-fit tripod with respect to $\mathcal{X}$ as the final result with best overall support from among k runs of an iterative fitting procedure, with each run distinctly initialized with a tripod corresponding to one of the k available constellations $\mathcal{C} \in \mathcal{K}$. The result with best support is the tripod $\mathcal{T}$ that, within N_2 iterations of initialization, yields the highest weight $\gamma_{\mathcal{T}}$. We present the pseudocode in Algorithm 2.

4. Evaluation

We examine our algorithm's performance by considering three Manhattan-world data sets: `opera`, `museum` and `ares`. We first demonstrate the outcome of a run of our algorithm on each of the three data sets by identifying the respective inlier segments of the vanishing points corresponding to the projection per view of the extracted pairwise-orthogonal scene orientations (cf. Figures 5, 6 and 7). On account of lack of space, not all images are shown here; the complete data sets have been made available on our website.[1] We then provide a depiction of the antipodal directions extracted across all views of each data set, and with them the corresponding best-fit tripods (cf. Figure 8). We compare these with the antipodal directions (note that they are not antipodal in their paper) extracted via the approach of Sinha *et al.* [21], numbering—as in their paper—eight

[1] www.vrvis.at/publications/PB-VRVis-2011-011/

Algorithm 2 Fitting a Tripod with Pairwise-Orthogonal Axes to the Antipodal Directions Extracted across k Views

```
$\mathcal{K} \leftarrow$ the set of $k$ constellations $\mathcal{C}$ obtained across $k$ views using Algorithm 1
$\mathcal{X} \leftarrow$ the set of antipodal unit vectors corresponding to the back-projection of each candidate vanishing point contained across all $k$ constellations in $\mathcal{K}$
for all $k$ constellations $\mathcal{C} \in \mathcal{K}$ do
    $\mathcal{T} \leftarrow$ the set of vectors corresponding to the back-projections of the pair or triplet of candidate vanishing points in the constellation $\mathcal{C}$
    if the set $\mathcal{T}$ contains only a pair of vectors then
        $\mathcal{T} \leftarrow \mathcal{T} \cup \{\mathbf{t}_1 \times \mathbf{t}_2\}$, where $\mathbf{t}_1, \mathbf{t}_2 \in \mathcal{T}$
    end if
    $\mathcal{T} \leftarrow \mathrm{orthogonalize}(\mathcal{T})$ {the tripod initialization}
    for $N_2$ iterations do
        for all 3 pairwise-orthogonal axes $\mathbf{t} \in \mathcal{T}$ do
            $\mathcal{X}_\mathbf{t} \leftarrow$ all $\mathbf{x} \in \mathcal{X}$ with $\cos^{-1}(\mathbf{x}^\top \mathbf{t}) < T_{\mathrm{axis}}$
            $\omega_\mathbf{t} \leftarrow \| \sum_{\mathbf{x} \in \mathcal{X}_\mathbf{t}} \mathbf{x} \|$
            $\boldsymbol{\mu}_\mathbf{t} \leftarrow \sum_{\mathbf{x} \in \mathcal{X}_\mathbf{t}} \mathbf{x} / \omega_\mathbf{t}$
            $\mathtt{R}_\mathbf{t} \leftarrow$ the matrix that rotates $\mathbf{t}$ into $\boldsymbol{\mu}_\mathbf{t}$
        end for
        $\mathtt{A} \leftarrow \sum_{\mathbf{t} \in \mathcal{T}} \omega_\mathbf{t} \mathtt{R}_\mathbf{t}$
        $\hat{\mathtt{R}} \leftarrow \mathrm{orthogonalize}(\mathtt{A})$
        $\mathcal{T} \leftarrow \bigcup_{\mathbf{t} \in \mathcal{T}} \{\hat{\mathtt{R}}\mathbf{t}\}$
    end for
    $\gamma_\mathcal{T} \leftarrow \sum_{\mathbf{t} \in \mathcal{T}} \sum_{\mathbf{x} \in \mathcal{X}_\mathbf{t}} \cos(\mathbf{x}^\top \mathbf{t})$
end for
return the tripod $\mathcal{T}$ with best support $\gamma_\mathcal{T}$
```

per view; to these, we likewise fit a tripod in our manner, since their clustering approach anyway does not guarantee orthogonality in its output tripod. Finally, we compare inlier proportions for the three data sets across three runs, respectively, for both our approach and for our tripod fitting applied to the directions—which we made antipodal—obtained via the approach of Sinha *et al.* (cf. Figure 9).

Remarks on Complexity and Parameters. Given n line segments extracted in a single view, there exist a total of $\binom{n}{2} \in O(n^2)$ candidate vanishing points from among which to choose. The complexity of obtaining the genuinely best-support triplet of vanishing points from an image thus calls for iterating through $\binom{\binom{n}{2}}{3} \in O(n^6)$ unique triplets. It is in order to overcome this crippling complexity that we opt instead to obtain our best-support result in Algorithm 1 from among a (typically) much smaller number N_1 of constellations chosen at random. In this respect, we note that we set the number N_1 to 1000 across all runs and for each of the three data sets in the evaluation of our algorithm. The parameters N_2 and T_{axis} in Algorithm 2 were kept the same at 25 and 15° (in radians), respectively, between fitting to the antipodal directions obtained via our method as well as to those obtained via that of Sinha *et al.* T_{Lieb} in Algorithm 1 was set to 15. Running time was within two minutes for each respective data set and run on a 2.4 GHz quad-CPU Windows box with 8 GB of RAM.

5. Conclusion

The problem of extracting vanishing points has remained an active field of research since the 1970's, owing primarily to problems of accuracy and robustness. Our method—tailored to urban reconstruction and most closely akin to that of Sinha *et al.* [21]—contributes to the literature by making use of a combination of a strong orthogonality criterion per view, optimal segment intersection estimation and a novel tripod fitting technique. Divergence in corresponding antipodal directions obtained per view via our approach (cf. Figure 8(a,c,e)) motivates the suggestion that small errors in extracted line segments can have a material effect on the accuracy of vanishing point estimates; fitting a tripod in our manner to those antipodal directions thus offers a way to integrate information obtained across multiple views to yield a close approximation of the three dominant pairwise-orthogonal orientations of a Manhattan-world scene. In our experiments, our algorithm gave results that correctly classified most image segments to respective corresponding vanishing points and was stable across runs, without parameter tuning between data sets (cf. Figure 9, top row). Performance of our tripod fitting approach on the directions obtained via the method of Sinha *et al.* (cf. Figure 9, bottom row) suggests that our algorithm could be robust even to more challenging data sets.

6. Acknowledgements

Christian Luksch and Janne Mihola contributed ideas that led to the tripod fitting approach presented in this paper. Irene Reisner-Kollmann provided invaluable help with recovery of camera parameters for the data sets. Thanks also goes to the entire Aardvark team at VRVis for building the framework within which this algorithm was implemented.

References

[1] M. E. Antone and S. Teller. Automatic Recovery of Relative Camera Rotations for Urban Scenes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2:282–289, 2000. 955

[2] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *Journal of Machine Learning*, 6:1345–1382, 2006. 957

[3] S. T. Barnard. Interpreting Perspective Images. *Artificial Intelligence*, 21(4):435–462, 1983. 954

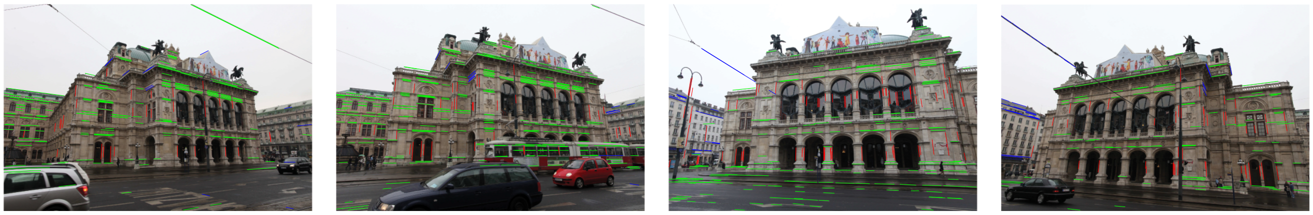

(a) (c) (e) (f)

Figure 5. The `opera` data set with its dominant three pairwise-orthogonal scene orientations extracted using our approach, with the inlier segments of their corresponding vanishing points shown in red, green and blue, respectively. Note the abundance of image segments corresponding to none of the dominant three pairwise-orthogonal scene orientations.

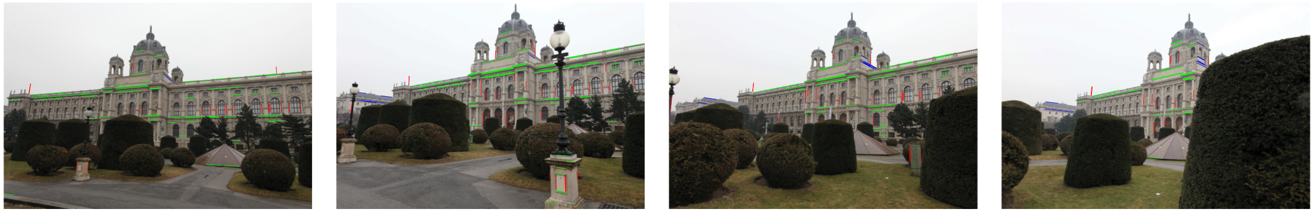

(a) (c) (d) (e)

Figure 6. The `museum` data set with its dominant three pairwise-orthogonal scene orientations extracted using our approach, with the inlier segments of their corresponding vanishing points shown in red, green and blue, respectively. Note the cupola and the prism in the foreground, which contribute image segments corresponding to none of the dominant three pairwise-orthogonal scene orientations.

[4] B. Caprile and V. Torre. Using Vanishing Points for Camera Calibration. *International Journal of Computer Vision*, 4(2):127–139, 1990. 955

[5] R. Cipolla and E. Boyer. 3D Model Acquisition from Uncalibrated Images. *IAPR Workshop on Machine Vision Applications*, pages 559–568, 1998. 955, 956

[6] R. T. Collins and R. S. Weiss. Vanishing Point Calculation as a Statistical Inference on the Unit Sphere. *Proceeding of the International Conference on Computer Vision*, 1990. 955

[7] J. M. Coughlan and A. L. Yuille. Manhattan World: Compass Direction from a Single Image by Bayesian Inference. *Proceedings of the International Conference on Computer Vision*, pages 941–947, 1999. 953

[8] 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. 955

[9] A. Irschara, C. Zach, and H. Bischof. Towards Wiki-based Dense City Modeling. *Proceedings of the International Conference on Computer Vision*, 2007. 955

[10] J. Košecká and W. Zhang. Efficient Computation of Vanishing Points. *Proceedings of the International Conference on Robotics and Automation*, pages 223–228, 2002. 954

[11] D. Liebowitz. *Camera Calibration and Reconstruction of Geometry from Images*. PhD thesis, 2001. 955

[12] D. Liebowitz and A. Zisserman. Combining Scene and Autocalibration Constraints. *Proceedings of the International Conference on Computer Vision*, 1:293–300, 1999. 954

[13] M. I. A. Lourakis. levmar: Levenberg-Marquardt Non-Linear Least Squares Algorithms in C/C++. 955

[14] E. Lutton, H. Maître, and J. Lopez-Krahe. Contribution to the Determination of Vanishing Points using Hough Transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 430–438, 1994. 954

[15] M. J. Magee and J. K. Aggarwal. Determining Vanishing Points from Perspective Images. *Computer Vision, Graphics, and Image Processing*, 26(2):256–267, 1984. 954

[16] R. Pflugfelder. *Self-Calibrating Cameras in Video Surveillance*. PhD thesis, 2008. 954, 955

[17] L. Quan and R. Mohr. Determining Perspective Structures using Hierarchical Hough Transform. *Pattern Recognition Letters*, 9(4):279–286, 1989. 954

[18] C. Rother. A New Approach to Vanishing Point Detection in Architectural Environments. *Image and Vision Computing*, 20:647–655, 2002. 954, 955

[19] P. H. Schönemann. A Generalized Solution of the Orthogonal Procrustes Problem. *Psychometrika*, 31(1):1–10, 1966. 957

[20] J. A. Shufelt. Performance Evaluation and Analysis of Vanishing Point Detection Techniques. *IEEE Transactions on Pattern Analysis and Machine*, 1999. 954

[21] S. N. Sinha, D. Steedly, R. Szeliski, M. Agrawala, and M. Pollefeys. Interactive 3D Architectural Modeling from Unordered Photo Collections. *ACM Transactions on Graphics*, 27(5):1, 2008. 953, 955, 957, 958

[22] F. A. van den Heuvel. Vanishing Point Detection for Architectural Photogrammetry. *International Archives of Photogrammetry and Remote Sensing*, 32:652–659, 1998. 954

[23] T. Werner and A. Zisserman. New Techniques for Automated Architectural Reconstruction from Photographs. *European Conference on Computer Vision*, pages 541–555, 2002. 954

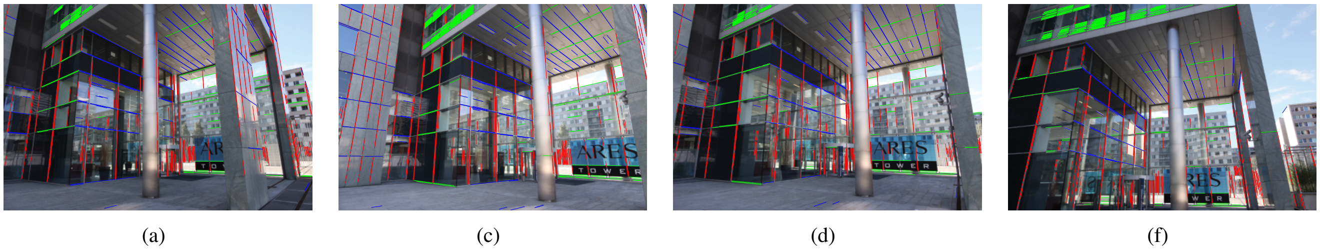
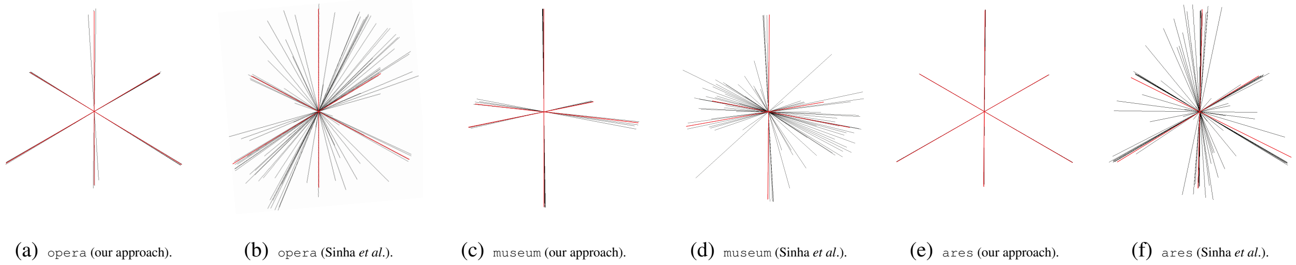


Figure 7. The ares data set with its dominant three pairwise-orthogonal scene orientations extracted using our approach, with the inlier segments of their corresponding vanishing points shown in red, green and blue, respectively. It is provided as a ‘toy’ data set, and serves to illustrate the relative quality of antipodal directions obtained using our method and that of Sinha *et al*. (cf. Figure 8(e,f)).



(a) opera (our approach). (b) opera (Sinha *et al*.). (c) museum (our approach). (d) museum (Sinha *et al*.). (e) ares (our approach). (f) ares (Sinha *et al*.).

Figure 8. Antipodal unit direction vectors extracted across all views of the given data set, with the corresponding best-fit tripod shown red. The tripods obtained using our approach correspond to Figures 5, 6 and 7, respectively.

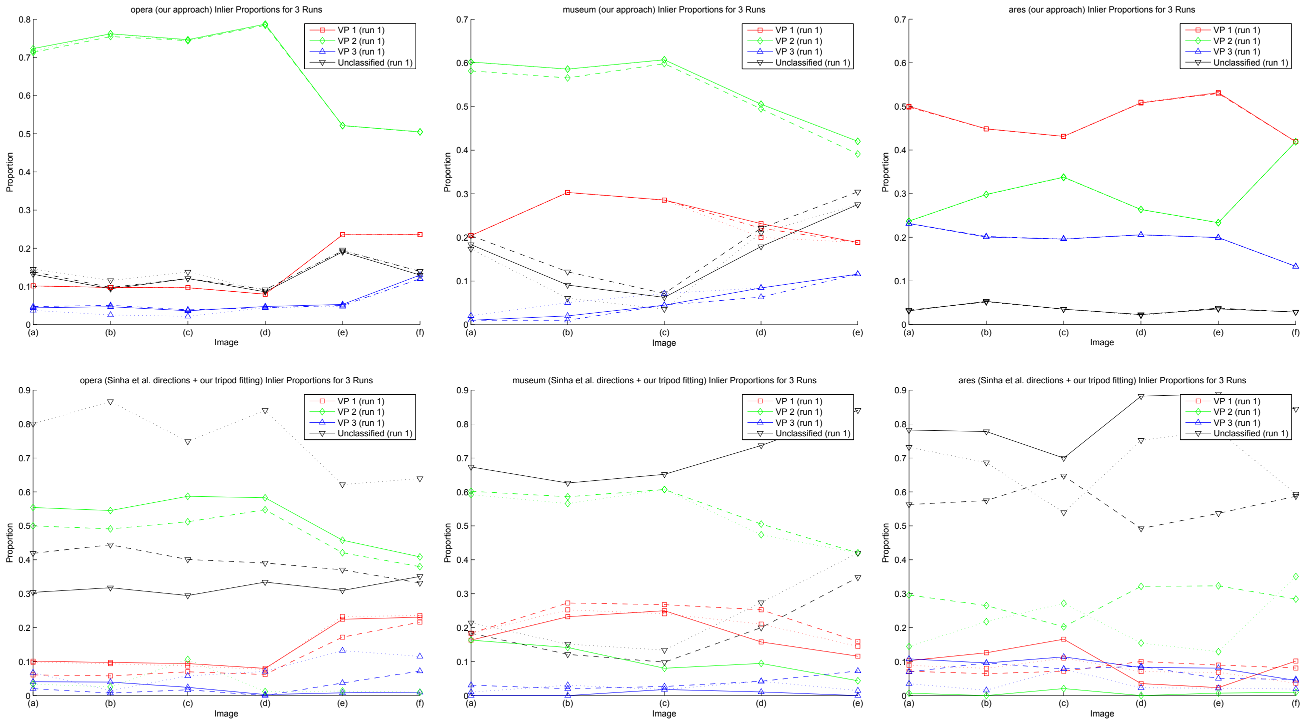


Figure 9. Inlier proportions (number of inlier extracted image segments with respect to a vanishing point divided by total number of extracted image segments) for the three data sets across three runs. The top row corresponds to the results obtained using our approach; the bottom, to our tripod fitting with respect to the antipodal directions obtained via the approach of Sinha *et al*. In both cases, the solid lines refer to the respective run that gave rise to the corresponding tripod in Figure 8; the dotted and dashed lines refer, respectively, to the two additional runs.