A Minimum Error Vanishing Point Detection Approach for Uncalibrated Monocular Images of Man-made Environments

Yiliang Xu Sangmin Oh Anthony Hoogs
Kitware Inc.
28 Corporate Drive, Clifton Park, NY 12065
{yiliang.xu,sangmin.oh,anthony.hoogs}@kitware.com

Abstract

We present a novel vanishing point detection algorithm for uncalibrated monocular images of man-made environments. We advance the state-of-the-art by a new model of measurement error in the line segment extraction and minimizing its impact on the vanishing point estimation. Our contribution is twofold: 1) Beyond existing hand-crafted models, we formally derive a novel consistency measure, which captures the stochastic nature of the correlation between line segments and vanishing points due to the measurement error, and use this new consistency measure to improve the line segment clustering. 2) We propose a novel minimum error vanishing point estimation approach by optimally weighing the contribution of each line segment pair in the cluster towards the vanishing point estimation. Unlike existing works, our algorithm provides an optimal solution that minimizes the uncertainty of the vanishing point in terms of the trace of its covariance, in a closed-form. We test our algorithm and compare it with the state-of-the-art on two public datasets: York Urban Dataset and Eurasian Cities Dataset. The experiments show that our approach outperforms the state-of-the-art.

1. Introduction

In man-made environments, structural objects such as building facades and road lane marks frequently present sets of parallel lines that intersect at points at infinity in the world, whose projections in an image are called *vanishing points* (VPs). For example, street curbs and building floor separation lines are all parallel and converge to horizontal VPs; all vertical lines, usually from buildings, converge to a common VP known as zenith.

Accurate detection of VPs is an important problem in computer vision because it provides a unique characterization of the geometric scene structure. In particular, VPs uniquely determine the orientations of parallel line clusters in the world. Therefore, VP detection has found many important real-world applications such as building facade detection [12], 3D geometric scene structure analysis [5], self-calibration [17], and robot navigation [7], among many others. Accordingly, many VP detection algorithms [1, 6, 9, 13–15, 17] have been developed.

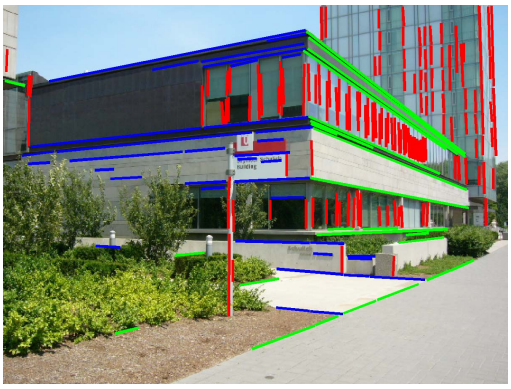


Figure 1. An illustration of vanishing point detection. Line segments with each color correspond to a vanishing point.

The challenge for VP detection arises from the inherent measurement error. Ideally, assuming perfect imaging condition and line segment extraction, parallel lines should intersect at the corresponding VPs. However, in the real-world, there exists pixel noise, image distortion, discretization error, and line segment extraction error, which make the problem much more challenging. The problem becomes even harder when camera parameters or motion cues are unavailable, or the scene becomes complex and does not satisfy the Manhattan world assumption [3] which considers only three mutually parallel line clusters.

In this paper, we present a novel minimum error vanishing point detection algorithm for uncalibrated monocular images of man-made environments, without the Manhattan world assumption. By error, we mean the uncertainty of the VP. We tackle the VP detection problem by modeling the measurement error in the line segment extraction and minimizing its impact on the ultimate error in VP estimation. The main novelty and contribution of our algorithm are twofold as follows.

First, a novel *consistency measure* is developed, which evaluates the consistency between a line segment and a hypothesized VP. Unlike existing hand-crafted models [4, 6, 13, 17], we formally derive a novel consistency measure, which captures the stochastic nature of the correlation between line segments and vanishing points due to the measurement error. This new consistency measure is used to improve the assignment of line segments to corresponding hypothesized VPs.

Second, a novel *minimum error vanishing point estimation method* is presented. Given a cluster of line segments, the extension lines of any line segment pair intersect at a hypothesized VP, establishing a minimal solution. Unlike existing works, our approach estimates the VP of the cluster by reconciling all minimal solutions. The error propagated from line segment endpoints towards the final VP estimation is minimized by optimally weighing all minimal solutions. Our algorithm provides an optimal solution that minimizes the uncertainty of the vanishing point in terms of the trace of its covariance, in a closed-form. The error analysis quantitatively indicates to what degree each line segment pair should be trusted.

In experiments, we compare our approach with the state-of-the-art [6, 13–15, 17] on two public datasets, namely, York Urban Dataset and Eurasian Cities Dataset, which have been widely used as benchmarks for VP detection algorithms. Our experiments showcase the strength of our algorithm where it outperforms the state-of-the-art with a significant margin.

2. Related Work

Over the last decade, there have been many previous works dedicated to VP detection from monocular images.

An early class of the VP detection algorithms, assuming known camera intrinsic parameters, relies on the Hough transform of the line segments on the Gaussian sphere [2] and clusters line segments in a bottom-up manner using their orientation votes. These approaches usually do not handle noise and outliers very well, and suffer from the sub-optimality, which leads to misclassification of line segments [10]. To address these issues, iterative approaches [15] such as EM [6] algorithm are usually used to refine the initial clustering results.

Many recent work assumes the Manhattan world [3], where only three mutually orthogonal vanishing directions are considered: one vertical and two horizontal. With known camera intrinsic parameters, [9] enforces this mutual orthogonality into the global optimization for simultaneous detection of all three VPs. With the Manhattan world assumption, a 4-line RANSAC algorithm [17], has been proposed. However, many man-made scenes do not satisfy the Manhattan world assumption. Also, for many applications such as robot navigation and surveillance, camera intrinsic parameters may change (e.g., zoom), hence, they are hard to be pre-calibrated and maintained. Our work does not require the Manhattan world or camera calibration.

Tardif [13] proposes to use a variant of RANSAC algorithm, called J-linkage, to generate line segment clusters. As described in Sec. 4.1, our algorithm is similar to [13] in the overall structure, but, uses our new consistency measure and minimum error VP estimation, which results in substantial improvement in accuracy.

Tretyak et al. [14] propose to jointly use different layers of geometric primitives and construct empirical energy functions for each layer with respect to hypothesized VPs. VPs are estimated by minimizing the overall energy across layers, which results in a non-convex optimization problem. In contrast, the optimization problem in our approach has a closed-form solution.

Bazin et al., [1] propose an algorithm to globally maximize the number of inlier line segments for all VPs. Our approach is different because line segment pairs are optimally weighted to minimize the VP error. Our reasoning is that the uniform use of more inlier segments [1] does not necessarily guarantee more accurate VP estimation (imagine a lot of short, noisy line segments versus a few long, certain line segments.)

Most aforementioned works require a consistency measure that assesses the degree to which geometric primitives such as edges and line segments are consistent with hypothesized VPs. For brevity, detailed discussions on existing approaches are deferred to Sec. 4.2 where their limitations are analyzed to provide the motivation and design principles for our new probabilistic consistency measure.

3. Problem Description

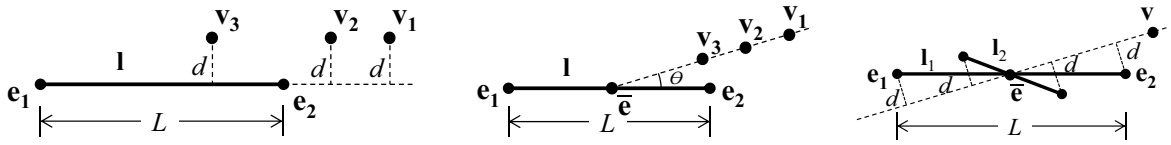
In this section, we formally define the VP detection problem, and enlist the underlying assumptions and notations to be used in the following sections.

The input of the problem is an uncalibrated image of a man-made scene. The output of the problem is a set of m VPs $\mathbb{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m\}$, which are the image projections of the intersections at infinity of parallel lines in the scene. The number of VPs m is unknown and needs to be automatically computed.

To facilitate the problem formulation, we assume:

- Man-made scenes that contain structural objects which present parallel lines.
- Radial distortion is either removed or reasonably small. This assumption is necessary for almost all geometry-based VP detection algorithms.

As a notation convention, we use bold font to denote vectors and matrices. Any point in the image is denoted by its coordinate $\mathbf{e} = [u, v]$. We denote any line segment by its two endpoints, $\mathbf{l} = [\mathbf{e}_1, \mathbf{e}_2] = [u_1, v_1, u_2, v_2]$. For any line



(a) Consistency measure in [6]. (b) Consistency measure in [4]. (c) Consistency measure in [13, 17].

Figure 2. Existing consistency measures.

segment $\mathbf{l}$, we define $\hat{\mathbf{l}}$ as its corresponding extension line in the homogeneous image coordinate system $\hat{\mathbf{l}} = [a, b, c]$.

4. Algorithm

In this section, we first describe the overview of our approach in Sec. 4.1, then, focus on the details of two of our main contributions in Sec. 4.2 and Sec. 4.3, respectively.

4.1. Approach Overview

Our algorithm consists of two steps for VP detection: initialization and refinement.

For initialization, first, line segments are extracted from images. Let us assume that total N line segments are extracted, which are denoted as $\{\mathbf{l}_1, ..., \mathbf{l}_N\}$. While any line segment extraction approach can be used, we used [16] in this paper. Then, an initial set of VPs are computed by clustering line segments using our *novel consistency measure* (see Sec. 4.2). In this work, the J-linkage algorithm [13] is used to cluster line segments to respective VP clusters in the following manner: first, M line segment pairs (e.g., $M = 3000$) are randomly sampled where each pair provides a potential hypothesized VP; for each line segment $\mathbf{l}$, we measure its consistency with respect to all hypothesized VPs and identify a set of VPs that associate with $\mathbf{l}$ above a pre-defined consistency threshold η; the identified set is called a "preference set" for $\mathbf{l}$; this step produces a binary $N \times M$ association matrix; then, clusters of line segments are obtained in a bottom-up manner using the Jaccard distance of their preference set; finally, we obtain $|\mathbb{V}_\mathbf{0}|$ number of line segment clusters and associated VPs where line segments across clusters do not share any hypothesized VP in their preference sets and $|\mathbb{V}_\mathbf{0}| \ll M$. After initialization, there are usually a number of noisy and small line segment clusters, which need to be corrected in the refinement step. For more details on the J-linkage clustering, readers are referred to [13].

In the refinement step, we refine the VP detection results through an iterative approach which is an EM-like algorithm. In the M-like step, for each cluster of line segments, we compute its corresponding VP using a *novel minimum error vanishing point estimation method* (see Sec. 4.3). In the E-like step, given a set of hypothesized VPs, we reassign each line segment to a VP such that the *consistency measure* (in Sec. 4.2) between the line segment and the VP is maximal and greater than a threshold (otherwise, assign to an outlier cluster). The resulting assignment forms a new line segment clusters for the M-like step. At the end of each EM-like iteration, similar VPs are merged and small outlier VPs are removed.

In the following sections, we describe both the justification and algorithmic details for the newly developed consistency measure and minimum error vanishing point estimation in Sec. 4.2 and Sec. 4.3, respectively.

4.2. Consistency Measure

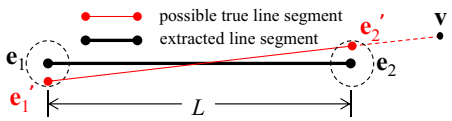
A consistency measure $c(\mathbf{l}, \mathbf{v})$, between a line segment $\mathbf{l}$ and a hypothesized VP $\mathbf{v}$, is one of the most important algorithmic component required by most VP detection algorithms. It is used to evaluate the degree of $\mathbf{l}$ being consistent with $\mathbf{v}$. While there are existing consistency measures [4, 6, 13, 17] which work fairly well in practice, however, most of them are hand-crafted; and do not encode the measurement error during line segment extraction and its impact on the VP estimation in a principled manner.

To provide the motivation for our new probabilistic consistency measure, we briefly review existing consistency measures and their potential limitations first.

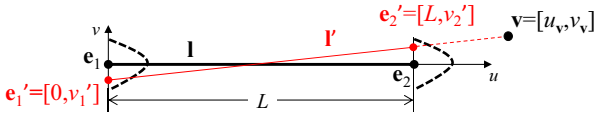
As shown in Figure 2(a), Kosecka and Zhang [6] model the consistency measure based on the orthogonal distance d from $\mathbf{v}$ to $\mathbf{l}$ in image. This formulation is biased against VPs far away from the line segment (along the direction of the line segment). Intuitively, in Fig. 2(a), $\mathbf{l}$ should have higher consistency with $\mathbf{v}_1$ than $\mathbf{v}_2$ and $\mathbf{v}_3$; and $c(\mathbf{l}, \mathbf{v}_3)$ should be zero since the projection of $\mathbf{v}_3$ on $\hat{\mathbf{l}}$ resides on $\mathbf{l}$ itself while any finite line segment in the world should not pass through its own VP. However, under the formulation in [6], $c(\mathbf{l}, \mathbf{v}_1) = c(\mathbf{l}, \mathbf{v}_2) = c(\mathbf{l}, \mathbf{v}_3) > 0$. This would lead to the frequent incorrect assignment of a line segment to a nearby VP; and degrade the detection of VPs at infinity or far away. Although a normalization step is applied in [6] with a rough guess on the camera parameters, it does not fully address the bias.

In another work illustrated in Figure 2(b), Denis et al. [4] use the angular deviation θ between the line segment and the line connecting the centroid $\bar{\mathbf{e}}$ of the line segment and the VP. Intuitively, similar to the case in Fig. 2(a), while $c(\mathbf{l}, \mathbf{v}_1) > c(\mathbf{l}, \mathbf{v}_2)$ and $c(\mathbf{l}, \mathbf{v}_3) = 0$, it is not the case under this formulation either, similar to [6].

In another approach shown in Fig. 2(c), Tardif [13] models the consistency measure based on the projection distance



(a) Consistency measure model.



(b) Approximation of consistency measure modeling.

Figure 3. An illustration of consistency measure modeling. (a) Error of endpoints is modeled as 2D Gaussian (ellipses); (b) Approximated modeling using 1D Gaussian (bell curves).

d from the endpoints of $\mathbf{l}$ to the line connecting the VP and the line segment centroid $\bar{\mathbf{e}}$. This formulation favors shorter line segments such as $\mathbf{l}_2$ in Fig. 2(c), which is intuitively less consistent with $\mathbf{v}$ than $\mathbf{l}_1$.

Another common issue with the aforementioned models is that they do not encode the impact of the length L of the line segments on the VP detection. For example, in Fig. 2(b), shortening the line segment without changing $\bar{\mathbf{e}}$ will not change its consistency measure. However, intuitively, very short line segments are frequently noisy, hence, their contribution should be constrained compared to the long segments. Though a weighting approach using weights proportional to length has been used [12], it is still empirical and does not capture the impact in a principled manner.

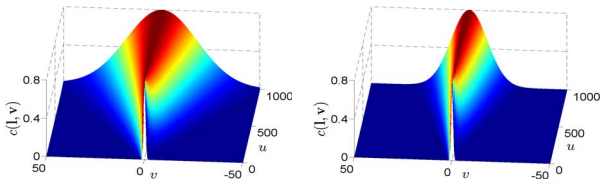
4.2.1 Probabilistic Consistency Measure

We propose a new consistency measure which models the measurement error in line segment extraction and its impact on the VP estimation. For each line segment, the proposed model builds a probabilistic consistency distribution over all possible VP locations, and explicitly captures the impact of segment length as well.

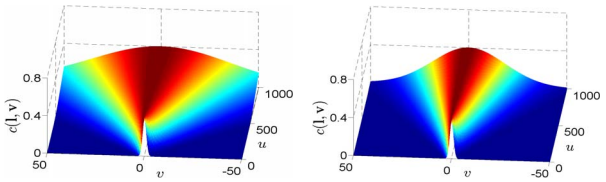
The overall model is illustrated in Fig. 3(a) in which an extracted (solid black) line segment $\mathbf{l} = [\mathbf{e}_1, \mathbf{e}_2]$ is shown with the measurement uncertainty in its endpoint locations illustrated with ellipses around them. Following the endpoint distribution, any possible true line segment $\mathbf{l}'$ collinear with a hypothesized VP $\mathbf{v}$ is shown as a red line. While the location error of $\mathbf{e}_1$ and $\mathbf{e}_2$ depends on the line segment extraction algorithm and sensor noise, in this work, we use a widely accepted isotropic Gaussian distribution[1] [8].

Let's define $\mathbb{L} = \{\mathbf{l}\}$ as the set of all possible true line segments, then, we define a subset $\mathbb{L}' = \{\mathbf{l}'\} \subset \mathbb{L}$ as the set of line segments which are collinear with $\mathbf{v}$. Our new probabilistic consistency measure is defined as $c(\mathbf{l}, \mathbf{v}) = p(\mathbb{L}')$, which can be computed as the integral over all lines in $\mathbb{L}'$. During integration, the probability of each possible true line segment is computed as the product of two probabilities of hypothesized end points with respect to two 2D uncertainty ellipses. For brevity, we derive the 1D approximation shown in Fig. 3(b), which also works well in practice. The rationale for the 1D approximation is that the collinearity between $\mathbf{l}'$ and $\mathbf{v}$ mostly depends on the endpoints' deviation along the normal direction of $\mathbf{l}$. Also, the standard deviation of the endpoint error is usually at the magnitude of sub pixel, and thus significantly smaller than L along the tangential direction of $\mathbf{l}$.

[1] Our algorithm can be easily extended to incorporate other error models such as uniform distribution.



(a) $L = 40,\ \sigma = 0.5$. (b) $L = 80,\ \sigma = 0.5$.



(c) $L = 40,\ \sigma = 1$. (d) $L = 80,\ \sigma = 1$.

Figure 4. Visualization of the consistency measure. Longer line segments and smaller σ produce narrower ridges, meaning more certainty about the location of the hypothesized vanishing point.

For the convenience of derivation and without loss of generality, as shown in Fig. 3(b), we make $\mathbf{l}$ align with the u-axis of the image, and endpoint $\mathbf{e}_1$ is at the origin. Then any possible endpoints of the true line segment can be presented as $\mathbf{e}_1' = [0, v_1']$ and $\mathbf{e}_2' = [L, v_2']$, with $v_1' \sim \mathcal{N}(0, \sigma)$ and $v_2' \sim \mathcal{N}(0, \sigma)$, where σ is the standard deviation. Define the hypothesized VP as $\mathbf{v} = [u_\mathbf{v}, v_\mathbf{v}]$. When $u_\mathbf{v} > L$ and $\mathbf{l}'$ is collinear with $\mathbf{v}$, we have, $v_2' = \frac{u_\mathbf{v}-L}{u_\mathbf{v}} v_1' + \frac{v_\mathbf{v} L}{u_\mathbf{v}}$. Accordingly, both v_1' and v_2' can be jointly parameterized with an auxiliary variable $t \in \mathbb{R} : v_1'(t) = \frac{u_\mathbf{v} t}{\sqrt{u_\mathbf{v}^2+(u_\mathbf{v}-L)^2}}$ and $v_2'(t) = \frac{(u_\mathbf{v}-L)t}{\sqrt{u_\mathbf{v}^2+(u_\mathbf{v}-L)^2}} + \frac{v_\mathbf{v} L}{u_\mathbf{v}}$. Therefore,

$$c(\mathbf{l}, \mathbf{v}) = p(\mathbb{L}') = \int_{-\infty}^{\infty} f\left(v_1'(t); 0, \sigma^2\right) f\left(v_2'(t); 0, \sigma^2\right) dt$$
$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{v_\mathbf{v}^2 L^2}{2\sigma^2(u_\mathbf{v}^2+(u_\mathbf{v}-L)^2)}}, \qquad (1)$$

where $f(\cdot; \mu, \sigma^2)$ is a Gaussian PDF.

From (1), we can observe that our formulation allows us to encode the measurement error σ explicitly and captures the intuitive impact of line length L, which provides a more sound framework compared to empirical set-ups by other approaches [4, 6, 13, 17]. In detail, Fig. 4 visualizes the consistency measure distributions with different L and

σ. It shows a ridge-like distribution across potential VP locations. Longer line segments produce narrower ridges, meaning more certainty about the location of the hypothesized VP. Smaller σ, i.e., more accurate line segment extraction and less image noise, produces higher confidence about the VP location.

During the E-like step described in Sec. 4.1, we associate each line segment with one of the hypothesized VPs or outlier class, using the consistency measure in (1). Given hypothesized VPs $\mathbb{V} = \{\mathbf{v}_1, ..., \mathbf{v}_m\}$ and line segment $\mathbf{l}$, we determine the VP that $\mathbf{l}$ belongs to by:

$$\zeta(\mathbf{l}; \mathbb{V}) = \begin{cases} \arg\max_{\mathbf{v} \in \mathbf{V}} c(\mathbf{l}, \mathbf{v}) & \text{if } \max_{\mathbf{v} \in \mathbb{V}} c(\mathbf{l}, \mathbf{v}) > \eta, \\ \text{outlier} & \text{otherwise}, \end{cases} \quad (2)$$

where η is a threshold.

4.3. Vanishing Point Estimation

4.3.1 Maximum Likelihood (ML) Estimation

By (2), we form clusters of line segments $\{\mathbb{S}_i\}$. For each cluster $\mathbb{S}_i$, a straightforward way to estimate its VP is,

$$\mathbf{v}_i^* \triangleq \psi(\mathbb{S}_i) = \arg\min_{\mathbf{v}} \sum_{\mathbf{l} \in \mathbb{S}_i} -\log\left(c(\mathbf{l}, \mathbf{v})\right). \quad (3)$$

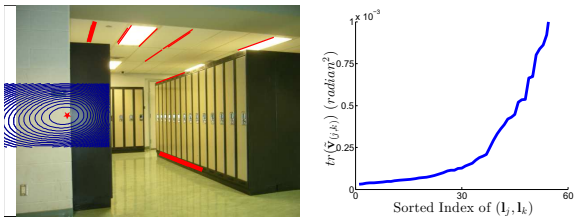
Although the minimization problem in (3) is not a strict convex optimization problem, as visualized in Fig. 5(a), there almost always exists a large neighborhood around $\mathbf{v}_i^*$ where $\sum_{\mathbf{l} \in \mathbb{S}_i} -\log\left(c(\mathbf{l}, \mathbf{v})\right)$ is quasiconvex (unimodal.) With reasonable initialization methods such as the least squares approach in [6], the initial point is almost always in this neighborhood. Therefore any local optimum solver can find the global optimal solution efficiently.

However, besides the possible sub-optimality issue, the VP estimation in (3) ignores the uncertainty of the resulting VP. By allowing all line segments uniformly contribute to the optimization, the surface of $\sum_{\mathbf{l} \in \mathbb{S}_i} -\log\left(c(\mathbf{l}, \mathbf{v})\right)$ around the maximum (as illustrated in Fig. 5(a)) could be very flat (i.e., large uncertainty) especially due to short line segments (see Fig 4) and subset of line segments that are close to parallel to each other. Then the solution is less reliable and even little numerical error from the optimizer could greatly compromise the solution quality.

4.3.2 Minimum Error Estimation

To resolve the aforementioned issues, we propose a novel minimum error VP estimation by optimally weighting minimal solutions produced by line segment pairs. Here, by error, we mean the uncertainty of the VP, in particular, the trace of the covariance matrix of the VP.

Given a line segment pair $(\mathbf{l}_j, \mathbf{l}_k) \in \mathbb{L}_i^2$, $j \neq k$, the intersection of corresponding $\hat{\mathbf{l}}_j$ and $\hat{\mathbf{l}}_k$ is a minimal solution as a hypothesized VP. In particular, given $\mathbf{l}_j = [u_{j1}, v_{j1}, u_{j2}, v_{j2}]$ and $\mathbf{l}_k = [u_{k1}, v_{k1}, u_{k2}, v_{k2}]$, the intersection $\mathbf{v}_{(j,k)}$ as a VP is shown in (4).



(a) (b)

Figure 5. An illustration of the maximum likelihood and minimum error VP estimation approaches. (a): The level-set visualization of the cost function $\sum_{\mathbf{l} \in \mathbb{S}_i} -\log\left(c(\mathbf{l}, \mathbf{v})\right)$, where the star indicates the maximum likelihood VP. (b): The visualization of the trace of $\tilde{\mathbf{v}}_{j,k}$ from different line segment pairs. The pair with the least trace is highlighted with a larger width in (a).

Define the Jacobian matrix of function $g(\cdot)$ in (4) with respect to $(\mathbf{l}_j, \mathbf{l}_k)$ as $\mathbf{G}_{(j,k)}$, which is a 2×8 matrix (omitted here for space, presented in supplementary material), then the covariance of the respective VP $\mathbf{v}_{(j,k)}$ in the image is,

$$\begin{aligned} \Sigma^{\mathbf{v}_{(j,k)}} &= \mathbf{G}_{(j,k)} \Sigma_{(j,k)} \mathbf{G}_{(j,k)}^T \\ &= \mathbf{G}_{(j,k)} \sigma^2 \mathbf{I}_8 \mathbf{G}_{(j,k)}^T = \sigma^2 \mathbf{G}_{(j,k)} \mathbf{G}_{(j,k)}^T, \end{aligned} \quad (5)$$

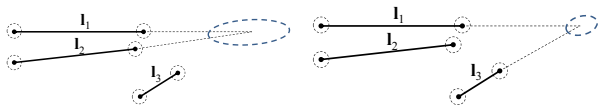
where $\Sigma_{(j,k)} = \sigma^2 \mathbf{I}_8$ is the covariance matrix of the endpoints of $\mathbf{l}_j$ and $\mathbf{l}_k$, and $\mathbf{I}_8$ is an 8×8 identity matrix.

Eq. (5) shows that the covariance of the hypothesized VP is a function of $\mathbf{l}_j$, $\mathbf{l}_k$, and σ. It encodes the relative location of line segments in the pair, line segment length, image noise level, segmentation error, etc. However, the covariance of the VP in the image space is less meaningful due to camera projective distortion. For example, a large covariance of a VP at infinity or very far away does not necessarily means an inaccurate estimation. What really matters is the vanishing direction, which can be represented as a point on a unit Gaussian sphere [2]. The center of the unit Gaussian sphere coincides at the camera center and its two rotation axes are parallel with image u-axis (tilt rotation) and v-axis (pan rotation), respectively. Here, we represent a VP $\mathbf{v}$ on the Gaussian spherical surface as a pair of pan and tilt angles $\tilde{\mathbf{v}} = [\theta, \phi]$. Following the coordinate system convention in [11], we project $\mathbf{v}_{(j,k)}$ to $\tilde{\mathbf{v}}_{(j,k)}$ as

$$\begin{aligned} \tilde{\mathbf{v}}_{(j,k)} &= [\theta_{(j,k)}, \phi_{(j,k)}] \triangleq h(\mathbf{v}_{(j,k)}) \\ &= \left[\tan^{-1}\left(\frac{u_{(j,k)}}{f}\right), -\tan^{-1}\left(\frac{v_{(j,k)}}{\sqrt{u_{(j,k)}^2 + f^2}}\right) \right], \end{aligned} \quad (6)$$

where f is a focal length. Here, we do not have the calibrated focal length. However, most cameras have a 15 to 60 degree field of view, which indicates $f \in [0.28W, 3.80W]$ with W being the size of image's longer side. We estimate the focal length as $f = 2W$. This rough approximation is unacceptable for approaches that rely on accurate

$$\mathbf{v}_{(j,k)} = \begin{bmatrix} u_{(j,k)} \\ v_{(j,k)} \end{bmatrix}^T \triangleq g(u_{j1}, v_{j1}, u_{j2}, v_{j2}, u_{k1}, v_{k1}, u_{k2}, v_{k2}) = \begin{bmatrix} \frac{(u_{k1}-u_{k2})(u_{j1}v_{j2}-u_{j2}v_{j1})-(u_{j1}-u_{j2})(u_{k1}v_{k2}-u_{k2}v_{k1})}{(u_{j1}-u_{j2})(v_{k1}-v_{k2})-(u_{k1}-u_{k2})(v_{j1}-v_{j2})} \\ \frac{(v_{k1}-v_{k2})(u_{j1}v_{j2}-u_{j2}v_{j1})-(v_{j1}-v_{j2})(u_{k1}v_{k2}-u_{k2}v_{k1})}{(u_{j1}-u_{j2})(v_{k1}-v_{k2})-(u_{k1}-u_{k2})(v_{j1}-v_{j2})} \end{bmatrix}^T. \tag{4}$$



(a) (b)

Figure 6. An illustration of the intuition behind the minimum error vanishing point estimation.

calibration. However, since our minimum error VP is reconciled from competing minimal solutions, which are projected (back and forth) with the same estimated focal length, this approximation is adequate for our purpose.

We define the 2×2 Jacobian matrix of $h(\cdot)$ in (6) as,

$$\mathbf{H}_{(j,k)} = \begin{bmatrix} \frac{f}{\rho^2_{(j,k)}}, & 0 \\ \frac{uv}{\left(\rho^2_{(j,k)}+v^2_{(j,k)}\right)\rho_{(j,k)}}, & \frac{-\rho_{(j,k)}}{\rho^2_{(j,k)}+v^2_{(j,k)}} \end{bmatrix},$$

where $\rho_{(j,k)} = \sqrt{u^2_{(j,k)} + f^2}$. Then the Jacobian from four endpoints of $(\mathbf{l}_j, \mathbf{l}_k)$ to the respective VP's pan-tilt angles is $\mathbf{J}_{(j,k)} = \mathbf{H}_{(j,k)}\mathbf{G}_{(j,k)}$. Then the covariance of $\tilde{\mathbf{v}}_{(j,k)}$ is,

$$\Sigma^{\tilde{\mathbf{v}}_{(j,k)}} = \mathbf{J}_{(j,k)}\Sigma_{(j,k)}\mathbf{J}^T_{(j,k)} = \sigma^2\mathbf{J}_{(j,k)}\mathbf{J}^T_{(j,k)}. \tag{7}$$

From (7), apparently, each line segment pair has a different impact on the covariance of the VP estimation. As illustrated in Fig. 6, even though the pair $(\mathbf{l}_1, \mathbf{l}_2)$ have longer line segments, since they are almost parallel, the covariance of the intersection is larger than that from the pair $(\mathbf{l}_1, \mathbf{l}_3)$, which has shorter line segments. Fig. 5(b) visualizes the sorted trace of $\tilde{\mathbf{v}}_{(j,k)}$ from different pairs of line segments (in the same cluster) shown in Fig. 5(a). This indicates that some line segment pairs are more important than others. For example, in Fig. 5(a), the line segment pair with the least trace (i.e., the most trustable) line segment pair is highlighted with a larger width. Uniformly using more line segments does not necessarily guarantee more accuracy. Accordingly, we derive an optimal weighting approach which computes the VP by reconciling all minimal solutions and weighs more on those with more certain VP estimations, so that the reconciled solution is the most certain one.

Given the line segment cluster $\mathbb{S}$ with n line segments, define the set of all line segment pairs, $\Gamma = \{(\mathbf{l}_j, \mathbf{l}_k) | \mathbf{l}_j, \mathbf{l}_k \in \mathbb{S}, j \neq k\}$. Define the minimal solution set produced by any pair in Γ as $\Psi = \{\tilde{\mathbf{v}}_{(j,k)} | (\mathbf{l}_j, \mathbf{l}_k) \in \Gamma\}$. Define the weighting vector for Ψ as $\mathbf{w} = [w_{(1,2)}, w_{(1,3)}, ..., w_{(n-1,n)}]^T$, subject to $\mathbf{1}^T_{(n-1)n/2}\mathbf{w} = 1$, where $\mathbf{1}_{(n-1)n/2}$ is a column vector with all elements being 1. We reconcile all the minimal solutions via linear weight combination, $\tilde{\mathbf{v}} = \sum_{j=1}^{n-1}\sum_{k=j+1}^{n} w_{(j,k)}\tilde{\mathbf{v}}_{(j,k)}$. Define the corresponding covariance of $\tilde{\mathbf{v}}$ as $\Sigma^{\tilde{\mathbf{v}}}$, our goal is to find the $\mathbf{w}$ that minimizes the trace of $\Sigma^{\tilde{\mathbf{v}}}$,

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} tr(\Sigma^{\tilde{\mathbf{v}}}), \quad \text{s.t. } -\mathbf{w} < 0, \quad \text{and } \mathbf{1}^T_{(n-1)n/2}\mathbf{w} = 1, \tag{8}$$

where function $tr(\cdot)$ returns the trace of any square matrix. Eq. (8) can be rewritten as

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}, \quad \text{s.t. } -\mathbf{w} < 0, \quad \text{and } \mathbf{1}^T_{(n-1)n/2}\mathbf{w} = 1, \tag{9}$$

where $\mathbf{A}$ is an $n(n-1)/2 \times n(n-1)/2$ matrix with its positive diagonal elements $\mathbf{A}(a,a) = tr(\Sigma^{\tilde{\mathbf{v}}_{(j,k)}})$, where a is the index of $\mathbf{w}$ corresponding to the line segment pair $(\mathbf{l}_j, \mathbf{l}_k)$. The off-diagonal elements of $\mathbf{A}$ model the correlation between two line segment pairs.

The minimization problem in (9) is a typical quadratic programming problem, which can be efficiently solved. In practice, very few observations dominate the contributions to the covariance as illustrated in Fig. 5(b). In other words, the correlation between observations, and in turn the resulting difference in contribution, is very weak (i.e., off-diagonal items in matrix $\mathbf{A}$ are small). Therefore we approximate the computation by assuming the independence between observations. Assuming $\mathbf{A}$ is a diagonal matrix, the problem in (9) has a closed-form solution,

$$w^*_{(j,k)} = w^*_a = \frac{(\mathbf{A}(a,a))^{-1}}{\sum_{a=1}^{(n-1)n/2}(\mathbf{A}(a,a))^{-1}}. \tag{10}$$

With $\mathbf{w}^*$, we compute the minimum error VP in the pan-tilt space as $\tilde{\mathbf{v}}^* = [\theta^*, \phi^*] = \sum_{j=1}^{n-1}\sum_{k=j+1}^{n} w^*_{(j,k)}\tilde{\mathbf{v}}_{(j,k)}$. The corresponding optimal VP in image is,

$$\mathbf{v}^* = [u^*, v^*] = \left[f\tan(\theta^*), -f\tan(\phi^*)\sqrt{\tan^2(\theta^*)+1}\right]. \tag{11}$$

With the minimum error VP estimation in (11) and the consistency measure in (1), we complete our algorithm described in Sec. 4.1.

5. Experiments

For experiments, the proposed algorithm is implemented in Matlab. For line segment detection, [16] with sub-pixel accuracy is used. Considering this accuracy and possible image distortion, we conservatively set $\sigma = 1$ pixel. We have implemented the J-linkage algorithm [13], with a modification where the line segment-VP association is

Figure 7. Sample qualitative results on York Urban Dataset (first rows) and Eurasian Cities Dataset (second rows.) For each image, it shows clusters of line segments (with different colors), and the computed horizon (solid purple) compared with the ground truth (dashed cyan.)

determined by thresholding the new consistency measure. We set the consistency measure threshold $\eta = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}}$, which is the consistency measure at one standard deviation away from the mean (see (1)). In J-linkage, we choose upto $M = 3000$ line segment pairs for VP initialization.

We have tested our algorithm on two public datasets: York Urban Dataset (YUD) [4], and Eurasian Cities Dataset (ECD) [14]. YUD consists of 57 outdoor and 47 indoor images of urban scenes on the campus of York University and in downtown Toronto, Canada. All the scenes satisfy the Manhattan world assumption. Manually annotated line segments, ground truth VPs and camera intrinsic parameters are provided. ECD consists of 103 outdoor images of European and Asian cities, covering different architectural styles and diverse scene types. Many scenes in ECD do not satisfy the Manhattan world assumption as more than two horizontal vanishing directions exist and/or buildings with irregular non-box shapes without clear straight lines are more common. Manually annotated line segments and ground truth VPs are provided but camera intrinsic parameters are missing. Overall, ECD is more challenging than YUD dataset.

We compare our algorithm with the "Self-Similar Sketch" algorithm [15], the 4-line RANSAC algorithm [17], the "Geometric Parsing" algorithm [14], the J-linkage+EM algorithm [13], and the well-known "Video Compass" algorithm [6]. To the best of our knowledge, "Geometric Parsing" [14] and "Self-Similar Sketch" [15] are the latest state-of-the-art algorithms without the Manhattan world assumption, and the 4-line RANSAC algorithm in [17] is the latest state-of-the-art with the Manhattan world assumption.

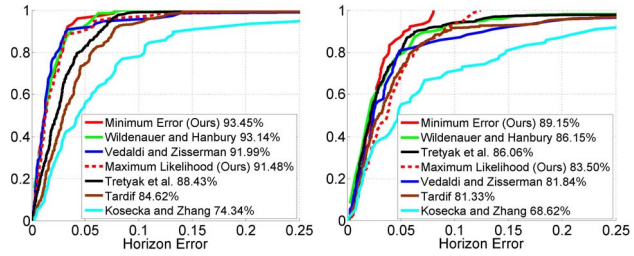
For each image in YUD and ECD, our algorithm takes a few seconds to finish on a moderate laptop. Considering its un-optimized Matlab implementation, our algorithm is fast, compared to non-Manhattan world works [14, 15], which take a few seconds to a few minutes. Following the protocol from [14, 15, 17], for each image, we ran 5 trials for the average performance. We also computed horizons using VPs, which are widely used for quantitative comparison.

Examples of qualitative results by our algorithm on YUD and ECD are shown in Fig. 7. It can be observed that our algorithm achieves very accurate line segment clustering and horizon detection for both indoor and outdoor scenes. On a small number of images, there are minor errors observed but they are reasonable, e.g., red cluster on bottom left-most and yellow cluster on top right-most images.

For quantitative evaluation, following recent state-of-the-art works [14, 15, 17], we use the error of horizon detection as the metric. We compute the horizon by first filtering out the zenith, which usually locates vertically and far from the image center. Since the horizon should be orthogonal with the line connecting the zenith and the principal point, by assuming the image center as the principal point and the focal length $f \in [0.28W, 3.8W]$, we further filter out VPs that are unlikely to be horizontal VPs, and the surviving VPs are treated as the horizontal VPs. We compute the horizon by enforcing orthogonality and linearly weighting horizontal VPs. This procedure is similar to [14], except that we use the inverse of the trace of each VP's covariance as the weight, instead of empirically choosing the number of associated line segments as the weight in [14]. Following [14, 15, 17], we define the horizon error as the maximum distance from the computed horizon to the ground truth horizon in the image, normalized by the image height.

The cumulative distributions for the horizon error for both datasets are shown in Fig. 8. The x-axis value is the horizon error and the y-axis value is the share of the images that have less horizon error than the corresponding x-value. The area under the curve (AUC) for each algorithm is sorted and depicted in the legend. Furthermore, to evaluate the impact of the minimum error VP estimation (Sec. 4.3.2) separately, two variations are used: our full approach is marked as "Minimum Error"; while another run "Maximum Likelihood" (ML) uses new consistency measure, but, with maximum likelihood estimation (Sec. 4.3.1).

On YUD, our "Minimum Error" algorithm achieves slightly higher AUC than the latest state-of-the-art [17]. Note that all scenes in YUD satisfy the Manhattan world assumption, which [17] utilizes to constrain the VP esti-



(a) York Urban Dataset (b) Eurasian Cities Dataset

Figure 8. Cumulative histograms for the horizon error. The x-axis value is the horizon error (see text for details). The y-axis value is the share of the images that have less horizon error than the corresponding x-value. The legends are sorted based on AUC.

mation. We do not assume the Manhattan world and still achieve slightly better performance. Furthermore, it shows that the maximum error by our "Minimum Error" algorithm is bounded very tightly. Our maximum error across YUD is 0.078 while the best maximum error from all competing algorithms is 0.158 (by [17]). Since the YUD is relatively "easy" compared to ECD, while our "Minimum Error" method shows the best results, almost all the latest algorithms perform well and our improvement is small.

On ECD, our "Minimum Error" algorithm outperforms all competing approaches with a much larger margin. The maximum error on ECD by our "Minimum Error" algorithm is 0.081 while the best maximum error from all competing methods is 0.532 (by [15]). In particular, the bottom right-most image in Fig. 7 shows a very challenging scene. The building facade is curvy and does not contain clear straight lines. However, our algorithm is still able to cluster small piecewise line segments and compute a fairly accurate horizon. The average error by our "Minimum Error" method on this image is only 0.034. In contrast, this image is a failure case in [14]. This shows that our "Minimum Error" algorithm is more robust to image noise/distortion and, to a certain extent, violation of the underlying assumptions.

Another important observation is that our ML algorithm outperforms Tardif's algorithm [13] on both datasets. In fact, our ML algorithm is very similar to [13] except the new consistency measure. Therefore, we attribute this improvement to the new consistency measure. Nevertheless, our full "Minimum Error" algorithm always outperforms ML version substantially, which showcases the clear benefits of the minimum error VP estimation.

6. Conclusion

We have presented a new algorithm for VP detection for uncalibrated monocular images without the Manhattan world assumption. Our approach advances the state-of-the-art using a new consistency measure and a minimum error VP estimation approach. Our experimental results on two public benchmark datasets are encouraging and shows the strength of the proposed approach.

Acknowledgement

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. W31P4Q-10-C-0214. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or the U.S. Government.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

References

[1] J. Bazin, Y. Seo, C. Demonceaux, P. Vasseur, K. Ikeuchi, I. Kweon, M. Pollefeys, et al. Globally optimal line clustering and vanishing point estimation in Manhattan world. In *CVPR*, 2012.

[2] R. Collins and R. Weiss. Vanishing point calculation as a statistical inference on the unit sphere. In *ICCV*, 1990.

[3] J. Coughlan and A. Yuille. Manhattan world: Orientation and outlier detection by bayesian inference. *Neural Computation*, 15(5):1063–1088, 2003.

[4] P. Denis, J. Elder, and F. Estrada. Efficient edge-based methods for estimating Manhattan frames in urban imagery. In *ECCV*, 2008.

[5] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In *CVPR*, 2009.

[6] J. Kosecka and W. Zhang. Video compass. In *ECCV*, 2002.

[7] H. Li, D. Song, Y. Lu, and J. Liu. A two-view based multilayer feature graph for robot navigation. In *ICRA*, 2012.

[8] D. Liebowitz and A. Zisserman. Metric rectification for perspective images of planes. In *CVPR*, 1998.

[9] F. Mirzaei and S. Roumeliotis. Optimal estimation of vanishing points in a Manhattan world. In *ICCV*, 2011.

[10] J. Shufelt. Performance evaluation and analysis of vanishing point detection techniques. *TPAMI*, 21(3):282–288, 1999.

[11] D. Song, Y. Xu, and N. Qin. Aligning windows of live video from an imprecise pan-tilt-zoom camera into a remote panoramic display for remote nature observation. *Journal of Real-Time Image Processing*, 5(1):57–70, 2010.

[12] R. Szeliski. *Computer vision: algorithms and applications*. Springer-Verlag New York Inc, 2010.

[13] J. Tardif. Non-iterative approach for fast and accurate vanishing point detection. In *ICCV*, 2009.

[14] E. Tretyak, O. Barinova, P. Kohli, and V. Lempitsky. Geometric image parsing in man-made environments. *IJCV*, 97:305–321, 2012.

[15] A. Vedaldi and A. Zisserman. Self-similar sketch. In *ECCV*, 2012.

[16] R. Von Gioi, J. Jakubowicz, J. Morel, and G. Randall. Lsd: A fast line segment detector with a false detection control. *TPAMI*, 32(4):722–732, 2010.

[17] H. Wildenauer and A. Hanbury. Robust camera self-calibration from monocular images of Manhattan worlds. In *CVPR*, 2012.