Planar Structure Matching under Projective Uncertainty for Geolocation

Ang Li, Vlad I. Morariu, and Larry S. Davis

University of Maryland, College Park
{angli,morariu,lsd}@umiacs.umd.edu

Abstract. Image based geolocation aims to answer the question: where was this ground photograph taken? We present an approach to geolocalating a single image based on matching human delineated line segments in the ground image to automatically detected line segments in ortho images. Our approach is based on distance transform matching. By observing that the uncertainty of line segments is non-linearly amplified by projective transformations, we develop an uncertainty based representation and incorporate it into a geometric matching framework. We show that our approach is able to rule out a considerable portion of false candidate regions even in a database composed of geographic areas with similar visual appearances.

Keywords: uncertainty modeling, geometric matching, line segments.

1 Introduction

Given a ground-level photograph, the image geolocation task is to estimate the geographic location and orientation of the camera. Such systems provide an alternative way to localize an image or a scene when and where GPS is unavailable. Visual based geolocation has wide applications in areas such as robotics, autonomous driving, news image organization and geographic information systems. We focus on a single image geolocation task which compares a single ground-based query image against a database of ortho images over the candidate geolocations. Each of the candidate ortho images is evaluated and ranked according to the query. This task is difficult because (1) significant color discrepancy exists between cameras used for ground and ortho images; (2) the images taken at different times result in appearance difference even for the same locations (e.g. a community before and after being developed); (3) the ortho image databases usually have a very large scale, which requires efficient algorithms.

Due to the difficulty of the geolocation problem, many recent works include extra data such as georeferenced image databases [9,14], digital elevation models (DEM) [1], light detection and ranging (LIDAR) data [16], etc. Whenever photographs need to be geolocated in a new geographic area, this side data has to be acquired first. This limits the expandability of these geolocation approaches. One natural question to ask is whether we can localize a ground photograph using only widely accessible satellite images.

D. Fleet et al. (Eds.): ECCV 2014, Part VII, LNCS 8695, pp. 265–280, 2014.

Fig. 1. Geolocation involves finding the corresponding location of the ground image (on the left) in ortho images (an example on the right) ©Google

We address this geolocation task with no side data by casting it as an image matching problem. This is challenging because the camera orientation of a ground image is approximately orthogonal to that of its corresponding ortho image. Commonly used image features are not invariant to such wide camera rotation. In addition, considering the presence of color and lighting difference between ground and ortho images, color-based and intensity-based image features become unreliable for establishing image correspondence. Therefore, structural information becomes the most feasible feature for this application. We utilize linear structures – line segments – as the features to be matched between ground and ortho images.

Both ground and ortho images are projections of the 3D world. The information loss between these two images becomes an obstacle even for matching binary line segments. Instead of inferring 3D structure, we extract and match the linear structures that lie on the ground a large subset of which is visible in both ground and ortho images. The ortho images can be regarded as approximately 2D planes and we use classic line extraction algorithms to locate the extended linear structures in them. The ground images are more challenging so we ask humans annotate the ground lines for these images. This is not a burdensome task. Additionally, the horizon line is annotated by the human so we can construct its corresponding aerial view with the camera parameters known.

Based on chamfer matching [15], we derive a criterion function for matching each ortho image with the ortho-rectified view of the ground image. However, the projection matrix for transforming the ground image to its ortho view is usually numerically ill-conditioned. Even a small perturbation to the annotated end points of a line segment may result in significant uncertainty in location and orientation of the projected line segments, especially those near the horizon line. Therefore, we propose a probablistic representation of line segments by modeling their uncertainty and introduce a model of geometric uncertainty into our matching criterion. Within each ortho image, the matching scores for possible pairs of camera locations and orientations are exhaustively evaluated. This sliding window search is speeded up by means of distance transforms [7] and convolution operations.

Contributions. The main contributions of this paper include (1) an uncertainty model for line segments under projective transformations (2) a novel distance transform based matching criterion under uncertainty (3) the application of geometric matching to single image geolocation with no side data.

2 Related Work

Image Geolocation. Previous work on image geolocation can be classified into two main streams: geotagged image retrieval and model based matching. Hays et al. [9] were among the first to treat the image geolocation as a data driven image retrieval problem. Their approach is based on a large scale geotagged image database. Those images with similar visual appearance to the query image are extracted and their GPS tags are collected to generate a confidence map for possible geolocations. Li et al. [13] devised an algorithm to match low level features from large scale database to ground image features in a prioritized order specified by likelihood. Similar approaches improve the image retrieval algorithms applied to ground level image databases [5,20,24,25]. Generally, data driven approaches assume all possible views of the ground images are covered in the database. Otherwise, the system will not return a reasonable geolocation.

Apart from the retrieval-style geolocation, the other track is to match the image geometry with 3D models to estimate the camera pose. Battz et al. [1] proposed a solution to address the geolocation in mountainous terrain area by extracting skyline contours from ground images and matching them to the digital elevation models. From the 3D reconstruction viewpoint, some other approaches estimate the camera pose by matching images with 3D point cloud [10,12,19].

Few works make use of the satellite images in the geolocation task. Bansal et al. [2] match the satellite images and aerial images by finding the facade of the building and rectifying the facade for matching with the query ground images. Lin et al. [14] address the out-of-sample generalization problem suffered by data-driven methods. The core of their method is learning a cross-view feature correspondence between ground and ortho images. However, their approach still requires a considerable amount of geo-tagged image data for learning.

Our work differs from all of the above work in that our approach casts the geolocation task as a linear geometric matching problem instead of reconstructing the 3D world, and it is relatively "low-cost" using only the satellite images without the need for large labeled training sets or machine learning.

Geometric Matching. In the geometric matching domain, our approach is related to line matching and shape matching. Matching line segments has been an important problem in geometric modeling. Schmid et al. [21] proposed a line matching approach based on cross correlation of neighborhood intensity. This approach is limited by its requirement on prior knowledge of the epipolar geometry. Bay et al. [4] match line segments using color histograms and remove false correspondences by topological filtering. In recent years, line segments have been shown to be robust to matching images in poorly textured scenes [11,23]. Most

of the existing works rely on local appearance-based features while our approach is completely based on matching the binary linear structures.

Our approach is motivated by chamfer matching [3], which has been widely applied in shape matching. Chamfer matching involves finding for each feature in an image its nearest feature in the other image. The computation can be efficiently achieved via distance transforms. A natural extension of chamfer matching is to incorporate the point orientation as an additional feature. Shotton et al. [22] proposed oriented chamfer matching by adding an angle difference term into their formulation and applied this technique in matching contour fragments for general object recognition. Another method for encoding the orientation is the fast directional chamfer matching proposed by Liu et al. [15]. They generalize the original chamfer matching approach by seeing each point as a 3D feature which is composed of both location and orientation. Efficient algorithms are employed for computing the 3D distance transform based on [7]. However, for geolocation, our problem is to match a small linear structures to fairly large structures that contain much noise, especially in ortho images. Our approach is carefully designed specifically for the needs of geolocation: it takes into account the projective transformations and line segments with uncertain end points as part of the matching criterion function.

Uncertainty Modeling. Uncertainty is often involved in various computer vision problems. Olson [17] proposed a probabilistic formulation for Hausdorff matching. Similar to Olsons work, Elgammal et al. [6] extended Chamfer matching to a probabilistic formulation. Both approaches consider only the problem of matching an exact model to uncertain image features, while our work handles the situation when the model is uncertain. An uncertainty model is proposed in [18] for projective transformations in multi-camera object tracking. They considered the case where the imaged point is sufficiently far from the line at infinity and provided an approximation method to compute the uncertainty under projective transformation. Our work differs in that (1) we provide an exact solution for projective uncertainty of line segments, and (2) we do not assume that line segments are far from the horizon line. To our knowledge, none of the previous work in geolocation were incorporated with uncertainty models.

3 Our Approach

A query consists of a single ground image with unknown location and orientation is provided. This ground image is then matched exhaustively to each candidate ortho images, and ortho images are ranked according to their matching scores. The ortho images are densely sampled by overlapped sliding windows over the candidate geographic areas. The scale of each ortho image can be around 10 centimeters per pixel. The ground images could be taken at any location within ortho images. Even in a 640×640 ortho image, there are over millions of possible discretized camera poses. The geolocation task is to localize the ground image into the ortho images, not necessarily the camera pose.

Fig. 2. Examples of line segments annotated in ground images ©Google

We have two assumptions here to simplify this problem. First, the camera parameter (focal length) for ground images is known, a reasonable assumption, since modern cameras store this information as part of the image metadata. Second, we assume the photographer holds the camera horizontally, i.e. the camera optical axis is approximately parallel to the ground. Camera rotation around the optical axis may happen and is handled by our solution. No restrictions assumed for the satellite cameras as long as satellite imagery is rectified to ensure linear structures remain linear, which is generally true.

3.1 Preprocessing

We reconstruct the aerial view of the ground image by estimating the perspective camera model from the manually annotated horizon line. In our matching approach, line segments are matched between ground and ortho images. Lines on the ground are most likely to be viewed in both ground and ortho images – most other lines are on the vertical surfaces that are not visible in satellite imagery – so we ask users to annotate only line segments on the ground plane in query images. Once the projection matrix is known, the problem becomes one of geometric matching between two planes.

Line Segment Labeling. Line segments in ground images are annotated by human users clicking pairs of ending points. It is affordable to incorporate such human labeling process into our geolocation solution since the annotation is inexpensive and each query image needs to be labeled only once. A person can typically annotate a query image in at most two minutes. Fig. 2 shows four ground image samples with superimposed annotated line segments.

Line segments in the ortho images are automatically detected using the approach of [8]. The detected line segments lie mostly on either the ground plane or some plane parallel to the ground, such as the roof of a building. We do not attempt to remove these non-ground lines. In fact, some of the non-ground plane lines prove useful for matching. For example, the rooflines of many buildings have the same geometry as their ground footprints. Human annotators label linear features around the bottoms of these buildings. Thus, the line segments lying on the edges of a building roof still contribute to the structure matching. Our geometric matching algorithm assumes a high level of outliers, so even if the rooflines and footprints are different the matching can still be successful.

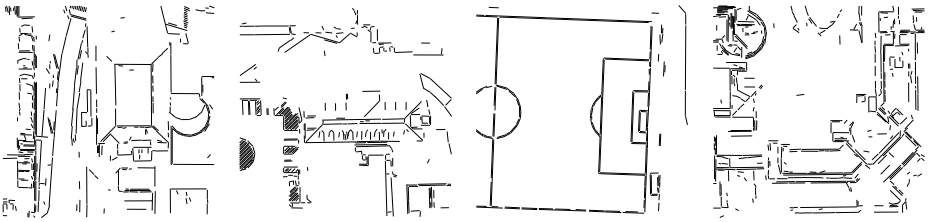


Fig. 3. Examples of line segments detected in ortho images ©Google

Aerial View Recovery. Using the computed perspective camera model, we transform the delineated ground photo line segments to an overhead view. Two assumptions are made for recovering the aerial view from ground images: (1) the camera focal length f is known, and (2) the optical axis of camera is parallel to the ground plane, i.e. the camera is held horizontally. These assumptions are not sufficient for reconstructing a complete 3D model but is sufficient for recovering the ground plane given the human annotated horizon line. The horizon line is located by finding two vanishing points, i.e. intersections of lines parallel in the real world.

Assuming the horizon line has slope angle θ, the ground image can be rotated clockwise by θ so that the horizon line becomes horizontal (the y-coordinate of rotated horizon line y'_0). The rotated coordinates are $(x', y')^\top = \mathbf{R}_\theta(x_g, y_g)^\top$ for every pixel (x_g, y_g) in the original ground image. In the world coordinate system (X, Y, Z), the camera is at the origin, facing the positive direction of the Y-axis, and the ground plane is $Z = -Z_0$. If we know pixel (x', y') is on the ground, then its corresponding world location can be computed by

$$x' = fX/Y, y' - y'_0 = fZ_0/Y \Rightarrow X = x'Z_0/(y' - y'_0), Y = fZ_0/(y' - y'_0) \quad (1)$$

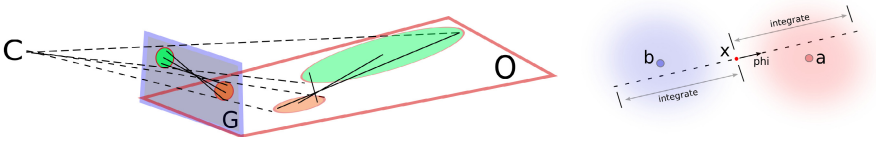
For the ortho image, a pixel location (x_o, y_o) can be converted to world coordinates by $(X, Y) = (x_o/s, y_o/s)$ where s is a scale factor with unit 1/meter relating the pixel distance to real world distance.

3.2 Uncertainty Modeling for Line Segments

User annotations on ground images are often noisy. The two hand-selected end points could easily be misplaced by a few pixels. However, after projective transformation, even a small pertubation of one pixel can result in significant uncertainty in the location and orientation of the line segment, especially if that pixel is close to the horizon (see Fig. 5(a)). Therefore, before discussing the matching algorithm, we first study the problem of modeling the uncertainty of line segments under projective transformation to obtain a principled probabilistic description for ground based line segments. We obtain a closed form solution by assuming that the error of labeling an end point on ground images be described by a normal distribution in the original image. We first introduce a lemma which is essentially the integration of Gaussian density functions over a line segment.

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

Fig. 4. Ortho view recovery: (a) the original ground image where the red line is the horizon line and the blue line is shifted 50 pixels below the red line so that the ortho-rectified view will not be too large. The blue line corresponds to the top line in the converted view (c); (b) is the same image with superimposed ground line segments; (c) is the ortho-rectified view; (d) is the corresponding linear features transformed to aerial view with field of view shown by dashed lines. The field of view (FOV) is 100 degrees which can be computed according to the focal length. ©Google



(a) (b)

Fig. 5. (a) G is the ground image, O is the ortho-view and C is the camera. The projection from G to O results in dramatic uncertainty (b) Let a and b are centers of normal distributions. If pixel location x and the slope angle φ of the line it lies on are known, then the two end points must be on the alternative directions starting from x.

Lemma 1. *Let* $\mathbf{a}, \mathbf{b}$ *be column vectors in* $\mathbb{R}^n$ *and* $\|\mathbf{a}\| = 1$*, then*

$$\int_{t_1}^{t_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\|\mathbf{a}t+\mathbf{b}\|^2}{2\sigma^2}} dt = e^{-\frac{\|\mathbf{b}\|^2-(\mathbf{a}^\top\mathbf{b})^2}{2\sigma^2}} \cdot \frac{1}{2}\left(erf\left(\frac{t_2+\mathbf{a}^\top\mathbf{b}}{\sqrt{2}\sigma}\right) - erf\left(\frac{t_1+\mathbf{a}^\top\mathbf{b}}{\sqrt{2}\sigma}\right)\right) \quad (2)$$

The proof of this lemma can be found in Appendix. Using this lemma, we derive our main theorem about uncertainty modeling. A visualization of the high level idea is shown in Fig. 5(b).

Theorem 1. *Let* ℓ *be a 2D line segment whose end points are random variables drawn from normal distributions* $N(\mathbf{a}, \sigma^2)$ *and* $N(\mathbf{b}, \sigma^2)$ *respectively. Then for any point* $\mathbf{x}$*, the probability that* $\mathbf{x}$ *lies on* ℓ *and* ℓ *has slope angle* φ *is*

$$p(\mathbf{x}, \varphi|\mathbf{a}, \mathbf{b}) = e^{-\frac{\|\mathbf{x}-\mathbf{a}\|^2-|\langle\mathbf{x}-\mathbf{a},\Delta_\varphi\rangle|^2+\|\mathbf{x}-\mathbf{b}\|^2-|\langle\mathbf{x}-\mathbf{b},\Delta_\varphi\rangle|^2}{2\sigma^2}} \cdot \frac{1}{2}\left(1 - erf\left(\frac{\langle\mathbf{x}-\mathbf{a},\Delta_\varphi\rangle}{\sqrt{2}\sigma}\right) erf\left(\frac{\langle\mathbf{x}-\mathbf{b},\Delta_\varphi\rangle}{\sqrt{2}\sigma}\right)\right) \quad (3)$$

where $\Delta_\varphi = (\cos\varphi, \sin\varphi)^\top$ *is the unit vector with respect to the slope angle* φ*.*

Proof. Let $p_n(\mathbf{x}; \boldsymbol{\mu}, \sigma^2)$ be the probability density function for normal distribution $N(\boldsymbol{\mu}, \sigma^2)$. The probability that $\mathbf{x}$ lies on the line segment equals the probability that random variables of the two ending points are $\mathbf{x} + t_a \Delta_\varphi$ and $\mathbf{x} + t_b \Delta_\varphi$ for some $t_a, t_b \in \mathbb{R}$ and $t_a \cdot t_b \leq 0$, therefore

$$p(\mathbf{x}, \varphi | \mathbf{a}, \mathbf{b}) = \int_{-\infty}^{0} p_n(\mathbf{x} + t\Delta_\varphi; \mathbf{a}, \sigma^2) dt \int_{0}^{\infty} p_n(\mathbf{x} + t\Delta_\varphi; \mathbf{b}, \sigma^2) dt + \int_{0}^{\infty} p_n(\mathbf{x} + t\Delta_\varphi; \mathbf{a}, \sigma^2) dt \int_{-\infty}^{0} p_n(\mathbf{x} + t\Delta_\varphi; \mathbf{b}, \sigma^2) dt \quad (4)$$

According to Lemma 1, Eq. 4 is equivalent to Eq. 3. □

Proposition 1. *Let ℓ' be a line segment transformed from line segment ℓ in 2D space by nonsingular 3×3 projection matrix $\mathbf{P}$. If the two ending points of ℓ are random variables drawn from normal distributions $N(\mathbf{a}, \sigma^2)$ and $N(\mathbf{b}, \sigma^2)$ respectively, then for any $\mathbf{x}$, the probability that $\mathbf{x}$ lies on ℓ' and ℓ' has slope angle φ is*

$$p_{proj}(\mathbf{x}, \varphi | \mathbf{P}, \mathbf{a}, \mathbf{b}) = p((x', \varphi') = proj(\mathbf{P}^{-1}, \mathbf{x}, \varphi) | \mathbf{a}, \mathbf{b}) \quad (5)$$

where $proj(\mathbf{Q}, \mathbf{x}, \varphi)$ is a function returns the corresponding coordinate and slope angle with respect to $\mathbf{x}$ and φ after projection transformation $\mathbf{Q}$.

The point coordinate transformed by $\mathbf{Q}$ can be obtained by homogeneous coordinate representation. For the slope angle, let $\mathbf{q}_i$ be the i-th row vector of projection matrix $\mathbf{Q}$, the transformed slope angle φ' at location $\mathbf{x} = (x, y)^\top$ is

$$\varphi' = \arctan \frac{f(\mathbf{q}_2, \mathbf{q}_3, x, y, \varphi)}{f(\mathbf{q}_1, \mathbf{q}_3, x, y, \varphi)} \quad (6)$$

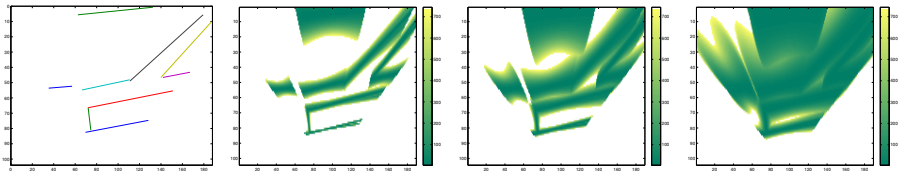
where

$$f(\mathbf{u}, \mathbf{v}, x, y, \varphi) = (u_2 v_1 - u_1 v_2)(x \sin\varphi - y \cos\varphi) + (u_1 v_3 - u_3 v_1) \cos\varphi + (u_2 v_3 - u_3 v_2) \sin\varphi \ . \quad (7)$$

According to the above, for each pixel location in the recovered view of a ground image, the probability that the pixel lies on a line segment given a slope angle can be computed in closed form. Fig. 6 shows an example probability distribution for line segments under uncertainty. It can be observed from the plot that more uncertainty is associated with line segments farther from the camera and is resulted from a larger σ value.

3.3 Geometric Matching under Uncertainty

Our approach to planar structure matching is motivated by chamfer matching. Chamfer matching efficiently measure the similarity between two sets of image



(a) (b) $\sigma = 0.5$ (c) $\sigma = 1$ (d) $\sigma = 2$

Fig. 6. Examples of uncertainty modeling: (a) the ortho-rectified line segments (b-d) the negation of probability log map for points on lines. The probability for each pixel location is obtained by summing up the probabilities for all discretized orientations. The camera is located in the image center and faces upward.

features by evaluating the sum of distances between each feature in one image and its nearest feature in the other image [3]. More formally,

$$D_c(\mathbf{A}, \mathbf{B}) = \sum_{\mathbf{a} \in \mathbf{A}} d(\mathbf{a}, \arg\min_{\mathbf{b} \in \mathbf{B}} d(\mathbf{a}, \mathbf{b})) \tag{8}$$

where $\mathbf{A}, \mathbf{B}$ are two sets of features, and $d(\cdot, \cdot)$ is the distance measure for a feature pair. Commonly, feature sets contain only the 2D coordinates of points, even if those points are sampled from lines that also have an associated orientation. Oriented chamfer matching (OCM) [22] makes use of point orientation by modifying the distance measure to include the sum of angle differences between each feature point and its closest point in the other image. Another way to incorporate orientation is directional chamfer matching (DCM) [15] which defines features to be, more generally, points in 3D space (x-y coordinates and orientation angle). This approach uses the same distance function as the original chamfer matching but has a modified feature distance measure. We follow the DCM method [15] to define our feature space. In our case, point orientation is set to the slope angle of the line it lies on.

Notations. All of the points in our formulation are in the 3D space. A point feature is defined as $\mathbf{u} = (\mathbf{u}_l, u_\phi)$ where $\mathbf{u}_l$ represents the 2D coordinates in real world and u_ϕ is the orientation associated with location $\mathbf{u}_l$. $\mathbf{G}_p$ is the set of points $\{\mathbf{g}\}$ in the ground image with uncertainty modeled by probability distribution $p(\cdot)$. $\mathbf{O}$ is the set of points in the ortho image. $\mathbf{L}_G$ is the set of annotated line segments in the ground image. A line segment is defined as $\boldsymbol{\ell} = (\mathbf{a}_\ell, \mathbf{b}_\ell)$ where $\mathbf{a}_\ell$ and $\mathbf{b}_\ell$ are the end points of $\boldsymbol{\ell}$. For any line segment $\boldsymbol{\ell}$ and an abitrary line segment $\hat{\boldsymbol{\ell}}$ in the feature space, $p(\hat{\boldsymbol{\ell}}|\boldsymbol{\ell})$ is the confidence of $\hat{\boldsymbol{\ell}}$ by observing $\boldsymbol{\ell}$.

Distance Metric. The feature distance for $\mathbf{u}, \mathbf{v}$ is defined as

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_g = \|\mathbf{u}_l - \mathbf{v}_l\|_2 + |u_\phi - v_\phi|_a \tag{9}$$

where $\|\mathbf{u}_l - \mathbf{v}_l\|_2$ is the Euclidean distance between 2D coordinates in meters and $|u_\phi - v_\phi|_a = \lambda \min(|u_\phi - v_\phi|, \pi - |u_\phi - v_\phi|)$ is the smallest difference between two

angles in radians. The parameter λ relates the unit of angle to the unit of world distance. We choose $\lambda = 1$ so that π angle difference is equivalent to around 3.14 meters in the real world. For this feature space definition, the chamfer distance in Eq. 8 can be efficiently computed by pre-computing the distance transform for the reference image (refer to [7,15] for more details) and convolving the query image with the reference distance transform.

Formulation. The distance function for matching ground image $\mathbf{G}_p$ to ortho image $\mathbf{O}$ is formulated as

$$D(\mathbf{G}_p, \mathbf{O}) = D_m(\mathbf{G}_p, \mathbf{O}) + D_\times(\mathbf{G}_p, \mathbf{O}) \tag{10}$$

where D_m is the probablistic chamfer matching distance and $D_\times$ is a term penalizing line segment crossings. The probablistic chamfer matching distance is defined as

$$D_m(\mathbf{G}_p, \mathbf{O}) = \frac{1}{|\mathbf{L}_G|} \sum_{\boldsymbol{\ell} \in \mathbf{L}_G} \int p(\hat{\boldsymbol{\ell}}|\boldsymbol{\ell}) \int p(\mathbf{g}|\hat{\boldsymbol{\ell}}) \left(\min_{\mathbf{o} \in \mathbf{O}} \|\mathbf{g} - \mathbf{o}\|_g \right) d\mathbf{g} d\hat{\boldsymbol{\ell}} \ . \tag{11}$$

The marginal distribution $\int p(\hat{\boldsymbol{\ell}}|\boldsymbol{\ell}) p(\mathbf{g}|\hat{\boldsymbol{\ell}}) d\hat{\boldsymbol{\ell}} = p(\mathbf{g}|\boldsymbol{\ell})$ is the probability that point $\mathbf{g}_l$ lies on line segment $\boldsymbol{\ell}$ with slope angle g_ϕ. Eq. 11 is equivalent to

$$D_m(\mathbf{G}_p, \mathbf{O}) = \frac{1}{|\mathbf{L}_G|} \sum_{\boldsymbol{\ell} \in \mathbf{L}_G} \int p(\mathbf{g}|\boldsymbol{\ell}) \left(\min_{\mathbf{o} \in \mathbf{O}} \|\mathbf{g} - \mathbf{o}\|_g \right) d\mathbf{g} \tag{12}$$

whose discrete representation is

$$D_m(\mathbf{G}_p, \mathbf{O}) = \sum_{\mathbf{g}} p'(\mathbf{g}|\mathbf{L}_G) \left(\min_{\mathbf{o} \in \mathbf{O}} \|\mathbf{g} - \mathbf{o}\|_g \right) \tag{13}$$

where $p'(\mathbf{g}|\mathbf{L}_G) = \frac{1}{|\mathbf{L}_G|} \sum_{\boldsymbol{\ell} \in \mathbf{L}_G} \frac{p(\mathbf{g}|\boldsymbol{\ell})}{\sum_{\mathbf{g}} p(\mathbf{g}|\boldsymbol{\ell})}$ is the probability of points lying on the structure and each line segment equally contributes to the distance value. In fact, Eq. 12 is equivalent to the original chamfer matching (Eq. 8) if no uncertainty is present.

Intersections between ortho line segments and ground line segments indicate low matching quality. Therefore, we add an additional term into our formulation to penalize camera poses that result in too many line segment intersections. The cross penalty for line segments is defined as

$$D_\times(\mathbf{G}_p, \mathbf{O}) = \frac{\sum_{\boldsymbol{\ell} \in \mathbf{L}_G} \int p(\hat{\boldsymbol{\ell}}|\boldsymbol{\ell}) \sum_{\mathbf{o} \in \mathbf{O}} \int p(\mathbf{g}|\hat{\boldsymbol{\ell}}) |g_\phi - o_\phi|_a \delta(\mathbf{g}_l - \mathbf{o}_l) d\mathbf{g} d\hat{\boldsymbol{\ell}}}{\sum_{\boldsymbol{\ell} \in \mathbf{L}_G} \int p(\hat{\boldsymbol{\ell}}|\boldsymbol{\ell}) \sum_{\mathbf{o} \in \mathbf{O}} \int p(\mathbf{g}|\hat{\boldsymbol{\ell}}) \delta(\mathbf{g}_l - \mathbf{o}_l) d\mathbf{g} d\hat{\boldsymbol{\ell}}} \tag{14}$$

where $\delta(\cdot)$ is the delta function. This function is is a normalized summation of angle differences for all intersection locations, which are point-wise equally weighted. Because $\int p(\hat{\boldsymbol{\ell}}|\boldsymbol{\ell}) p(\mathbf{g}|\hat{\boldsymbol{\ell}}) d\hat{\boldsymbol{\ell}} = p(\mathbf{g}|\boldsymbol{\ell})$, the function is equivalent to

$$D_\times(\mathbf{G}_p, \mathbf{O}) = \frac{\sum_{\boldsymbol{\ell} \in \mathbf{L}_G} \int p(\mathbf{g}|\boldsymbol{\ell}) \sum_{\mathbf{o} \in \mathbf{O}} |g_\phi - o_\phi|_a \delta(\mathbf{g}_l - \mathbf{o}_l) d\mathbf{g}}{\sum_{\boldsymbol{\ell} \in \mathbf{L}_G} \int p(\mathbf{g}|\boldsymbol{\ell}) \sum_{\mathbf{o} \in \mathbf{O}} \delta(\mathbf{g}_l - \mathbf{o}_l) d\mathbf{g}} \tag{15}$$

whose equivalent discrete formulation is

$$D_{\times}(\mathbf{G}_p, \mathbf{O}) = \frac{\sum_{\mathbf{g}} p'(\mathbf{g}|\mathbf{L}_G) \sum_{\mathbf{o}\in\mathbf{O}} |g_\phi - o_\phi|_a \delta[\mathbf{g}_l - \mathbf{o}_l]}{\sum_{\mathbf{g}} p'(\mathbf{g}|\mathbf{L}_G) \sum_{\mathbf{o}\in\mathbf{O}} \delta[\mathbf{g}_l - \mathbf{o}_l]} \tag{16}$$

where $p'(\mathbf{g}|\mathbf{L}_G)$ is defined in Eq.3.3 and $\delta[\cdot]$ is the discrete delta function.

Hypothesis Generation. Given a ground image $\mathbf{G}_p$, the score for ortho image $\mathbf{O}_i$ corresponds to one of the candidate geolocations. is evaluated as the minumum possible distance, so the estimated fine camera pose within ortho image $\mathbf{O}_i$ is

$$\hat{\mathbf{x}}_i = \hat{\mathbf{x}}(\mathbf{O}_i, \mathbf{G}_p) = \arg\min_{\mathbf{x}_l, x_\phi} D(\mathbf{R}_{x_\phi}\mathbf{G}_p + \mathbf{x}_l, \mathbf{O}_i) \tag{17}$$

where $\mathbf{R}_\alpha$ is the rotation matrix corresponded to angle α.

3.4 Implementation Remarks

The two distance functions can be computed efficiently based on distance transforms in which the orientations are projected into 60 uniformly sampled angles and the location of each point is at the pixel level. Firstly, probability $p(\mathbf{g}|\boldsymbol{\ell})$ can be computed in closed form according to Proposition 1. So the distribution $p'(\mathbf{g}|\mathbf{L}_G)$ can be pre-computed for each ground image. Based on 3D distance transform [15], Eq. 13 can be computed with a single convolution operation. The computation of Eq. 16 involves delta functions, which is essentially equivalent to a binary indicator mask for an ortho image: $M_{\mathbf{O}}(\mathbf{x}) = 1$ means there exists a point $\mathbf{o} \in \mathbf{O}$ located at coordinate $\mathbf{x}$ and 0 means there is no feature at this position. Such indicator mask can be directly obtained. So we compute for every orientation φ and location $\mathbf{x}$ a distance transform $A_\varphi(\mathbf{x}) = \sum_{\mathbf{o}\in\mathbf{O}\wedge\mathbf{o}_l=\mathbf{x}} |\varphi - o_\phi|_a$. The denominator of Eq. 16 can be computed directly by convolution, while the numerator needs to be computed independently for each orientation. For a discretized orientation θ, a matrix is defined $W(\mathbf{g}) = p'(\mathbf{g}|\mathbf{L}_G) M_{\mathbf{O}}(\mathbf{g}_l)$ for all $\mathbf{g}$ such that $g_\phi = \theta$ and otherwise $W(\mathbf{g}) = 0$. Convolving matrix W with the distance transform A_θ will achieve partial summation of Eq. 16. Summing them up for all orientations gives the numerator in Eq. 16.

4 Experiment

4.1 Experimental Setup

Dataset. We build a data set from Google Maps with an area of around $1km \times 1km$. We randomly extract 35 ground images from Google Street View together with their ground truth locations. Each ground image is a 640×640 color image. Field of view information is retrieved. A total of 400 satellite images are extracted using a sliding window within this area. Each ortho photo is also a 640×640 color image. The scale of ortho images is 0.1 meters per pixel. We use 10 ground images for experiments on the uncertainty parameter σ and the remaining 25 ground images are used for testing. Example ground and satellite images are shown in Fig. 7. Geolocation in this dataset is challenging because most of the area share highly similar visual appearance.

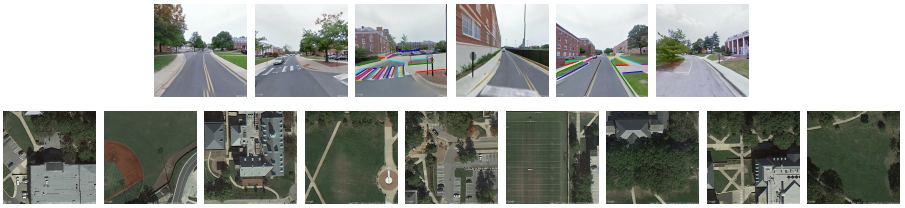


Fig. 7. Example ground images (upper) and ortho images (lower) from our dataset. The ground image can be taken anywhere within one of the satellite images. ©Google

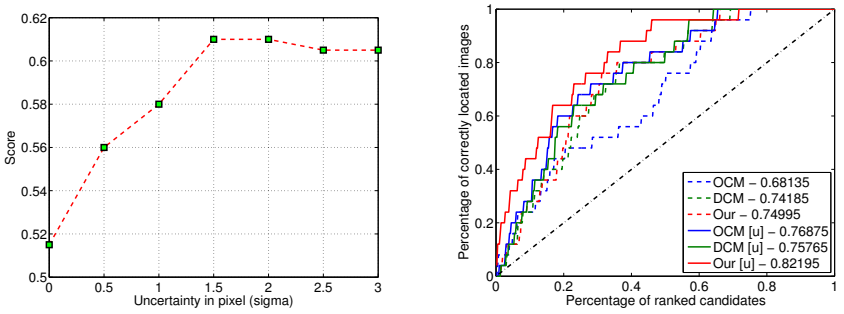
Evaluation Criterion. Three quantitative criteria are employed to evaluate the experiments. First, we follow previous work [14] by using curves on *percentage of ranked candidate* vs. *percentage of correctly localized images*. By ranking all the ortho images in descending order of their matching scores, *percentage of ranked candidates* is the percentage of top ranked images in all of the ortho images and *percentage of correctly located images* is the percentage of all the queries whose ground truth locations are among the corresponding top ranked candidate images. Second, we obtain a overall score by counting the area under this curve (AUC). A higher overall score generally means more robustness in the algorithm. Third, we look into the *percentage of correctly localized images* among 1%, 2%, 5% and 10% top ranked locations.

Parameter Selection. Intuitively, σ represents the pixelwise variance of the line segment end points, so it should not be more than several pixels. We randomly pick 10 ground images and 20 ortho images including all ground-truth locations to compose training set for tuning σ. The geolocation performance over a set of σ values ranged from 0 to 3 with a step 0.5 are evaluated and shown in Fig. 8(a) where $\sigma = 0$ means no uncertainty model is used. The peak is reached when the σ is between 1.5 and 2. Therefore, we fix $\sigma = 2$ in all of the following experiment.

4.2 Results

Our geometric matching approach returns distance values densely cover every pixel and each of the 12 sampled orientations in each ortho image. The minimum distance is picked as the score of an ortho image. Therefore, our approach not only produces ranking among hundreds of ortho images but also shows possible camera locations and orientations.

We compare our approach with two existing matching methods i.e. oriented chamfer matching [22] and directional chamfer matching [15]. To study the effectiveness of our uncertainty models, we also evaluate these methods with uncertainty model embedded. DCM is equivalent to the first term D_m in our formulation. OCM is to find the nearest feature in the other image and compute the sum of pixel-wise distance and the angle differences to the same pixel. We apply our uncertainty model into their formulation in a similar way as the probablistic chamfer matching distance does. Thus, in total we have six approaches



(a) (b)

Fig. 8. (a) Geolocation AUC score under different uncertainty variances σ where $\sigma = 0$ represents the approach without uncertainty modeling. (b) Performance curve for six approaches: the ortho images are ranked in ascending order. The x-axis is the number of selected top ranked ortho images and the y-axis is the total number of ground image queries whose true locations are among these selected ortho images. The overall AUC scores are shown in the legend where "[u]" means "with uncertainty modeling". The black dash-dot line indicates chance performance.

Table 1. Comparison among oriented chamfer matching [22], directional chamfer matching [15] and our approach. The uncertainty model is evaluated for each method. For each evaluation criterion, the highest score is highlighted in red and the second one highlighted in blue. Our uncertainty based formulation is top among all these methods. Both of the three methods can be improved by our uncertainty model. OCM boosts its performance when incorporated with our probablistic representation.

	w/o uncertainty			**w/ uncertainty**		
Method	OCM	DCM	our	OCM	DCM	our
Top 1%	0.08	0.00	0.00	0.04	0.00	0.12
Top 2%	0.08	0.04	0.08	0.04	0.04	0.20
Top 5%	0.16	0.12	0.12	0.20	0.12	0.32
Top 10%	0.24	0.24	0.28	0.28	0.28	0.44
Score(AUC)	0.6814	0.7419	0.7500	0.7688	0.7577	0.8219

in our comparison. Their performance curves are shown in Fig. 8(b). Over 90% of the ground queries can be correctly located when half of the ortho images are rejected. Numerical results are in Table 1. While our approach significantly outperforms at any percentage of retrieved images, our performance improvement is particularly large for top ranked images.

Four successfully localized queries are shown in Fig. 9. For these ground images, the ground truth locations are included in the top 5 ranked candidate ortho images out of 400. From this visualization, few labeling errors can be noticed from miss-alignment between ortho images and rectified line segments. Among these top responses, most false alarms are building roofs. A common property is that they have relatively denser line features. Another issue is the line detection

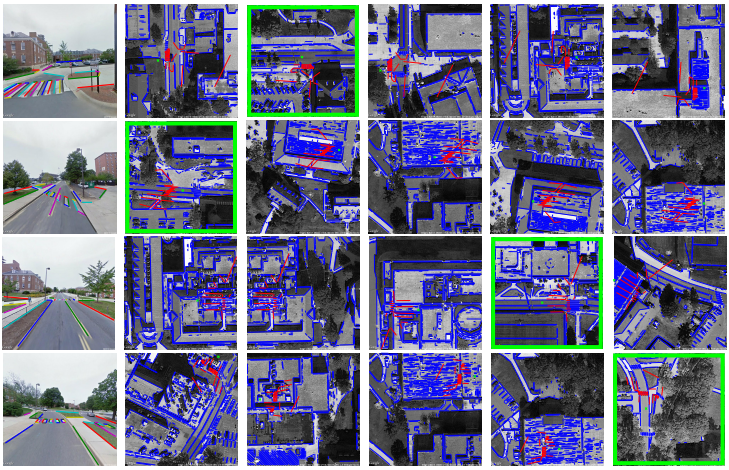


Fig. 9. Four queries successfully geolocated within top five candidates are shown. The leftmost column is the ground image with annotated line segments. For each query, top five scoring ortho images are shown in ascending order of their rank. Ground-truths are highlighted by green bounding boxes. For each ortho image, blue lines are automatically detected and red lines are parsed from ortho-rectified ground images. Green cross indicates the most probable camera location within that ortho image.

in ortho images does not handle shadows well. Most linear structures in these shadow areas are not detected.

5 Conclusion

We investigated the single image geolocation problem by matching human annotated line segments in the ground image to automatically detected lines in the ortho images. An uncertainty model is devised for line segments under projective transformations. Using this uncertainty model, ortho-rectified ground images are matched to candidate ortho images by distance transform based methods. The experiment has shown the effectiveness of our approach in geographic areas with similar local appearances.

Acknowledgement. This material is based upon work supported by United States Air Force under Contract FA8650-12-C-7213 and by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

References

1. Baatz, G., Saurer, O., Köser, K., Pollefeys, M.: Large scale visual geo-localization of images in mountainous terrain. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 517–530. Springer, Heidelberg (2012), http://dx.doi.org/10.1007/978-3-642-33709-3_37
2. Bansal, M., Sawhney, H.S., Cheng, H., Daniilidis, K.: Geo-localization of street views with aerial image databases. In: ACM Int'l Conf. Multimedia (MM), pp. 1125–1128 (2011), http://doi.acm.org/10.1145/2072298.2071954
3. Barrow, H.G., Tenenbaum, J.M., Bolles, R.C., Wolf, H.C.: Parametric correspondence and chamfer matching: Two new techniques for image matching. In: Proceedings of the 5th International Joint Conference on Artificial Intelligence, IJCAI 1977, vol. 2, pp. 659–663. Morgan Kaufmann Publishers Inc., San Francisco (1977), http://dl.acm.org/citation.cfm?id=1622943.1622971
4. Bay, H., Ferrari, V., Van Gool, L.: Wide-baseline stereo matching with line segments. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 329–336 (June 2005)
5. Chen, D., Baatz, G., Koser, K., Tsai, S., Vedantham, R., Pylvanainen, T., Roimela, K., Chen, X., Bach, J., Pollefeys, M., Girod, B., Grzeszczuk, R.: City-scale landmark identification on mobile devices. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 737–744 (November 2011)
6. Elgammal, A., Shet, V., Yacoob, Y., Davis, L.: Exemplar-based tracking and recognition of arm gestures. In: Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis, ISPA 2003, vol. 2, pp. 656–661 (September 2003)
7. Felzenszwalb, P.F., Huttenlocher, D.P.: Distance transforms of sampled functions. Theory of Computing 8(19), 415–428 (2012), http://www.theoryofcomputing.org/articles/v008a019
8. von Gioi, R., Jakubowicz, J., Morel, J.M., Randall, G.: Lsd: A fast line segment detector with a false detection control. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 32(4), 722–732 (2010)
9. Hays, J., Efros, A.A.: im2gps: estimating geographic information from a single image. In: IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2008)
10. Irschara, A., Zach, C., Frahm, J.M., Bischof, H.: From structure-from-motion point clouds to fast location recognition. In: IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 2599–2606 (June 2009)
11. Kim, H., Lee, S.: Wide-baseline image matching based on coplanar line intersections. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1157–1164 (October 2010)
12. Li, Y., Snavely, N., Huttenlocher, D., Fua, P.: Worldwide pose estimation using 3D point clouds. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 15–29. Springer, Heidelberg (2012), http://dx.doi.org/10.1007/978-3-642-33718-5_2
13. Li, Y., Snavely, N., Huttenlocher, D.P.: Location recognition using prioritized feature matching. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 791–804. Springer, Heidelberg (2010), http://dl.acm.org/citation.cfm?id=1888028.1888088
14. Lin, T.Y., Belongie, S., Hays, J.: Cross-view image geolocalization. In: IEEE Conf. Computer Vision and Pattern Recognition (CVPR). Portland, OR (June 2013)

15. Liu, M.Y., Tuzel, O., Veeraraghavan, A., Chellappa, R.: Fast directional chamfer matching. In: IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2010)
16. Matei, B., Vander Valk, N., Zhu, Z., Cheng, H., Sawhney, H.: Image to lidar matching for geotagging in urban environments. In: IEEE Workshop on Applications of Computer Vision (WACV), pp. 413–420 (January 2013)
17. Olson, C.: A probabilistic formulation for hausdorff matching. In: Proceedings of 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 150–156 (June 1998)
18. Sankaranarayanan, A.C., Chellappa, R.: Optimal multi-view fusion of object locations. In: Proceedings of the 2008 IEEE Workshop on Motion and Video Computing, WMVC 2008, pp. 1–8. IEEE Computer Society, Washington, DC (2008), http://dx.doi.org/10.1109/WMVC.2008.4544048
19. Sattler, T., Leibe, B., Kobbelt, L.: Fast image-based localization using direct 2d-to-3d matching. In: IEEE Int'l Conf. Computer Vision (ICCV), pp. 667–674 (November 2011)
20. Schindler, G., Brown, M., Szeliski, R.: City-scale location recognition. In: IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 1–7 (2007), http://www.cs.bath.ac.uk/brown/location/location.html
21. Schmid, C., Zisserman, A.: Automatic line matching across views. In: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR 1997), pp. 666–. IEEE Computer Society, Washington, DC (1997), http://dl.acm.org/citation.cfm?id=794189.794450
22. Shotton, J., Blake, A., Cipolla, R.: Multiscale categorical object recognition using contour fragments. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 30(7), 1270–1281 (2008)
23. Wang, L., Neumann, U., You, S.: Wide-baseline image matching using line signatures. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1311–1318 (September 2009)
24. Zamir, A., Shah, M.: Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) (2014)
25. Zheng, Y.T., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T.S., Neven, H.: Tour the world: Building a web-scale landmark recognition engine. In: IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 1085–1092 (2009)