Epipolar Geometry Estimation for Wide-baseline Omnidirectional Street View Images

Tomokazu Sato[1], Tomas Pajdla[2], and Naokazu Yokoya[1]

[1] Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[2] Center of Machine Perception, Czech Technical University in Prague

Abstract

This paper presents a new robust method of epipolar-geometry estimation for omnidirectional images in wide-baseline setting, e.g. with Google Street View images. The main idea is to learn new statistical geometric constraints that are derived from the feature descriptors into the model verification process of RANSAC. We show that these constraints provide more reliable matches, which can be used to retrieve correct epipolar geometry in very difficult situations. Robustness of epipolar-geometry estimation is quantitatively evaluated for omnidirectional image pairs with variable baseline. The performance of the proposed method is demonstrated using the complete pipeline of structure-from-motion with real dataset of Google Street View images.

1. Introduction

3D information of large urban environments is useful for mobile applications, e.g. image based localization and augmented reality on smartphones. In order to estimate 3D structure of large environments from images, SfM and dense reconstruction was intensively investigated in recent years with impressive results [4, 10, 11, 12, 16, 22, 23, 25]. Epipolar geometry estimation is one of the critical steps of many SfM pipelines.

The bundler [4, 22] is one of well-known SfM engines that have been employed in many recent works for unordered photo collections. The robustness of the bundler and its capability for wide-baseline conditions have been validated in numerous demonstrations and the publically available open source software. However, omnidirectional street view images, that are currently available through the internet-based map services, e.g. Google Street View [1] and Microsoft Streetside [2], pose another challenge. Automatic computation of camera motion and 3D scene reconstruction of large part of cities is beyond capabilities of current SfM pipelines. As shown in Figure 1, extreme change in scene appearance, dominant occlusions, repetitive and confusing texture patterns, changing of lighting condition, moving object, are all existing in the street view imagery. Only two good matches were found by SIFT [14] in the image pair shown in Figure 1(top), and completely wrong epipolar geometry was thus computed by RANSAC [9] (Figure 1(middle)). Although there are many works for street view images [7, 16, 20, 21, 23, 25], there are no works that evaluate their performance with real Google Street View images whose average distance reaches over 10 meters.

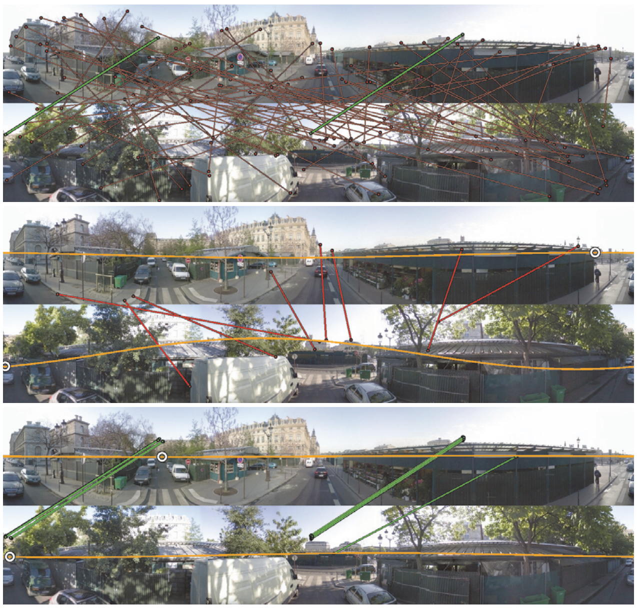


Figure 1. Example of epipolar geometry estimation for Google Street View imagery. Top: Tentative matches (green: good matches, red: bad matches). Middle: Epipolar lines (orange), epipoles (white) and a few selected incorrect matches after RANSAC. All the selected matches are wrong and are violating our constraints. Bottom: Matches consistent with our constraints after harvesting additional matches.

For successful estimation of epipolar geometry from street view imagery in a wide-baseline situation, the following approaches are employed in this work; (1) Avoiding mismatches by finding and using more valid constraints; (2) Harvesting additional matches by guided matching; (3)

Balancing model complexity and flexibility in RANSAC to avoid invalid models. Figure 1(bottom) demonstrates the successful result by our new pipeline. The contributions of this paper are as follows. (1) Introduction of new constraints using scale and orientation of feature descriptors into model verification process of RANSAC. (2) Verification of effectiveness of guided matching in epipolar geometry estimation and SfM. (3) Suggestion of an effective way to avoid wrong estimates that uses a constraint concerning the direction of the epipole.

2. Related works

There are many works on large scale SfM, and the epipolar geometry estimation is an essential part of them. These works on SfM can be categorized into two groups; (1) methods designed for unordered image collections and (2) for sequential images. The photo tourism [22], which is the origin of the bundler, is designed for unordered image collections and it constructs the graph in which images are connected topologically in order to harvest an estimated 3D structure using bundle adjustment. For making the graph, SIFT features on each pair of images are matched by ANN library and epipolar geometry for each pair of images is computed by RANSAC. If the number of matches after RANSAC is too small, the pair is removed and one of the other candidates is tested. Several extensions of this framework were proposed mainly to increase the scalability of the SfM pipeline [10, 11, 12]. One advantage of using large unordered data set is that they can find many good easy pairs to harvest the graph rather than to stick to difficult pairs of images.

Many works on omnidirectional street view images have already been reported [7, 16, 20, 21, 23, 25]. Unlike methods for unordered data set, these methods often rely on image sequences and cannot find alternative pairs in many cases. Thus, the robustness of the epipolar geometry estimation for each pair of images is very important in these works. Tardif et al. [23] demonstrated the performance of SLAM based SfM on street view images. It was pointed out that separate estimation of rotation and translation is better for the short-baseline conditions. The guided matching was employed in their work by assuming that epipolar geometry is similar among neighboring images. Unlike their method, our method for guided matching does not assume any motion model. Torii et al. [25] has proposed the SfM pipeline for street view images with loop closing. The work [25] demonstrated the 3D structure of the long sequence with low accumulation errors by detecting closed loop of a route using the global similarity of images. Micusik et al. [16] recovered 3D models with piecewise planar structure constraints. Our new constraints are mainly focused on the model verification process of RANSAC, and these constraints can be easily incorporated to these conventional methods [16, 23, 25].

One constraint that has often been employed mainly in the robotics field is the assumption of planar motion [7, 20, 21]. It works well in short-baseline conditions by reducing degree of freedom (DOF) to 2 in epipolar geometry of calibrated camera. It is also effective for reducing computational cost. This paper will show that this planar assumption does not work in wide-baseline conditions where banking of the camera cannot be ignored.

Other possibility for improving robustness of epipolar geometry estimation is by using many alternative feature detectors and descriptors (operators). Although there are many feature operators [5, 14, 15, 17, 26], selection of the best feature operator is out of the scope of this research. It should be noted that our new constraints is applicable with any feature operator that computes the scale and orientation of feature points.

Using line segments [6, 27] and introducing constraint on local topology of feature positions are proposed in [24]. Unlike these methods, we focus on already existing scale and orientation information that have not been used in conventional works except for making tentative matches.

3. Epipolar geometry estimation using constraints

The proposed method can be easily incorporated to the common RANSAC-based framework employed in conventional works for omnidirectional SfM [16, 25]. In this section, our pipeline for epipolar geometry estimation is introduced and details of new constraints are then described.

3.1. Pipeline of epipolar geometry estimation using RANSAC

A common RANSAC-based pipeline for estimating epipolar geometry is as follows.

1) Make tentative matches $\mathbf{V}$ for image pairs.

2) Sample t matches from $\mathbf{V}$ randomly.

3) Estimate epipolar geometry using selected matches.

4) Count the number of inliers u.

The estimated epipolar geometry that maximizes u is selected by iterating steps 2 to 4. Feature points whose distances from corresponding epipolar-lines are under given threshold are counted as inliers in the step 4. If essential matrix is decomposed to rotation matrix and translation vector up to scale, the sign of depth for each feature point can be used to avoid mismatches [25]. We follow this RANSAC-based pipeline but effective ways to avoid invalid estimation are newly incorporated into the pipeline.

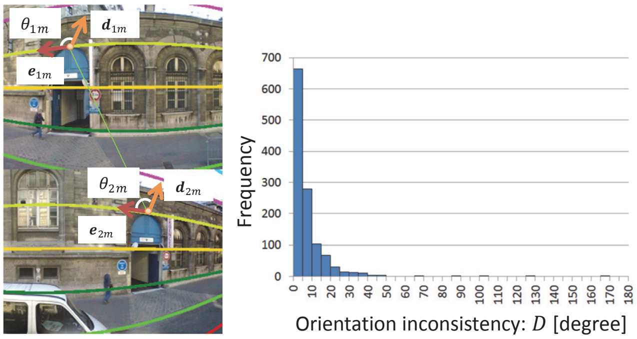


Figure 2. Left: Feature orientation in panoramic image. Right: Histogram of orientation inconsistency D for correct matches in Google Street View images.

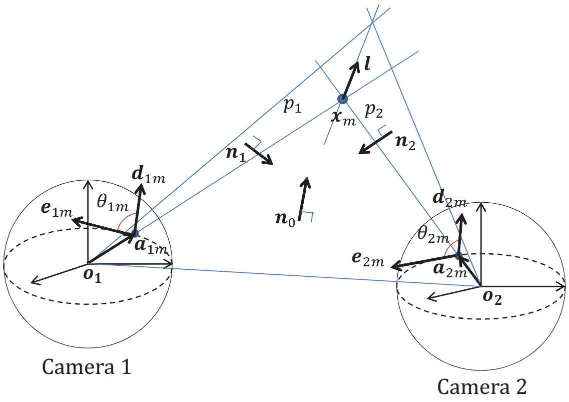


Figure 3. Orientation constraint for matched features.

3.2. Three new constraints and guided matching in RANSAC

In order to avoid mismatches that cannot be removed by conventional methods, e.g. shown in Figure 1(middle), three constraints and guided matching are employed in RANSAC. Two additional constraints, that are evaluated based on the relationship between a putative epipolar solution and parameters of feature descriptor, are employed in the step 4. Another constraint concerning the direction of the epipole can immediately reject invalid models after the step 3 by assuming that camera is moving roughly on a plane. We have confirmed that harvesting additional matches in the step 1 before starting RANSAC iteration is also effective to increase robustness of the estimation. It should be noted that we are not suggesting to use guided matching in RANSAC iteration because it drastically slows down the iterating processes.

3.3. Statistical orientation constraint on matched features

In order to achieve rotation independent matches for feature points, common feature descriptors, e.g. SIFT [14] and SURF [5], compute the orientation parameter for each feature point. By aligning rotation of texture patterns (feature vectors) using these orientation parameters, rotation independent matching for feature points has been achieved.

We use this orientation parameter to decrease the number of mismatches. Here, we consider two matched features with angles θ_{1m}, θ_{2m} that are defined between the orientation vectors $\mathbf{d}_{1m}$, $\mathbf{d}_{2m}$ from the feature descriptor and corresponding epipolar curves passing through the features, as shown in Figure 2(left). We have observed that a large number (88%) of correctly matched features have very similar angles $|\theta_{1m} - \theta_{2m}| \leq 15°$ as shown in Figure 2(right). This statistical evidence suggests to form a new statistical orientation constraint on matched features.

We define the orientation inconsistency D_m for the tentative match m in camera 1 and 2 as follows.

$$D_m = \min(|\theta_{1m} - \theta_{2m}|, 2\pi - |\theta_{1m} - \theta_{2m}|), \tag{1}$$

$$\theta_{im} = angle(\mathbf{e}_{im}, \mathbf{d}_{im}),\ \mathbf{e}_{im} = \mathbf{n}_0 \times \mathbf{a}_{im}, \tag{2}$$

where $i = (1, 2)$ is the camera index, $\mathbf{n}_0$ is the normal vector of the epipolar plane for the match m, $\mathbf{a}_{im}$ is the unit vector from the camera projection center $\mathbf{o}_i$ to the feature point in the image plane of the spherical projection camera i as shown in Figure 3. The function $angle$ returns the angle between two vectors in the range $[-\pi, \pi]$, the operator $\times$ means the cross product of vectors. All the vectors are defined in the world coordinate system that coincides with the local coordinate system of the camera 1. The rays $\mathbf{a}_{1m}$ and $\mathbf{a}_{2m}$ from cameras are assumed to cross at 3D point $\mathbf{x}_m$.

Ideally when a normal vector of a local shape on $\mathbf{x}_m$ is almost parallel to $(\mathbf{a}_{1m} + \mathbf{a}_{2m})$, and which holds true in many situations, the inconsistency D_m becomes very small if estimated epipolar geometry and orientation parameters given from feature descriptor are consistent. In this paper, the match m, whose orientation inconsistency D_m, is larger than threshold T_{ori} is removed from the inliers in the step 4 of the RANSAC iteration.

We next examine the orientation inconsistency D_m. First, as shown in Figure 3, we define the plane p_i through three points $\mathbf{o}_i$, $\mathbf{x}_m$, and $\mathbf{o}_i + \mathbf{a}_{im} + \mathbf{d}_{im}$. Next, we assume that pixels on the arc around $\mathbf{a}_{im}$, which are on the intersection of the plane p_i and the sphere of the camera i, are projected from the line where p_1 and p_2 intersect. We analyze the ideal condition for the inconsistency D_m by checking how this intersecting line moves under the condition θ_{1m}=θ_{2m}.

In order to understand how can the intersection line move under the ideal condition, the line direction vector $\mathbf{l}$ is considered as shown in Figure 3. It follows

$$\mathbf{n}_0 = \mathbf{a}_{1m} \times \mathbf{a}_{2m}, \mathbf{n}_1 = \mathbf{a}_{1m} \times \mathbf{l}, \mathbf{n}_2 = \mathbf{l} \times \mathbf{a}_{2m}. \tag{3}$$

$$|\mathbf{a}_{1m}| = |\mathbf{a}_{2m}| = |\mathbf{l}| = 1, \tag{4}$$

$$\cos\theta_{1m} = \frac{\mathbf{n}_0 \cdot \mathbf{n}_1}{|\mathbf{n}_0||\mathbf{n}_1|}, \cos\theta_{2m} = -\frac{\mathbf{n}_0 \cdot \mathbf{n}_2}{|\mathbf{n}_0||\mathbf{n}_2|}, \tag{5}$$

where $\mathbf{n}_i$ is the normal vector of the plane p_i. From Eq. (5) and $\theta_{1m} = \theta_{2m}$, there follows

$$\left(\frac{\mathbf{n}_0 \cdot \mathbf{n}_1}{|\mathbf{n}_0||\mathbf{n}_1|}\right)^2 - \left(\frac{\mathbf{n}_0 \cdot \mathbf{n}_2}{|\mathbf{n}_0||\mathbf{n}_2|}\right)^2 = 0. \tag{6}$$

By substituting Eqs. (3) and (4) to Eq. (6), the following equation is obtained.

$$\frac{(\mathbf{l} \cdot (\mathbf{a}_{1m} + \mathbf{a}_{2m}))(\mathbf{l} \cdot (\mathbf{a}_{1m} - \mathbf{a}_{2m})) \sin^2\theta_{1m}}{((\mathbf{l} \cdot \mathbf{a}_{1m})^2 - 1)((\mathbf{l} \cdot \mathbf{a}_{2m})^2 - 1)} = 0. \tag{7}$$

Here, $\theta_{1m} = \theta_{2m} = 0$ or π are special cases when no unique intersecting line of the planes p_1 and p_2 exists. Except for the special cases, the condition in Eq. (7) can be simplified as

$$(\mathbf{l} \cdot (\mathbf{a}_{1m} + \mathbf{a}_{2m}))(\mathbf{l} \cdot (\mathbf{a}_{1m} - \mathbf{a}_{2m})) = 0. \tag{8}$$

This equation means that vector $\mathbf{l}$ must be perpendicular to the vector $(\mathbf{a}_{1m} + \mathbf{a}_{2m})$ or $(\mathbf{a}_{1m} - \mathbf{a}_{2m})$. In practice, the latter condition will never be satisfied because rays $\mathbf{a}_{1m}$ and $\mathbf{a}_{2m}$ must see opposite sides of the object in this case. It should be noted that the normal vector $(\mathbf{a}_{1m} + \mathbf{a}_{2m})$ of the local plane in the former case is exactly equal to the normal vector of the ellipse on the epipolar plane through the point $\mathbf{x}_m$ whose foci are $\mathbf{o}_1$ and $\mathbf{o}_2$. It means that if textures are drawn on the ellipsoid (rotated version of the ellipse) whose size depends on $\mathbf{x}_m$, D_m becomes 0 with true epipolar geometry and true orientation parameters.

In order to absorb the difference from the ideal condition in real street view scenes, a right threshold T_{ori} on D should be found. For this purpose, an experiment is carried out using 10 pairs of Google Street View images from Paris. Here, the distribution of the orientation consistency is checked by using only correctly matched pairs of SIFT points and epipolar geometry estimated from these matches. Figure 2(right) shows the histogram of the orientation inconsistency D. Here, 99% of matches satisfy $D < 40°$. From this test, we have determined the threshold T_{ori} as $40°$.

3.4. Statistical constraint on the scale of matched features

The scale parameter given from feature detectors is also containing information which can be used to avoid mismatches. This parameter has been used to determine the region size of texture patterns from which feature vector is extracted. It should be noted that if feature points of the match m are correctly corresponding with correct scale parameters (s_{1m}, s_{2m}) as shown in Figure 4(left), the image regions of these features should be capturing the same object of the same size r on the 3-D space.

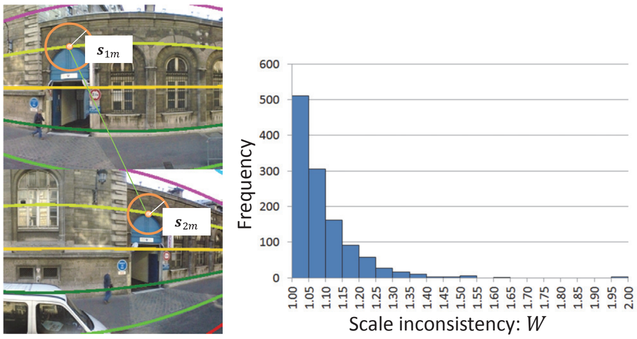


Figure 4. Left: Scale of features in panoramic image. Right: Histogram of scale inconsistency W for correct matches in Google Street View images.

From the nature of perspective imagery, the size s_{im} on an image can be defined as follows:

$$s_{im} = \frac{r}{fc_{im}}, \tag{9}$$

where f is the focal length of the perspective camera and c_{im} is the depth of the object w.r.t. the camera i. In the epipolar geometry estimation, although c_{im} cannot be determined due to unknown scale, we can compute the ratio of c_{1m} and c_{2m}. From Eq. (9),

$$\frac{s_{1m}}{s_{2m}} = \frac{c_{2m}}{c_{1m}}. \tag{10}$$

This equation is trivial in perspective imagery but can be used to remove outliers.

Concretely, we define the scale inconsistency W_m as follows.

$$W_m = max(\frac{s_{1m}}{s_{2m}}\frac{c_{1m}}{c_{2m}}, \frac{s_{2m}}{s_{1m}}\frac{c_{2m}}{c_{1m}}) \tag{11}$$

Again, the absolute scale of depths is not necessary to compute this consistency. W_m has positive values and becomes 1 if all the parameters are consistent. In the proposed method, the match m, whose inconsistency W_m is over T_{sc}, is removed from the count of the inliers in the RANSAC iteration.

As with the orientation constraint, the threshold T_{sc} is determined experimentally. Figure 4(right) shows the histogram of the scale inconsistency W given in the same condition shown in Section 3.3. From the result that 99% of matches satisfy $W < 1.4$, we have determined the threshold T_{sc} as 1.4.

3.5. Constraint of the direction of epipole

In many applications using omnidirectional images, camera is mounted in a fixed position on a vehicle whose height from the ground is almost constant. In this case, the model of camera motion can be simplified to facilitate stable estimation using e.g. 2DOF camera motion (1 parameter for the horizontal rotation, 1 parameter for the horizontal direction of the epipole). However, we have found that the assumption of this planar camera motion is often violated for Google Street View imagery due to banking of the vehicle and changing the camera settings (which was apparently mounted at different heights). On the other hand, when there are only a few valid pairs of tentative matches, full 5-DOF estimator often gives us wrong models as shown in Figure 1(middle) where the direction of the epipole is far from the horizon in the panoramic image.

In order to balance the model complexity and the flexibility, we employ the full 5-DOF estimator with the constraint on the direction of the epipole. Concretely, models that do not satisfy the following condition are immediately removed from candidates after estimating the model with 5-DOF

$$\max(\frac{|\mathbf{z}_1 \cdot \mathbf{t}_1|}{|\mathbf{z}_1||\mathbf{t}_1|}, \frac{|\mathbf{z}_2 \cdot \mathbf{t}_2|}{|\mathbf{z}_2||\mathbf{t}_2|}) < cos(T_{hor}), \tag{12}$$

where $\mathbf{z}_i$ is a pre-defined vertical axis of the camera i, $\mathbf{t}_i$ is the direction of the epipole from the camera i. We have observed that $T_{hor} = 87^\circ$ (3° margin for planar camera motion) was adequate for Google Street View imagery.

3.6. Guided matching using discriminative matches

In order to increase the chance to select many good matches in a sample of RANSAC, one straightforward way is to increase the number of good matches. We use matching guided by discriminative matches to harvest promising tentative matches before starting sampling processes of RANSAC. We use the ratio of the first to the second closest feature descriptor as the measure of the discriminativeness [14].

The following steps 1 to 5 are iterated I times.

1) Select the most discriminative pair m from the yet unselected group of tentative matches.

2) If feature positions for the pair m are near to one of already selected pairs, return to the step 1.

3) For each camera i, select group of feature points $\mathbf{F}_i$ which satisfy $\mathbf{a}_{im} \cdot \mathbf{a}_{iq} < cos(T_{gui})$ for any q.

4) Make tentative matches $\mathbf{T}$ for the selected groups $\mathbf{F}_1$ and $\mathbf{F}_2$.

5) If the number of matches in $\mathbf{T}$ is smaller than N_{min}, discard these matches.

6) Otherwise add top N_{max} discriminative matches from $\mathbf{T}$ to the tentative matches $\mathbf{V}$ for RANSAC.

4. Experiments

In this section, we report two experiments to show the robustness of the proposed method. First, estimated epipole directions are evaluated for image pairs of variable baseline lengths using experimental dataset of street view images. Camera positions acquired by using complete SfM pipeline for panoramic images downloaded from Google Street View are then evaluated using GPS positions associated with images.

4.1. Epipolar geometry estimation for variable length of baseline

The proposed method is quantitatively evaluated using a 5,000 image Google Street View Pittsburg Experimental data set [3]. The length of the route for this image sequence is 4.8 kilometers and the average baseline length for successive image pairs is 0.96 meters. Since the original panoramic images have large distortions at the top and bottom, we cropped the original image by 115 pixels from the top and by 205 pixels from the bottom to obtain $1,664 \times 512$ pixel images.

On this data, we have compared the estimated results of the following six methods w.r.t. variable length of baseline ranging from 8 to 120 frame separations, which is roughly corresponding to 7 to 110 meters.

1) **Baseline**: 5-point estimator with RANSAC (e.g. in [16, 25]).

2) **Planar**: 2-DOF estimator under assumption of planar motion (e.g. in [20]).

3) **Ori.&Sca.**: **Baseline** with orientation and scale constraints.

4) **Epipole**: **Baseline** with constraint for direction of epipole.

5) **Guided**: **Baseline** with guided matching.

6) **Proposed**: **Baseline** with all extensions.

As the 5-point estimator, we have used Nister's solver [19]. For **Planar**, instead of random sampling, the model that maximizes the number of inliers was found by exhaustive search for 2 parameters by enumerating the model on all pairs. We have implemented all methods and the code was the same for all methods except for selections of solvers and extensions. For the methods without guided matching, top 200 discriminative pairs of SIFT features were used as tentative matches. For **Guided** and **Proposed**, top 100 discriminative pairs and maximum 100 guided matches with the parameters ($I = 5, T_{gui} = 20^\circ, N_{min} = 7, N_{min} = 20$) were used as tentative matches. All these matches were found by FLANN [18]. For speeding-up the RANSAC process,

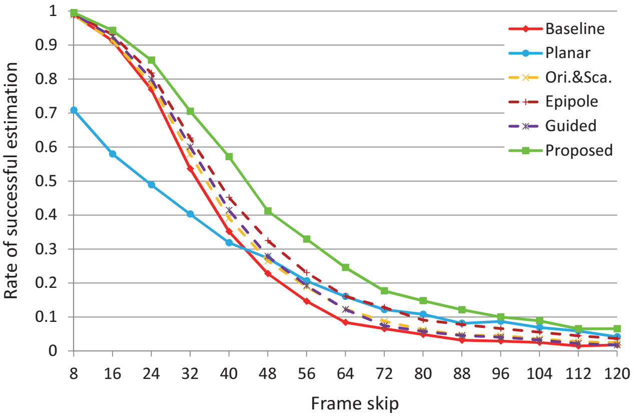


Figure 5. Rate of successful estimation for varying skips of image sequence.

ordered sampling suggested in [8] was employed with features ordered by the discriminativeness of SIFT [14].

In order to evaluate estimated epipolar geometry quantitatively, we have compared estimated directions of epipoles with the ground truth. In this experiment, the ground truth is computed from the camera positions and orientations provided by the authors of the state-of-the-art omnidirectional SfM pipeline with loop closing and bundle adjustment [25]. As shown in paper [25], accumulation of errors is sufficiently small for the purpose of judging the quality of our epipolar geometry estimation. By using this ground truth, we have judged the estimation as successful when angle between the direction of estimated epipole and the ground truth was smaller than 5 degrees.

Figure 5 shows the success rate of estimation for variable length of the baseline measured in the number of skipped frames, i.e. roughly in meters. The rate is computed using around $1,250$ pairs for each baseline length. For all baselines, **Proposed** is the best among all the compared methods. We see that each extension improved the result. In the case of the shortest baseline (8 frame skip, average 7.6 meters baseline), all methods except for **Planar** achieve near 100% success but the proposed method is still the best (**Baseline**: 98.9%, **Proposed**: 99.5%). We can see that the planar motion assumption does not work for around 30% of image pairs in this case. Our constraint on the direction of epipole is working effectively even for these image pairs. We have confirmed from the other experiment that the success rate of the estimation is stable for wide range of threshold with maximum at $T_{ori} = 35°$ and $T_{sc} = 1.4$ and plateau from 20° to 115° in T_{ori} and 1.2 to 2.5 in T_{sc} with drop less than 1%. Figure 6 shows an example of compared results for **Baseline** and **Proposed** where result is improved by new constraints.

Table 1. Comparison of the average of the maximum baseline lengths and the corresponding computational costs. Time is measured using a PC (CPU: Core i7-920, 2.7GHz) with single thread computing.

	Average of maximum baselines [m]	Average computational time for a pair [sec]		
		FLANN	RANSAC	total
Baseline	35.7	3.30	1.91	5.21
Planar	42.1		1.31	4.61
Ori.&Sca.	39.7		2.03	5.32
Epipole	44.5		1.78	5.08
Guided	38.9	3.93	1.96	5.89
Proposed	50.6		1.82	5.75

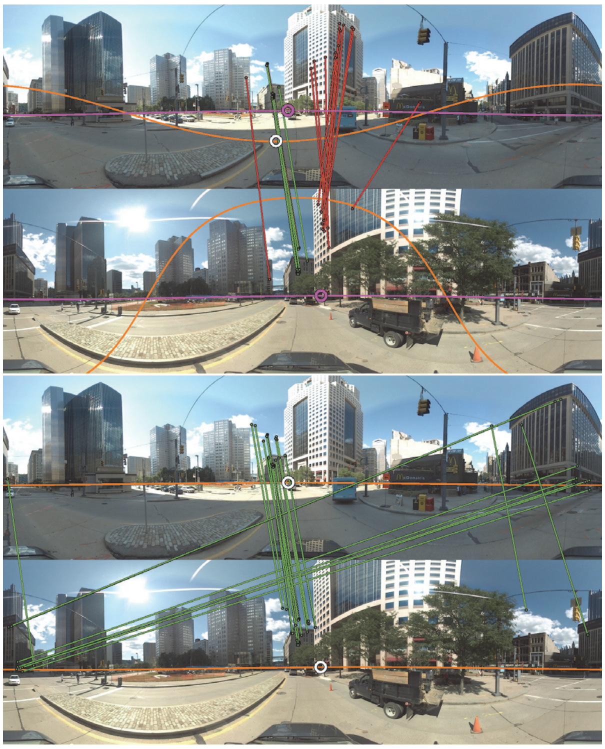
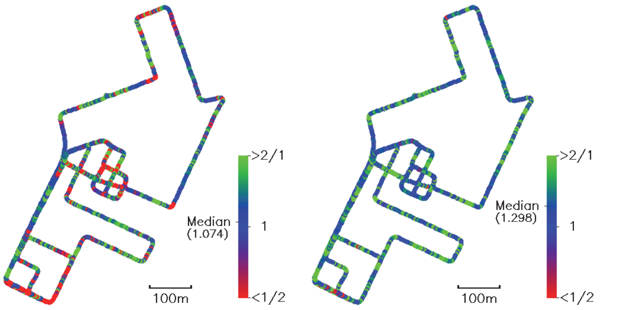


Figure 6. Example of matched features and epipolar geometry estimated from them. Top: Ground truth of epipolar line, epipole (purple) and result by **Baseline**. Red lines are matches that violate orientation and scale constraints with estimated geometry. Bottom: Result by **Proposed** with consistent matches.

Next, for each camera position of the input sequence, the maximum baseline length for which the estimation was succeeded was evaluated. Table 1 shows the average of the maximum baseline lengths and the computational costs for each method. The average of the maximum baseline lengths for **Proposed** reaches 50.6 meters and this is 42% longer than that of **Baseline**. Figure 7 shows the ratio of maximum baseline lengths for each camera position for (a) **Planar / Baseline** and (b) **Proposed / Baseline**. From (a), we can see that the planar motion assumption does not work around corners of the route. This is due to the banking of the camera mounted on a car. From (b), we can see that the proposed method extends the maximum baseline length in many parts of the route.



(a) **Planar / Baseline** ratios, (b) **Propsed / Baseline** ratios.

Figure 7. Ratio of maximum baseline lengths for which estimation were successful.

Let we discuss the cost of computation. Although the constraint on the direction of epipole reduces the cost by avoiding testing clearly invalid models, additional costs are incurred by checking the consistency of orientation and scale. As the result of using the all three constraints, the cost of the RANSAC has totally been reduced. It should be noted that the cost of FLANN in **Proposed** and **Guided** is larger than that of the other methods. This is caused by the initialization cost of FLANN in our current implementation. In the current implementation, after selecting feature groups for finding matches, KD-trees in FLANN are reconstructed every time for different group of guided match.

4.2. Structure from motion for real Google Street View images

In this experiment, performance of our constraints is evaluated in a SfM pipeline using real Google Street View images. We downloaded 189 omnidirectional images of Pittsburg from the site of Google Street View [1] along the route of 2.15 kilometers. Figure 8 shows GPS positions associated to downloaded images. From the downloaded images, we cropped the original image with the same condition as ones in Section 4.1. Average baseline length of successive viewpoints for this route is 11.5 meters.

First, epipolar geometry for all the successive pairs of 189 images was computed by **Baseline** and **Proposed** under the same conditions as in the previous experiment. Feature matches in image pairs were then chained into triplets of images in order to determine the scale of epipolar geometry for successive pairs. Here, a RANSAC-based method for chaining that finds the scale maximizing the number of inliers for chains was employed to determine the scales [25]. Sparse bundle adjustment library [13] was used as the bundle adjustment engine after the chaining process.

It should be noted that in so extreme wide-baseline situation (for triplet of images, baseline length becomes 23 meters), chaining often failed because no common match existed among detected features for triplets in some places. For successful chaining, here, we have therefore employed additional guided matching using estimated epipolar lines. Simply, groups of feature points existing around corresponding epipolar lines in pairs of images are matched by SIFT before starting the chaining process.

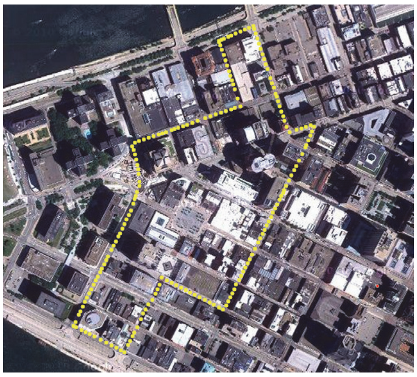


Figure 8. GPS positions of downloaded street view images.

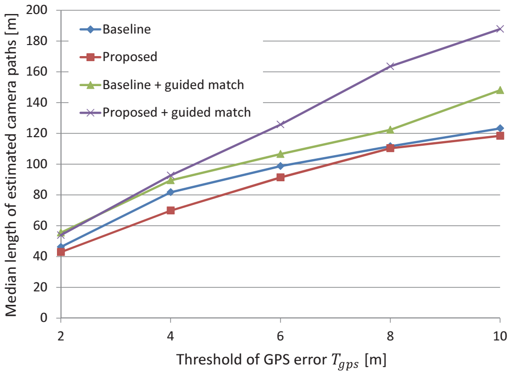


Figure 9. Comparison of median lengths of successfully and accurately reconstructed chains of camera trajectories.

In order to evaluate estimated camera positions after bundle adjustment, GPS positions associated with images are used as for reference. Unfortunately, all these GPS positions are aligned to the street grid on the map in Google Street View, and that introduces additional position errors especially around corners. Thus, estimated results by SfM pipeline are compared with GPS positions allowing variable size of errors T_{gps} for GPS positions. Concretely, from every position of the street view images, sequential SfM pipeline is started and continued until the estimation is judged as failure. We judge the result as failure if the maximum of distances between camera positions and corresponding GPS positions reaches over T_{gps} after fitting estimated camera positions to corresponding GPS positions by minimizing the sum of squared distances of them.

Median lengths of the estimated camera paths for variable T_{gps} are shown in Figure 9. We can confirm that **Proposed** with additional guided matching gives the longest paths among all the methods. Especially, when T_{gps} is larger than 4 meters, we observe clear advantage of our method. We observed that GPS error is sometimes reaching up to 8 to 10 meters at street junctions. On the other hand, notice that for all T_{gps}, **Baseline** without guided matching is better than **Proposed** without guided matching. This is due to fewer chance of successful chaining of **Proposed** because **Proposed** has fewer matches than **Baseline** due to new constraints on feature matches. This fact is supporting our claim that **Proposed** can compute more accurate epipolar geometry than that by **Baseline** because we can see that additional guided matching in this stage is working better for **Proposed** than for **Baseline**.

5. Conclusion

Three new statistical geometric constrains were proposed for epipolar geometry estimation of omnidirectional image pairs with wide-baseline configuration. The proposed constraints are simple and can be incorporated into most of RANSAC-based pipelines for omnidirectional epipolar geometry estimation. The performance of these constraints and the guided matching in epipolar geometry estimation is demonstrated using variable baseline between street view images. We have confirmed advantage of the proposed method in the SfM stage where the length of camera path is enhanced by using proposed constraints and additional guided matching. We have enlarged the median length of accurately (up to 5 meter error) reconstructed camera trajectory from 80 to 110m. For complete recovery of 3D structures for large environment, merging of separately estimated structures suggested in [28] will be applicable. Development of a strategy for determining good routes for loop closing and harvesting structures is a next challenge.

Acknowledgments This research was supported by ME CR MSM 6840770038, CTU SGS10/190/OHK3/2T/13 and FP7-SPACE-241523 PROVISCOUT grants, and MEXT (Japan) KAKENHI 23240024, 23700208.

References

[1] http://maps.google.com/help/maps/streetview/.

[2] http://www.bing.com/maps/.

[3] Google street view pittsburgh experimental data set. in Google Cityblock Research Dataset V1.7., 2008.

[4] S. Agarwal, N. Snavely, I. Simon, S. Seitz, and R. Szeliski. Building rome in a day. *in ICCV*, pages 72–79, 2009.

[5] H. Bay, A. Ess, T. Tuytelaars, and L. Gool. Speeded up robust features (surf). *CVIU*, 110(3):346–359, 2008.

[6] H. Bay, V. Ferrari, and L. Gool. Wide-baseline stereo matching with line segments. *in CVPR*, pages I:329–336, 2005.

[7] S. Choi, J. Lee, J. Joung, M. Ryoo, and W. Yu. Numerical solutions to relative pose problem under planar motion. *in Int. Conf. on Ubiquitous Robots and Ambient Intelligence*, 2010.

[8] O. Chum and J. Matas. Matching with prosac - progressive sample consensus. *in CVPR*, pages I:220–226, 2004.

[9] M. Fischler and R. 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.

[10] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Towards internet-scale multi-view stereo. *in CVPR*, pages 1434–1441, 2010.

[11] M. Havlena, A. Torii, J. Knopp, and T. Pajdla. Randomized structure from motion based on atomic 3d models from camera triplets. *in CVPR*, pages 2874–2881, 2009.

[12] Y. Jeong, D. Nister, D. Steedly, R.Szeliski, and I. Kweon. Pushing the envelope of modern methods for bundle adjustment. *in CVPR*, pages 1474–1481, 2010.

[13] M. Lourakis and A. Argyros. Sba: A software package for generic sparse bundle adjustment. *ACM Trans. Math. Software*, 36(1):1–30, 2009.

[14] D. Lowe. Distinctive image features from scale invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[15] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[16] B. Micusik and J. Kosecka. Piecewise planar city 3d modeling from street view panoramic sequences. *in CVPR*, pages 2906–2912, 2009.

[17] K. Mikolajajczyk and C. Shmid. Scale & affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004.

[18] M. Muja and D. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. *in VISAPP*, pages 221–340, 2009.

[19] D. Nister. An efficient solution to the five-point relative pose problem. *PAMI*, 26(6):756–770, 2004.

[20] D. Scaramuzza, F. Fraundorfer, and R. Siegwart. Real-time monocular visual odometry for on-road vehicles with 1-point ransac. *in Int. Conf. on Robotics and Automation*, pages 1015–1026, 2009.

[21] D. Scaramuzza and R. Siegwart. Appearance-guided monocular omnidirectional visual odometry for outdoor ground vehicles. *IEEE Trans. on Robotics*, 24(5):1015–1026, 2008.

[22] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3d. *ACM Trans. Graphics*, pages 835–864, 2006.

[23] J. Tardif, Y. Pavlidis, and K. Daniilidis. Monocular visual odometry in urban environment using an omnidirectional camera. *in Int. Conf. on Robots and Systems*, pages 2531–2538, 2008.

[24] D. Tell and S. Carlsson. Combining appearance and topology for wide baseline matching. *in ECCV*, pages I:68–81, 2002.

[25] A. Torii, M. Havlena, and T. Pajdla. From google street view to 3d city models. *in OMNIVIS*, 2009.

[26] T. Tuytelaars and L. Gool. Matching widely separated views based on affine invariant regions. *IJCV*, 59(1):61–85, 2004.

[27] L. Wang, U. Neumann, and S. You. Wide-baseline image matching using line signatures. *in ICCV*, pages 1311–1318, 2009.

[28] C. Wu, B. Clipp, X. Li, J. Frahm, and M. Pollefeys. 3d model matching with viewpoint invariant patches (vips). *in CVPR*, pages 1–8, 2008.