Fast estimation of relative poses for 6-DOF image localization

Yafei Song, Xiaowu Chen*, Xiaogang Wang, Yu Zhang and Jia Li
State Key Laboratory of Virtual Reality Technology and Systems
School of Computer Science and Engineering, Beihang University, Beijing, China

***Abstract*—The 6-DOF (Degrees Of Freedom) image localization, which aims to calculate the spatial position and rotation of a camera, is a challenging task for location-based services. In existing approaches, this problem is often tackled by finding the matches between 2D image points and 3D structure points so as to derive the location information using direct linear transformation (DLT). However, these approaches may fail to localize images when the 3D structure points are not available, especially for massive data. To address this problem, this paper presents a novel data-driven approach for 6-DOF image localization. In this approach, we propose to localize an image according to the position and rotation information of multiple similar images retrieved from a large reference dataset. From the reference images, a fast relative pose estimation algorithm is proposed to derive a set of candidate poses for the input image. Since each candidate pose actually encodes the relative rotation and direction of the input image with respect to a specific reference image, we can thus fuse all these candidate poses so that the 6-DOF location of the input image can be efficiently derived through least-square optimization. Experimental results show that our approach performs comparable with GPS devices in image localization. In addition, the proposed relative pose estimation algorithm is much faster than existing work.**

***Index Terms*—image localization; relative pose estimation; one-sided radial fundamental matrix estimation**

I. Introduction

6-DOF (Degrees Of Freedom) image localization is an important problem for many multimedia applications such as location-based services. The problem is to calculate 6-DOF location information of the camera when it captures an image. The 6-DOF location contains 3-DOF spatial position and 3-DOF rotation of the image. As soon as obtain this location information, it has no difficulty to implement many applications, such as augmented reality [1], [2], autonomous navigation [3]–[5], organize the on-line sharing photos in 3D space [6] and other location-based services [7]. However, this problem is still a challenge especially how to utilize massive and increasing geo-information tagged data.

Previous methods usually perform 6-DOF image localization task based on 3D point cloud model and formulate the process as 2D-to-3D registration problem [3], [8]–[12]. They first find the matches between 2D image points and 3D structure points. The point matches can be utilized to localize input image via applying direct linear transformation (DLT) algorithm [13], which can be embedded into a RANdom SAmple Consensus (RANSAC) [14] process for robustness. So far, these methods can perform the task well through exploiting the information of 3D point cloud model. However, these model-based methods are not flexible with increasing geo-information tagged data. And it is very time-consuming to construct the 3D point cloud model which is usually obtained applying structure from motion (SfM) algorithm from the plentiful data [15]. Moreover, since SfM does not always work well especially for massive data, previous methods will face difficulties when 3D point cloud model is not available.

Fig. 1. Our basic idea is to localize an image via transferring the location information of references to it. To this end, we first estimate the relative pose between input image (with red lines) and each of its references (with green lines), and then fuse all these candidate poses to obtain the final 6-DOF location. As a result we can perform the 6-DOF localization task without 3D point cloud (the point cloud in the figure is only for visualization).

Besides 3D point cloud model based methods, some researchers [16], [17] first recognize the landmark in input image, then transfer the position of landmark to input image. Others [18]–[21] usually retrieve or select the nearest neighbours of input image from the database through measuring the similarity. Then the final position of input image can be obtained by fusing the position of retrieved neighbours.

*Xiaowu Chen is the corresponding author. E-mai: chen@buaa.edu.cn

These methods can be scalable benefiting from scalable image retrieval methods [22]. However, they lack the ability to get the accurate 6-DOF location of input image, which limits their practical applicability.

In this paper, we are devoted to propose a data-driven method to localize an image, which can be flexible for massive and increasing data. Inspired by recent image label transfer work [23], [24] on scene parsing. Our basic idea is to transfer the 6-DOF location information of reference images to input image, which is intuitively illustrated in Figure 1. This idea requires a premise that there are abundant images with 6-DOF location information. With the development of differential GPS and gyroscope, it is feasible to construct such image datasets. Actually, this type of image datasets have appeared, such as Google Street View. However, these data lack 3D point cloud which makes it difficult to localize an image for existing 3D point cloud model based methods. Therefore, we propose a novel data-driven method to localize an image. Given an input image, we first retrieve its nearest neighbours using a bag-of-visual-words based algorithm [22] from a large reference dataset. From the nearest neighbours, we propose a fast relative pose estimation algorithm to obtain a set of candidate poses for input image. Since each candidate pose actually encodes the relative rotation and direction of input image with respect to a specific neighbour image, we can thus fuse all these candidate poses so that the final 6-DOF location can be efficiently derived through least-square optimization.

Our main contributions include: (1) we first perform the 6-DOF image localization task without 3D point cloud to the best of our knowledge; (2) we propose a fast algorithm to estimate the relative pose between a calibrated and an uncalibrated image.

II. Related work

There are mainly three classes of methods related to our work, including image localization methods based on 3D point cloud model, landmark recognition and localization, relative pose estimation.

Image localization based on 3D point cloud model. These methods commonly formulate the localization task as a 2D-to-3D registration problem. They first find a set of point matches between 2D feature points in input image and 3D structure points in the point cloud model which is usually reconstructed by SfM systems. Then the accurate 6-DOF location of input image can be estimated using DLT algorithm [13]. The key of these methods is to find abundant and robust matches fast. Donoser and Schmalstieg [9] formulate the matching process as a discriminative classification problem, while most researchers use a distance metric to measure the similarity between 2D and 3D points features. Lim *et al.* [3] use inexpensive binary feature descriptors instead of scale-invariant features which enables the method to run in real-time. Li *et al.* [10] utilize co-occurrence prior and bidirectional matching to get the camera pose fast in worldwide scale. Sattler *et al.* [11] evaluate the performance of direct 2D-to-3D matching method and propose a direct matching framework based on visual vocabulary quantization and a prioritized correspondence search. Irschara *et al.* [12] apply image retrieval techniques to reduce searching space via finding the nearest views of input image. The nearest views are generated using the 3D point cloud. State-of-the-art methods can perform 6-DOF image localization task well via exploiting 3D point cloud, however, they lack the ability to perform the localization task when 3D point cloud is not available.

Landmark recognition and localization. These methods usually first recognize the landmark in input image or retrieve the nearest neighbours of input image, then transfer the landmark or neighbours position information to input image. Li *et al.* [17] formulate the task as a classification problem. They can recognize 500 categories of landmarks using SVM on a large dataset which contains 30 million of images. Chen *et al.* [19] publish a city scale street view dataset and incorporate user's position priors to improve the recall rates on mobile devices. Zamir *et al.* [20] address the localization problem through querying the input image's scale-invariant feature transform (SIFT) descriptors in the indexed tree of reference images. Then an associated voting scheme is utilized to determine input image's final position. Zhang and Kosecka [21] also localize an input image through estimating relative pose to each of its references, however, they can only obtain the position of input image. Though these methods are always good at handling large scalar data, they are incapable of obtaining the accurate 6-DOF location of input image.

Relative pose estimation. In order to transfer 6-DOF location of neighbour images to input image, it is necessary to estimate the relative pose between input image and each of its neighbours. Given a pair of images, different algorithms have been proposed to estimate their relative pose corresponding to different configurations. For a pair of calibrated images, the relative pose can be estimated using 5-point algorithm [25], which has been well-studied. For a pair of images composed of one calibrated and one uncalibrated image, we must first estimate the intrinsic parameters, then the problem is transformed to the configuration of two calibrated images. Bujnak *et al.* [26] use Gröebner basis method to address calibrated-uncalibrated setting and apply it to 3D reconstruction, who assume that the uncalibrated image only has focal intrinsic parameter. Brito *et al.* [27] suppose the uncalibrated image has focal and one radial distortion intrinsic parameters, which is same with our configuration. However they solve a higher-order polynomial system to obtain the relative pose and intrinsic parameters, which is time-consuming and slower than our SVD based algorithm.

III. Overview

In this paper, we aim at tackling the 6-DOF image localization problem using a data-driven method given massive data. Inspired by label transfer work [23], we localize an input image via transferring 6-DOF location information of reference images to input image. The work-flow of our method is illustrated in Figure 2. First of all, we retrieve several nearest neighbour images of input image which have duplicated

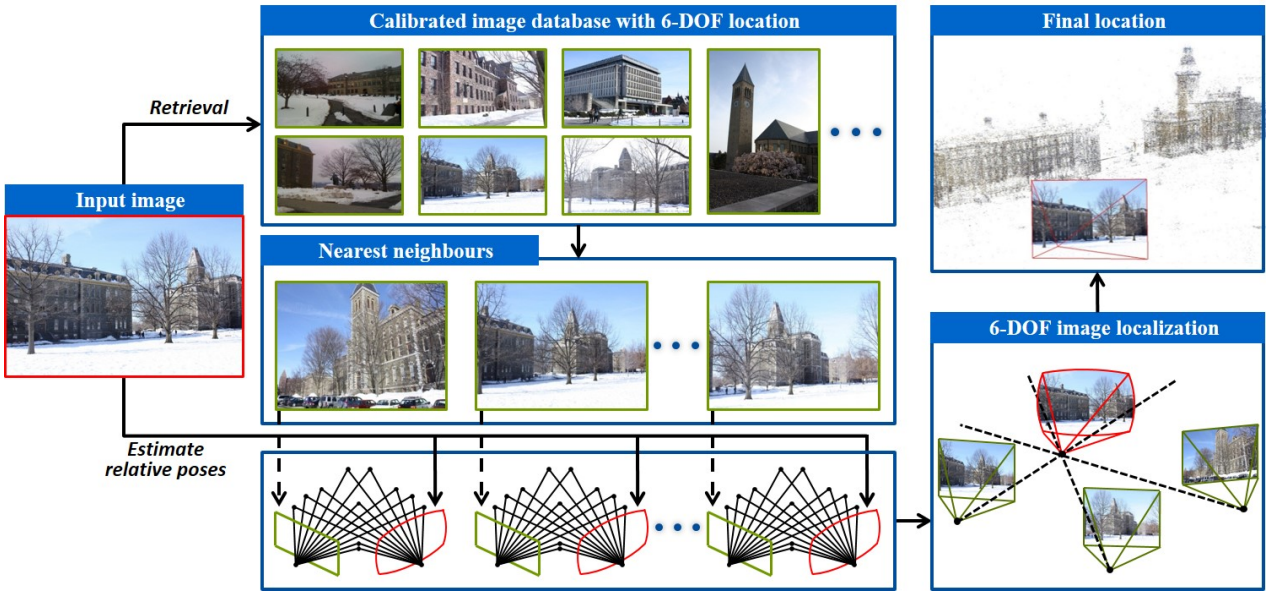


Fig. 2. The framework of our method. Given an input image, first, we retrieve its nearest neighbours from the database using a bag-of-visual-words method. Each neighbour has 6-DOF location information. Then, we estimate relative pose between the input image and each of its neighbours, and achieve several pose candidates. At last we can get the final optimal 6-DOF location via integrating all pose candidates.

content with input image. This step can be performed using a bag-of-visual-words based image retrieval algorithm [22].

One-sided radial (OSR) fundamental matrix estimation. After the nearest neighbours are retrieved, we should estimate relative pose of input image to each of its neighbours. This is actually the classical relative pose estimation problem in multiple view geometry. We assume that all the images in dataset have been calibrated during building the dataset and that input image is not calibrated. Moreover according to the influence of camera intrinsic parameters, we assume that the input image has two intrinsic parameters including focal length and one radial distortion parameter. Under this configuration, one-sided radial fundamental matrix can capture the epipolar geometry constrain which is presented by Brito *et al.* [27]. We propose a fast algorithm based on SVD to estimate the OSR fundamental matrix in Section 4.2.

Intrinsic parameters extraction and relative pose estimation. Based on the OSR fundamental matrix, we can extract the intrinsic parameters, i.e., focal length and one radial distortion, of input image according to the properties of fundamental matrix. Therefore the problem is transferred to calibrated relative pose estimation problem. Then we can obtain the pose of input image relative to each of its neighbours same as in 5-point algorithm [25], which has been well studied. After that we get a set of pose candidates of input image.

Final location determination. Based on a pose candidate we can get a rotation candidate of input image. The final rotation of input image can be obtained via averaging all rotation candidates. The last problem now is that each pose candidate only encodes relative direction of input image to each of its neighbours but has no scale information. In other words, we can only know that the position of input image is in a line according to a pose candidate. Fortunately two or more lines can determine a point, which is exactly the basic idea of triangulation theory. As the position of input image should be in all lines, we can use least square to calculate the final optimal position of input image.

IV. Image Localization Based on Relative Poses

In this section we present the details of each step in our method. In Section 4.1, we retrieve nearest neighbours of input image from the database using a content based image retrieval algorithm. Then, we estimate the fundamental matrix between input image and each neighbour in Section 4.2. Based on the fundamental matrix we can extract the intrinsic parameters of input image and obtain a set of pose candidates in Section 4.3. At last, we can figure out final optimal 6-DOF location of input image via integrating all the pose candidates in Section 4.4.

A. Nearest neighbours retrieval

As our basic idea is to estimate the 6-DOF location of an image from its nearest neighbours, which must be retrieved from the database at first. The similarity between images is usually measured by computing the distance between global features or local features along with bag-of-visual-words. In our case, the input image and its neighbours should have some duplicated areas, which makes relative pose estimation feasible. Thus it is appropriate to use bag-of-visual-words based algorithms.

Benefiting from the progress of content based image retrieval, there are many algorithms satisfying our requirement. LIRe [22] provides a library of basic and advanced functions for visual information retrieval, which is used in our method

without loss of generality. In consideration of the time efficiency, we use SURF [28] as the local feature which is faster than classical SIFT [29]. We apply k-means to learn visual words from the local features extracted from the images in dataset. For each image the visual words can be used to establish a histogram which can be used to measure the similarity between images. Some researchers [23] usually re-rank the retrieval results aiming at more robust. However, our method can automatically pick out robust neighbours from all the retrieved images subsequently via fundamental matrix estimation. Thus, we have no re-rank step.

B. OSR fundamental matrix estimation

For a pair of images capturing the same scene, the fundamental matrix $\mathbf{F}$ encapsulates the geometry relationship between the two images, which only depends on the cameras' intrinsic parameters and relative pose [13]. In our situation, the input image is uncalibrated and its intrinsic parameters contain focal length and one radial distortion parameter, while its neighbours are calibrated. Therefore before estimating their relative pose, we should estimate the fundamental matrix $\mathbf{F}$ in this section and extract intrinsic parameters of input image from $\mathbf{F}$ in Section 4.3. Then we can estimate the relative pose same as in classical 5-point algorithm [25].

Given a pair of images composed of one calibrated and one uncalibrated image, we first find the initial point matches between the two images through *ratio test* algorithm using Euclidean distance of SIFT features [29]. For each feature point $\mathcal{Q}$ in input image, the algorithm first find the closest point $\mathcal{P}$ and the second closest point $\mathcal{P}_2$ in its neighbour image. Then the nearest point $\mathcal{P}$ is taken as a point that matches $\mathcal{Q}$ if the ratio of Euclidean distance from the closest point to second closest point is less than a threshold

$$\frac{dist\left(\mathcal{P},\mathcal{Q}\right)}{dist\left(\mathcal{P}_2,\mathcal{Q}\right)} < T_h, \tag{1}$$

where T_h is a predefined threshold. Given a point match $\langle \mathcal{P}, \mathcal{Q} \rangle$, we denote their image coordinate as $\mathbf{p} = (x_p, y_p, 1)^T$ and $\mathbf{q} = (x_q, y_q, 1)^T$, respectively. As we suppose that input image has one radial distortion parameter λ, the undistorted image coordinate of point $\mathcal{Q}$ can be given as $\mathbf{q}_u \propto \left(x_q, y_q, 1+\lambda r^2\right)^T$, where r is the distance between the point and the distortion center (u, v) which can be computed as

$$r^2 = \left(x_q - u\right)^2 + \left(y_q - v\right)^2, \tag{2}$$

Furthermore we assume that the distortion center (u, v) is in image center.

Obviously a point match $< \mathcal{P}, \mathcal{Q} >$ satisfies the epipolar constraint

$$\mathbf{p}^T\mathbf{F}\mathbf{q}_u = 0, \tag{3}$$

which can be written as

$$\mathbf{p}^T\mathbf{F}\begin{pmatrix} x_q \\ y_q \\ 1+\lambda r^2 \end{pmatrix} = \mathbf{p}^T[\mathbf{f}_1\ \mathbf{f}_2\ \mathbf{f}_3\ \lambda\mathbf{f}_3]\begin{pmatrix} x_q \\ y_q \\ 1 \\ r^2 \end{pmatrix} = 0\ , \tag{4}$$

where $\mathbf{f}_i$ is the $i\text{-}th$ column of $\mathbf{F}$. We denote matrix $[\mathbf{f}_1\ \mathbf{f}_2\ \mathbf{f}_3\ \lambda\mathbf{f}_3]$ as $\mathbf{V}$ and call it the one-sided radial (OSR) fundamental matrix as in [27] in order to distinguish from the original matrix $\mathbf{F}$. Moreover it can be found in [27] that the detailed derivation of OSR fundamental matrix.

Similar to the 8-point algorithm to fundamental matrix estimation [13], we first estimate $\mathbf{V}$ using 11 pairs of point matches, then enforce the low-rank constrains on $\mathbf{V}$ using singular value decomposition (SVD). Given 11 point matches, we can obtain the linear equations

$$\mathbf{A}\mathbf{v} = \mathbf{0}, \tag{5}$$

according to the epipolar geometry constrain, where $\mathbf{A}$ is an 11×12 coefficient matrix and $\mathbf{v}$ is the vector version of $\mathbf{V}$ in row major order. Moreover, each row of $\mathbf{A}$ is $[x_p x_q, x_p y_q, x_p, x_p r^2, y_p x_q, y_p y_q, y_p, y_p r^2, x_q, y_q, 1, r^2]$ corresponding to a point match $\langle \mathcal{P}, \mathcal{Q} \rangle$. Then SVD algorithm can be applied to solve (5). We decompose $\mathbf{A}$ via SVD, and obtain the initial estimation of $\mathbf{v}$ through picking out the right-singular vector corresponding to the smallest singular value.

As $\mathbf{F}$ is a low-rank matrix and $\mathbf{V} = [\mathbf{f}_1\ \mathbf{f}_2\ \mathbf{f}_3\ \lambda\mathbf{f}_3]$, we should enforce these constraints on $\mathbf{V}$. As we know, SVD can be used to calculate the low-rank matrix which is closest to the original matrix measured by Frobenius norm. For a given matrix, we can get its closest matrix with rank of c via retaining the c largest singular value and setting the others as zero. Therefore we can apply SVD twice to estimate $\mathbf{F}$ and $\mathbf{V}$. First, we get the matrix with rank one which is closest to the last two columns of $\mathbf{V}$ via SVD, and get the λ by direct division on the two columns. Then, we figure out the matrix with rank two, which is closest to the first three columns of $\mathbf{V}$, as final estimated $\mathbf{F}$. Moreover all the above is embedded in an RANSAC process for robustness.

C. Intrinsic parameters extraction

As mentioned before, we suppose that the input image has two intrinsic parameters, focal length f and one radial distortion λ. We have obtained λ in Section IV-B during estimating fundamental matrix, and we will present the details to extract focal length f from $\mathbf{F}$ matrix in this section. Since the images in database are fully calibrated, their intrinsic parameter matrix can be regarded as an identity matrix. Then the essential matrix can be written as

$$\mathbf{E} = \mathbf{F}\mathbf{K}, \tag{6}$$

where $\mathbf{E}$ is the essential matrix, $\mathbf{F}$ is the fundamental matrix achieved in Section IV-B and $\mathbf{K}$ is the intrinsic parameter matrix of input image. As λ is known, input image can be assumed to be captured by ideal pin-hole camera. So $\mathbf{K}$ has the formulation as

$$\mathbf{K} = diag\left(f, f, 1\right). \tag{7}$$

An essential matrix has rank two and has two equal non-zero singular values [13]. That is to say, a real non-zero 3×3

matrix $\mathbf{E}$ is an essential matrix if and only if it satisfies the equation

$$2\mathbf{E}\mathbf{E}^T\mathbf{E} - \mathbf{tr}\left(\mathbf{E}\mathbf{E}^T\right)\mathbf{E} = \mathbf{0}. \tag{8}$$

We substitute (6) into (8) and obtain

$$2\mathbf{F}\mathbf{K}\mathbf{K}^T\mathbf{F}^T\mathbf{F}\mathbf{K} - \mathbf{tr}\left(\mathbf{F}\mathbf{K}\mathbf{K}^T\mathbf{F}^T\right)\mathbf{F}\mathbf{K} = \mathbf{0}. \tag{9}$$

As $f > 0$ and $\mathbf{K}$ is invertible, we right-multiply (9) with $\mathbf{K}^{-1}$, and obtain

$$2\mathbf{F}\mathbf{K}\mathbf{K}^T\mathbf{F}^T\mathbf{F} - \mathbf{tr}\left(\mathbf{F}\mathbf{K}\mathbf{K}^T\mathbf{F}^T\right)\mathbf{F} = \mathbf{0}. \tag{10}$$

For clarity, we set $\mathbf{\Omega} = \mathbf{K}\mathbf{K}^T = diag\left(f^2, f^2, 1\right)$, then (10) can be written as

$$2\mathbf{F}\mathbf{\Omega}\mathbf{F}^T\mathbf{F} - \mathbf{tr}\left(\mathbf{F}\mathbf{\Omega}\mathbf{F}^T\right)\mathbf{F} = \mathbf{0}. \tag{11}$$

We can get nine linear equations of f^2 from (11) and correspondingly obtain nine estimated values of f. The process from (11) to f is trivial and the detailed derivation is not presented for space reason. Here we take the first row and first column as an example and directly give its result as

$$f = \sqrt{\frac{F_{11}\mathbf{f}_3^T\mathbf{f}_3 - 2F_{13}\mathbf{f}_1^T\mathbf{f}_3}{2F_{12}\mathbf{f}_1^T\mathbf{f}_2 + F_{11}\left(\mathbf{f}_1^T\mathbf{f}_1 - \mathbf{f}_2^T\mathbf{f}_2\right)}}, \tag{12}$$

where F_{ij} is the i-th row and j-th column value of $\mathbf{F}$. We take the mean value of nine estimated f as the final result. Moreover, we discard the equations in which the coefficient of f^2 is close to zero for computational stability.

After extract the intrinsic parameters, we can obtain the essential matrix $\mathbf{E}$ applying essential matrix equation (6). Then the relative pose can be uniquely estimated same as in the 5-point algorithm [25]. We denote a projection matrix of an image as $\mathbf{P} = [\mathbf{R}, \mathbf{t}]$, where $\mathbf{R}$ is the 3×3 rotation matrix and $\mathbf{t}$ is the translation vector. For clarity, we denote $\mathbf{P}_n = [\mathbf{R}_n, \mathbf{t}_n]$ as the rotation matrix and translation vector of one neighbour relative to the world coordinate, and denote $\mathbf{P}_{rn} = [\mathbf{R}_{rn}, \mathbf{t}_{rn}]$ as the pose of input image relative to its neighbour image. After applying our relative pose estimation algorithm, we can obtain a set of $\mathbf{P}_{rn}$.

D. Final location determination

Based on the relative poses between input image and its nearest neighbours, we can figure out the final 6-DOF location information of input image. Given the rotation matrix of one neighbour relative to world coordinate $\mathbf{R}_n$ and the rotation matrix of input image relative to the neighbour $\mathbf{R}_{rn}$, we can get the rotation matrix of input image relative to world coordinate as:

$$\mathbf{R}_r = \mathbf{R}_{rn}\mathbf{R}_n. \tag{13}$$

For a given rotation matrix $\mathbf{R}_r$, it can be decomposed into three Euler angles $\theta_z, \theta_x, \theta_y$ corresponding to z-x-y coordinate axis in turn. We can average all $\theta_z, \theta_x, \theta_y$ candidates respectively to obtain final rotation angles. From three basic rotation matrix corresponding to $\theta_z, \theta_x, \theta_y$, the final rotation matrix can be figured out by matrix multiplication.

Now the last problem is that the relative pose estimation algorithm can only get the direction of $\mathbf{t}_i$ but no length. Moreover when there are more than one relative pose candidates, we can resort to triangulation theory because two or more rays can determine a point in 3D space. If there are only one relative pose candidate unfortunately, we simply suppose that $\mathbf{t}_i$ has unit length. Given input image and its one neighbour, although the 3D position of input image cannot be obtained, the relative pose $[\mathbf{R}_{rn}, \mathbf{t}_{rn}]$ and the 3D position of one neighbour (x_l, y_l, z_l) determine a straight line, which can be denoted as:

$$\frac{x - x_l}{x_d} = \frac{y - y_l}{y_d} = \frac{z - z_l}{z_d}. \tag{14}$$

The 3D position of input image is in this line. As $\mathbf{P}_n$ and $\mathbf{P}_{rn}$ are known, $\mathbf{l} = (x_l, y_l, z_l)^T$ can be obtained from:

$$\mathbf{l} = \begin{bmatrix} x_l \\ y_l \\ z_l \end{bmatrix} = -\mathbf{R}_n^{-1}\mathbf{t}_n, \tag{15}$$

and $\mathbf{d} = [x_d, y_d, z_d]^T$ can be obtained as

$$\mathbf{d} = \begin{bmatrix} x_d \\ y_d \\ z_d \end{bmatrix} = -\mathbf{R}_n^{-1}\mathbf{R}_{rn}^{-1}\mathbf{t}_{rn}. \tag{16}$$

Obviously, (14) contains only two different equations as:

$$\begin{cases} z_d x - x_d z = z_d x_l - x_d z_l \\ z_d y - y_d z = z_d y_l - y_d z_l \end{cases}. \tag{17}$$

Therefore given n $(n \geqslant 2)$ relative poses, we can get $2n$ linear equations according to (17). We can solve these $2n$ linear equations using least square algorithm to get the final position (x, y, z) of input image.

V. EXPERIMENTS

In order to verify the feasibility and effectiveness of our proposed method, we test our image localization method on Cornell Arts Quad dataset. We also quantitatively compare our calibrated-uncalibrated relative pose estimation algorithm with previous work on synthetic and real data.

A. Image localization

Cornell Arts Quad dataset [15] is originally created for 3D reconstruction and also be used by 3D point cloud model based localization algorithms. The dataset contains not only the cameras' position and rotation information but also 3D structure point cloud. While in our experiments, we only use the cameras' location information. There are 6514 images in Quad dataset, we random select 348 images as query images while the rest 6166 images as references. In the retrieval step, we first create a visual vocabulary via clustering SURF features which are extracted from reference images. Most features should appear more than twice in all reference images as duplicated image areas are common in this dataset. So we random selected 3000 images from all the reference images for computational efficiency. We set the number of visual words

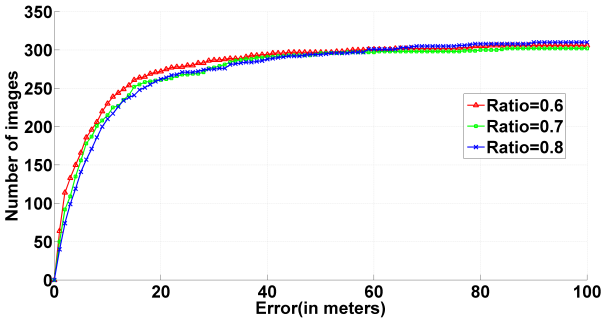


Fig. 3. The localization error on Quad dataset using different ratio value.

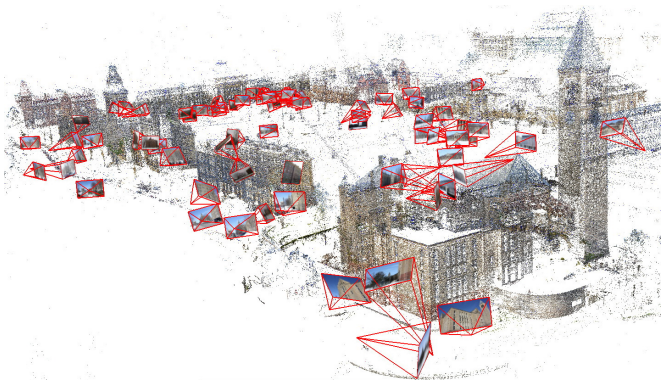


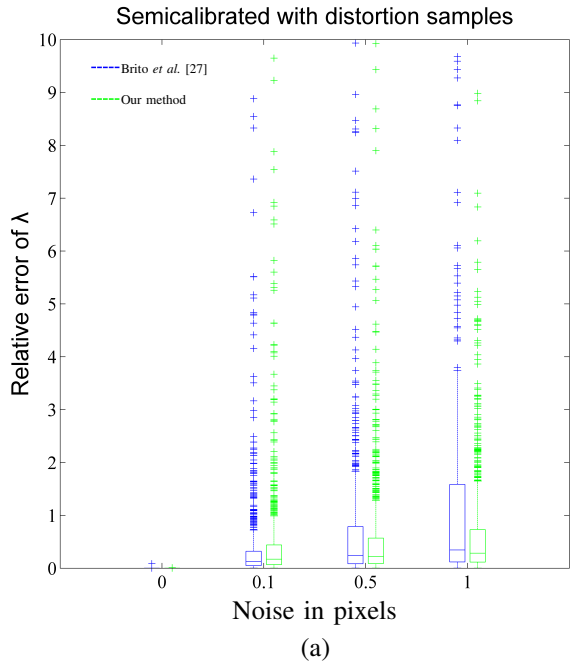
Fig. 4. Localization results of 100 random selected images on Quad dataset by our method.

as 30000. For each query image, we retrieve its top 20 nearest neighbour images.

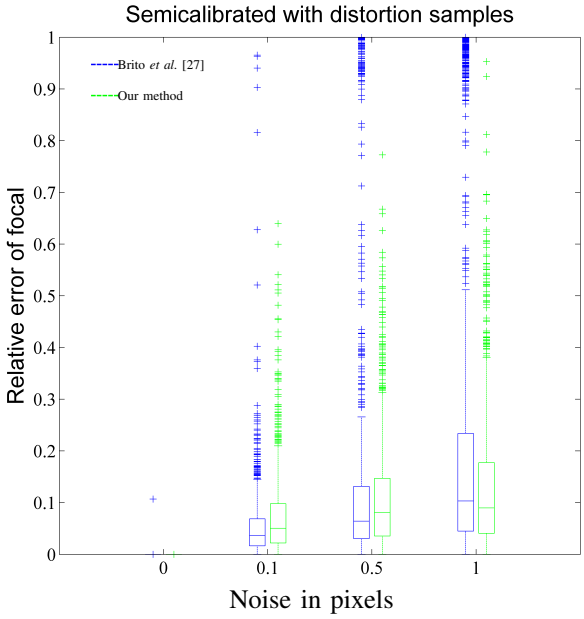
We find the initial point matches between input image and each of its neighbours using ratio test algorithm and set the ratio threshold r as 0.6, 0.7, 0.8 respectively. As presented in Figure 3, the localization accuracy of our method is not influenced obviously by the ratio test parameter. Our method can localize most query images with high accuracy and more than 60% images can be localized under ten meters. The localization error has a median of 5.38m, 1st quartile of 1.28m and 3rd quartile of 14.98m when the ratio is set as 0.6, which is comparable with civilian GPS. Figure 4 visualize one hundred localization results on Quad dataset. The 3D point cloud is only used for visualization and not used in localization process. We perform 6-DOF image localization task without 3D point cloud and achieve comparable accuracy with civilian GPS, which is a promising method as SfM algorithm usually takes a long time to reconstruct 3D point cloud and does not always succeed.

B. Relative pose estimation

In our task, we must estimate the relative pose between input image and each of its neighbours. Input image is uncalibrated and has two intrinsic parameters including focal length f and one radial distortion parameter λ. Under the same configuration with our, Brito *et al.* [27] have estimated



(a)



(b)

Fig. 5. Comparisons with Brito *et al.* [27] on synthetic data. (a) Relative error of distortion parameter λ compared with Brito *et al.* [27]. (b) Relative error of focal length compared with Brito *et al.* [27]. The horizontal axis represents the variance of Gaussian noise added on points and the vertical axis represents relative error of result.

the intrinsic parameters. Therefore we compare our algorithm with Brito *et al.* [27] on synthetic and real image data. We first test our algorithm using synthetic data, which has a set of random 3D points and 1000 generated random camera poses. The results are shown in Figure 5, where the horizontal axis represents the variance of Gaussian noise added on points and the vertical axis represents relative error of result. Figure 5(a) shows the relative error of distortion parameter λ compared

TABLE I
OUR METHOD NEED LESS THAN HALF TIME TO ESTIMATE OSR FUNDAMENTAL MATRIX AND RELATIVE POSE COMPARED WITH [27].

	Our method	Brito *et al.* [27]
Estimate OSR fundamental matrix	3.846 s	9.768 s
Intrinsic parameters extraction and estimate relative pose	0.415 s	0.435 s
Total	4.261 s	10.203 s

with Brito *et al.* [27], Figure 5(b) shows the relative error of focal length. The experiments show that, the accuracy of our result is comparable with Brito *et al.* [27] and more robust with increasing noisy.

We also test our algorithm on 4 sets of real images which are used in [27]. Our results' mean inlier ratios are 85.2%, 74.5%, 81.8% and 90.7% respectively on these 4 datasets, while Brito *et al.* [27] 84.5%, 77.4%, 83.8% and 90.1%. Our mean inlier ratio is higher than Brito *et al.* [27] on dataset 1 and 4, while Brito *et al.* [27] higher on dataset 2 and 3. There are some undistorted results using the λ estimated by our algorithm and Brito *et al.* [27] respectively in Figure 6, which have little difference. Though the accuracy of our algorithm is not better than [27], our algorithm is faster. As shown in Table I, our algorithm need less than half time to estimate OSR fundamental matrix and to estimate relative pose excluding feature extracting and matching time. Note that, we embed the fundamental matrix estimation algorithm in an RANSAC process with 200 iterations. All these experiments are done using Matlab 2013a on the same PC with I7-3770 CPU and 4GB RAM. The reason for our algorithm's faster is that, our algorithm mainly applies SVD algorithm three times which can be performed fast, while Brito *et al.* [27] have to solve a higher-order polynomial system which needs more time.

VI. CONCLUSION AND DISCUSSION

In this paper, we propose a data-driven method to perform 6-DOF image localization task without 3D point cloud and present a fast algorithm to estimate the relative pose between a calibrated and an uncalibrated image. It has practical significance to use data-driven methods to perform 6-DOF image localization task because SfM algorithm does not always succeed in reconstructing 3D point cloud especially with the massive and increasing data. In order to perform this task, we first retrieve the nearest neighbours of input image from a dataset. Then we estimate the relative pose between the uncalibrated input image and each of its calibrated neighbours. Finally we figure out the final optimal location via integrating all relative poses. Our localization accuracy is comparable with civilian GPS. At the same time, compared with state-of-the-art 3D point cloud model based methods, our method need more time in localization process, as it takes a long time to estimate the relative poses. We could accelerate our method via using multiple CPU cores in the future work, as it can be easily parallelized to estimate each relative pose.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful suggestions in improving this paper. This work was partially supported by 863 Program (2013AA013801), NSFC (61325011), SRFDP (20131102130002) and Funded by Lenovo Outstanding Young Scientists Program(LOYS).

REFERENCES

[1] D. Wagner, G. Reitmayr, A. Mulloni, T. Drummond, and D. Schmalstieg, "Real-time detection and tracking for augmented reality on mobile phones," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 3, pp. 355–368, 2010.

[2] G. Takacs, V. Chandrasekhar, N. Gelfand, Y. Xiong, W.-C. Chen, T. Bismpigiannis, R. Grzeszczuk, K. Pulli, and B. Girod, "Outdoors augmented reality on mobile phone using loxel-based visual feature organization," in *ACM International Conference on Multimedia Information Retrieval*, 2008, pp. 427–434.

[3] H. Lim, S. Sinha, M. Cohen, and M. Uyttendaele, "Real-time image-based 6-dof localization in large-scale environments," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1043–1050.

[4] L. Meier, P. Tanskanen, F. Fraundorfer, and M. Pollefeys, "Pixhawk: A system for autonomous flight using onboard computer vision," in *IEEE International Conference on Robotics and Automation*, 2011, pp. 2992–2997.

[5] E. Royer, M. Lhuillier, M. Dhome, and J.-M. Lavest, "Monocular vision for mobile robot localization and autonomous navigation," *Int. J. Comput. Vision*, vol. 74, no. 3, pp. 237–260, 2007.

[6] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3d," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, Jul. 2006.

[7] J. Kopf, B. Neubert, B. Chen, M. Cohen, D. Cohen-Or, O. Deussen, M. Uyttendaele, and D. Lischinski, "Deep photo: Model-based photograph enhancement and viewing," *ACM Trans. Graph.*, vol. 27, no. 5, pp. 116:1–116:10, Dec. 2008.

[8] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt, "Scalable 6-dof localization on mobile devices," in *European Conference on Computer Vision*, ser. Lecture Notes in Computer Science, 2014, vol. 8690, pp. 268–283.

[9] M. Donoser and D. Schmalstieg, "Discriminative feature-to-point matching in image-based localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 516–523.

[10] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua, "Worldwide pose estimation using 3d point clouds," in *European Conference on Computer Vision*, ser. Lecture Notes in Computer Science, 2012, vol. 7572, pp. 15–29.

[11] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2d-to-3d matching," in *IEEE International Conference on Computer Vision*, 2011, pp. 667–674.

[12] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2599–2606.

[13] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[14] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[15] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher, "Discrete-continuous optimization for large-scale structure from motion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3001–3008.

[16] Q. Hao, R. Cai, Z. Li, L. Zhang, Y. Pang, and F. Wu, "3d visual phrases for landmark recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3594–3601.

[17] Y. Li, D. Crandall, and D. Huttenlocher, "Landmark classification in large-scale image collections," in *IEEE International Conference on Computer Vision*, 2009, pp. 1957–1964.

[18] S. Cao and N. Snavely, "Graph-based discriminative learning for location recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 700–707.

Fig. 6. Some undistorted results using the distortion parameter λ estimated by our method and Brito *et al.* [27].

[19] D. Chen, G. Baatz, K. Koser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk, "City-scale landmark identification on mobile devices," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 737–744.

[20] A. Zamir and M. Shah, "Accurate image localization based on google maps street view," in *European Conference on Computer Vision*, ser. Lecture Notes in Computer Science, 2010, vol. 6314, pp. 255–268.

[21] W. Zhang and J. Kosecka, "Image based localization in urban environments," in *Third International Symposium on 3D Data Processing, Visualization, and Transmission*, June 2006, pp. 33–40.

[22] M. Lux and S. A. Chatzichristofis, "Lire: lucene image retrieval: an extensible java cbir library," in *ACM International Conference on Multimedia*, 2008, pp. 1085–1088.

[23] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing via label transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2368–2382, 2011.

[24] X. Chen, Q. Li, Y. Song, X. Jin, and Q. Zhao, "Supervised geodesic propagation for semantic label transfer," in *European Conference on Computer Vision*, ser. Lecture Notes in Computer Science, 2012, vol. 7574, pp. 553–565.

[25] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 6, pp. 756–770, 2004.

[26] M. Bujnak, Z. Kukelova, and T. Pajdla, "3d reconstruction from image collections with a single known focal length," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1803–1810.

[27] J. H. Brito, C. Zach, K. Koeser, M. Ferreira, and M. Pollefeys, "One-sided radial-fundamental matrix estimation," in *the British Machine Vision Conference*. BMVA Press, 2012, pp. 96.1–96.12.

[28] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *European Conference on Computer Vision*, ser. Lecture Notes in Computer Science, 2006, vol. 3951, pp. 404–417.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.