Extrinsic Camera Parameter Estimation from a Still Image based on Feature Landmark Database

Mitsutaka Susuki Tomoka Nakagawa Tomokazu Sato Naokazu Yokoya

Nara Institute of Science and Technology 8916-5 Takayama, Ikoma-shi, Nara, 630-0192 Japan

abstract: In this paper, we propose a 6-DOF extrinsic camera parameter estimation method for a still image input. In the proposed method, first, we construct feature landmark database that contains the multi-view and multi-dimensional data. In our method, feature landmark database is constructed by applying structure from motion approach to the image sequences acquired by a moving omnidirectional multi-camera system. After database construction, camera position and posture are estimated from a still image by using pairs of corresponding image features and feature landmarks. However, it is not easy to detect correct correspondences because a large number of visually similar landmarks are registered in the database. To achieve robust matching between image features and feature landmarks, we gradually limits searching range for landmark database by using GPS position, SIFT distance, and consistency of camera position and posture. The validity of the proposed method has been shown through experiments for an outdoor environment.

1 Introduction

Several human navigation services are currently available on the cellular phones that uses embedded GPS and 2-D map. However, 2-D map based human navigation is not always easy to understand for users because that is not intuitive. To realize more intuitive human navigation, AR (Augmented Reality) based navigation where guiding information is overlaid in the real image is expected to be the next generation navigation system. For AR navigation, the key problem is how to acquire the accurate position and posture of the embedded camera on the cellular phone. Many researchers have intensively tackled to the camera parameter estimation problem for AR in recent years. However, most of these methods cannot be applied to the current mobile devices due to several problems.

To realize AR based navigation system on the consumer mobile devices, at least, following four requirements should be satisfied.

(1) Equipment should be simple.
(2) Method should work in various environments.
(3) 6-DOF position and posture should be estimated.
(4) Computational costs for mobile device should be low.

Based on these requirements, the availability of conventional localization methods for ubiquitous AR is discussed in the follows.

To realize position and posture estimation of mobile devices, sensor based approach [1, 2, 3], image based approach [4, 5, 6, 7, 8, 9, 10, 11] have been investigated. Sensor based approach usually uses combination of several sensors to acquire both position and posture. However most of these methods relies for GPS, the raw accuracy of the position given by the embedded GPS on mobile devices is 10m order and it is not sufficient to realize AR navigation. Although there is possibility to acquire high accurate position and posture by combining several sensors such as RTK-GPS, optical fiber gyro, and magnetic sensors, equipment becomes complex and it is not suitable for mobile devices.

Image based approach uses captured images to estimate camera position and posture. Methods in this approach can be categorized as four groups; (a) Marker based methods that use number of artificial markers and their 3-D positions [4, 5], (b) image retrieval based methods that estimate camera position and posture by searching the image database for similar image [6, 7], (c) 3-D model based methods where position and posture are determined in such a way that the visibility of 3-D model becomes similar to the input image [8, 9], (d) feature landmark based methods that use pre-registered 3-D points of image features instead of artificial markers [10, 11]. All these methods can estimate position and posture of a camera without sensors by using pre-constructed database that models the target environment or object. Thus, equipment of the mobile device can be simple if the image based approach is employed.

However, each group still has problems to realize AR navigation on mobile devices. In the marker based methods (a), the problem is cost for marker arrangement and maintenance. It is almost impossible to set up and maintain these artificial markers in large outdoor environments. Image re-

trieval based methods (b) cannot estimate 6-DOF position and posture because these image databases do not contain 3-D data for target environment. 3-D model based methods (c) also have a problem that construction of accurate 3-D model for complex and large environment is difficult. On the other hand, feature landmark based methods (d) do not have such problems. In the feature landmark based methods, 3-D database for an environment is constructed automatically by applying structure-from-motion method to the video sequences that capture the target environment. By using a high-resolution omni-directional camera, a large number of feature landmarks in the target environment can be effectively collected and registered to the feature landmark database. After database construction, camera position and posture are estimated by using corresponding pairs of feature points and landmarks.

The feature landmark based method potentially satisfies all the requirements (1) to (4) if we assume that camera parameter estimation is done with server-client scheme where the client (mobile device) captures a still image and transmits to the server (service provider) through the internet. After camera parameter estimation on the server, the server returns estimated position and posture to the client. However this approach is suitable for the current cellular phone services, video sequence has been assumed as an input in the conventional works and it will easily consume the bandwidth of the network. In the conventional works, searching range of the database is extremely limited by assuming that movement of the camera between successive video frames is very small. Thus, video based methods can easily find the correct correspondences even if there are a large number of similar landmarks in the database. To apply feature landmark approach for a still image input, we must develop the method that can find correct correspondences without good initial parameters for camera position and posture.

In this paper, based on the feature landmark approach, we propose a novel method that can estimate 6-DOF extrinsic camera parameters from a still image input. As shown in Figure 1, our method is constructed of two phases. In the offline phase, landmark database is developed by applying structure from motion method to the omni-directional video streams. In the online phase, to find the correct matches from a large number of visually similar landmarks, we gradually discard the candidates of landmarks. First, rough position of the user (100m order accuracy is assumed) given by embedded GPS or strength of electric waves of mobile phone is used to select the database from a mount of landmark database that may be developed in every site of the world. Next, visually similar landmarks with image features on the input image are selected using SIFT descriptor and LoG based scale detector. Spatially consistent landmarks are then selected by voting approach for camera position and posture. Finally, 6-DOF camera position and posture

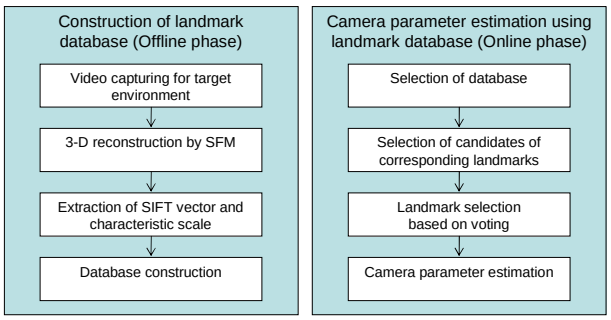


Figure 1: Flow diagram of proposed method.

are estimated by solving PnP problem with RANSAC based outlier elimination.

This paper is organized as follows. In Section 2, we describe the first phase of our method, which specifies the construction of the feature landmark database. Section 3 describes the position and posture estimation method for a still image using the feature landmark database. Section 4 shows the experimental results in an outdoor environment, and Section 5 summarizes this paper.

2 Database construction

Before camera parameter estimation for mobile devices, feature landmark database must be constructed for the target environment where localization service will be provided. Basically, the feature landmark database is constructed by 3-D positions of feature points and their visual information. In the following sections, first, elements of feature landmark database are described. The way of database construction is then detailed.

2.1 Elements of Feature Landmark Database

As shown in Figure 2, feature landmark database consists of a number of landmarks. Each landmark retains the 3-D coordinate of itself (A), and several information acquired from different observation positions (B). Information for different observation positions consists of three elements: (a) observed position, (b) characteristic scale, (c) SIFT vector. In offline phase, these elements are acquired by analyzing video images that captures target environment.

2.2 3-D reconstruction for environment

The target environment is captured as video sequences at first. In this paper, moving an omni-directional multi-camera system is assumed to be used as a scanner for the target environment. For the acquired video sequences, structure from motion method for multi-camera system [12] is

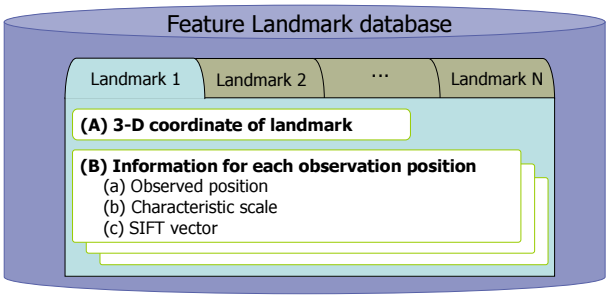


Figure 2: Elements of feature landmark database.

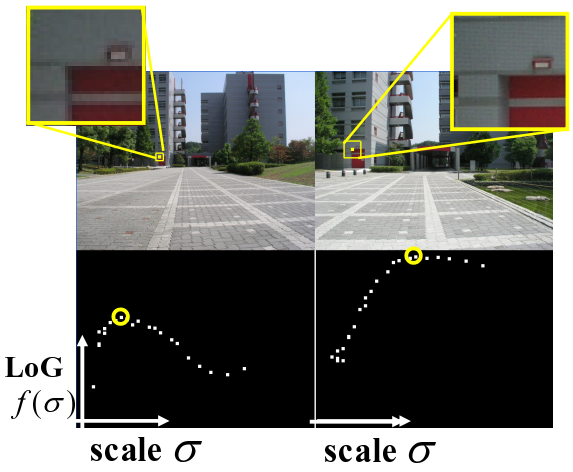


Figure 3: Determination of caracteristic scale by Laplacian of Gaussian.

applied to estimate 3-D positions of feature points and camera parameters. In the method [12], image features detected by Harris operator [13] are tracked trough the input video sequences and 3-D positions of image features and camera parameters of the camera system are estimated based on the bundle adjustment [14]. Note that general structure from motion cannot recover the absolute 3-D position and scale by itself and all the recovered 3-D positions are relative. To arrange the recovered 3-D points for an absolute coordinate, some control points of known 3-D positions should be given to the bundle adjustment process [12], or hybrid approach should be employed that uses both motion of image features and absolute positions measured by GPS attached with the camera system [15]. In the experiment that is described in Section 4, the former approach has been employed.

2.3 Extraction of characteristics for landmarks

In this research, to achieve rotation and distortion invariant matching, SIFT [16] is employed as feature descriptor. However SIFT also includes characteristic scale determination scheme that uses DoG (Difference of Gaussian), in this paper, we have employed LoG (Laplacian of Gaussian) instead of DoG because DoG is used as an approximation of LoG in the article [16] and LoG based scale detector is compatible with Harris detectors that is used as feature detector in the structure from motion process [17].

To remove visual differences caused by camera posture and lens distortion, first, all the acquired images by the omni-directional multi-camera system are warped to the sphere whose radius is infinite. Next, for each image feature whose 3-D position is estimated, characteristic scale is determined by using LoG function. As shown in Figure 3, responses of the LoG function for different images become similar if there exist same image structure. Thus, characteristic scale is determined by searching for scale parameter σ that maximizes the LoG value for the image feature on the sphere. The LoG value is computed by applying LoG filter to the image on the sphere. The LoG filter is defined as follows:

$$f(r,\sigma) = -\frac{r^2 - 2\sigma^2}{2\pi\sigma^6}\exp(-\frac{r^2}{2\sigma^2}), \qquad (1)$$

where the scale parameter σ is identical to the radius of this filter and r is a distance from the center of the filter to the target pixel. After characteristic scale is determined, image pattern inside of the circle whose radius is σ and center is the position of the feature point is coded by SIFT descriptor [16].

3 Camera parameter estimation using landmark database

In this section, the processes for the online phase are described. In this research, we assume that feature landmark database constructed by the offline phase is stored in the server and all the processes described in this section are carried out in the server. To estimate the camera position of the client, the server needs both an input image and rough position that is measured by embedded GPS or strength of electric waves from cellular phones. The rough position is necessary to select the appropriate database from a mound of database for every site. After landmark selection, candidates of corresponding landmarks are firstly selected based on SIFT distance. Spatially consistent landmarks are then selected by voting for observable position and posture. Finally, camera position and posture are estimated by solving PnP problem with RANSAC based outlier elimination.

3.1 Selection of similar landmarks

In this process, first, feature points are detected for the input image by using Harris operator [13]. Next, characteristic

scale and SIFT vector for each feature point is computed by the same manner with the landmark construction process. For all the pairs of registered landmarks and feature points, SIFT distance S defined in Eq. (2) is computed using SIFT vector $\mathbf{v}_{LDB}(p)$ of landmark p and SIFT vector $\mathbf{v}_{IN}(q)$ of feature point q.

$$S(p,q) = |\mathbf{v}_{LDB}(p) - \mathbf{v}_{IN}(q)| \quad (2)$$

The distance S indicates pattern similarity and lower value means similar pattern. By using the distance S, we select the top-similar α landmarks per each feature point and the rest of landmarks are discarded. The landmarks whose distance S is below threshold are also rejected in this step.

3.2 Landmark selection based on voting

The candidates selected in the previous step include a lot of outliers. However outliers are visually similar to the feature points on the input image, most of them are placed in quite different positions from the correct landmarks. In this step, outliers are eliminated by verifying the space consistency based on the fact that the input image is captured from a unique position and a unique posture. To verify the space consistency, we employ the voting approach.

As illustrated in Figure 4, the ground plane around the roughly specified client position is divided into $(2h+1) \times (2h+1)$ grid boxes. Each grid box has l ballot boxes that are for horizontal orientation. Each landmark throws a vote for the m-th ballot box at the grid (i,j); $(-h \leq i \leq h, -h \leq j \leq h)$ if the box satisfies following two conditions.

Angular condition: The angle ϕ is under threshold T. ϕ is the angle of following two lines; (Line 1) the line connecting 3-D position $\mathbf{Q}_p$ of the landmark p and the observed position $\mathbf{C}_p$ for the landmark p, (Line 2) the line connecting $\mathbf{Q}_p$ and the position $\mathbf{W}_{ij} = (w_i, w_j, c_z)$ where w_i, w_j are the horizontal position of the grid box (i,j) and c_z is altitude element of $\mathbf{C}_p$.

Scale condition: The ratio of the characteristic scale ω_p of the landmark p and the characteristic scale ω_q of the corresponding image feature q on the input image is agree with the distance ratio of $|\mathbf{W}_{ij} - \mathbf{Q}_p|$ and $|\mathbf{C}_p - \mathbf{Q}_p|$: $1 - \alpha < \frac{\omega_p |\mathbf{W}_{ij} - \mathbf{Q}_p|}{\omega_q |\mathbf{C}_p - \mathbf{Q}_p|} < 1 + \alpha$ (where α is a threshold).

The ballot box m at the grid (i,j) is determined by $m = [\theta l / 2\pi]$ where $[a]$ returns the integer value of a, and θ is the horizontal angle (radian) between the x-axis and the vector $\mathbf{W}_{ij} - \mathbf{Q}_p$. After voting from all the landmarks, the landmarks that vote for the box of maximum count are selected. These selected landmarks are considered to be visible from a unique position and a unique posture. However

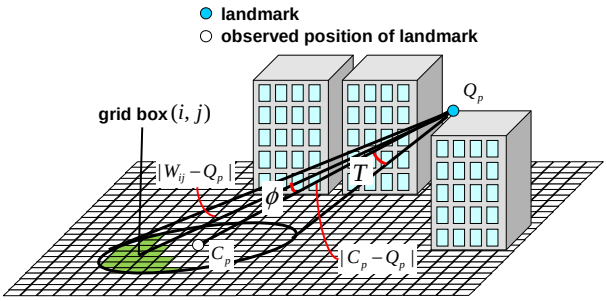


Figure 4: Voting from a landmark.

position and posture can be determined by (w_i, w_j) and θ in this step, they are still 3-DOF and they cannot be used for AR navigation. Next section describes how to upgrade these parameters to 6-DOF.

3.3 6-DOF extrinsic camera parameter estimation

In this step, by using selected landmarks in the previous step, 3-DOF camera position and posture are upgraded to 6-DOF. By using the 3-DOF parameters as initial parameter, the re-projection errors are minimized with non-linear optimization. The re-projection error is defined by distances between 2-D position of feature point on the input image and projected position of the 3-D position of the corresponding landmark. To remove outliers from the selected landmarks, in this process, we employ RANSAC approach that can be used when outliers are fewer than inliers.

Note that after final estimation result is given, the system can judge the validity of the estimated camera parameter based on the re-projection errors. In this research, if an average re-projection error is larger than the given threshold R, the system judges that the estimated result is not accurate. If the system judges that a result is in-accurate, the system re-tries the camera parameter estimation by using other landmark groups that are voting for the box of the next maximum count in the previous step.

4 Experiment

To show the validity of the proposed method, the feature landmark database is actuary constructed for a real outdoor environment and the accuracy of camera position and posture by the proposed method is evaluated by comparing it with the ground truth. In this experiment, first, we have captured the target environment as two omni-directional video sequences by using an omni-directional multi-camera system (Pointgrey Ladybug) that has six camera units of XGA resolution. In this experiment, about 12,500 landmarks are registered to the database. Each input image is taken as

(a) Sampled input images for direction 1.

(b) Sampled input images for direction 2.

Figure 5: Sampled input images.

Table 1: Thresholds that is used in experiment.

Dimension for SIFT vector	128
Angle threshold T (degree)	10
Scale ratio threshold α	0.2
Maximum re-projection error R (pixel)	5.0

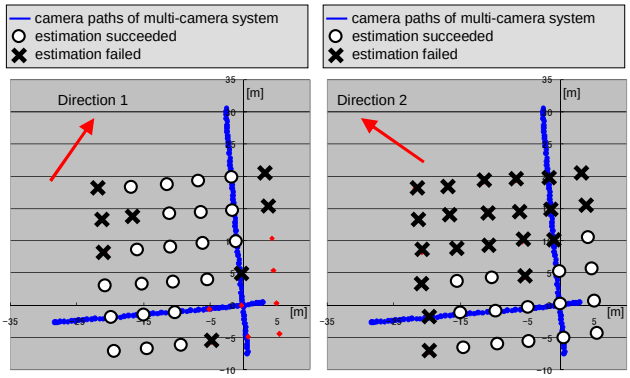
VGA sized digital photograph from the 6×6 positions that are arranged by the 5m grid using the digital camera embedded in the cellular phone (CASIO GzOne W42CA). In this environment, we take images for two directions. For direction 1, buildings that contains many landmarks are visible from most of the positions as shown in Figure 5(a). In contrast, for direction 2, buildings are blinded by trees from many positions as shown in Figure 5(b).

In this experiment, although server-client system is assumed, the prototype system has not been developed yet. Thus, input images were once stored into the mobile phone and camera position and posture for each image were estimated after all photographs were taken. Table 1 shows thresholds that are used in this experiment. The ground truth of the camera position and postures are given by solving PnP problem using manually specified correspondences for landmarks and image features. For 7 of 72 images, we cannot compute the ground truth due to lack of visible landmarks and thus we didn't use these 7 images for this evaluation. Thus, we have evaluated with 65 images whose ground truths are available.

Table 2 shows success rate of estimation and average and standard deviation of estimation errors. Figure 6 illustrates judged results for each input. For direction 1, except the positions where trees severely blind buildings, most of estimated results is judged as successful by the system. As shown in Table 2, average position error and posture error are 1.4m and 1.4degree and it is considered as sufficient level for AR navigation. For direction 2, for more than

Table 2: Success rate and accuracy of estimated results.

	Dir. 1	Dir. 2
Success rate of estimation (%)	72.4	41.7
Average position error (m)	1.4	6.8
Std.dev. position error (m)	2.5	9.1
Average posture error (deg.)	1.4	3.9
Std.dev. posture error (deg.)	2.0	4.5
Average re-projection error (pix.)	2.0	2.0
Std.dev. re-projection error (pix.)	0.9	1.1



(a) for direction 1 (b) for direction 2

Figure 6: Succeeded and failed positions for each direction.

half positions, estimated result was judged as failed (re-projection error was over 5 pixels). For these images, even when enough number of landmarks are visible in the input image, most of correct landmarks had been rejected because image patterns around landmarks on the buildings are overlaid by branches of trees and they affect characteristic scale determination. To realize more stable estimation for various environments, introduction of more robust ways for characteristic scale determination and corresponding point detection is necessary. For the environment where landmarks are hardly detected, user interactive scheme is also one of the solutions where the system suggests good directions for landmark detection based on the current estimation.

5 Conclusion

In this paper, we propose novel method for extrinsic camera parameter estimation using a still image and feature landmark database that contains multi-dimensional and multi-view data. In this method, feature landmark database is constructed in advance by using structure from motion method for video sequences that capture the target environment. From the registered landmarks, corresponding points of image features in the input image are searched by gradually limiting candidates. Finally, 6-DOF camera position and

posture are estimated by using pairs of image features and landmarks. By the experiment, the accuracy of the position and posture estimation are reached to the level that AR navigation can be realized on the mobile devices. However, for the environment where most of the scene is covered by nature, it is difficult to estimate the position and posture stably. In the future work, robustness for camera parameter estimation will be improved by investigating robust scale determination and matching for occluders. User interactive scheme will also be explored to develop more useful system.

Acknowledgment: This paper is supported by SCOPE (Strategic Information and Communications R&D Promotion Programme) of Ministry of Internal Affairs and Communication Japan.

References

[1] S. Feiner, B. MacIntyre, T. Höller and A. Webster: "A touring machine: Prototyping 3d mobile augmented reality systems for exploring the urban environment,", Proc. Int. Symp. on Wearable Computers (ISWC 1997), pp. 74–81, 1997.

[2] T. Höllerer, S. Feiner and J. Pavlik: "Situated documentaries: Embedding multimedia presentations in the real world,", Proc. Int. Symp. on Wearable Computers (ISWC 1999), pp. 79–86, 1999.

[3] R. Tenmoku, M. Kanbara and N. Yokoya: "A wearable augmented reality system using positioning infrastructures and a pedometer,", Proc. Int. Symp. on Wearable Computers (ISWC 2003), pp. 110–117, 2003.

[4] D. Wagner and D. Schmalstieg: "First steps towards handheld augmented reality,", Proc. Int. Symp. on Wearable Computers (ISWC 2003), pp. 21–23, 2003.

[5] M. Möhring, C. Lessig and O. Bimber: "Video see-through ar on consumer cell-phones,", Proc. Int. Symp. on Mixed and Augmented Reality (ISMAR 2004), pp. 252–253, 2004.

[6] R. Cipolla, D. Robertson and B. Tordoff: "Image-based localization,", Proc. Int. Conf. on Virtual Systems and Multimedia (VSMM 2004), pp. 22–29, 2004.

[7] J. Sato, T. Takahashi, I. Ide and H. Murase: "Change detection in streetscapes from gps coordinated omnidirectional image sequences,", Proc. Int. Conf. on Pattern Recognition (ICPR 2006), Vol. 4, pp. 935–938, 2006.

[8] L. Vacchetti, V. Lepetit and P. Fua: "Combining edge and texture information for real-time accurate 3d camera tracking,", Proc. Int. Symp. on Mixed and Augmented Reality (ISMAR 2004), pp. 48–57, 2004.

[9] E. Rosten and T. Drummond: "Fusing points and lines for high performance tracking,", Proc. Int. Conf. on Computer Vision (ICCV 2005), Vol. 2, pp. 1508–1515, 2005.

[10] I. Skrypnyk and D. G. Lowe: "Scene modelling, recognition and tracking with invariant image features,", Proc. Int. Symp. on Mixed and Augmented Reality (ISMAR 2004), pp. 110–119, 2004.

[11] M. Oe, T. Sato and N. Yokoya: "Estimating camera position and posture by using feature landmark database,", Proc. Scandinavian Conf. on Image Analysis (SCIA 2005), pp. 171–181, 2005.

[12] T. Sato, S. Ikeda and N. Yokoya: "Extrinsic camera parameter recovery from multiple image sequences captured by an omni-directional multi-camera system,", Proc. European Conf. on Computer Vision, Vol. 2, pp. 326–340, 2004.

[13] C. Harris and M. Stephens: "A combined corner and edge detector,", Proc. Alvey Vision Conf., pp. 147–151, 1988.

[14] B. Triggs, P. McLauchlan, R. Hartley and A. Fitzgibbon: "Bundle Adjustment a Modern Synthesis," Proc. Int. Workshop on Vision Algorithms, pp. 298–372, 1999.

[15] S. Ikeda, T. Sato and N. Yokoya: "Camera recovery of an omnidirectional multi-camera system using gps positions,", Proc. First Korea-Japan Joint Workshop on Pattern Recognition, pp. 91–96, 2006.

[16] D. G. Lowe: "Distinctive image features from scale-invariant keypoints,", Int. Journal of Computer Vision, Vol. 60, No. 2, pp. 91–100, 2004.

[17] K. Mikolajczyk and C. Schmid: "Scale & affine invariant interest point detectors,", Int. Journal of Computer Vision, Vol. 60, No. 1, pp. 63–86, 2004.