Geometric Registration in Outdoor Environments Using Landmark Database Generated from Omnidirectional Videos and GPS Positions

Sei IKEDA[*] Tomokazu SATO[†] and Naokazu YOKOYA[‡]
Nara Institute of Science and Technology

ABSTRACT

This paper describes a geometric registration method for superimposing CG objects onto a real video acquired in outdoor environments. The method is applicable to various augmented reality systems such as pre-visualization tools which provide a director and other stuffs with information of the positional relation among real and virtual actors in preproduction of filmmaking. We have already developed a registration method using landmark database which stores positions of sparse feature points with their view-dependent image templates. The database is generated in advance by using a structure-from-motion technique which requires measuring some absolute positions of feature points manually in the environment. To skip the measurement process, we newly propose a method to construct a landmark database from omnidirectional videos and GPS positions. This paper also reports some experimental results with the proposed method.

Keywords: Geometric Registration, Landmark Database, 3-D Reconstruction.

1 INTRODUCTION

This paper describes a model-based geometric registration method to superimpose CG objects onto a real video acquired in outdoor environments. This kind of technique is applicable to various augmented reality systems including pre-visualization tools of filmmaking [1, 2] and guidance systems equipped with wearable computers [3, 4]. Pre-visualization tools using augmented reality are usable in substitution for traditional storyboards. In preproduction of filmmaking, they provide directors and other stuffs with information of the positional relation among real and virtual actors/objects by presenting CG objects superimposed on a captured video through a monitor. Accurate positions and postures of a camera are required to display CG objects at geometrically correct positions. While camera pose information sometimes can be obtained from other sensors equipped with a dolly or a crane, there are many scenes where such large-sized facilities are not suitable for outdoor environments. Wearable guidance systems based on augmented reality are used by people in daily life. The systems present annotations overlayed on a user's view image through a head-mounted display in real-time. Displaying annotations enable users to grasp positions of objective buildings and spots intuitively in their field of view. The annotations should be displayed at correct positions even if users walk around in a wide area.

In these applications, the main problem is real-time geometric registration so that superimposed objects do not drift with passage of time without any markers even if the camera moves widely. For this problem, a number of real-time registration methods using 3-D models such as wire frame models [5–8] were proposed. These methods require accurate 3-D modes to estimate accurate positions and postures of a camera. However, it is difficult to reconstruct complex scenes such as natural environments with hands.

On the other hand, registration methods which do not require any artificial markers and 3-D models have already been developed [9–11]. Such methods require a landmark database which stores 3-D positions of sparse feature points with their view-dependent visual features. Construction of the database is done semi-automatically by using a structure-from-motion technique which requires some absolute 3-D positions of feature points and their correspondences between the world and image coordinate systems.

In this paper, we propose a registration method which is based on using landmark database generated from omnidirectional videos and GPS positions. We newly improve the method for constructing a landmark database so as to skip the manual measurement process. The construction of landmark database is basically an extension of the structure-from-motion algorithm using GPS positions [12, 13] for an omnidirectional multi-camera system (OMS). Using this method enables us to obtain absolute positions of natural feature points and absolute camera positions and postures in the geodetic coordinate system full-automatically. In the remainder of this paper, the registration method using landmark database is first briefly described in Section 2. The generation of landmark database using an omnidirectional video and GPS positions is then described in Section 3. In Section 4, the validity of the method is demonstrated by experiments with a real outdoor scene. Finally, we give conclusion and future work in Section 5.

2 GEOMETRIC REGISTRATION USING LANDMARK DATABASE

This section describes the registration method using the feature landmark database. First, the elements of the feature landmark database are defined. How to use that information to estimate positions and postures of a camera is then described.

2.1 Elements of Feature Landmark Database

Feature landmark database consists of a number of landmarks as shown in Figure 1. Each landmark retains the 3-D position of itself (1), and multiple view-dependent image templates and their geometric information (2). The former is used with 2-D position of a feature point detected in an input image in order to estimate position and posture of the camera. The latter is used for a robust matching between the landmark image template and input image acquired from various directions. These database elements are generated from omnidirectional videos by a 3-D reconstruction method described later.

(1) 3-D position of landmark
3-D coordinate of each landmark is estimated by 3-D reconstruction of the environment and is obtained in the world coordinate system. The X and Y axes of the world coordinate system are aligned to the ground and the Z axis is vertical to the ground.

(2) Information for view-dependent image template
This information is used to find correspondences between feature points in an input image and the landmarks.

[*] e-mail: sei-i@is.naist.jp
[†] e-mail: tomoka-s@is.naist.jp
[‡] e-mail: yokoya@is.naist.jp

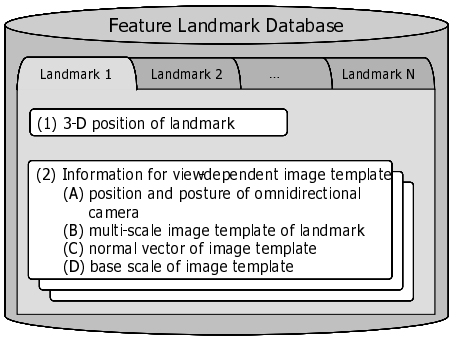


Figure 1: Elements of feature landmark database.

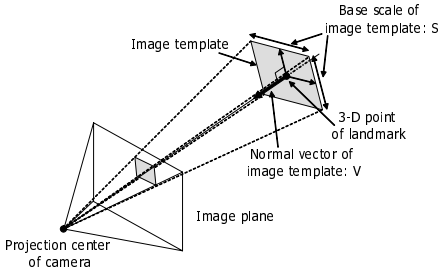


Figure 2: Landmark and its image template.

(A) Position and posture of omnidirectional camera
Position and posture of omnidirectional camera are retained in the world coordinate system. They are used to select landmarks from the database to match with the input image.

(B) Multi-scale image template of landmark
Image template is created by rectifying the omnidirectional image so as to be vertical to the line passing through the landmark's 3-D position and the omnidirectional camera's position, as shown in Figure 2.

(C) Normal vector of image template
As shown in Figure 2, the normal vector of image template is the normal vector of the plane which is vertical to the line passing through the landmark's 3-D coordinate and the omnidirectional camera's position. It is used to select an image template for matching from multi-directional image templates taken by different camera positions.

(D) Base scale of image template
As shown in Figure 2, the scale of image template is the size of the plane used to create the image template. The scale size is retained in the world coordinate system, and the base scale is determined so that the resolution of the omnidirectional image and the image template becomes nearly equal.

2.2 Algorithm of Registration

This section describes a camera position and posture estimation algorithm based on the feature landmark database. This algorithm assumes that the initial camera position and posture are estimated by another method. In the subsequent frames, first, landmarks are selected from the landmark database by using the previous camera position and posture. Detecting natural features from the input image and matching them with the landmark image templates, the correspondence between landmark and input image is then established. Lastly, camera position and posture are estimated from the correspondences between landmarks and input image. The following sections describe these steps.

2.2.1 Selecting Landmark from Landmark Database

To find a correspondence with a feature point in the input image, several landmarks are selected from numerous landmarks in the landmark database. Furthermore, to handle partial occlusions and aspect changes, an image template with the nearest appearance to the input image is chosen from a number of image templates. Considering the appearance, it is ideal if the image template and input image are taken in the same position. However, the camera position and posture of the input image are not yet estimated. We use the camera position and posture of the previous frame as a replacement. Landmarks satisfying the following requirements are selected to make correspondence with the input image.

(requirement 1) Landmark has to be in the image when projecting its 3-D coordinate using the previous camera position and posture: We project the landmark's 3-D position on the input image by using previous camera position and posture. Only the landmarks projected in the input image frame are selected.

(requirement 2) Distance between the camera position where the landmark was taken and the camera position where the input image was taken should be small: We actually calculate the distance between the camera position when the landmark was taken and the camera position of the previous frame, and select landmarks under the threshold.

(requirement 3) Angle between the normal vector of the image template and the vector from landmark to camera position when the input image was taken should be smallest of all the image templates of the landmark: We select the image template if angle θ between the normal vector of the image template and the vector from landmark to previous camera position is the minimum for all image templates of the same landmark. If the angle θ of the selected image template is over the threshold, that landmark is not selected.

(requirement 4) Landmark must not be adjacent to already selected landmarks: First, the input image is divided in a grid. We project the landmarks on the input image by using the previous camera position and posture, and only select one landmark per each grid.

Landmarks that satisfy the requirement 1 are selected first. Then, the selected landmarks are narrowed down to a fixed number of landmarks by the ascending order of the distance mentioned in the requirement 2. From the list of landmarks, landmarks with smaller angles in the requirement 3 are picked up one by one, and are repeated until a fixed number of landmarks that satisfy the requirement 4 are chosen.

2.2.2 Determining Correspondence between Landmark and Input Image Feature

The next step is to find the correspondences between selected landmarks and features in an input image. Natural features are detected from the input image, and are corresponded with the selected landmarks.

Detecting Natural Feature from Input Image To find the correspondence between landmarks and input image, natural feature points are detected from the input image by Harris operator[14]. In this step, a landmark is projected to the input image, using previous camera position and posture. On the assumption that the corresponding point for the landmark exists near the projected point, natural feature points are detected within a fixed window surrounding the projected point. The detected feature points are listed as correspondence candidates of the landmark.

Matching Between Landmark Image Template and Input Image In this step, each landmark is compared with its correspondence candidates. First, an image pattern is created for each natural feature point listed as a correspondence candidate. Next, the landmark image template is compared with each image pattern by normalized cross correlation. Then, the feature point with the most correlative image pattern is selected, and its neighboring pixels are also compared with the landmark as correspondence candidates. Lastly, the most correlative feature point is corresponded with the landmark.

2.2.3 Camera Position and Posture Estimation Based on Established Correspondences

Camera position and posture are estimated from the list of 2-D and 3-D correspondences acquired from the matching between landmarks and input image. First of all, outliers are eliminated by RANSAC[15]. Then, camera position and posture are estimated, using only the correspondences that are supposed to be correct. As a result, camera position and posture with the minimum re-projection error becomes the answer. Here it should be noted that more than five correspondences are required in order to determine camera position ans posture uniquely.

3 Generation of Landmark Database from Omnidirectional Video and GPS Positions

The requirements to obtain the elements of the landmark database described in the previous section are positions and postures of OMS and 3-D positions of natural feature points. This section describes a 3-D reconstruction method which enables us to estimate these parameters. In our 3-D reconstruction method, the general structure-from-motion algorithm is enhanced to treat multiple videos acquired with OMS and GPS position information. In the general structure-from-motion algorithm, re-projection error that is observation error is minimized to obtain camera path parameters and 3-D positions of feature points. In the proposed method, an error function combining re-projection error and the error concerning GPS is simultaneously minimized. First, in this section, we explain omnidirectional multi-camera system. We then define a new error function combining re-projection error and the error function concerning GPS. Finally, the algorithm to minimize the new error function is described.

Note that the following conditions are assumed: (i) OMS and GPS are correctly synchronized; (ii) the geometrical relation between all the cameras and the GPS receiver is always fixed; (iii) the distance between the GPS receiver and the representative camera of the OMS is known, and the direction of GPS receiver in camera coordinate system is unknown. In this paper, it is also assumed that OMS has been calibrated in advance [16] and the intrinsic camera parameters (including lens distortion, focal length and aspect ratio) of each element camera of OMS are known.

3.1 Omnidirectional Multi-camera System

Omnidirectional multi-camera system is constructed of a set of element cameras such as Ladybug (Point Grey Research) which can obtain omnidirectional videos as shown in Figure 3. As mentioned above, we assume that position and posture relations among element cameras are known and fixed in this paper. The positions and postures of all the cameras can be relatively expressed as a pair of position and posture of a representative camera. In the i-th frame, the transformation from the world coordinate system to the camera coordinate system of each element camera c can be expressed by the following matrix N_{ic} by using the transformation M_c from the world coordinate system of a calibration process to the camera coordinate system of the camera c $(=0,1,2,3...n)$.

$$\mathsf{N}_{ic} = \mathsf{M}_c(\mathsf{M}_0)^{-1}\mathsf{N}_{i0} = \begin{bmatrix} \mathsf{R}_{ic} & \mathbf{t}_{ic} \\ 0 & 1 \end{bmatrix}, \tag{1}$$

Figure 3: A sampled frame of an acquired omnidirectional video. Right bottom is an image of vertical element camera. Others are horizontal ones.

where $\mathbf{t}_{ic}$ and R_{ic} represent the translation and the rotation from the world coordinate system of the i-th frame to the camera coordinate system of the camera c. This problem is treated as estimation of position ($\mathsf{R}_i = \mathsf{R}_{i0}$) and posture ($\mathbf{t}_i = \mathbf{t}_{i0}$) of the representative camera (c=0).

3.2 Error Function for Optimization Process

Re-projection Error Re-projection error is generally used for extrinsic camera parameter recovery based on feature tracking. The method for minimizing the sum of squared re-projection error is usually referred to as bundle adjustment. The re-projection error Φ_{ijc} for the feature j in the i-th frame of the camera c is defined as follow.

$$\Phi_{ijc} = |\mathbf{q}_{ijc} - \hat{\mathbf{q}}_{ijc}|, \tag{2}$$

where $\hat{\mathbf{q}}$ represents the 2D projected position of the feature's 3D position and $\mathbf{q}$ represents the detected position of the feature in the image. The 2D projected position $\hat{\mathbf{q}}$ of the 3-D position $\mathbf{p}_j$ of the feature j whose depth is z is calculated by the following equation.

$$\begin{bmatrix} z\hat{\mathbf{q}}_{ijc} \\ z \\ 1 \end{bmatrix} = \mathsf{N}_{ic}\mathbf{p}_j, \tag{3}$$

Error of GPS positions Generally, if GPS positions and estimated parameters do not contain any errors, the following equation is satisfied in the i-th frame among the parameters (position $\mathbf{t}_i$, posture R_i), GPS position $\mathbf{g}_i$ and the position of GPS receiver $\mathbf{d}$ in the camera coordinate system.

$$\mathsf{R}_i\mathbf{g}_i + \mathbf{t}_i = \mathbf{d} \quad (i \in \mathscr{F}), \tag{4}$$

where $\mathscr{F}$ denotes a set of frames in which GPS positions are obtained. However, unfortunately GPS position $\mathbf{g}_i$ and the parameters $\mathbf{t}_i$ and R_i usualy contain some errors. We introduce the following error function Ψ_i as an error of measured GPS position, which means the distance between the measured position of the GPS receiver and the predicted one.

$$\Psi_i = |\mathsf{R}_i\mathbf{g}_i + \mathbf{t}_i - \mathbf{d}|. \tag{5}$$

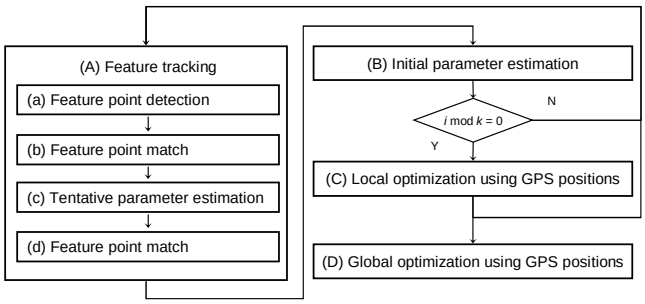


Figure 4: Overview of the proposed algorithm.

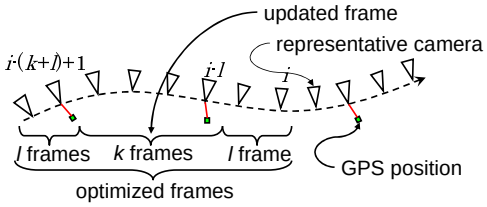


Figure 5: Optimized frames in the process (C).

Error Function Concerning Feature and GPS The new error function E is defined as follows:

$$E = \frac{\omega}{|\mathscr{F}|}\sum_{i\in\mathscr{F}}\Psi_i^2 + \frac{1}{\sum_i\sum_c|\mathscr{S}_{ic}|}\sum_i\sum_c\mu_i\sum_{j\in\mathscr{S}_{ic}}w_j\Phi_{ijc}^2,$$

where ω means a weight for Ψ_i, and $\mathscr{S}_{ic}$ denotes a set of feature points detected in the i-th frame of the camera c. The coefficients μ_i and w_j mean the confidences for frame and feature, respectively. w_j represents the confidence coefficient of feature point j, which is computed as an inverse variance of re-projection error Φ_{ij}. The coefficient μ_i denotes the confidence of the i-th frame. Two terms in the right-hand side in Eq. (6) is normalized by $|\mathscr{F}|$ and $\sum_i\sum_c|\mathscr{S}_{ic}|$ so as to set ω as a constant value independent of the number of features and GPS positioning points.

3.3 Algorithm of 3-D Reconstruction

The proposed method basically consists of feature tracking and optimization of camera parameters as shown in Figure 4. First, two processes of (A) feature tracking and (B) initial parameter estimation are performed in order. At constant frame intervals, the local optimization process (C) is then carried out to reduce accumulative errors. Finally, estimated parameters are refined using many tracked feature points in the global optimization process (D). In the processes (C) and (D), a common optimization is performed. The difference in both processes is the range of optimized frames. In the process (C), the range of optimization is a small part of the input frames because future data cannot be treated in sequential process. On the other hand, in the process (D), all the frames are optimized and updated.

(A) Feature tracking : The purpose of this step is to determine corresponding points between the current frame i and the previous frame $(i-1)$. The main strategy to avoid mismatching in this process is that feature points are detected at corners of edges by Harris operator [14] and detected feature points are tracked robustly with a RANSAC [15] approach. Note that feature point tracking is carried out in intra- and inter-camera images of OMS.

In the first step (a), natural feature points are automatically detected by using the Harris operator for limiting feature position candidates in the images. In the next step (b), every feature in the $(i-1)$-th frame is tentatively matched with a candidate feature point in the i-th frame by using a standard template matching. In the third step (c), tentative parameters are then estimated by selecting correct matches using a RANSAC approach [15]. In the final step (d), every feature is re-tracked within a limited searching area in image frames of all the element cameras, which can be computed by the tentative parameters and 3D positions of the features.

(B) Initial parameter estimation : This processes computes 3D positions of feature points and position and posture parameters of cameras which minimize the sum of squared re-projection errors. In this process, the parameters of the current frame i are computed by using the tracked feature points. The error function E_{init} defined by Eq. (6) is minimized to optimize both the parameters $\mathbf{t}_i$ and R_i of all the frames and 3D positions of all the feature points.

$$E_{init} = \sum_{j\in\mathscr{S}_{ic}} w_j\Phi_{ijc}^2. \quad (6)$$

(C) Local optimization : In this process, the frames from the $(i-(k+l)+1)$-th to the current frame are used to refine the camera parameters from the $(i-(k+l)+1)$ to the $(i-l)$-th frames, as illustrated in Figure 5. This process is designed to use feature points and GPS positions obtained in the frames around the updated frames. To reduce computational cost, this process is performed every k frames. Note that the estimation result is insensitive to the value of l if it is large enough. The constant l is set as tens of frames to use a sufficient number of feature points reconstructed in the process (B). The constant k is set as several frames, which is empirically given so as not to accumulate errors in the initial parameters estimated in the process (B). The weight μ_i $(\in\mathscr{F})$ in which GPS positions are obtained is set as larger number than other frames.

(D) Global optimization : The optimization in the process (C) dose not provide sufficient accuracy for a final output because it is performed for a part of frames and GPS positions for feedback to feature tracking process (A). The purpose of this process is to refine parameters by using tracked features and GPS positions in all the frames. The algorithm of this process is the same as the narrow optimization process (C) when l and k are set as several hundred frames except that divided ranges are independent of each other.

4 Experiments

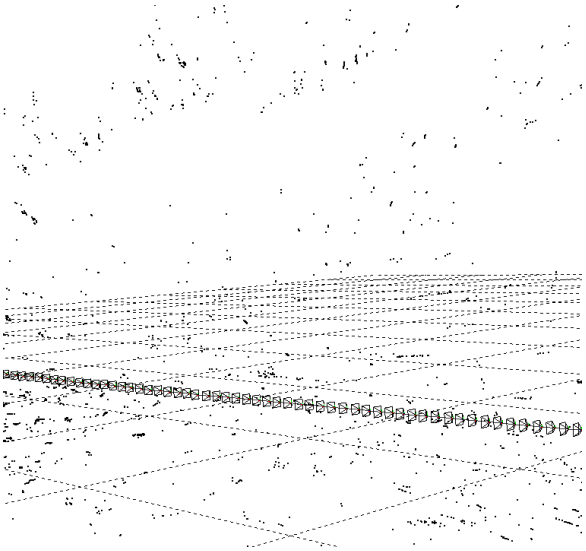
To construct the landmark database, we used Ladybug (resolution of element camera 768x1024, 15fps) and a GPS receiver (Nikon LogPakII, horizontal accuracy $\pm$3.0 cm, vertical accuracy $\pm$4.0 cm) mounted on a car as shown in Figure 6. Captured image sequence consists of 100 frames long with 6 images per each frame (totally 600 images). The distance between the first frame and the last frame is about 40m. Then, the landmark database is created by estimating camera path and 3-D coordinates of natural features. For every landmark, multi-scale image template with three different scales of 15×15 pixels each, is created per each camera position. The number of landmarks created in this experiment is about 10,000, and the number of image templates created per each landmark is 8 on average. Figure 7 shows a part of estimated camera path and 3-D positions of natural feature points in constructing the landmark database.

Next, we have captured a 300 frames long monocular video image sequence (720×480 pixels, progressive scan, 15fps) with a video camera (SONY DSR-PD-150) and camera position and posture are sequentially estimated using the landmark constructed earlier. To give the initial position and posture of the camera, image coordinates of six landmarks are manually specified in the first frame of the input sequence. The maximum number of landmarks selected from the database to correspond with input image is 100 per frame, the window size for detecting natural features from input image is 120×60 pixels, and the number of RANSAC iterations is 500. As a result, processing time for a frame was about 3.4 seconds

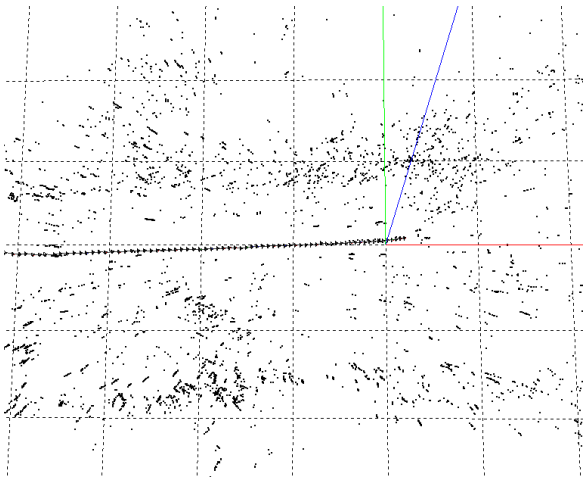
Acquisition vehicle.

GPS receiver (left) and OMS (right).

Figure 6: Equipments for acquisition of images and GPS positions.



Bird's-eye view



Top view.

Figure 7: Estimated camera path of the OMS and 3-D positions of feature landmarks.

with a PC (Intel Pentium4 3GHz CPU $\times$2, 1.5GB RAM). Figure 8 also shows the result of match move; matching virtual 3-D objects to the camera movements using the estimated camera position and posture as shown in Figure 9. It can be observed that the CG person and tank are drawn in geometrically correct positions throughout the sequence.

5 Conclusion

This paper describes a geometric registration method using landmark database generated from omnidirectional videos and GPS positions. The method enables us to obtain absolute positions and postures of a camera. In construction of landmark database, any manual measurement processes are not required. Experiments indicate that geometric registration of a video sequence captured with a general single camera is successfully achieved.

We will experiment the proposed method in a larger environment than the current one. As our future works, for use in augmented reality applications, we must investigate the method to estimate camera position and posture in real-time and to estimate the initial position and posture automatically.

Acknowledgements

This research is partially supported by CoreResearch for Evolutional Science and Technology (CREST) Program "Foundation of Technology Supporting the Creation of Digital Media Contents" of Japan Science and Technology Agency (JST).

References

[1] MR PreViz Project: http://www.rm.is.ritsumei.ac.jp/MR-PreVizProject/top.html.

[2] M. Shin, B.-S. Kim and J. Park: "AR storyboard: An augmented reality based storyboard authoring tool", Proc. 4th IEEE and ACM Int. Symp. on Mixed and Augmented Reality, pp. 98–99 (2005).

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

[4] P. Daehne and J. Karigiannis: "ARCHEOGUIDE: System architecture of a mobile outdoor augmented reality system", Proc. 1st Int. Symp. on Mixed and Augmented Reality, pp. 263–264 (2002).

[5] A. I. Comport, É. Marchand and F. Chaumette: "A real-time tracker for markerless augmented reality", Proc. 2nd IEEE and ACM Int. Symp. on Mixed and Augmented Reality, pp. 36–45 (2003).

[6] E. Rosten and T. Drummond: "Fusing points and lines for high performance tracking", Proc. 10th IEEE Int. Conf. on Computer Vision, Vol. 2, pp. 1508– 1515 (2005).

[7] H. Wuest, F. Vial and D. Stricker: "Adaptive line tracking with multiple hypotheses for augmented reality", Proc. 4th IEEE and ACM Int. Symp. on Mixed and Augmented Reality, pp. 62–69 (2005).

Figure 8: CG person and object superimposed onto captured frames.

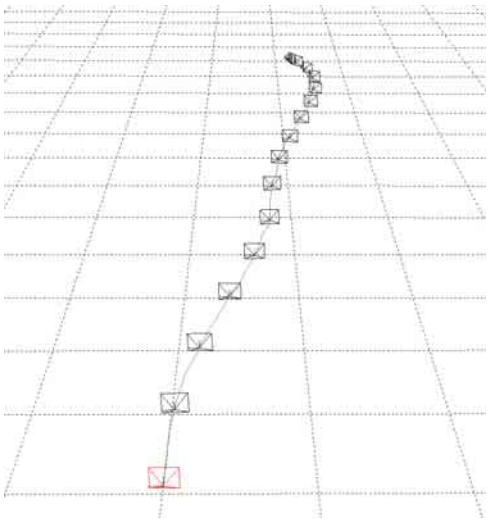


Figure 9: Estimated positions and postures of a handy camera.

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

[9] D. Burschka and G. D. Hager: "V-GPS (SLAM): Vision-based inertial system for mobile robots", Proc. the 2004 IEEE Int. Conf. on Robotics and Automation, pp. 409–415 (2004).

[10] I. Gordon and D. G. Lowe: "Scene modelling, recognition and tracking with invariant image features", Proc. 3rd IEEE and ACM Int. Symp. on Mixed and Augmented Reality, pp. 110–119 (2004).

[11] M. Oe, T. Sato and N. Yokoya: "Estimating camera position and posture by using feature landmark database", Proc. 14th Scandinavian Conf. on Image Analysis, pp. 171–181 (2005).

[12] Y. Yokochi, S. Ikeda, T. Sato and N. Yokoya: "Extrinsic camera parameter estimation based on feature tracking and GPS data", Proc. 7th Asian Conf. on Computer Vision, Vol. 1, pp. 369–378 (2006).

[13] D. Nistér, O. Naroditsky and J. Bergen: "Visual odometry", Proc. 2004 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Vol. 2, pp. 964–971 (2004).

[14] C. Harris and M. Stephens: "A combined corner and edge detector", Proc. 4th Alvey Vision Conf., pp. 147–151 (1988).

[15] M. A. Fischler and R. C. Bolles: "Random sample consensus: A pradigm for model fitting with applications to image analysis and automated cartography", Communications of the ACM, **24**, 6, pp. 381–395 (1981).

[16] S. Ikeda, T. Sato and N. Yokoya: "High-resolution panoramic movie generation from video streams acquired by an omnidirectional multi-camera system", Proc. IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent System, pp. 155–160 (2003).