Extrinsic Camera Parameter Estimation Based-on Feature Tracking and GPS Data

Yuji YOKOCHI[1*], Sei IKEDA[1], Tomokazu SATO[1], and Naokazu YOKOYA[1]

Nara Institute of Science and Technology, Graduate School of Information Science,
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192 Japan
{sei-i,tomoka-s,yokoya}@is.naist.jp
http://yokoya.naist.jp/

Abstract. This paper describes a novel method for estimating extrinsic camera parameters using both feature points on an image sequence and sparse position data acquired by GPS. Our method is based on a structure-from-motion technique but is enhanced by using GPS data so as to minimize accumulative estimation errors. Moreover, the position data are also used to remove mis-tracked features. The proposed method allows us to estimate extrinsic parameters without accumulative errors even from an extremely long image sequence. The validity of the method is demonstrated through experiments of estimating extrinsic parameters for both synthetic and real outdoor scenes.

1 Introduction

Extrinsic camera parameter estimation from an image sequence is one of important problems in computer vision, and accurate extrinsic camera parameters are often required for a widely moving camera in an outdoor environment to realize outdoor 3D reconstruction and new view synthesis [1, 2]. In this field, accumulative errors in estimated camera parameters often cause un-desired effects for each application. This problem is unavoidable as long as we use only relative constraints among multiple frames [3, 4].

To avoid the accumulative error problem, some kinds of prior knowledge about surroundings and external position and posture sensors have been often used in the literatures [5–9]. As prior knowledge about surroundings, known 3D-positions [5, 6] (called feature landmarks) and wire frame of CAD models [7, 8] are used. The method using feature landmarks [5, 6] is based on the feature tracking approach. Extrinsic camera parameters and 3D positions of feature points are estimated by minimizing the re-projection error of feature landmarks and image feature points tracked in each frame. The method described in [7, 8] is based on matching silhouettes of CAD models with edges in input images. Such image based methods do not require any other sensors. However, the acquisition of these kinds of prior knowledge requires much human cost in a large scale

---

[*] Presently at Tochigi R&D Center, Honda R&D Co., Ltd.

outdoor environment. On the other hand, in the method using a sensor combination [9], an RTK-GPS (Real Time Kinematic GPS), a magnetometer and a gyro sensor are sometimes integrated to obtain position and posture data without accumulative errors. However, it is difficult to reconstruct high frequency component in motion by only these sensors because the acquisition rate of position information from a general GPS receiver is 1Hz and is significantly lower than video rate. Moreover, highly accurate calibration and synchronization among sensors is needed but this problem has hardly been treated in the literature.

The most hopeful solution for the accumulative error problem is combination of camera and GPS [10, 11]. In this paper, we propose a method to estimate extrinsic parameters for a widely moving camera using both video sequence and GPS position data. To estimate accurate parameters, our method is based on structure-from-motion with extrinsic parameter optimization using the whole of GPS positions and video frames as an offline process; this is the main difference from the conventional methods described in [10, 11]. In the proposed method, tentative extrinsic parameters are estimated from GPS position data and are used to avoid mismatching in feature tracking. In the optimization process, a new error function defined by using GPS position data and re-projection error is minimized to determine some calibration parameters between camera and sensor. In our method, the following conditions are assumed. (i) Camera and GPS have been already synchronized. (ii) Position relation between camera and GPS receiver is always fixed. (iii) Distance between camera and GPS receiver is known, and direction of GPS receiver in camera coordinate system is unknown. In this paper, it is also assumed that cameras have been calibrated in advance and the intrinsic camera parameters (including lens distortion, focal length and aspect ratio) are known.

In the remainder of this paper, we firstly describe the proposed method that handle GPS position data for estimation of extrinsic parameters in Section 2. In Section 3, the validity of the proposed method is demonstrated through experiments of estimating extrinsic parameters for both synthetic and real outdoor scenes. Finally, we present conclusion and future work in Section 4.

2 Extrinsic Camera Parameter Estimation Using Features and GPS

The goal of this research is to obtain extrinsic camera parameters and a direction of GPS receiver from camera when multiple video frames and GPS positions are given. The main topic described in this section is how to integrate GPS position data to the structure-from-motion problem. In the proposed method, the general structure-from-motion algorithm is enhanced to treat GPS position information.

This method basically consists of feature tracking and optimization of camera parameters as shown in Figure 1. Tow process of (A) feature tracking and (B) initial parameter estimation are performed in order. At constant frame intervals, the local optimization process (C) is done to reduce accumulative errors. Finally, estimated parameters are refined using the tracked feature points and feature
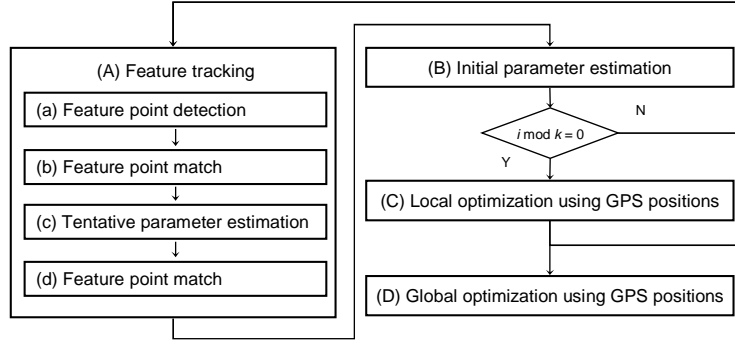
Fig. 1. Procedure of the proposed algorithm.

landmarks 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 limited in a part of the input image sequence because future data cannot be treated in sequential process. On the other hand, in the process (D), all the frames are simply optimized and updated.

In the following sections, we firstly define a new error function that treats both re-projection errors and GPS position errors. After that, each process is also detailed.

2.1 Formulation of Error Function with GPS Position

In this section, we define a new error function E which is combination of the error function concerning GPS and the re-projection error. The way of error minimization will be also mentioned. First, re-projection error is briefly explained as an error function of general structure-from-motion problem. Then, error function concerning GPS is also defined by modeling geometric relation between camera and GPS. Finally, we describe a new error function combining re-projection error and the error function concerning GPS.

Re-projection Error :   Re-projection error is generally used for extrinsic camera parameter estimation based on feature tracking. The method minimizing the sum of squared re-projection error is called bundle adjustment. This error Φ_{ij} is defined as $|\mathbf{q}_{ij} - \hat{\mathbf{q}}_{ij}|$ for feature j in the i-th frame, 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.

Error of GPS :   Generally, if GPS positions and estimated extrinsic parameters do not contain any errors, the following equation is satisfied in the i-th frame among the extrinsic camera parameters (position $\boldsymbol{t}_i$, posture $\boldsymbol{R}_i$), GPS position

$\boldsymbol{g}_i$ and the position of GPS receiver $\boldsymbol{d}$ in the camera coordinate system.

$$\boldsymbol{R}_i\boldsymbol{g}_i + \boldsymbol{t}_i = \boldsymbol{d} \ (i \in \mathcal{F}), \tag{1}$$

where $\mathcal{F}$ denotes a set of frames in which GPS position is obtained. However, if GPS position $\boldsymbol{g}_i$ and extrinsic parameters $\boldsymbol{R}_i$ and $\boldsymbol{t}_i$ contain some errors, we must introduce an error vector $\boldsymbol{n}_i$.

$$\boldsymbol{R}_i\boldsymbol{g}_i + \boldsymbol{t}_i = \boldsymbol{d} + \boldsymbol{n}_i. \tag{2}$$

In this paper, we introduce an error function Ψ_i related to GPS receiver by using the length of the error vector $\boldsymbol{n}$: $\Psi_i = |\boldsymbol{n}_i|$. This function means the distance between the measured position of the GPS receiver and the predicted position of the receiver using the extrinsic parameters $\boldsymbol{R}_i$ and $\boldsymbol{t}_i$ and GPS position. Next, we describe a new error function E which is a combination of the error function Ψ_{ij} related to GPS receiver and the re-projection error Φ.

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

$$E = \frac{\omega}{|\mathcal{F}|} \sum_{i \in \mathcal{F}} \Psi_i^2 + \frac{1}{\sum_i |\mathcal{S}_i|} \sum_i \mu_i \sum_{j \in \mathcal{S}_i} w_j \Phi_{ij}^2, \tag{3}$$

where ω means a weight for Ψ_i, and $\mathcal{S}_i$ denotes a set of feature points detected in the i-th frame. 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. (3) is normalized by $|\mathcal{F}|$ and $\sum_i |\mathcal{S}_i|$ each other so as to set ω as a constant value independent of the number of feature and GPS positioning points.

Note that it is difficult to obtain a global minimum solution because there are a large number of local minima in the error function E. In order to avoid this problem, we currently adopt a method to change the weight μ_i in the iteration of the optimization, which is experimentally derived from computer simulations. In this method, the weight is changed whenever optimization process is converged. We expect that local minima can be avoided because the global minimum does not move largely even if local minima move by changing the weight μ_i.

2.2 Implementation of Each Process

(A) Feature tracking : The purpose of this process 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 [12] and detected feature points are tracked robustly with RANSAC approach. In the first process (a), natural feature points are automatically detected by using the Harris operator for limiting feature position candidates on the images. In the next process (b), every
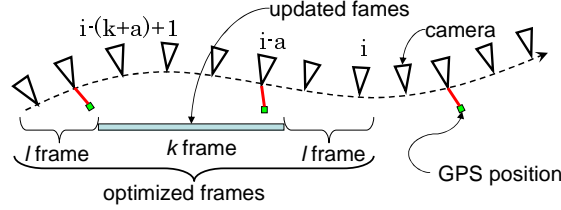
Fig. 2. Optimization frames in the process (C).

feature in the $(i-1)$-th frame is tentatively matched with the candidate feature points in the i-th frame by using a standard template matching. Then, in the third process (c) Tentative extrinsic parameters are then estimated by selecting correct matches using RANSAC approach [13]. In the final process (d), every feature is re-tracked within a limited searching area that can be computed by the tentative extrinsic parameters and 3D positions of the features.

(B) Initial parameter estimation : This procedure computes 3D position of feature points and extrinsic parameters which minimize the sum of squared re-projection errors. In this process, extrinsic parameters of all the frames are refined to reduce the accumulated errors by the bundle adjustment using feature points. The error function E_{init} defined by Eq. (4) is minimized to optimize both extrinsic camera parameters of all the frames and 3D positions of all the feature points.

$$E_{init} = \sum_{h=1}^{i} \mu_h \sum_j w_j \Phi_{hj}^2. \tag{4}$$

(C) Local optimization : In this process, the frames from the $(i-(k+2l)+1)$-th to the current frame are used to refine the camera parameters from $(i-(k+2l)+1)$ to $(i-l)$-th frame, as illustrated in Figure 2. 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 extrinsic parameters estimated in the process (B).

(D) Global optimization : The optimization in the process (C) dose not provide enough accuracy as the final output because it is performed for a part of whole of frames and GPS positions for feedback to feature tracking process (A). The purpose of this process is to refine extrinsic camera parameters by using whole of tracked features and GPS positions. The algorithm of this process is the same as the local optimization process (C) when l is set as zero and k is set as the total number of frames.

3 Experiment

In this section, we demonstrate experiments for both synthetic and real outdoor scenes. First, the experiment for synthetic data is carried out to evaluate the accuracy of extrinsic parameters estimated by the proposed method when the correspondences of feature points are given. The experiment for real data is then demonstrated to confirm the validity of the whole proposed method.

Note that some parameters used in the optimization process (C) and (D) are set as follows. The weight coefficient ω in the error function E defined by Eq. (3) was set as 10^{-9}. When a GPS position was obtained, the weight μ_i of the corresponding frame is always set as 1.0. When it was not obtained, 1.0 and 2.0 were alternately set as the weight μ_i whenever the optimization step was converged. In the local optimization process (C), we set the number of updated frames $k = 5$ and the number of optimized frames 49 ($l = 22$). The positions of the first and 15th frames were set as GPS positions. The postures of these frames were set as the true value for synthetic scene, and as the design value of the car system for real scene.

3.1 Synthetic data

The purpose of this simulation is to evaluate extrinsic parameters estimated in the global optimization process (D). In addition, the validity of the proposed method is confirmed by comparison with the conventional method [6]. We gave a point set as a virtual environment that was used to generate 2D feature positions in synthetic input images. The virtual camera takes 990 images by moving in the virtual environment. The intrinsic parameters of the virtual camera are set the same as the real camera described in the next section. The position of GPS receiver in the camera coordinate system is set as (600,600,600)[mm]. We added errors to input data as follows. The GPS positions with Gaussian noise ($\sigma = 30$ mm) are given every 15 frames. The feature points are projected to the virtual camera, and detected with Gaussian noise ($\sigma = 0.6$ pixel) and quantization error. The initial extrinsic parameters $\boldsymbol{R}_i$ and $\boldsymbol{t}_i$ are generated by adding Gaussian noise (position: $\sigma = 500$ mm, posture: $\sigma = 0.020$ rad) to the ground truth. In the compared method, all the frames is set as key frames in which more than 15 feature landmarks appear. The landmarks are given as feature points whose confidence coefficient is set as large enough, and the 2D positions of the landmarks in each frame are given without any errors. In this simulation, 200 feature points are observed on average in each frame.

Position and posture errors in the simulation result for the synthetic data are shown in Figure 3. In the compared method, the position error is 39.8 mm, and the postures error is 0.0019 rad on average. In the proposed method, the position error is 32.9 mm, and the posture error is 0.0036 rad on average. We have also confirmed this extrinsic parameters obtained in this experiment are not converged to local minima in this simulation.

These results indicate that the proposed method enable us to obtain extrinsic parameters in the same order precision as the conventional method without
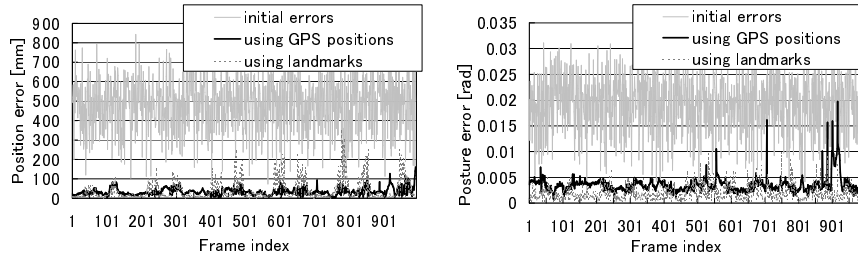
Fig. 3. Position and posture errors of estimated extrinsic parameters.

any manual acquisitions of surrounding information. The difference of the accuracy between the proposed method and the compared one can be caused by the difference of behavior of the given absolute position information such as GPS positions and landmarks. Concretely, we consider that posture errors of the compared method becomes smaller than the proposed one because landmark position information obtained from images is more sensitive to postures of camera than GPS position information.

3.2 Real scene

The purpose of this experiment using real data is to confirm the validity of the proposed method which includes the feature tracking and the error models of feature point detection. In this section, first, we describe the condition of this experiment. After that, two kinds of experimental results are shown.

In the first experiment, we used a video camera (Sony DSR-PD-150, 720x480 pixel, 14.985fps, progressive scan) with a wide conversion lens (Sony DSR-PD-150) and a GPS receiver (Nikon LogPakII, accuracy ±3.0 cm) that were mounted on a car. We acquired 3600 frames and GPS positions while the car was moving 1.1km distance at 16.5km/h. The acquired frames and GPS positions were manually synchronized. Intrinsic parameters are estimated by Tsai's method [14]. The distance between camera and GPS receiver is 1020 mm which is manually measured.

First, to confirm the effect to the process (C), we compared the result of the sequential process of camera parameter estimation using the fully activated proposed method and the proposed method without the process (C). In both methods, the same extrinsic parameters of the first frame and the 15th frame are manually given.

The two comparison of the result of both methods are shown in Figure 4. In the method not using GPS position, the process has been terminated at the 1409th frame because tracked feature points decrease. On the other hand, 300 of feature points on average are tracked at all the frames in the method using GPS positions. This result indicates that the performance of the feature tracking is improved by using GPS positions.

Figure 8 shows the result of extrinsic parameters estimation after the global optimization process (D). In this figure, the camera path is smoothly recovered
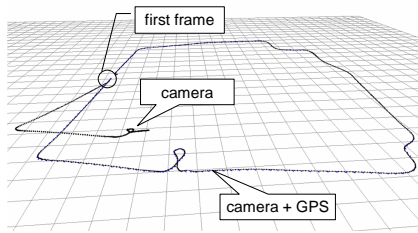
Fig. 4. Accumulative errors of extrinsic parameters.
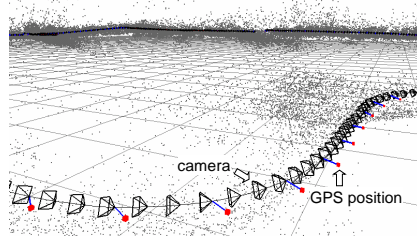
Fig. 5. Result of estimated extrinsic parameters.

even at the frames where GPS positions are not obtained. The match move using the estimated extrinsic parameters is also demonstrated in Figure 6. The virtual objects were inserted to the input images. We have confirmed that estimated extrinsic parameters do not contain fatal errors because the virtual objects seem to be located at the same position in the real environment in most part of the input sequence (`http://yokoya.naist.jp/pub/movie/yokochi/match_move.mpg`).

However, the virtual objects are drifted from the 995th to the 1030th frames as shown in Figure 7. This position drift is due to the multi-path effect of GPS, which is the corruption of the direct GPS signal by one or more signals reflected from the local surroundings. The standard deviation as a degree of confidence of GPS positioning are also obtained from our RTK-GPS receiver. It increases from the 995th to the 1030th frames as shown in Figure 8. To detect the occurrence of the multipath effect, we will explore to design an estimation method using a degree of confidence of GPS positioning.

4   Conclusion

In this paper, we have proposed a method to estimate extrinsic camera parameters of a video sequence without accumulative errors by integrating feature tracking with GPS positions. In the proposed method, GPS position information is used for both feature tracking and optimization of extrinsic parameters.

We have confirmed that the proposed method allows us to obtain extrinsic parameters in the same order precision as the conventional shape-from-motion method using a large number of landmarks in every frame through experiments using both synthetic and real outdoor data. However, the multipath error of GPS is not acceptable for the proposed method. To detect the occurrence of the multipath effect, we will explore to design an estimation method using a degree of confidence of GPS positioning.
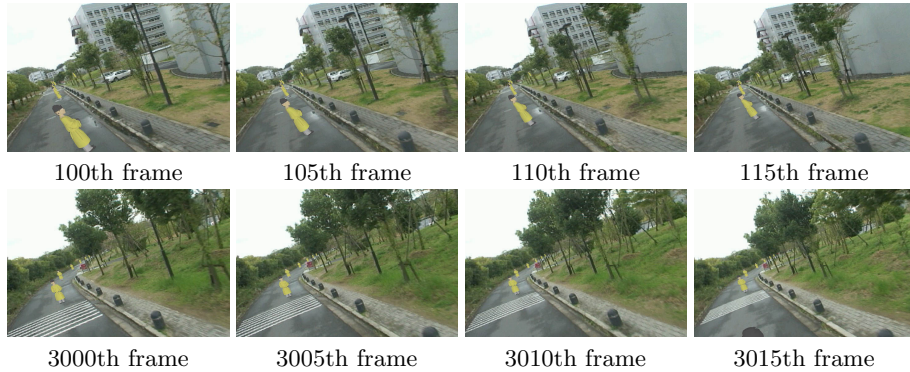
Fig. 6. Match move using estimated extrinsic parameters.
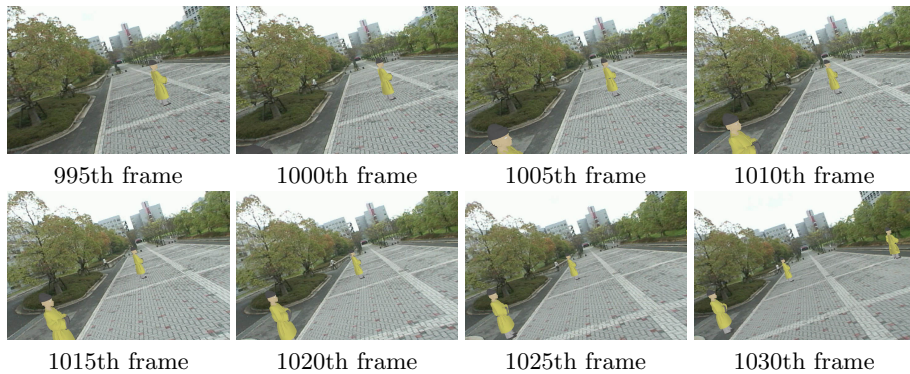
Fig. 7. Examples of incorrect match move.

References

1. Feiner, S., MacIntyre, B., Höllerer, T., Webster, A.: A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment. In: Proc. 1st IEEE Int. Symp. on Wearable Computers. (1997) 208–217
2. D. Kotake, T. Endo, F. Pighin, A. Katayama, H. Tamura, M. Hirose: Cybercity Walker 2001 : Walking through and looking around a realistic cyberspace reconstructed from the physical world. In: Proc. 2nd IEEE and ACM Int. Symp. on Mixed Reality. (2001) 205–206
3. Fitzgibbon, A.W., Zisserman, A.: Automatic camera recovery for closed or open image sequences. In: Proc. 5th European Conf. on Computer Vision. Volume I. (1998) 311 – 326
4. Pollefeys, M., Koch, R., Vergauwen, M., Deknuydt, B., Gool, L.V.: Three-dimentional scene reconstruction from images. In: Proc. SPIE. Volume 3958. (2000) 215–226
5. Davison, A.J.: Real-time simultaneous localisation and mapping with a single camera. In: Proc. 9th IEEE Int. Conf. on Computer Vision. Volume 2. (2003) 1403–1410
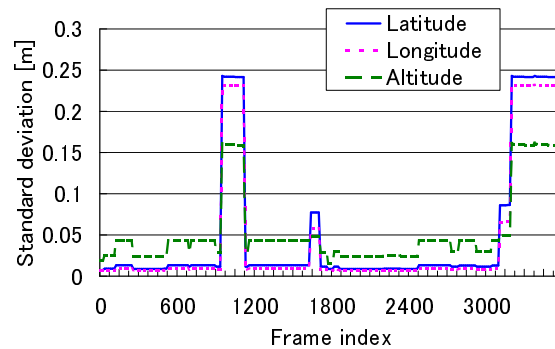
Fig. 8. Standard deviations in GPS data.

6. Sato, T., Kanbara, M., Yokoya, N., Takemura, H.: Dense 3-D reconstruction of an outdoor scene by hundreds-baseline stereo using a hand-held video csamera. Int. Jour. of Computer Vision **47** (2002) 119–129
7. Comport, A.I., Marchand, É., Chaumette, F.: A real-time tracker for markerless augmented reality. In: Proc. 2nd ACM/IEEE Int. Symp. on Mixed and Augmented Reality. (2003) 36–45
8. Vacchetti, L., Lepetit, V., Fua, P.: Combining edge and texture information for real-time accurate 3D camera tracking. In: Proc. 3rd IEEE and ACM Int. Symp. on Mixed and Augmented Reality. (2004) 48–57
9. Güven, S., Feiner, S.: Authoring 3D hypermedia for wearable augmented and virtual reality. In: Proc. 7th IEEE Int. Symp. on Wearable Computers. (2003) 118–126
10. Nistér, D., Naroditsky, O., Bergen, J.: Visual odometry. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. Volume 2. (2004) 964–971
11. Hu, Z., Keiichi, U., LU, H., Lamosa, F.: Fusion of vision, 3D gyro and GPS for camera dynamic registration. In: Proc. 17th Int. Conf. on Pattern Recognition. (2004) 351–354
12. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. Alvey Vision Conf. (1988) 147–151
13. Fischler, M.A., Bolles, R.C.: Random sample consensus: A pradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24** (1981) 381–395
14. Tsai, R.Y.: An efficient and accurate camera calibration technique for 3D machine vision. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (1986) 364–374