Fine-Motion Estimation Using Ego/Exo-Cameras

Taeyoung Uhm, Minsoo Ryu, and Jong-Il Park

Robust motion estimation for human–computer interactions played an important role in a novel method of interaction with electronic devices. Existing pose estimation using a monocular camera employs either ego-motion or exo-motion, both of which are not sufficiently accurate for estimating fine motion due to the motion ambiguity of rotation and translation. This paper presents a hybrid vision-based pose estimation method for fine-motion estimation that is specifically capable of extracting human body motion accurately. The method uses an ego-camera attached to a point of interest and exo-cameras located in the immediate surroundings of the point of interest. The exo-cameras can easily track the exact position of the point of interest by triangulation. Once the position is given, the ego-camera can accurately obtain the point of interest's orientation. In this way, any ambiguity between rotation and translation is eliminated and the exact motion of a target point (that is, ego-camera) can then be obtained. The proposed method is expected to provide a practical solution for robustly estimating fine motion in a non-contact manner, such as in interactive games that are designed for special purposes (for example, remote rehabilitation care systems).

Keywords: Image motion analysis, pose estimation, human–computer interaction, fine-motion estimation.

I. Introduction

Recently, vision-based interfaces, which are widely used in many smart devices [1], are being designed with the help of motion estimation from camera images; for example, a depth camera, in comparison to an RGB camera, can offer different and more useful information for human–computer interaction (HCI) [2].

Motion estimation using either a single camera or multiple cameras has been used in electronic devices to achieve natural HCI. However, the vision-based methods in [1]–[2] are not precise enough for estimating fine motion (for example, breath-induced motion or wrist motion in home environments). To solve this problem, this paper presents a hybrid vision-based pose estimation method for fine-motion estimation that takes advantage of both an ego-camera and exo-cameras, as shown in Fig. 1. The proposed method can be roughly divided into ego-camera-based and exo-camera-based pose estimation. Pose estimation using an ego-camera is mainly employed for object self-motion (for example, robots, vehicles, and so on) [3]–[4]. The ego-motion-based pose estimation methods in [3]–[4] for estimating self-position use a large number of

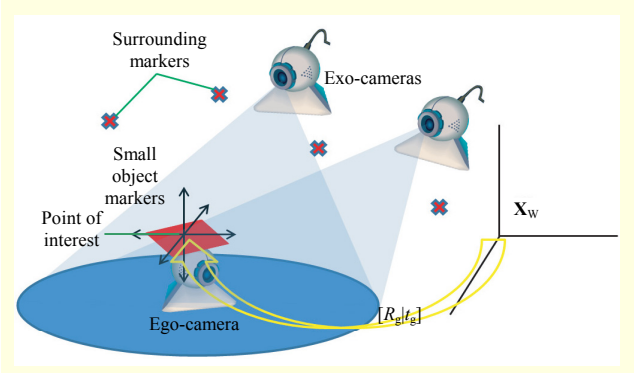

Fig. 1. Example configuration of ego-camera and exo-cameras.

Manuscript received May 7, 2014; revised Mar. 23, 2015; accepted Apr. 1, 2015.

Taeyoung Uhm (uty02@mr.hanyang.ac.kr), Minsoo Ryu (msryu@hanyang.ac.kr), and Jong-Il Park (corresponding author, jipark@hanyang.ac.kr) are with the Department of Computer Science and Engineering, Hanyang University, Seoul, Rep. of Korea.

background markers or features.

The estimation of ego-motion using a vision method that uses sequential stereo images [5] is limited by the inherent ambiguity of rotation and translation for fine motion [6].

Exo-camera-based pose estimation by fixed external observation cameras is popular for HCI [7]. This method of estimating a user's pose involves using attached body markers or depth information with trained body part locators (for example, limbs, torso, head, and so on). Nevertheless, the fact still remains that it is difficult to detect and estimate the exact rotations of a small target object due to the difficulty in distinguishing the object's fine motions. Moreover, accurate pose estimation often requires complex optimization using non-linear equations [8]–[10].

As the ego-camera-based and exo-camera-based approaches have distinct advantages and disadvantages, it is necessary to develop an effective method that utilizes the advantages of both approaches. The main contribution of this paper lies in developing an efficient algorithm that incorporates the advantages of both systems.

In this paper, we propose a hybrid vision-based pose estimation method for fine-motion estimation that can provide a high degree of accuracy by combining both of the aforementioned approaches. Therefore, in the proposed method, as it has such advantages, there is no need to solve complex non-linear equations to ensure greater accuracy for real-time estimation. The proposed method first estimates the exact position of an ego-camera from two or more given exo-cameras. Then, it accurately estimates the fine rotation of the ego-camera by simple computation using the determined position. For this investigation, the proposed method was compared with a state-of-the-art depth camera by error analysis tasks.

The remainder of this paper is organized as follows. Section II describes a robust pose estimation method. Section III demonstrates our experimental results and analysis. Finally, Section IV presents our conclusion and describes future work.

II. Robust Hybrid Vision-Based Pose Estimation Method

Camera pose estimation using a calibrated camera involves finding the six external parameters of a point of interest; that is, the relative position and orientation of an ego-camera with respect to a world coordinate system.

Figure 1 illustrates a configuration of the hybrid vision method. External parameters corresponding to the ego-camera are denoted by R_g (rotation) and t_g (translation). The rotation and translation of the ego-camera can be calculated from the coordinates of the surrounding markers seen by the ego-camera. However, these motions are vulnerable to noise (observation error of a marker position in an ego-camera image). Thus, this paper introduces exo-cameras that observe target objects from the immediate surroundings of a point of interest. Therefore, it is a relatively simple procedure to extract the exact position of the ego-camera by triangulation using two or more of these exo-cameras. The value of translation (t_g^e) is then used to calculate the exact rotation of the ego-camera. Since t_g^e is highly accurate, the exact rotation of the ego-camera can be easily calculated, without ambiguity. Thus, this kind of hybrid system allows for the accurate attainment of the positions and orientations of small objects.

In this paper, we use markers (feature points with known positions) for pose estimation without a loss of generality. However, it is possible to apply the proposed method to any point of interest in an image space.

The projective image coordinate of the ith point, x_{g_i}, is related to the world coordinate of the ith point, $\mathbf{X}_{Wi} = [X_i, Y_i, Z_i, 1]^T$, as follows:

$$x_{g_i} = \mathbf{P}\mathbf{X}_{Wi} = \mathbf{K}[R_g \mid t_g]\mathbf{X}_{Wi}, \tag{1}$$

where $\mathbf{K}$ denotes the camera matrix belonging to the ego-camera. Assuming that $\mathbf{K}$ is calibrated, ego-motion can be estimated using the following minimization:

$$\min_{R_g, t_g} \sum_i \left\| x_{g_i} - \mathbf{K}[R_g \mid t_g]\mathbf{X}_{Wi} \right\|. \tag{2}$$

In the above problem, the obtained R_g and t_g tend to have considerable error due to the inherent ambiguity of rotation and translation. However, if a translation (t_g^e) is obtained accurately by using exo-cameras, then the forthcoming rotation (R_g') obtained from the ego-camera will be more accurate than R_g. The aforementioned forthcoming rotation can be calculated from the following:

$$x_{g_i} = \mathbf{K}[R_g \mid t_g]\mathbf{X}_{Wi} = \mathbf{K}[R_g' \mid t_g^e]\mathbf{X}_{Wi}, \tag{3}$$

where R_g' indicates the value of the updated rotation based on the exact translation (t_g^e) given by the exo-cameras. Suppose that we have n marker observations. Then, the solution of R_g' can be obtained by the following minimization:

$$\min_{R_g'} \left\| R_g' C - D \right\|, \tag{4}$$

where

$$C = [C_1, C_2, \ldots, C_n]; C_i = [I \mid 0]\mathbf{X}_{Wi} = [X_i, Y_i, Z_i]^T;$$

$$D = [D_1, D_2, \ldots, D_n]; D_i = (\mathbf{K}^{-1} x_g - t_g^e).$$

The solution of (4) can be found as follows [11]. First, we define a 4 by 4 matrix, $\mathbf{B}$, by

$$\mathbf{B} = \sum_{i=1}^{3} \mathbf{B}_i^T \mathbf{B}_i, \tag{5}$$

where

$$\mathbf{B}_i = \begin{bmatrix} 0 & (C_i - D_i)^{\mathrm{T}} \\ D_i - C_i & [D_i + C_i]_X \end{bmatrix}.$$

Here, $[.]_X$ denotes a mapping from a three-dimensional vector to a 3 by 3 matrix, which is the solution of the rotation matrix in (4).

III. Experimental Results

1. Environments

Figure 2 shows an experimental environment, whereby the featured ego-camera is able to move without error. From the figure, an ego-camera can be seen located on a rail. Two exo-cameras are fixed on opposite sides of the ego-camera. The baseline of the exo-cameras is 110 mm, and the distance between the ego-camera and the exo-cameras is 0.8 m.

For comparison between a single and hybrid method, experiments were performed with an optical board and instruments (see Fig. 2(d)). Images from the ego-camera and exo-cameras were captured simultaneously (see Fig. 3). For convenience, we used markers in our experiments, but our method can be used in conjunction with any points of interest

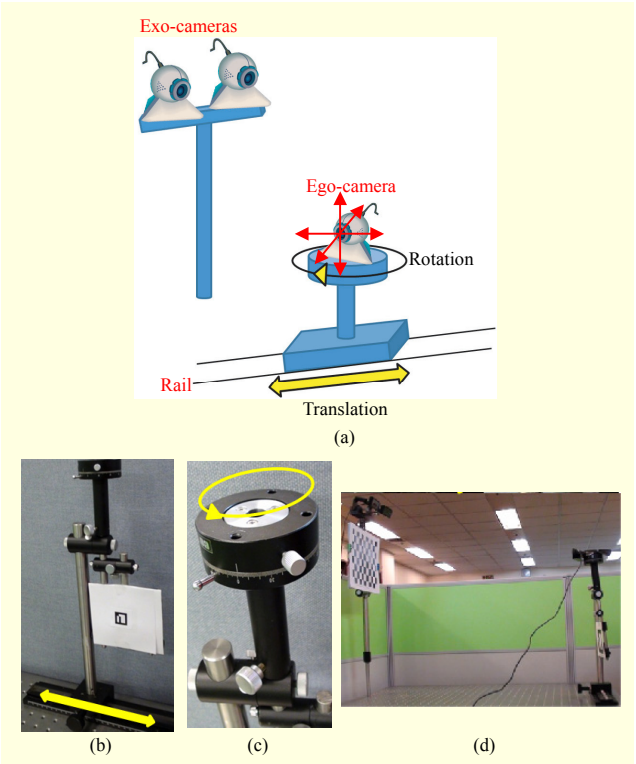

Fig. 2. Experimental environment: (a) overview; (b) ego-camera and marker on rail; (c) ego-camera tumblers for rotation; and (d) experimental image.

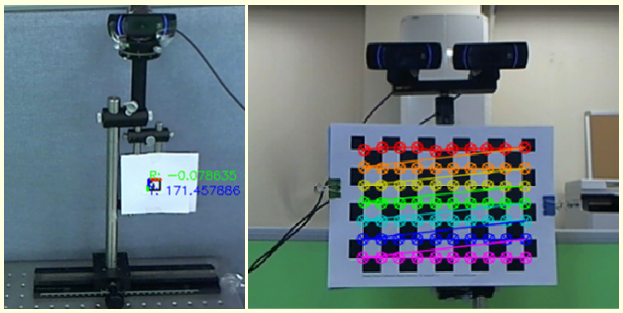

Fig. 3. Camera images from: (a) exo-camera and (b) ego-camera.

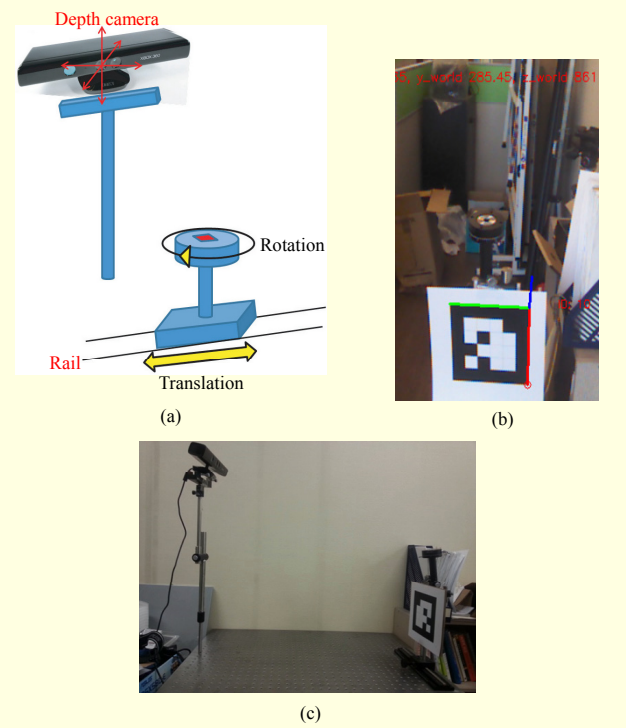

Fig. 4. Depth camera images from: (a) overview, (b) exo-camera, and (c) experimental image.

in an image space.

Next, Fig. 4 shows another experimental environment, whereby a depth camera is fixed and markers are moved in the same way as above. A depth camera is featured as opposed to exo-cameras; there is no ego-camera. The purpose of experiments carried out under this environment is to be able to investigate any systematic errors found to be in the depth camera and to then compare the findings with the results obtained from the experiments carried out under the proposed method in the experimental environment featured in Fig. 2. The results from the experiments carried out in the experimental environment featured in Fig. 4 were averaged.

ETRI Journal, Volume 37, Number 4, August 2015
http://dx.doi.org/10.4218/etrij.15.0114.0525

2. Analysis

For error analysis, four cases of estimating the motion of the ego-camera were compared: *exo*, in which a single exo-camera observes a small square marker (65 mm by 65 mm); *ego*, in which an ego-camera observes a relatively large square marker (150 mm by 150 mm); *depth*, in which a depth camera observes a small square marker position and estimates its orientation from a single exo-camera and depth information; and *hybrid*, in which a stereo exo-camera system calculates a marker's position by triangulation and an ego-camera estimates its orientation using this position data (This is the proposed hybrid method.). First, to compare the accuracy and noise sensitivity of each method, we estimated the external parameters of each case after adding Gaussian random noise (GRN) with mean zero and standard deviation (σ = 0.3 to 0.5) to the image coordinates of the markers, as shown in Table 1. The rotation errors were calculated by root-mean-square error (RMSE). The results show that the single exo-camera case, exo, gives a poor performance. The depth case shows noise-sensitive results, and the hybrid case demonstrates the best performance, as expected. Figure 5 shows the depth and hybrid systems' improvement gains as the rotation error of the exo system is varied. To evaluate the similarity of these two systems, we use a normalized sum of RMSE as follows:

$$P = \frac{1}{N} 10 \log \sum \left(\sqrt{\frac{(\text{exo})_e^2}{(.)_e^2}} \right), \quad (6)$$

where $(.)_e$ is the error for each method and N is the total number of frames.

Consequently, the results show that the use of a depth camera for fine-motion estimation is an improvement over using the correction based on depth information; however, a depth camera is more sensitive to noise. The improvement gain was calculated by dividing the depth and hybrid results over those from the exo case.

This paper next analyzes the accuracy and noise sensitivity of the translation parameters (see Table 2). The ego case shows a poorer performance than the depth and hybrid cases, which demonstrate good performance, as expected. Furthermore, the depth camera has measurement resolution and error properties (for example, depth hole) [12], but we display the best result by GRN. Thus, the accuracy of a translation is not guaranteed in this case due to the inherent ambiguity of the ego-motion. Figure 6 shows that a depth camera can practically reduce the translation error, of a fine motion, in the *z*-axis, but the hybrid method has the best performance over all three axes (*x*-, *y*-, and *z*-axis).

To demonstrate the effect of using an exact translation when estimating the fine motion (for example, 50 mm movement along *x*-axis) of the ego-camera, the ego, depth, and hybrid

Table 1. Results of RMSE for GRN (σ = 0.3 to 0.5).

Camera	σ	Exo	Depth	Hybrid
Rotation error (rad)	0.3	0.027405	0.001876	0.001166
	0.4	0.037292	0.015975	0.001847
	0.5	0.056390	0.052769	0.001620

Table 2. Results of RMSE for GRN (σ = 0.5).

Camera	Axis	Ego	Depth	Hybrid
Translation error (mm)	*x*	0.128101	0.097565	0.020732
	y	0.110062	0.046205	0.033204
	z	0.520179	0.077950	0.038168

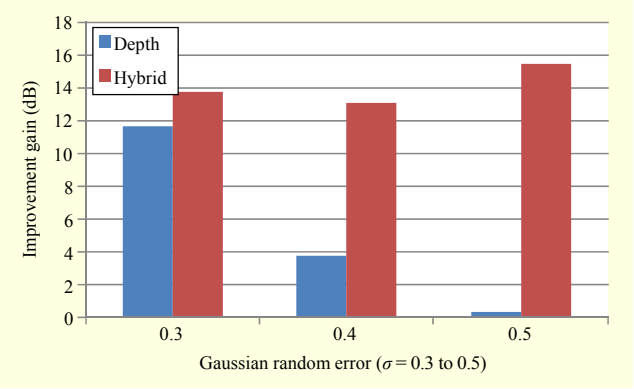

Fig. 5. System improvement gain in comparison with single exo-camera and varying GRN: σ = 0.3 to 0.5 (depth/hybrid).

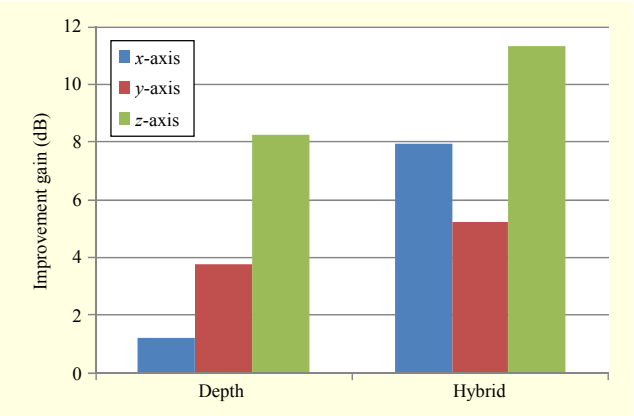

Fig. 6. System improvement gain in comparison with ego-camera and GRN σ = 0.5 (depth/hybrid): depth camera has noise model, but we display only best result [12].

Table 3. RMSE comparison for estimating fine motion movement.

Camera	Axis	Ego	Depth	Stereo [5]	Hybrid
Rotation error (rad)	x	0.010448	0.006430	0.002803	0.000684
	y	0.036996	0.019655	0.014920	0.004468
	z	0.002765	0. 002249	0.001141	0.001233

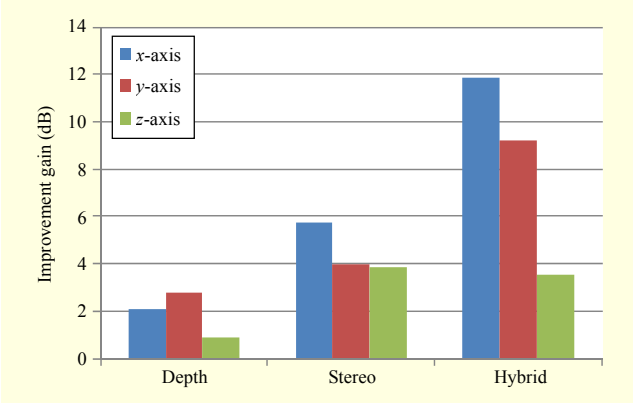

Fig. 7. System improvement rate comparison: ego-camera is moved 50 mm along x-axis (depth/stereo/hybrid).

cases were compared (see Table 3). It is difficult to determine the ground truth of the fine motion (that is, external camera parameters). Thus, this paper bypasses the problem by introducing controllable relative motion. The implications for motion are estimated by the external parameters with the ego-camera at a given position by moving the camera 50 mm in the x-direction. Since there is no rotation, the rotation matrix in each of the three cases should remain the same. The angle difference before and after motion represents the amount of rotation error. As shown in Table 3, exchanging the translation by exo-cameras has a dramatic effect. The rotation error about the y-axis is dominant in this case. The horizontal movement seems to be misinterpreted as rotation about the vertical axis, which is a typical example of a motion ambiguity. Furthermore, the hybrid case shows the best performance, as expected.

The proposed method shows the rotation error results in less than 0.005 rad, and the average system improvement gain is 8.6 dB, as shown in Fig. 7.

This study experimentally investigates the influence of marker size on the accuracy of motion estimation when ground truth data is obtained by movement without rotation, as shown in Fig. 8. As the marker size decreases, the rotation error grows in each case: the stereo, depth, and hybrid cases.

The noise model derived from the errors of the depth camera considers both axial and lateral noise distributions, which are calculated from the distances and angles of the camera in

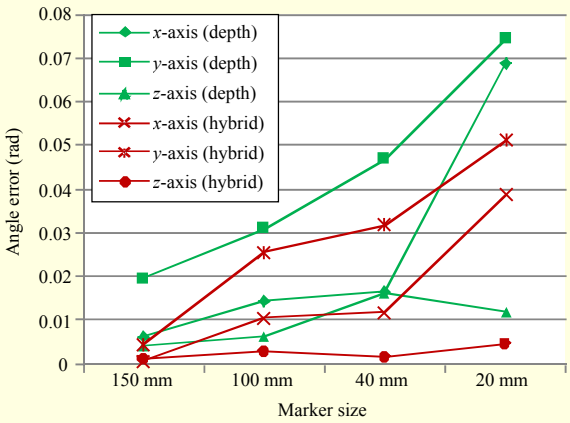

Fig. 8. Influence of marker size on rotation accuracy: markers move 50 mm along x-axis (depth/hybrid).

relation to the markers. [13]. Due to the noise model, the stereo vision-based ego-motion estimation method produced slightly better results than the depth camera method. In realistic situations where large markers are available, the proposed hybrid method demonstrates considerable enhancement compared with the stereo and depth camera cases.

The results confirm that the proposed method is more accurate for pose estimation than the depth camera–based method. Moreover, the computational complexity of the hybrid method is very low because the entire problem is linear. Thus, the proposed hybrid method is well suited to extracting accurate motion information in real time.

IV. Conclusion and Future Work

This paper presents a hybrid pose estimation method based on an ego-camera and exo-cameras without the need for complex computation of non-linear equations. The hybrid system takes advantage of both ego-camera and exo-camera systems, demonstrating superior performance with reduced computational complexity. The method is applied to a typical rehabilitation application that requires accurate body motion, and demonstrates its usefulness. This study expects the method can be applied to a variety of applications that require fine and accurate motion estimation.

We are currently conducting rigorous performance analyses using the Cramer-Rao lower bound. In the future, the generalization of this framework to other scenarios will be explored, as well as expansion to incorporating different visual cues, such as optical flow and shading.

References

[1] S. Wang, H. Yu, and R. Hu, "3D Video Based Segmentation and

Motion Estimation with Active Surface Evolution," *J. Signal Process. Syst.*, vol. 71, no. 1, Apr. 2013, pp. 21–34.

[2] K. Khoshelham and S.O. Elberink, "Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications," *Sensors*, vol. 12, no. 2, Feb. 2012, pp. 1437–1454.

[3] C. Lundquist and T.B. Schon, "Joint Ego-Motion and Road Geometry Estimation," *Inf. Fusion*, vol. 12, no. 4, Oct. 2011, pp. 253–263.

[4] J. Weng, Y. Cui, and N. Ahuja, "Transitory Image Sequences, Asymptotic Properties, and Estimation of Motion and Structure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 5, May 1997, pp. 451–464.

[5] A. Seki and M. Okutomi, "Ego-Motion Estimation by Matching Dewarped Road Regions Using Stereo Images," *IEEE Int. Conf. Robot. Autom.*, Orlando, FL, USA, May 15–19, 2006, pp. 901–907.

[6] G. Adiv, "Inherent Ambiguities in Recovering 3D Motion and Structure from a Noisy Flow Field," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 5, May 1989, pp. 477–489.

[7] R. Poppe, "Vision-Based Human Motion Analysis: An Overview," *Comput. Vis. Image Understanding*, vol. 108, no. 1, Oct. 2007, pp. 4–18.

[8] J. Weng, T.S Huang, and N. Ahuja, "Motion and Structure from Two Perspective Views: Algorithms, Error Analysis, and Error Estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 5, May 1989, pp. 451–476.

[9] J. Weng, N. Ahuja, and T.H. Huang, "Optimal Motion and Structure Estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, Sept. 1993, pp. 864–884.

[10] G. Schweighofer and A. Pinz, "Robust Pose Estimation from a Planar Target," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, Dec. 2006, pp. 2024–2030.

[11] J. Weng, T.S. Huang, and N. Ahuja, "*Motion and Structure from Image Sequences*," New York, NY, USA: Springer Publishing Company, 1993, pp. 1–444.

[12] A. Fossati et al., "*Consumer Depth Cameras for Computing Vision: Research Topics and Applications*," Google eBook, 2012.

[13] C.V. Nguyen, S. Izadi, and D. Lovell, "Modeling Kinect Sensor Noise for Improved 3D Reconstruction and Tracking," *Int. Conf. 3D Imag., Modeling, Process., Vis. Transmission*, Zurich, Switzerland, Oct. 13–15, 2012, pp. 524–530.

Taeyoung Uhm received his BS degree in electrical engineering from Kyonggi University, Suwon, Rep. of Korea, in 2004 and his MS degree in electronics and computer engineering from Hanyang University, Seoul, Rep. of Korea, in 2006. He is currently pursuing a PhD degree in electronics and computer engineering from the Division of Electronics and Computer Engineering, Hanyang University, Seoul, Rep. of Korea. His current research interests include multimodal 3-D motion estimation, emotion recognition, and human–computer interaction.

Minsoo Ryu received his BS degree in control engineering from Seoul National University, Rep. of Korea, in 1995. He received his MS and PhD degrees in electrical and computer engineering from Seoul National University in 1997 and 2002, respectively. He is currently an associate professor with the Department of Computer Science and Engineering at Hanyang University, Seoul, Rep. of Korea. His research has been centered on real-time embedded systems. Specific research areas include real-time system design and analysis methodology; real-time operating systems and middleware; software engineering for embedded software; and multicore embedded systems. He has served on a number of program committees including IEEE, RTAS, and IEEE RTCSA.

Jong-Il Park received his BS, MS, and PhD degrees in electronics engineering from Seoul National University, Rep. of Korea, in 1987, 1989, and 1995, respectively. From 1996 to 1999, he was a researcher with the Advanced Telecommunications Research Institute International Media Integration and Communication Research Laboratories, Kyoto, Japan. In 1999, he joined the Department of Electrical and Computer Engineering, Hanyang University, Seoul, Rep. of Korea, where he is currently a professor. His research interests include computational imaging, augmented reality, 3-D computer vision, and human–computer interaction.