Upper Body Gesture Recognition for Human-Robot Interaction

Chi-Min Oh, Md. Zahidul Islam, Jun-Sung Lee, Chil-Woo Lee, and In-So Kweon

Chonnam National University, Korea,
Korea Advanced Institute of Science and Technology, Korea
{sapeyes,zahid,aliesim}@image.chonnam.ac.kr,
leecw@chonnam.ac.kr,
iskweon@kaist.ac.kr

Abstract. This paper proposes a vision-based human-robot interaction system for mobile robot platform. A mobile robot first finds an interested person who wants to interact with it. Once it finds a subject, the robot stops in the front of him or her and finally interprets her or his upper body gestures. We represent each gesture as a sequence of body poses and the robot recognizes four upper body gestures: “Idle”, “I love you”, “Hello left”, and “Hello right”. A key pose-based particle filter determines the pose sequence and key poses are sparsely collected from the pose space. Pictorial Structure-based upper body model represents key poses and these key poses are used to build an efficient proposal distribution for the particle filtering. Thus, the particles are drawn from key pose-based proposal distribution for the effective prediction of upper body pose. The Viterbi algorithm estimates the gesture probabilities with a hidden Markov model. The experimental results show the robustness of our upper body tracking and gesture recognition system.

1 Introduction

For a long time, gesture recognition has been a practical and useful research for human-robot interaction (HRI). Implementing the gesture recognition method, a robot system firstly has to segment the pose of user action and then infer a meaning of the sequential pose as a gesture.

The pose recognition can perform in two ways. The first way is so called feature-based method [1] and it obtains the visual features with about pose manifolds from an image sequence. Many interesting algorithms such as principle component analysis (PCA) and independent component analysis (ICA) have been applied for feature analysis, and nonlinear and strong pattern classifiers such as multi-layer perceptron (MLP) and support vector machine (SVM) have been used for pose recognition. However, feature-based method has limitations of generosity; it cannot be applied for arbitrary poses which did not participate in training procedure, and the poses of different people not adjusted to the body structure of a specific-sized person.

By contrast, the second way is model-based method [2,3,4] and it adjusts to any pose of inter persons with a 2D parts-based body model which is made for a specific-sized person but can accept the some variance of inter-person poses. This model

J.A. Jacko (Ed.): Human-Computer Interaction, Part II, HCII 2011, LNCS 6762, pp. 294–303, 2011.
© Springer-Verlag Berlin Heidelberg 2011

describes the body pose as a configuration vector. Estimating a correct configuration vector of the body model in an image determines the pose of an acting user but is nontrivial due to the high degree of freedom (DOF). Therefore, some of robust tracking methods are used; so called linear tracking method, the Kalman filter, as in [5,6], but results in inconsistent tracking for pose estimation. Nonlinear tracking method, the particle filter, as in [7,8] have been applied for this issue. Adopting particle filter can cover the problem of high DOF but the computational cost increases because it needs the great number of particles (hypotheses) to cover the high DOF.

Most of tracking methods usually predict multiple hypotheses based on the assumption of Markov process which is the critical reason for creating many particles. To overcome the problem of high DOF with Markov chain model, key poses can be one of the efficient clues for prediction with smaller particles and we assume that the similarities between key poses and the input image can be a better proposal distribution, as in [9] where we usually draw particles in the prediction step of particle filtering.

This paper proposes a new proposal distribution which is based on key poses and Makov chain model which is still used to recycle the good particles from the previous tracking step. By using both ways for the proposal distribution, we track upper body poses robustly and obtain the pose sequence as a result. In the pose sequence, each tracked pose is named with pose number determined from the number of the closest key pose. For gesture recognition we use HMM which represents the relationship between gesture states and pose events. We use four gestures already explained before, so this means four gesture states are defined in HMM, however, we cannot understand the gestures directly from body tracking. We need to transform the pose sequence into a gesture sequence using HMM and the Viterbi algorithm.

This paper is organized as follows: Section 2 describes our mobile robot platform for HRI. Thus, it describes the HRI operation with two steps: user selection and gesture recognition stages. In the user selection stage, from the omnidirectional view of its omni-camera, the robot is eager to find a person who wants to interact with the robot. In the gesture recognition stage, robot interprets the upper body motion as a gesture. After that, we explain the gesture recognition algorithm. Section 3 describes the model of upper body pose. Section 4 presents the upper body tracking algorithm using particle filtering. We define our key pose-based proposal distribution method in this section. Section 5 briefly describes the gesture recognition process. Section 6 shows the experimental results and section 7 concludes our work.

2 HRI in Mobile Robot Platform

In this work we use *NRLAB02 mobile platform* [10] which is developed by RedOne Initiative Inc. The aim of our work is to construct a prototype of an intelligent HRI system with visual interaction. Therefore the upper row of mobile robot is equipped with visual sensors and a computer. We installed two visual sensors: Microsoft's Kinect camera [11](TOF camera) and an ordinary Omni-camera as shown in Fig. 1. The Kinect camera is used to get the silhouette image of nearby people.

Our upper body tracking algorithm utilizes the silhouette image. The other camera, omni-camera implemented with a hyperbolic mirror made by Eizoh Inc., captures the panoramic images with the FOV of 360 degrees. The robot selects the interested person from the panoramic images.

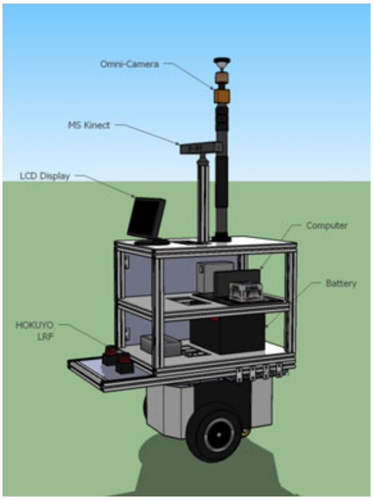


Fig. 1. Our mobile robot platform has omnidirectional camera for interaction-user selection and MS Kinect for gesture recognition

The scenario of HRI is as follows. At first, the robot finds the interested person who wants to interact with it. This is called “User Selection Stage”. Then, the tobot interprets the intention of the person’s motion with the meaning of tracked body poses and we call this procedure “Gesture Recognition Stage”.

2.1 User Selection Stage

We assume that users are pedestrians appearing in panoramic images of omni-camera. HOG-SVM detector, as in [12] detects the nearby pedestrians but it is quiet slow to find the pedestrians from a whole image. As pedestrians usually move around the robot, we assumed that moving area can possibly be the candidate areas of pedestrians to reduce the search area. However, the egomotion of the robot distorts the segmentation of moving region. So we use KLT tracker to estimate the egomotion of the robot [13] and obtain the egomotion-compensated frame difference. Therefore we can detect the pedestrians from only moving areas in a short time.

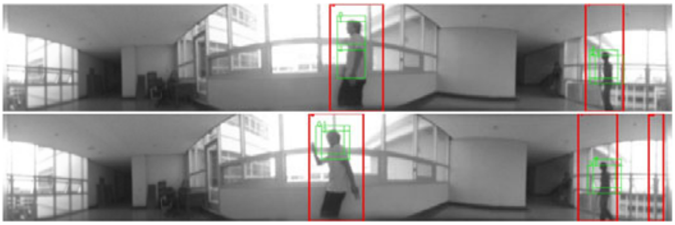


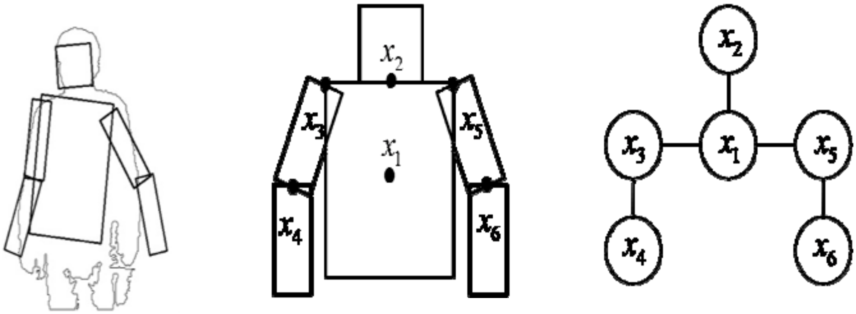
Fig. 2. Pedestrian detection in moving object areas

2.2 Gesture Recognition Stage

After user selection among pedestrians, the user can show the intention by upper body gestures. We recognize four upper body gestures: “Idle”, “I love you”, “Hello left” and, “Hello right”. The gestures are recognized from the pose sequence. To get the pose sequence we use our model-based upper body tracking. In this model-based approach, we define a 2D-parts-based upper body model.

3 Upper Body Model

We define an upper body model based on Pictorial Structures (PS) [2]. PS is known for an efficient method for tracking deformable objects. PS describes the deformable objects with 2D parts and each part is connected to its parent part with joint location, as shown in Fig. 3(b). Additionally the connected parts can be represented as a graphical structure, as shown in Fig. 3(c). Around joint locations all parts can translate and rotate. The benefit of PS model is that any pose of upper body can be made with 2D parts but 3D information is not regarded in PS. PS model can represents a sufficient number of 2D poses for gesture recognition. In addition to the benefit of PS, each pose of PS can be considered as a 2D template image. As shown in Fig. 3(a), by overlaying PS model on the silhouette image, we can simply measure the similarity between the silhouette image and 2D template image of PS model using frame difference or chamfer matching, as in [14].



(a) (b) (c)

Fig. 3. The upper body model: (a) An overlaid model on a silhouette, (b) the joint locations of all parts, and (c) the representation of graphical relation between parts

Fig. 4 shows the body part of upper body model as $x_i=\{(x,y),(dx,dy),\theta,(w,h)\}$ parameters: joint location *(x,y)*, spring-like displacement *(dx,dy)*, orientation θ and rectangular sizes *(w,h)*. Among the parameters, the joint location *(x,y)* cannot change by its part and only can change by parent part’s location. This constraint makes each part connected with its parent. On the other hand, the only way of moving around its parent is related to *(dx,dy)* displacement vector. This is so called spring-like displacement determined by in Gaussian distribution with the mean *(x,y)* and a variance which is a control parameter. In addition to the displacement, the part can rotate with the orientation parameter θ based on the rotation origin, *(x,y)*.

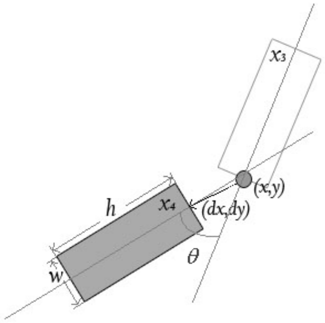


Fig. 4. The parameters of child part x_4 which is connected to parent part x_3

4 Upper Body Tracking

4.1 Particle Filtering

The assumption of most tracking methods is that the current information must be similar to and dependent on the previous information at each time step. Particle filter estimates a proper posterior distribution by updating the posterior distribution at previous tracking time step. The proposal distribution predicts the current posterior distribution from previous posterior distribution with discrete and weighted particles. In Eq. (1), the posterior distribution $p(x_{t-1}|y_{t-1})$ at previous time step represents itself with discrete and weighted particles, where $x^{(i)}_{t-1}$ is ith particle, $w^{(i)}$ is the weight of ith particle, and N is the total number of particles.

$$p(x_{t-1} \mid y_{t-1}) \approx \{(x_{t-1}^{(i)}, w_{t-1}^{(i)})\}_{i=1}^{N} \tag{1}$$

As in [8], particle filter has two steps: prediction and update. In the prediction step, the previous posterior distribution is marginalized to eliminate x_{t-1} and to be updated to x_t based on transition model $p(x_t|x_{t-1})$, Markov chain model.

$$p(x_t \mid y_{t-1}) = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid y_{1:t-1}) dx_{t-1} \tag{2}$$

In the update step, the posterior distribution is reformulated to adjust to the current observation y_t. Based on Baye's rule, posterior distribution $p(x_t|y_t)$ is represented with the likelihood $p(y_t|x_t)$ and prior distribution $p(x_t|y_{t-1})$.

$$p(x_t \mid y_{1:t}) = \frac{p(y_t \mid x_t) p(x_t \mid y_{1:t-1})}{p(y_t \mid y_{1:t-1})} \tag{3}$$

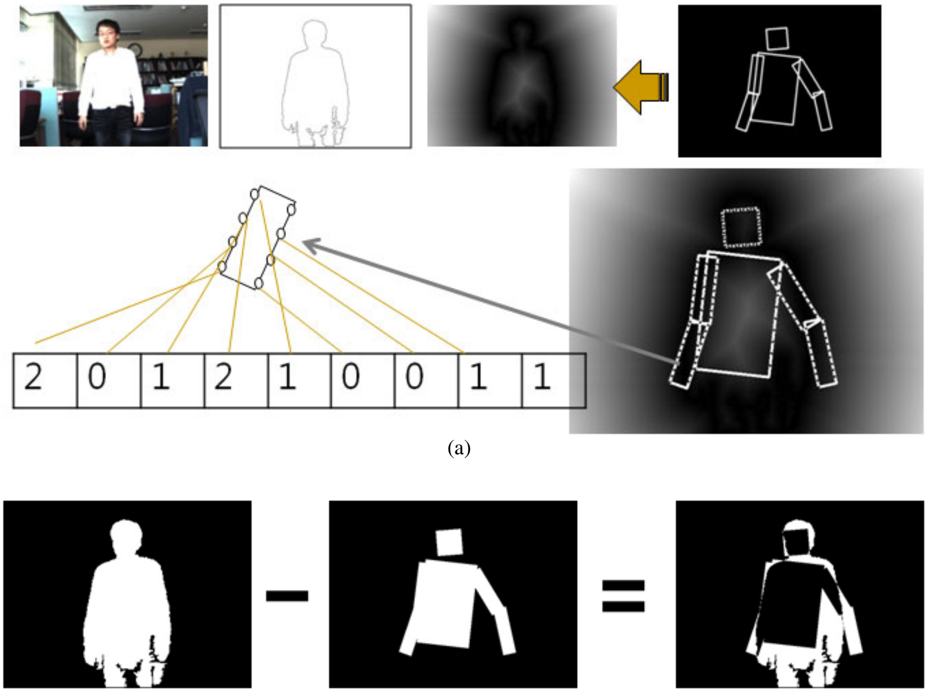
In addition to the prediction step, the particles for x_t prediction are drawn from the proposal distribution $q(x_t|x_{t-1},y_t)$ and in the update step the weights of particles are determined by below:

$$w_t^{(i)} = w_{t-1}^{(i)} \frac{p(y_t \mid x_t^{(i)}) p(x_t^{(i)} \mid x_{t-1}^{(i)})}{q(x_t^{(i)} \mid x_{t-1}^{(i)}, y_t)} \tag{4}$$

In the process of weighting particles, likelihoods of particles are measured. Our likelihood $p(y_t|x^{(i)}_t)$ is a joint likelihood with edge and silhouette likelihoods.

$$p(y_t \mid x_t^{(i)}) = p(I_S \mid x_t^{(i)})\, p(I_E \mid x_t^{(i)}) \tag{5}$$

$p(I_S|\ x^{(i)}_t)= exp(-||I_S-I_{SM}||)$ is the likelihood of silhouette matching and $p(I_E|\ x^{(i)}_t) =exp(-d(I_E, x^{(i)}_t))$ is the likelihood of chamfer matching as in Fig. 5.



(a)

(b)

Fig. 5. The joint likelihood of **(a)** silhouette likelihood and (b) edge likelihood

4.2 Key Pose-Based Proposal Distribution

As mentioned in introduction section, the formal proposal distribution with first order Markov chain, $q(x^{(i)}_t|x_{t-1})=p(x^{(i)}_t|x_{t-1})$, could not cover abrupt motion. The occurrence of abrupt motion is mainly up to the time leaping of the slow system. The increased particles for covering high DOF are the reason of low speed. Therefore, key poses are useful for prediction of abrupt motion and reducing the number of particles.

We define a key pose library (KPL) which consists of 13 key poses. Fig. 6 shows all the key poses as PS models and silhouette images. Each key pose is represented as $k_i=(I_i, PS_i, f_i)$; I_i is the pose image of key pose, PS_i is PS model, and f_i is visual feature of key pose, as in [9]. Visual features are used for measuring key pose similarity. Based on the KPL, the key pose-based proposal distribution is defined as

$$q(x_t \mid x_{t-1}, y_t, KPL) = \alpha p(x_t \mid x_{t-1}) + (1-\alpha) p(x_t \mid y_t, KPL)\,. \tag{6}$$

This proposal distribution is a combined version of Markov chain-based proposal distribution and KPL-based prediction model which is defined as

$$p(x_t \mid y_t, KPL) = \frac{1}{13}\sum_{k=1}^{13} p(x_t \mid PS_k)\, p(y_t \mid PS_k)\,. \tag{7}$$

KPL-based prediction model predicts the particles to be similar to some of key poses in KPL. The latest observation y_t is referred in KPL-based prediction model to measure how much key poses are similar to the current observation using $p(y_t|PS_k)$. $p(x_t|PS_k)$ is the probability for similarity to key pose k of the predicted state x_t.

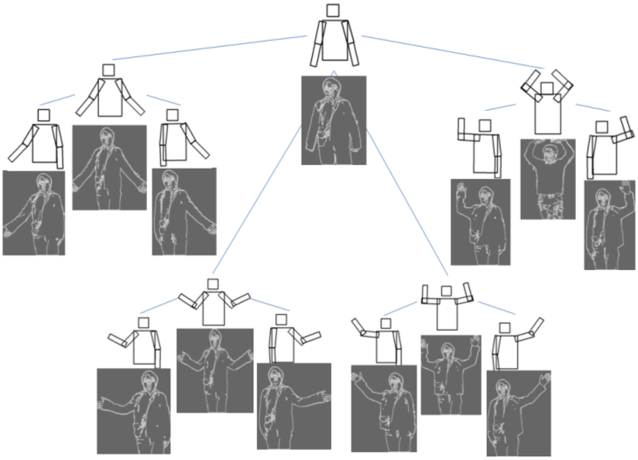


Fig. 6. Key Pose Library with 13 key poses (PS and Silhouette)

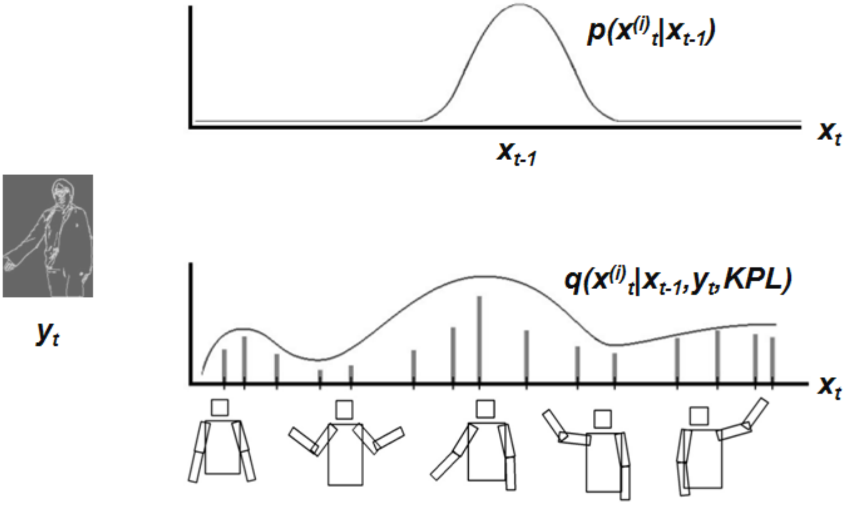


Fig. 7. Conceptual difference between our proposal distribution and Markov chain model

In Fig.7, we conceptually describe the difference of two proposal distributions. Consequently, our proposal distribution benefits from sparsely selected key poses which make it possible to predict particles globally. On the contrary, Markov chain model-based proposal distribution locally predicts particles around the previous state. When abrupt motion happens, our proposal distribution will predict better.

5 Gesture Recognition

From the pose sequence, we recognize the upper body gestures. We have 4 gestures: “Idle”, “I Love You”, “Hello Left” and “Hello Right”. The gesture recognition considers poses as events in the HMM structure. The events are seeable results of the pose tracking system but gesture states cannot be directly estimated from the system. HMM defines the gesture state transition matrix and the event emission matrix. Using Viterbi algorithm and HMM, we obtain the probability of hidden gesture states.

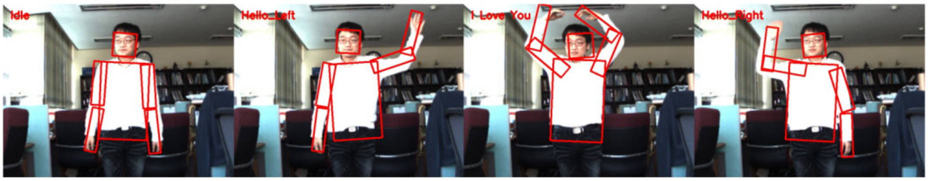


Fig. 8. Upper body gestures: “Idle”, “Hello Left”, “I Love You”, and “Hello right”

6 Experimental Result

For evaluation, we use our ground-truth database. By comparing with bootstrap, we found that our key pose-based proposal distribution overcomes the weakness of Markov chain model. Markov chain model has affected the results getting slow and failed from abrupt motion. With only 100 particles, KPL efficiently predicts smaller particles and overcomes the failures from abrupt motion, as shown in Fig. 9. Fig 10 shows the evaluated resulting images with several people. Our system adaptively tracks the different genders and heights.

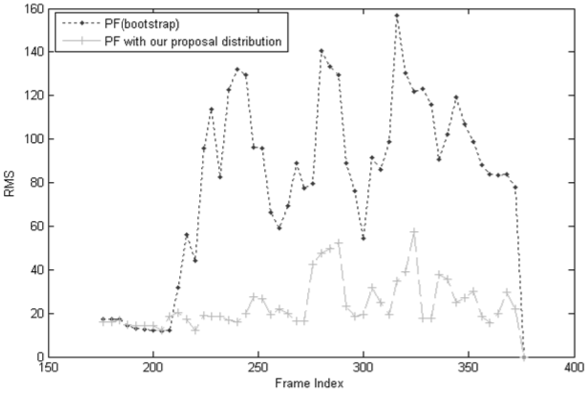


Fig. 9. Root Mean Square Error of bootstrap and our method

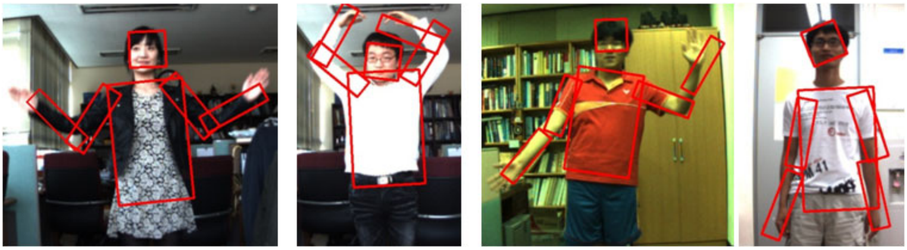


Fig. 10. Tracking results from several people

7 Conclusion

We have argued that key poses can be useful to improve the upper body tracking. As shown in Fig. 9, our system outperformed with only 100 particles using key poses. In prediction of particles and in recovering errors, key pose-based proposal distribution has been a great role in our system.

In the future, we will investigate how the key poses are mathematically working in particle filtering framework. Additionally we will try to build a regressive model of KPL to cover whole pose spaces not only 13 key poses. Finally we will test our system in the developing mobile robot platform.

Acknowledgement

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the Human Resources Development Program for Convergence Robot Specialists support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2010-C7000-1001-0007).

References

[1] Thurau, C., Halavac, V.: Pose Primitive based Human Action Recognition in Videos or Still Images. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (2008)

[2] Felzenszwalb, P., Huttenlocher, D.: Pictorial Structures for Object Recognition. International Journal of Computer Vision 61(1), 55–79 (2005)

[3] Oh, C.M., Islam, M.Z., Lee, C.W.: A Gesture Recognition Interface with Upper Body Model-based Pose Tracking. In: International Conference on Computer Engineering and Technology, vol. 7, pp. 531–534 (2010)

[4] Andriluka, M., Roth, S., Schilele, B.: Pictorial Structures Revisited:People Detection and Articulated Pose Estimation. In: International Conference on Computer Vision and Pattern Recognition, pp. 1014–1021 (2009)

[5] Barker, A.L., Brown, D.E., Martin, W.N.: Bayesian Estimation and the Kalman Filter. Computer & Mathematics with Applications 30(10), 55–77 (1995)

[6] Weng, S.K., Kuo, C.M., Tu, S.K.: Video Object Tracking using Adaptive Kalman Filter. Journal of Visual Communication and Image Representation 17(6), 1190–1208 (2006)

[7] Merwe, R.A., Freitas, N.D., Wan, E.: The Unscented Particle Filter. In: Advances in Neural Information Processing Systems, vol. 13 (2001)

[8] Islam, M.Z., Oh, C.M., Lee, C.W.: Real Time Moving Object Tracking by Particle Filter. In: International Symposium on Computer Science and Its Application, pp. 347–352 (2008)

[9] Oh, C.M., Islam, M.Z., Lee, C.W.: Pictorial Structures-based Upper Body Tracking and Gesture Recognition. In: Korea-Japan Joint Workshop on Frontiers of Computer Vision (2011)

[10] RedOne Technology (February 2011), `http://urc.kr/`

[11] MS Kinect (February 2011), `http://www.xbox.com/en-US/kinect`

[12] Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian Detection: A Benchmark. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (2009)

[13] Oh, C.M., Setiawan, N.A., Aurahman, D., Lee, C.W., Yoon, S.: Detection of Moving Objects by Optical Flow Matching in Mobile Robots using an Omnidirectional Camera. In: The 4th International Conference on ubiquitous Robots and Ambient Intelligence (2007)

[14] Barrow, H.G., Tenenbaum, J.M., Bolles, R.C., Wolf, H.C.: Parametric Correspondence and Chamfer Matching: Two New Technique for Image Matching. In: The 5th International Joint Conference on Artificial Intelligence, pp. 1175–1177 (1997)