Human and Robot Tracking Using Histogram of Oriented Gradient Feature

[1]Jeong-eom Lee, [2]Chong-ho Yi, [*3]Dong-won Kim
[1, First Author]*Robotics Team, Hyundai Motor Company, lee.jeongeom@hyundai.com*
[2,] *Inha Technical College, Digital Electronics, ycm@inhatc.ac.kr*
[*3, Corresponding Author] *Inha Technical College, Digital Electronics, dwnkim@inhatc.ac.kr*

Abstract

This paper describes a real-time human and robot tracking method in Intelligent Space with multi-camera networks. The proposed method detects candidates for humans and robots by using the histogram of oriented gradients (HOG) feature in an image. To classify humans and robots from the candidates in real time, we apply cascaded structure to constructing a strong classifier which consists of many weak classifiers as follows: a linear support vector machine (SVM) and a radial-basis function (RBF) SVM. By using the multiple view geometry, the method estimates the 3D position of humans and robots from their 2D coordinates on image coordinate system, and tracks their positions by using stochastic approach. To test the performance of the method, humans and robots are asked to move according to given rectangular and circular paths. Experimental results show that the proposed method is able to reduce the localization error and be good for a practical application of human-centered services in the Intelligent Space.

***Keywords**: Intelligent Space, human and robot tracking, histogram of oriented gradients, AdaBoost learning, feature*

Ⅰ. Introduction

As we called Intelligent Space as a well configured environment to provide various services for humans and robots, it is a room or an area that is equipped with distributed networked sensors, which enable the space to perceive the state of human and objects[1]. Many researchers have gone forward study for human-centered services by using multi-camera networks in the Intelligent Space[2-4]. For this, mobile robots have been applied networked as physical agents to the Intelligent Space. In this case, tracking for human and robots is a fundamental and important issue, in order to keep the distance between human and robot and avoid a collision between them. In this paper, we propose a method for tracking human and robot in the Intelligent Space with multi-camera networks. The objective of tracking humans and robots in the Intelligent Space is to localize their position in the world coordinates. In case of this environment, it becomes an easier task to estimate the three dimensional (3D) position of humans and robots. However, tracking humans and robots by using vision sensors is still a difficult and challenging task. The proposed method detects humans and robots in images captured by each camera. From the positions of humans and robots on image coordinates, the Intelligent Space estimates their 3D positions and tracks their positions by using stochastic approach. Fig. 1 shows a process of the proposed method. For finding humans and robots in images, the first need is a robust feature set that allows humans and robots form to be discriminated clearly, even in cluttered backgrounds under various illumination[5-6]. So, we adopt the template based histogram of oriented gradients (HOG) feature, since it is shown that the feature of locally normalized HOG provides excellent performance[7-9]. The extracted feature is classified by the support vector machine (SVM) using linear and radial-basis function (RBF) kernels. For a real-time process, by using a cascaded structure, we configure the best SVM classifiers, which are chosen by adaptive boosting (AdaBoost) learning technique[10]. Since a relation between 3D point and image point can be interpreted by a pin-hole camera model, rays from

* Corresponding Author
Received: Nov. 26, 2018, Revised: Dec. 18, 2018, Accepted: Dec. 26, 2018

the focus of camera to objects (humans or robots) position detected on image in each camera can be calculated. Thus, we obtain 3D position by intersection of the rays and then track them by using the particle filters.

We briefly introduce our system configuration in section 2. Section 3 provides more details of human and robot detection algorithm. And section 4 describes a 3D position estimation and tracking method. Result of experiments is described in section 5. Finally, we summarize our results in section 6.

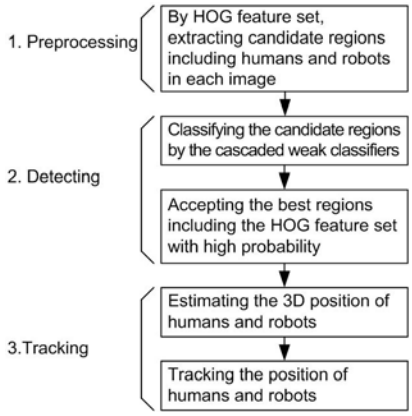


Figure 1. Tracking process for humans and robots

II. System configuration

Like Fig. 2, we have set up vision-based Intelligent Space. In our Intelligent Space, we use 4 DINDs (Distributed Intelligent Network Devices)[1]. A DIND is composed of three basic elements: a vision sensor (camera), a processor and a communication device. For detecting humans and robots, 4 cameras were mounted at approximately 2.1m above the floor. These are able to cover the entire space. Each of the DINDs is connected to a management agent and can send and receive data packet including contents of detected information. The management agent estimates 3D positions of humans and robots. In this work, we use OpenRTM-aist [11] for the system integration.

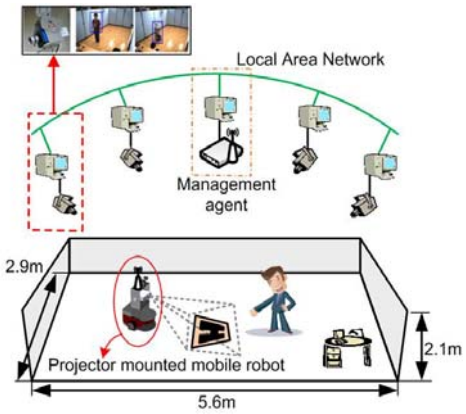


Figure 2. System configuration

III. Human and robot detection

Firstly, the position of human and the UD on image coordinates should be detected in each DIND. In this section, we describe the Histogram of Oriented Gradients (HOG) feature method for detecting

human and the UD. We have used a template-based HOG feature set. For real-time detection, we use a cascade structure and use AdaBoost to select suitable features and to construct a strong classifier consisting of weak classifiers. In order to detect humans and robots in an image, we adopt the template-based HOG feature method[4]. Following works presented in [7], each of HOG features is represented by a histogram of local image measurements within a region. We compute orientation θ and magnitude γ from local image gradient at each point in the region. To construct HOG features, we discretize the orientation θ into m bins, which include the magnitude γ. The HOG feature in normalized to the L1-norm. This normalization results in better invariance to changes in illumination or shadowing. For reducing noise and minimizing quantization error, we utilize the HOG feature with the histogram operating convolution to the Gaussian filter. Furthermore, in order to speed up the HOG computation, we adopt the "Integral Histogram Image" (IHI) [8-9].

Since we use template-based HOG features, there are many rectangular templates with fixed size considering the region of humans and robots in an image. For example, there are 1,024 templates for 64x64 image[9]. In this paper, we utilize AdaBoost learning to construct a strong classifier with many weak classifier[10]. Given f_1, f_2, ... , f_M for a set of learned weak hypotheses, thus the ensemble hypothesis is as following:

$$F(\mathbf{x}) = E\{c|x\} = \sum_{i=1}^{M} f_i(\mathbf{x}) \tag{1}$$

where $E\{c|x\}$ represents the expectation of the predicting label c from the sample $\mathbf{x}$. The Gentle Adaboost (GAB) is utilized, since it gives higher performance and requires a few iterations to train.

A cascaded structure of the weak classifiers achieves increased accuracy for detection performance and reducing computation time. The cascaded boosting classifiers are able to reject most of the negative sub-windows while detection. In addition, simple classifiers are used to reject the majority of sub-windows before more complex classifiers are called upon to achieve low false positive rates. In this paper, we construct the cascaded classifiers with 16 stages: four preceding stages are trained by the linear SVM, and the rest stages are trained by the RBF-SVM. We use total 80 weak classifiers. The number of classifiers in each stage is defined by the value of sigmoid function.

IV. Human and robot tracking

A management agent receives data packets from four DINDs and estimate position of humans and robots. By using multiple view geometry[12], it estimates 3D positions of humans and robots from their 2D positions on image coordinates. Generally, a relation between 3D point and image point can be interpreted by a pin-hole camera model. Using a pin-hole camera model, the rays from the focus of camera to position of humans or robots detected in image at each camera can be calculated. Thus, we obtain 3D position by intersection of the rays as shown in Fig. 3. We assume that all cameras are calibrated.

In this paper, we use a particle filter to approximate the position of humans and robots. The particle filter is a Sequential Importance Sampling (SIS) based on the Bayesian inference in a Hidden Markov Model (HMM)[13-14]. The key idea of the particle filter is to approximate the probability distribution by a weighted sample set $S_k = \{(\mathbf{x}_k^i, \pi_k^i),\ i \in \{1,...,N\}\}$ where π_k^i is the weight of sample $\mathbf{x}_k^i$ at time k. The sum of weights is normalized as $\sum_{i=1}^{N} \pi_k^i = 1$. Thus, $p(\mathbf{x}_k|z_{0:k})$ can be approximated as

$$p(\mathbf{x}_k|z_{0:k}) \approx \sum_{i=1}^{N} \pi_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i) \tag{2}$$

where $\delta(\square)$ is a Dirac-delta function. The iterative algorithm to approximate $p(\mathbf{x}_k|z_{0:k})$ is described as following.

Initialization: N samples with uniformly distribution probability $\pi_0^i = N^{-1}$ are generated.

Transition: Each sample from the weighted sample set S_{k-1} is propagated by a dynamic state equation as following:

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{w}_{k-1} \tag{3}$$

where $\mathbf{A}$ is a deterministic model, and $\mathbf{w}_{k-1}$ is Gaussian white noise at time $k-1$.

Observation: There are more than one 3D position estimated in the world coordinates. Thus, we consider the distance between measured positions and predicted position at time $k-1$. The likelihood is defined by the Euclidian distance as following:

$$\pi_k^i = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\mathbf{x}' - \mathbf{x}_{k-1}}{2\sigma^2}\right) \tag{4}$$

where $\mathbf{x}'$ is a measured states and $\mathbf{x}_{k-1}$ is a predicted state at time $k-1$.

Resampling: A common problem of particle filter is that after a few iterations, most samples have negligible weights, called '*degeneracy*'. We apply the stochastic universal sampling to the resampling scheme to avoid the degeneracy of samples[14].

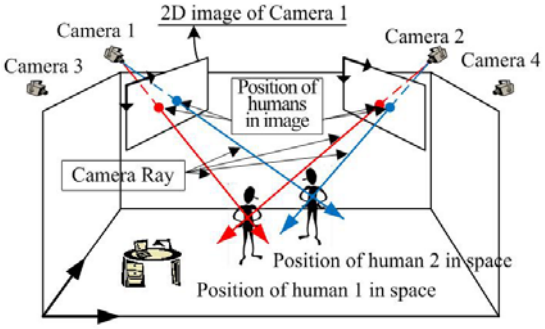


Figure 3. Estimating the 3D position by intersection of the ray

V. Experimental results

Experiments were performed in a 5.6m x 2.9m space which has four cameras mounted at 2.1m height. The cameras have a specification with 320 x 240 resolution and 30 fps. For tracking experiments in this environment, we considered two types of paths: a rectangular-shaped path (Type-1) and a circular-shaped path (Type-2). Three persons walked three laps according to both of the two paths. A mobile robot moved one lap according to the Type-1 path. For tracking, the hidden state of humans and robot is denoted by $\mathbf{x}_k = [x_k, \dot{x}_k, y_k, \dot{y}_k]^T$. Initially, the samples of particle filter randomly are scattered in the space. Each sample has a window to tract humans and robots. The parameters of particle filter are described in Table 1.

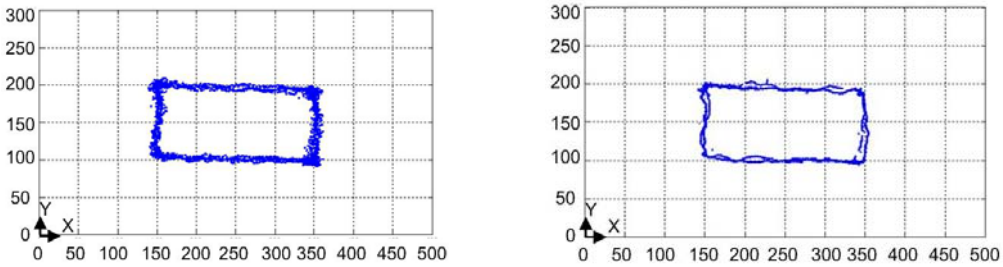
Table 1. Parameters of particle filter

Parameter	Value
Number of particles	100
Standard deviation	$\sigma_x, \sigma_y = 20cm$
Uniform distribution	$u \square U[0,1]$
σ in equation (4)	1

To test the performance of the method, a person is asked to track two type trajectories. Fig. 4 and Fig. 6 show the results. In Fig. 4-(a) and 6-(a), the results without tracking show that the estimated positions are distributed in a wide range. Fig. 5 and Fig. 7 show an accumulated error for both type-1 and type-2 path. Tracking humans by using stochastic approach has the high accuracy for the estimated positions. The results of average error for estimated positions between without tracking and with tracking are summarized in Table 2.

Table 2. Result of average error between given paths and estimated human positions

Type of walking path	Type-1 (rectangle)	Type-1 (circle)
Average error without tracking	3.1 cm	21.7 cm
Average error with tracking	2.2 cm	12.9 cm



(a) Type-1 path by human walking without tracking (b) Type-1 path by human walking with tracking

Figure 4. Plots of estimated positions of a person for the Type-1 path

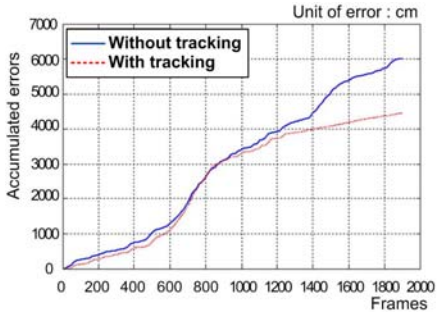
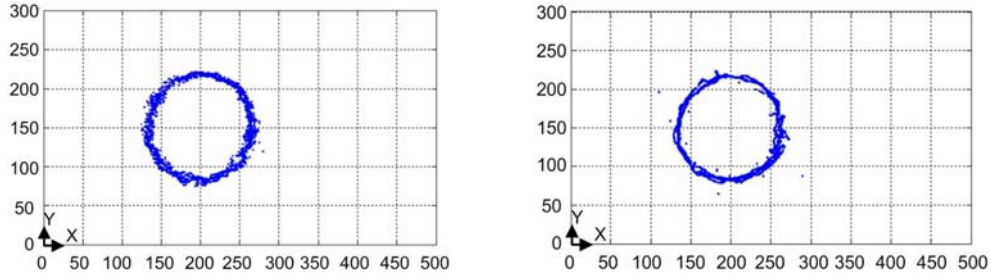


Figure 5. The accumulated error for the Type-1 path



(a) Type-2path by human walking without tracking (b) Type-2path by human walking with tracking

Figure 6. Plots of estimated positions of a person for the Type-2path

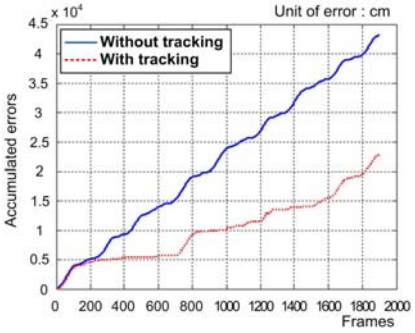
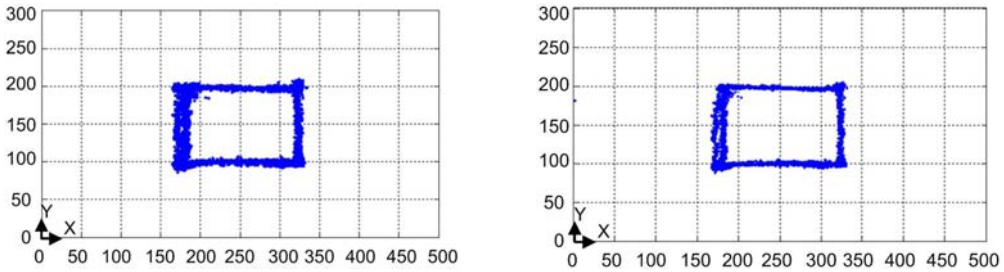


Figure 7. The accumulated error for the Type-2 path

For tracking a robot, we used a steerable projector mounted mobile robot which is called Ubiquitous Display (UD) [4]. The moving speed of the UD is 10cm/sec.

Fig. 8 shows the trajectories of estimated positions of a robot for the Type-1 path. Like the result of human tracking, while Fig. 8-(a) shows that the estimated positions are distributed in a wide range, the result of Fig. 8-(b) shows a smaller range. Fig. 9 shows the result of accumulated error for Type-1 path. There is no almost the difference of accumulated error between without tracking and with tracking, since the moving speed of the robot is slower than human's one, and the movement of robot is very straightforward. In this case, a robot is easy to track, compared with human.



(a) Type-1 path by a robot moving without tracking (b) Type-1 path by a robot moving with tracking

Figure 8. Plots of estimated positions of a robot for the Type-1 path

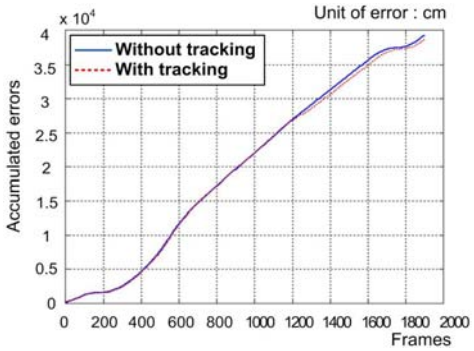


Figure 9. Estimating the 3D position by intersection of the ray

VI. Conclusions and future works

In this paper, we propose a vision-based humans and robots tracking method in Intelligent Space with multi-camera networks. The proposed method for tracking humans and robots used the Histogram of Oriented Gradients(HOG) feature to extract them in each image, and apply SVM with linear and RBF kernel as a classifier for high accuracy by using AdaBoost learning technique. And then, we

estimated the 3D position of humans and robots from their positions on image coordinates by using the multiple view geometry and tracked their 3D positions by using stochastic approach. Experimental results showed that the proposed tracking method reduces the error of localization for humans and robots.

We have been applying this tracking method to our active information display system based on Intelligent Space and the UD. The system is able to provide a user with relevant information by projecting it on where the user is looking at. Although the proposed tracking method is used, there still exists localization error. Since this error affects the visual information generated by the Intelligent Space for projection, we need to evaluate influences of this error on the system. We conducted experiments to examine how the error of the position of the user and the UD affects the perception of the projected image. The subjects are asked whether the projected image is similar to the original image. In this case, the original image is the same image as we were trying to transfer. As a result, we could say that the position of the UD is more critical than the position of the user. We could also say that a permitted limit of the error for the position of the UD is 3cm and a permitted limit of the error for the position of the user is 5cm, although the variance is large. In the near future, we will report the evaluation of the active information display system and user test.

VII. Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No. 2017R1D1A1B03031467).

VIII. References

[1] J. H. Lee and H. Hashimoto, “Intelligent Space — concept and contents,” Adv. Robotics., vol. 16, pp. 265-280, 2002
[2] J. H. Lee and H. Hashimoto, “Controlling mobile robots in distributed intelligent sensor network ,” IEEE Trans. Ind. Electron., vol. 50, pp. 890-902, 2003
[3] J. H. Kim, D. W. Kim, B. J. Yoo and G. T. Park, “A Design of Framework for Smart Services of Robots in Intelligent Environment,” Int. J. Control Autom.,vol. 2, pp. 1-12, 2009.
[4] J. E. Lee, J. H. Kim, S. J. Kim, Y. G. Kim, J. H. Lee and G. T. Park, Human and Robot Localization using Histogram of Oriented Gradient(HOG) Feature for an Active Information Display in Intelligent Space, Proceedings of the First International Conference on Engineering and Technology Innovation, 2011, November 11-15, Kenting, Taiwan
[5] S. W. Ha and Y. H. Moon, “Multiple object tracking using sift features and location matching,” Int. J. Smart Home, vol. 5, no. 4, 17-26, 2011
[6] M. Stommel and O. Herzog, "Binarising SIFT descriptors to reduce the curse of dimensionality in histogram-based object recognition," International Journal of Signal Processing, vol. 3, pp. 25–36, 2010.
[7] N. Dalal and B. Triggs, Histogram of Oriented Gradients for Human Detection, Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, June 20-25; San Diego, CA, USA
[8] Q. Zhu, S. Avidan, M. Yeh and K. Cheng, Fast Human Detection using a Cascade of Histogram of Oriented Gradients, Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, June 17-22; New York, NY, USA
[9] H. Wang, P. Li and T. Zhang, “Histogram feature-based Fisher linear discriminant for face detection,” Neural Comput. Appl., vol. 17, pp. 49-58, 2008
[10] P. Viola and M. Jones, “Robust Real-Time Face Detection ,” Int. J. Comput.Vision, vol. 57, pp. 137-154, 2004
[11] OpenRTM-aist : www.openrtm.org
[12] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, New York, 2003
[13] S. Arulampalam, S. Maskell, N. Gordon and T. Clapp, “A tutorial on particle filters for online

nonlinear/non-Gaussian Bayesian tracking," IEEE Trans. Signal Process., vol. 50, pp. 174-188, 2002

[14] M. Z. Islam, C. M. Oh and C. W. Lee, "Video Based Moving Object Tracking by Particle Filter ," International Journal of Signal Processing, Image Processing and Pattern, vol. 2, pp. 119-132, 2009