Computing and Informatics, Vol. 27, 2008, 719–730

MOVING OBJECT DETECTION AND TRACKING IN OPEN-AIR TEST BED

Tatsuya Yamazaki, Tetsuo Toyomura
Kentaro Kayama, Seiji Igi

National Institute of Information and Communications Technology
3-5 Hikaridai Seika-cho Soraku-gun Kyoto
619-0289, Japan
e-mail: yamazaki@nict.go.jp

Revised manuscript received 10 January 2008

Abstract. In mobile and ubiquitous computing environments, acquisition of contextual information about a user situation is necessary to provide useful services. Although the definition of user context may change according to the situation or the service used, contextual information about who, where, and when are considered to be essential. We have built a test bed with multiple sensors: floor pressure sensors, RFID (radio frequency identification) tag systems, and cameras, to carry out experiments to detect the positions of users and track their movement. The conventional background subtraction method by using cameras was used for moving object detection and tracking. In this paper, we propose knowledge application and parameter adaptation in the background subtraction method. The results are presented to show that the proposed method decreases the detection errors.

Keywords: Human tracking, background substruction, parameter adaptation, test bed

1 INTRODUCTION

Progress of wireless communication, microchips, and sensing technologies, for example, is accelerating, which is enabling the achievement of mobile and ubiquitous computing environments. In mobile and ubiquitous computing environments, people can enjoy the benefit of high-speed Internet access networks at any time, com-

municating from and to anywhere. Thus, we are at the beginning of the ubiquitous network society era. To provide useful services to people in the ubiquitous network society, we have to consider to whom and how the service needs to be provided. In other words, service personalization on the basis of each person's ability or preference and an interface for service provision have significance as research and development themes in the next step of the construction of a ubiquitous network society.

In particular, the method of detecting and utilizing contextual information about each user's situation, which is user context, is an issue. The definition of user context or the necessary contextual information may change according to the situation or the service used. In [1], Dey and Abowd defined context as 'any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves'. In addition, they defined context-aware computing as follows: 'A system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task'. By referring to [1], Jang and Woo [2] proposed a unified context that describes a user-centric situation independently of the purpose of any service, in terms of 5W1H (who, what, where, when, why, and how). In a unified context, the meanings of who, where, and when are explicit. 'What' means the entity that a user is paying attention to; 'how' means the manner in which a user is making an expression with gestures or action, and 'why' means the reason a user is going to trigger a service. With these definitions, one finds that recognizing what, why, and how is different from recognition of who, where, and when. More precise sensing and wider fields of science and technology such as psychology are needed.

Therefore, we consider that who, where, and when are usually the main items of contextual information. There are several ways to detect these main items of contextual information, and they are categorized into subgroups, for example, passive, active, wearable, or embedded. From the viewpoint of a user, not wearing or carrying any device nor performing any active action is desirable. Therefore, we are focusing on technologies to detect users' positions and track their movement using environment-embedded sensors in a passive way.

Cameras are the most common environment-embedded sensors. Satake and Shakunaga [3] proposed an appearance-based condensation tracker, which is composed of a condensation tracker and a sparse template matching method to detect the movement of people with a camera. The template-based condensation tracker is stabilized for tracking even in the case of object occlusion. Thonnat and Rota [4] used low-level image processing techniques to detect and track mobile objects. Their aim was rather to understand images, namely, to generate alarms automatically for operators when interesting scenarios had been recognized by the system. In [3] and [4], they applied their methods to actual situations. In particular, subway (metro) station applications are shown in [4]. However, they only used one camera [3] and [4]. In a ubiquitous network society, we have to consider collaboration

of several distributed cameras embedded in the environment. As a study of real-time synchronization of several cameras, Matsuyama [5] proposed a protocol for negotiation among agents linked to their respective cameras.

By using floor pressure sensors and RFID (Radio Frequency Identification) tags in addition to the cameras, we have built a test bed to carry out experiments of detecting and tracking people in an actual situation. The test bed is situated in the entrance of our research center, and the sensors are set in the environment so as not to hinder the behavior of subjects in the experiments.

In this paper, firstly, the test bed in our research laboratory is introduced as an open-air environment for detecting and tracking people. Secondly, among the various sensors in the test bed, cameras are used to detect and track moving objects. The conventional background subtraction method by using multiple cameras is introduced. Especially, we propose knowledge application and parameter adaptation in the background subtraction method. The results are presented to show that the proposed method decreases the detection errors.

2 NICT ENTRANCE OPEN-AIR TEST BED FOR DETECTING AND TRACKING PEOPLE

Since home-style test beds have been built throughout the world, in-home test beds are available in a relatively easy way, and data have been collected to analyze human behavior. However, there are few open-air test beds, and developing a test bed to collect data on detecting and tracking people in an open area is necessary.

We constructed such an open-air test bed in the entrance of our research laboratory. Hereafter, it is called the NEO (NICT Entrance Open-air) test bed. Although the detection area of the NEO test bed includes the outside of the entrance, we only present the detection area inside the entrance in this paper. Floor pressure sensors are installed throughout the floor. The floor pressure sensors are covered by carpet, so subjects are not aware of the existence of the sensors. The floor pressure sensors contain binary detection units and are used to track the movement of subjects. The distance of units is 5 mm and the minimum detectable pressure is $200\,\mathrm{g/cm^2}$ under the best condition. In the ceiling of the NEO test bed, there are five cameras, and the area covered by the cameras is the meshed region in Figure 1. The camera can take a 768×494 pixel image and has remote pan, tilt, and zoom functions. Three RFID tag readers are installed above the ceiling of the NEO test bed. The RFID tag system uses active RFID tags. The tags emit electromagnetic waves at 315 MHz, so they can be detected at the distance of 10 m. Each tag reader can read 60 tags per second.

The total system architecture is depicted in Figure 2. The floor pressure sensors, the cameras, and the RFID tag system are connected via Ethernet and the collected data are stored in a common database server.

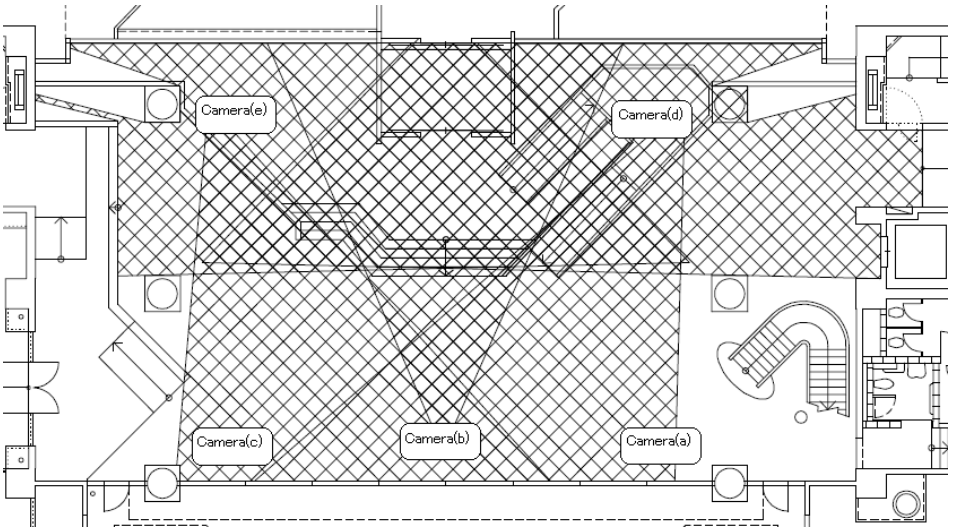

Fig. 1. Area monitored by cameras

3 DETECTION OF PEOPLE BY BACKGROUND SUBTRACTION METHOD

Identifying moving objects from a video sequence is a fundamental and critical task in many computer-vision applications. A common approach is background subtraction, which detects objects moving in the foreground as the difference between

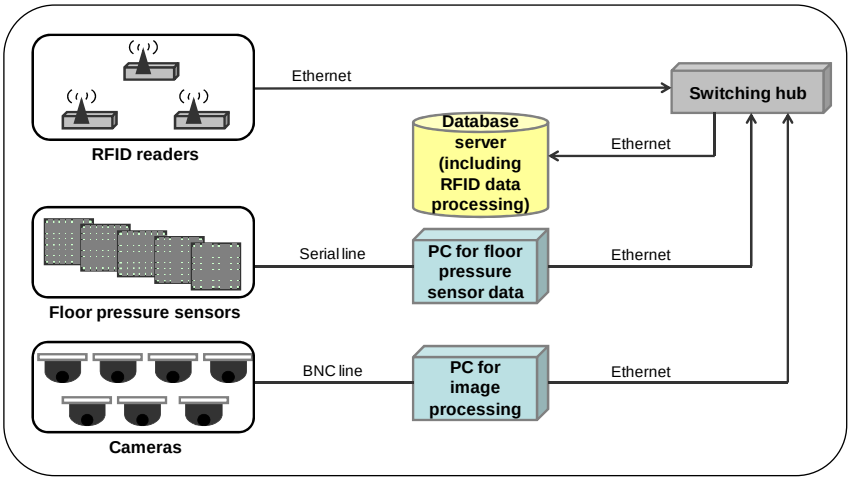

Fig. 2. Total system architecture

the current frame and an image of the scene's static background. Although the background subtraction method sometimes suffered from illumination changes and background object changes such as tree branches, it is a simple and effective method to detect moving objects.

In order to cope with the illumination changes, we adopt the normalized distance method. Here, a unit vector is defined as the projection to the unit sphere of a vector whose elments are intensity values of pixels in a target region. The normalized distance is defined as a distance between two unit vectors. Let τ and β be vectors consiting of intensity values of pixels in an observed image and background images, respectively. Then the distance δ and normalized distance δ' are shown as follows:

$$\delta = |\tau - \beta| \tag{1}$$

$$\delta' = |\frac{\tau}{|\tau|} - \frac{\beta}{|\beta|}|. \tag{2}$$

Suppose that we process image frames during a period of T_{int}. We calculate δ for each frame in T_{int}, then $\max(\delta)$, $\min(\delta)$, and $Ave(\delta)$ are calculated as maximum value, minimum value, and averaged value of δ. In the same way, $\max(\delta')$ and $\min(\delta')$ are calculated as maximum value and minimum value of the normalized distance δ'. Consequently, the following three discriminant functions are defined:

1. discriminant function for scene change

$$\max(\delta) - \min(\delta) > Th_{scene} \tag{3}$$

2. discriminant function for background change

$$\text{ave}(\delta) > Th_{bs} \tag{4}$$

3. discriminant function whether environment change or illumination change

$$\max(\delta') - \min(\delta') > Th_{ill} \tag{5}$$

where Th_{scene}, Th_{bs}, and Th_{ill} are thresholds to be determined.

Using these discriminant functions, whether there is a moving object detection or a background, updating in T_{int} can be judged as follows:

- If both (3) and (5) are true, there is a scene change by a moving object.
- If (3) is true and (5) is false, there is a scene change by an illumination change.
- If (3) is false and (4) is true, there is a background change.
- If both (3) and (4) are false, there is nothing. It is a normal background.

A challenging point in this method is adaptively setting the threshold value to differentiate foreground objects from the background image in spite of environmental changes.

To determine the threshold value, Wren et al. [6] modeled the background using a Gaussian distribution and estimated the parameters adaptively. Grimson et al. [7] also set up parameters according to the statistical analysis of training samples of the background images. Stauffer and Grimson used a mixture model of Gaussian distributions of the images to cope with multimodal background distributions [8].

4 KNOWLEDGE APPLICATION AND PARAMETER ADAPTATION

We apply two techniques to the original background subtraction method in order to cope with unexpected moving objects and adaptive threshold parameter setting.

Our aim is to detect and track people as moving objects. There are, however, other unexpected moving objects in the scene, such as an automatic door. To avoid detection of such an unexpected moving object, we introduce knowledge about special spots as the first technique. The positions of special spots are assumed to be known and masking is applied not to detect the unexpected moving object. This is a simple but effective technique.

To set the threshold values adaptively, we introduce a kind of steepest descent method as the second technique. The algorithm is depicted in Figure 3.

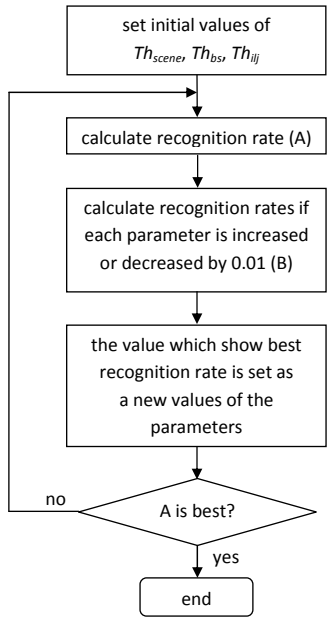

Fig. 3. Flowchart of threshold adaptation by steepest descent

In the first step of the algorithm shown in Figure 3, initial values of Th_{scene}, Th_{bs}, and Th_{ill} are set. Then the recognition rate A with the current threshold values, that

are the same as the initial values at the very first stage of the algorithm. After the threshold values are increased or decreased by a small value, the new recognition rates B are calculated. 0.01 is set as the small value in Figure 3. B is a set of the recognition rates because increasing and decreasing of each threshold value are tried. A is compared with the best recognition rate in B. If A is superior to the best recognition rate, the algorithm is terminated. Otherwise, the best threshold values that correspond to the best recognition rate are substituted to the current threshold values and the same steps are carried out.

5 EXPERIMENTAL RESULTS AT NEO TEST BED

In the NEO test bed, we tried to detect moving objects using camera images and the background subtraction method. In the background subtraction method, first of all, we selected the initial image without any moving objects as the background image to be subtracted. Then, the difference between the current and the background images is calculated and pixels that have a difference larger than the threshold are registered as candidate pixels of an image of moving objects. This difference calculation is operated for each small image block of 80×60 pixels. The judgement described in Section 3 is applied. The adjacent small image blocks of candidate pixels are merged into a larger image block. To track the moving objects, a two-dimensional histogram with hue and saturation values in an HSV color space is constructed to calculate the correspondence between objects in the current and previously captured images. When the difference between two images in the two-dimensional histogram is smaller, the probability that objects belong to the same object is higher.

We applied the above background subtraction method to a ten-second video captured in an actual situation. The number of frames captured from each camera was different because the time consumed for image compression was different.

Two experiments were carried out. The first experiment was moving object detection by the background subtraction method with knowledge application only. The second one was moving object detection by the background subtraction method with parameter adaptation as well as knowledge application. The first and the second are referred to as Experiment I and Experiment II, respectively. The knowledge applied is that the position of the automatic door is known.

In Experiment I, the threshold parameters were fixed as $Th_{scene} = 0.10$, $Th_{bs} = 0.25$, and $Th_{ill} = 0.05$. One result of detecting a moving object is presented in Figure 4. The people in the image should be recognized as moving objects, and rectangles are drawn as a result. Two larger rectangles are drawn in Figure 4; one is for a group of four persons on the left and the other is for a single man on the right. These larger rectangles are hand-made markings that indicate correct answers. Several smaller rectangles in the left larger rectangle (the group of four persons) indicate detected results obtained by the background subtraction method. In the right larger rectangle, no object was detected by the background subtraction

method. We call the larger rectangles ground truth rectangles and the smaller rectangles are called detected rectangles.

Although an identical match between ground truth and detected rectangles is desirable, detected rectangles are almost always included in or overlapped on ground truth rectangles. Here, we define two kinds of error: the type one error and the type two error. The type one error is that in which no detected rectangle is drawn where there was a ground truth rectangle. The type two error is that in which detected rectangles appeared where there was no ground truth rectangle. The total error rate can be calculated by averaging these two types of error. The rates of occurrence of two types of error and the total error rate are shown in Table 1 for cameras (a)–(e). The positions of the cameras correspond to those of the images shown in Figure 1.

Fig. 4. Result of detecting moving objects in Experiment I

Next, we applied threshold parameter adaptation presented in Figure 3 as well as the knowledge application. This is Experiment II. The initial threshold parameters were set as $Th_{scene} = 0.10$, $Th_{bs} = 0.25$, and $Th_{ill} = 0.05$. One moving object detection result with the parameter adaptation is presented in Figure 5, that corre-

	Type one error rate (%)	Type two error rate (%)	Total error rate (%)
Camera (a)	29.38	0.32	14.85
Camera (b)	54.35	27.94	41.15
Camera (c)	28.39	0.31	14.35
Camera (d)	10.23	16.19	13.21
Camera (e)	10.06	8.38	9.22

Table 1. Error rates with knowledge application and fixed parameters

ponds to Figure 4. By comparing two images, it is found that the person in the right side was detected in Experiment II, who was missed in Experiment I. The rates of occurrence of two types of error and the total error rate are shown in Table 2.

Fig. 5. Result of detecting moving objects in Experiment II

Almost all error rates were improved. Especially improvement for camera (b) was splendid. The final adapted parameters are in Table 3.

	Type one error rate (%)	Type two error rate (%)	Total error rate (%)
Camera (a)	14.69	0.27	7.48
Camera (b)	17.39	0.00	8.97
Camera (c)	19.36	1.97	10.67
Camera (d)	1.14	14.70	7.92
Camera (e)	10.06	0.08	5.07

Table 2. Error rates with knowledge application and parameter adaptation

	Th_{scene}	Th_{bs}	Th_{ill}
Camera (a)	0.10	0.21	0.04
Camera (b)	0.10	0.20	0.04
Camera (c)	0.10	0.21	0.04
Camera (d)	0.07	0.23	0.04
Camera (e)	0.12	0.26	0.04

Table 3. The final threshold parameters in Experiment II

6 CONCLUSION

Detecting contextual information is an important issue for providing a personalized, adaptive, situation-aware service in a ubiquitous network society. To approach this issue, we defined essential contextual information, such as who, where, and when, and constructed a test bed with multiple sensors including floor pressure sensors, RFID tag systems and cameras. We tracked the movement of people just by using camera sensors.

The background subtraction method was used and its performance was improved by introducing knowledge application and parameter adaptation. We used a kind of the steepest descent method to adjust the threshold parameters, which is one of the supervised learning schemes. Therefore the ground truth data are necessary and shifting to the unsupervised learning is a urgent and important further study. As a technical issue, how to set the small value in Figure 3 is an interesting problem.

In future studies, combining information from other types of sensors to improve detection accuracy and image understanding is worthwhile.

REFERENCES

[1] Dey, A. K.—Abowd, G. D.: Towards a Better Understanding of Context and Context-Awareness. In: Workshop on The What, Who, Where, When, and How of Context-Awareness, as part of the 2000 Conference on Human Factors in Computing Systems (CHI 2000), The Netherlands, April 2000.

[2] Jang, S.—Woo, W.: Unified Context Describing User-Centric Situation: Who, Where, When, What, How and Why. 1st Korea-Japan JointWorkshop on Ubiquitous Computing and Networking Systems (UbiCNS 2005), No. 32, 2005, pp. 175–180.

[3] Satake, J.—Shakunaga, T.: Multiple Target Tracking by Appearance-Based Condensation Tracker Using Structure Information. The 17th International Conference on Pattern Recognition (ICPR 2004), Vol. 3, August 2004, pp. 537–540.

[4] Thonnat, M.—Rota, N.: Image Understanding for Visual Surveillance Applications. Third International Workshop on Cooperative Distributed Vision (CDV-WD '99), No. 3, 1999, pp. 51–82.

[5] Matsuyama, T.: Dynamic Memory: Architecture for Real Time Integration of Visual Perception, Camera Action, and Network Communication. Third International Workshop on Cooperative Distributed Vision (CDV-WD '99), No. 1, 1999, pp. 1–30.

[6] Wren, C.—Azarbayejani, A.—Darrell, T.—Pentland, A.: Pfinder: Real-Time Tracking of the Human Body. IEEE Trans. on Patt. Anal. and Machine Intell., Vol. 19, 1997, No. 7, pp. 780–785.

[7] Grimson, W. E.-L.—Stauffer, C.—Romano, R.—Lee, L.: Using Adaptive Tracking to Classify and Monitor Activities in a Site. Proc. 1998 Conference on Computer Vision and Pattern Recognition (CVPR '98), pp. 22–29.

[8] Stauffer, C.—Grimson, W. E. L.: Adaptive Background Mixture Models for Real-Time Tracking. Proc. 1999 Conference on Computer Vision and Pattern Recognition (CVPR '99), pp. 246–252.

Tatsuya Yamazaki works in National Institute of Information and Communications Technology. He received the B. Eng., M. Eng. and Ph. D. degrees in information engineering from Niigata University in 1987, 1989 and 2002, respectively. From 1992 to 1993 and 1995 to 1996 he was a visiting researcher at the National Optics Institute, Canada. Since 1997 he has been with ATR Adaptive Communications Research Laboratories. His research interests are adaptive QoS management, statistical image processing and ubiquitous environmental information processing. He is a member of the IEEE, and a member of the Information Processing Society of Japan.

Tetsuo Toyomura works in National Institute of Information and Communications Technology since 2005. He is managing several equipments in NICT test beds.

Kentaro Kayama received the B. Eng. degree from the Department of Mechano-Informatics at the University of Tokyo in 1996 and the M. Eng. degree and Ph. D. degree from the Information Engineering School ofthe University of Tokyo in 1998 and 2001, respectively. In 2001 he joined the Communications Research Laboratory (now National Institute of Information and Communications Technology) and engaged in Robotic Communication Terminals Project. His research interests include computer vision (especially in outdoor environment recognition), outdoor mobile robots, and computer Shogi. He is a member of IEEE, the Robotics Society of Japan, and the Japan Society of Artificial Intelligence.

Seiji Igi is a managing director of International Alliance Division in Research Promotion Department, the National Institute of Information and Communications Technology. He received the B. Eng. and M. Eng. degrees in Nagoya Institute of Technology in 1973 and 1975, respectively. He has been the group leader of Human-computer Intelligent Interaction Group from 1995 to 2006. He has been the Director of Keihanna Human Info-Communication Research Center from 2005 to 2006. His research interest is human computer interaction and human interface.