Fourth-person sensing for a service robot

Kazuto Nakashima, Yumi Iwashita, Pyo Yoonseok, Asamichi Takamine, Ryo Kurazume
Kyushu University, Fukuoka, Japan
Email: k_nakashima@irvs.ait.kyushu-u.ac.jp

***Abstract*—This paper proposes a new concept named "fourth-person sensing" for a service robot. The proposed concept combines wearable sensors (the first-person viewpoint), sensors mounted on robots (the second-person viewpoint), and sensors embedded in the environment (the third-person viewpoint), and the disadvantages in individual sensors are compensated by combining sensory information from the first, second, and third-person viewpoints. The fourth-person sensing is effective to understand a user's intention and a context of the scene, thus it enables to provide a proper service by a service robot. As one of applications of the fourth-person sensing, we develop a HCI system combining the first-person and the third-person sensing, and show the effectiveness of the proposed concept through robot service experiments.**

I. Introduction

Due to the rapid aging of the population, a labor shortage in hospitals or care facilities is becoming a serious problem. To mitigate the impact on this problem, the development of a service robot which coexists with human in a daily-life environment is an urgent challenge. On the other hand, since a high level of safety is required for a service robot, the robot needs to acquire a wide variety of surrounding information, and plans and executes a proper service task. However, these functions are quite hard to be implemented in a single robot due to a limitation of payload or processing and sensing capabilities.

To tackle this problem, we have been developing an informationally structured environment and its architecture named Town Management System (TMS). In TMS, a variety of sensors such as a laser range finder or a camera are embedded in an environment and a distributed sensor network is organized [1], [2]. The captured information is integrated and stored to a TMS database. A service robot accesses the TMS database and obtains required information anytime and anywhere. In other words, the service robot is able to acquire a high sensing performance which cannot be realized by itself. We also have started the development of a new TMS named ROS-TMS, which adopts Robot Operating System (ROS) as a middleware so that it makes more flexible to add and replace sensors and robots [3].

Here, let us classify the environmental information managed by the ROS-TMS in terms of "person". The viewpoint of the user can be regarded as the first-person information, and the one of the service robot which provides a service task is the second-person information. In addition, the information obtained by the embedded sensors in the environment can be regarded as the third-person information. From the second and third-person viewpoints, a wide variety of information of the environment can be obtained, and the systems which combine second and third-person information have been reported [4], [5]. However, in the area near a user which is important to provide an appropriate service task, it is often difficult to obtain sufficient information in terms of the resolution and the accuracy due to the occlusions or the distance from sensors.

This paper proposes a new concept named "the fourth-person sensing" for a service robot, which combines conventional second and third person sensing with the first-person sensing obtained by wearable devices. In addition, to show the effectiveness of the fourth-person sensing, we focus on an ambiguous verbal communication between a user and a service robot, and show a robot service experiment triggered by user's ambiguous voice commands.

II. The fourth-person sensing

A. Concept of the fourth-person sensing

The fourth-person sensing is regarded as an imaginary viewpoint from which we can understand the correct situation of the environment objectively by combining the first, second, and third-person information. Let us show an example for correct understanding about this concept with a "novel". In a novel, the viewpoint of the main character provides the first-person information and his partner's viewpoint gives the second-person information. Moreover, the viewpoints of people surrounding them provide the third-person information. In real life, there is no way of obtaining the second and the third-person information directly. However, when we read a novel, we can know these information explicitly from the sentences, imagine the story, and forecast the next scenario. This reader's viewpoint, which we call the fourth-person information, will be quite useful to understand the situation of the world in the novel correctly. The extreme target of the fourth-person sensing is to understand the user's intention or the context correctly by integrating there information from different viewpoints.

As explained in Section I, each viewpoint information has pros and cons for a HCI system. The first-person information is quite useful to recognize user's action and estimate his/her intension. However, the measurement area is narrow and local and fragmented information tends to be obtained. The second-person information has high degrees of freedom in terms of data acquisition comparing with the third-person information obtained by fixed sensors, since the robot equipped with second-person sensors is able to move freely in an environment. However, the payload and the processing performance of the robot is limited and it is almost impossible to acquire sufficient information completely to perform a proper task safely. Though the third-person information is able to measure the target, the user, and the environment comprehensively, sensors tend to be located away from the user and the robot in the environment and sometimes it is hard to obtain accurate information due to occlusion or low resolution. Therefore, the correct recognition of the user's intension from the third-person information is sometimes very difficult.

978-1-4799-8203-5/15/$31.00 ©2015 IEEE

On the other hand, by fusing all-person information in a complementary way, the following advantage will be expected for the instruction to a service robot. Correct understanding for an ambiguous instruction will be realized. The verbal communication is effective for requesting a robot service since it is induced by a user intentionally and can be an explicit trigger for a service request. However in natural conversation, it is often ambiguous and user's intention or request are not clearly expressed. On the other hand, for instance, the first-person images taken by a wearable camera contain a rich information such as what the user is gazing or what the user is doing now. These gaze information or action information taken from the first-person images may make an ambiguous instruction more clear.

In the following sections, we introduce the first, second, and third-person sensors we utilized in the experiments.

B. The first-person sensing

In recent years, several high performance wearable cameras are provided in the market. Especially, smart glasses equipped with on-board cameras are very popular as a HCI device and enable to capture first-person images. In this research, we adopted Moverio BT-200AV (Epson) smart glass as shown in Fig.1. First-person images have been used to sense the environment and the user's activities from the user's view point, for various purposes of activity recognition in daily life, sports scene, and so on [6].

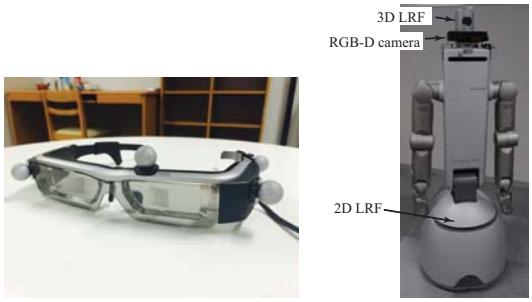


Fig. 1. (Left) the first-person sensor in a wearable device, (right) the second-person sensors on a service robot.

C. The second-person sensing

A service robot needs to acquire user or environmental information correctly by on-board sensors for executing a proper task safety. The obtained information is regarded as the second-person information. The service robot used in the experiments (SmartPAL V, Yaskawa, Fig.1) is equipped with a 3D laser range finder (HDL-32e, Velodyne) and a RGB-D camera (Xtion, ASUS) on a head, and a 2D laser range finder (TopUrg, Hokuyo) on a body.

D. The third-person sensing

As introduced in Section I, we are developing the distributed sensor network and information management architecture named ROS-TMS. Figure 2 shows the experimental room which is managed by ROS-TMS. In this room, laser range finders (UTM-30LX-EW, Hokuyo), RGB-D cameras (Xtion, ASUS), the position tracking system (Vicon MX, Vicon), RFID tag, intelligent cabinet/refrigerator system consisting of RFID-tag readers and load cells, etc. are embedded and environmental information is collected and stored in the TMS database. Position of objects, robots, and humans are tracked by LRF, Vicon MX and the intelligent cabinet/refrigerator. Especially the intelligent cabinet/refrigerator is able to detect not only the object name by reading RFID tag attached on the object but also the position of the object using the force distribution measured by the load cells. These informations are obtained from the embedded sensors in the environment, thus we regard these sensors as the third-person sensors.

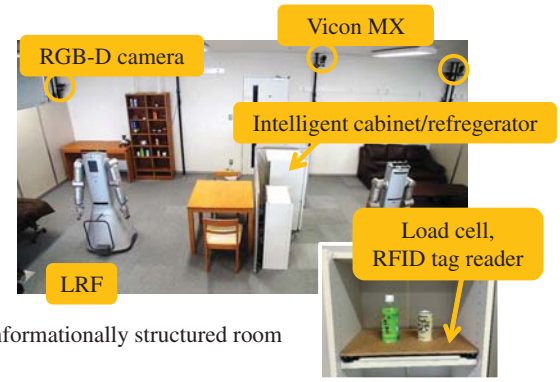


Fig. 2. The third-person sensors embedded in an environment

III. Applications of the Fourth-Person Sensing

The most promising application of the fourth-person sensing is the correct understanding of user's instruction from an ambiguous verbal communication. In this section, we focus on a fetch-and-carry task by a service robot, and show an application of the fourth-person sensing which performs a service task appropriately even if the user's instruction is ambiguous and the system cannot recognize it correctly by the conventional second or third-person sensors.

The typical scenario is shown in Fig.3. There are three objects (”pet bottle”,”bucket”, and ”watering pot”) in the environment, which are all related to ”water”. We can imagine that ”pet bottle”, ”bucket”, and ”watering pot” are related to ”food”, ”cleaning”, and ”gardening”, respectively.

At first, the user during a meal instructs the robot to bring ”water” by voice. However, since there are plural objects related to ”water” in this environment as described above, the robot cannot determine the proper one which fits the user's preference from such an ambiguous oral information. These ambiguous instructions are often encountered in our daily life.

On the other hand, if the information about the user's action, for instance ”eating” in this case, can be obtained, the robot recognizes that the user might ask to take something related to food, and is able to choose a pet bottle for drink from plural candidates. As obviously shown in this example, human action is quite important to understand the user's intension.

To recognize a user's action, the first-person vision is a powerful information source comparing the second and third-person sensors. Though the second and third-person sensors are also helpful for action recognition, the first-person vision can give a more reliable information since the human action is tightly related to the image he/she has seen and the ego-motion which might be closely related to the action can also be captured directly from the first-person vision.

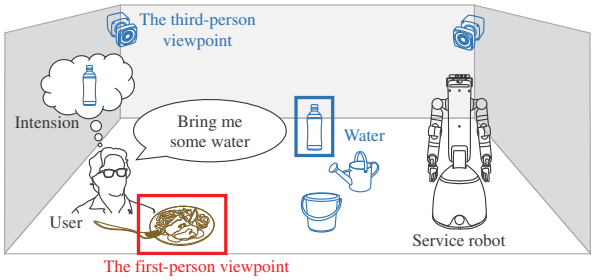
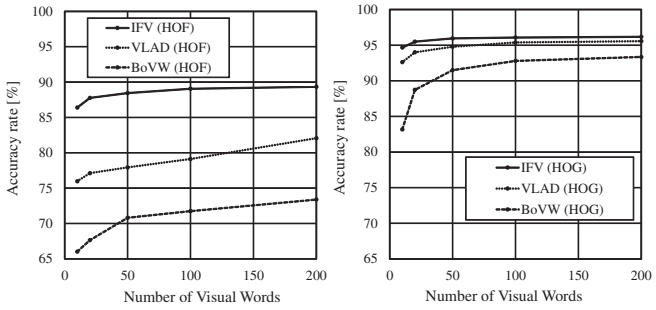


Fig. 3. An example of service scenarios.



(a) HOF (b) HOG

Fig. 4. Accuracy rate for various numbers of Visual Word k: (a) Maximum rate is 89.3% with fisher vector, k=200 (b) Maximum rate is 96.2% with fisher vector, k=200.

In this section, we assume the scenario for taking a "water-related" object via oral instruction described above, and demonstrate the effectiveness of the fourth person sensing combining the first and third-person sensors.

A. System configuration

We built a distributed processing system for a forth person sensing consisting of a smart glass and a processing server.

1) Smart glass: We utilized a Moverio BT-200AV (Epson) smart glass for capturing a first-person images. Moverio is equipped with Android OS, an overlay display, a camera, and a microphone, and is able to capture first-person images and voice at the same time. Captured images and voice are compressed and transmitted to a processing server with a frame rate of captured video periodically. The information such as the server status and the recognized action is overlayed to the display of the Moverio to confirm the processing results.

2) Processing server: The processing server recognizes the user's action periodically, and it suggests candidate objects after the server receives user's voice. Both processes are explained in the following section.

IV. Understanding a User's Request

In this section, firstly we explain several methods, which are commonly used for activity recognition [8], [9]. Secondly, we describe the way how to select the correct one from multiple objects, which are candidates based on the user's voice commands.

A. Activity recognition using the first person vision

In this section, first we describe methods to extract local features from videos, followed by three feature encoding methods for more efficient representation of motion information in videos. Then we show experimental results of the activity recognition. In the video datasets there are five categories: (i) reading a book, (ii) eating, (iii) watching a tree, (iv) watching a robot, and (v) looking around.

1) Feature extraction: We utilize the Space-Time Interest Points (STIP), which was proposed by Laptev [7], as a local feature detector, and we describe feature vectors using Histogram of Optical Flow (HOF) and Histogram of Oriented Gradients (HOG). After extracting feature vectors, we apply the principal component analysis (PCA) to original feature vectors, while keeping 95 % contribution rate, for dimension reduction.

For a more efficient representation of motion information in videos, we employ the concept of feature encoding. More specifically, we use three encoding methods: (i) visual words [10], (ii) vector of locally aggregated descriptors (VLAD) [12], and (iii) fisher vector [11].

2) Activity recognition: We collected first-person videos using the camera of the Moverio. Each class has 50 sequences, thus totally there are 250 (=50×5) sequences. Each sequence is 10 seconds with 30 fps frame rate, and image resolution is 320 × 240. After we obtain encoded feature vectors from all sequences, we utilize Linear Support Vector Machine (Linear SVM) for activity recognition. Here, we used 2-folds for evaluations. The mean classification accuracy was obtained by repeating this random training-test splits for 100 times. The number of clusters in visual words, which is used for the three encoding methods, is changed as 10, 20, 50, 100, and 200.

Figure 4 shows the results of each feature encoding method and each feature vector. From these results, we can see that the results of HOG descriptor showed better performance than those of HOF descriptor. Moreover, Fisher Vector shows the best performance among all encoding methods. Thus in the following experiments, we use HOG as a feature descriptor and Fisher Vector as a feature encoding method.

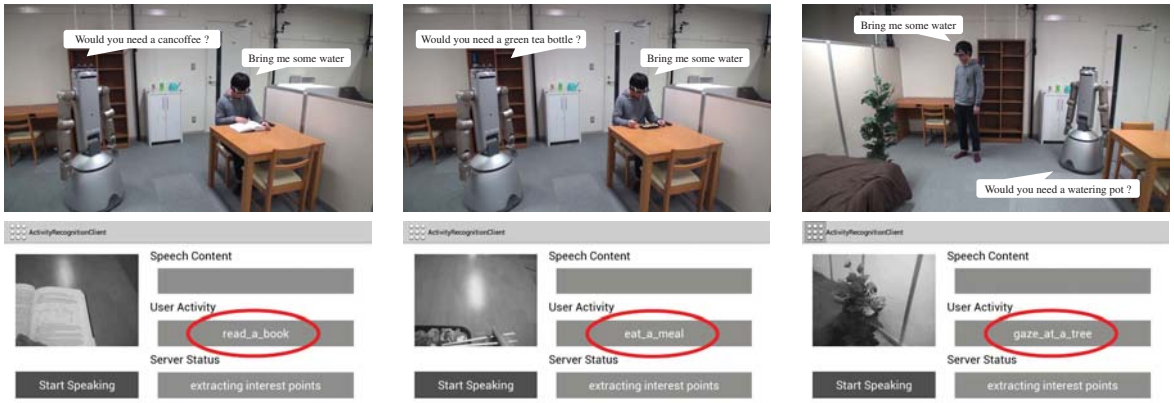
B. Understanding a user's request

In this system we take the advantage of "tag" information, which is assigned to each object stored in TMS database. Tag information of an object consists of keywords which explain the object. For example tags of "drink", "tea", and "water" are assigned to a bottle of tea. Each activity also has tag information, so that we can count the number of tag information, which is the same tag information of the activity, at each candidate object. We then give priority to each candidate object, which is proportional to the number of counted tag.

V. Experiment

A. Experimental settings

We explain experimental settings to evaluate the proposed system. A user wearing the Moverio asks to a robot to give him water, while doing three different activities, "reading a book", "eating" and "watching a tree". A service robot gives him the correct one he wants, depending on his request. We assume that a canned coffee is the right one for "reading a book", a bottle of tea for "eating", and a watering pot for "watching a tree". Table I shows a list of tag information of each activity, and Table II shows object information related to water.



(a) Reading a book (b) Eating a meal (c) Gazing at a tree

Fig. 5. Experiment: Figures on upper row shows actual images and a user did some activities. Figures on lower row shows the screen of wearable camera. Recognized results are shown as a "User Activity" (red circles)

TABLE I. TAGS ASSOCIATING TO ACTIVITIES

Activity	Tag
read a book	drink, coffee
eat a meal	drink, tea
gaze at a tree	pot

TABLE II. OBJECTS STORED IN THE DATABASE

Category	Name	Tag
Coffee	cancofee	drink, coffee, water
Tea	greentea_bottle	drink, tea, water
Watering Pot	watering_pot	pot, water

B. Experimental results

We explain experimental results of the three activities. Figure 5 (a) shows an actual image of "reading a book" and the system estimated the user's activity as "read a book", which is shown in "User Activity" with red circle. After the user asked the service robot to bring water for him, the robot successfully chose a canned coffee. Figure 5 (b) shows the actual image of "eating", and his activity was estimated as "eat a meal". After he asked the service robot to bring water for him, the robot successfully chose a bottle of green tea. Figure 5 (c) shows the actual image of "watching a tree", and his activity was estimated as "gaze at a tree". After he asked the service robot to bring water for him, the robot successfully chose a watering pot. These experimental results show that the system could successfully understand the user's request based on his activity, even though there are some ambiguities and multiple candidates in the scene.

VI. CONCLUSIONS

This paper introduced a new concept of the fourth-person sensing, which combines conventional second and third person sensing for a service robot with the first-person sensing. As an example scenario of the fourth-person sensing, we focused on an ambiguous verbal communication with a service robot, and we developed a system which combines first and third-person sensing. Experimental results showed that the effectiveness of the proposed system and we confirmed that the fourth-person sensing contributes for more accurate understandings of user's requests.

The proposed system for the fourth-person sensing does not include the second-person sensing yet. Thus the future work includes combining first, second, and third-person sensing, and developing a system for much more accurate understanding of a user's intention and a context of the scene.

ACKNOWLEDGMENT

The present study was supported by a Grant-in-Aid for Exploratory Research (26630099).

REFERENCES

[1] R. Kurazume, Y. Iwashita, K. Murakami, and T. Hasegawa, Introduction to the robot town project and 3-d co-operative geometrical modeling using multiple robots, in 15th International Symposium on Robotics Research (ISRR 2011), 2011.

[2] O.M. Mozos, T. Tsuji, H. Chae, S. Kuwahata, T. Hasegawa, K. Morooka, and R. Kurazum, The intelligent room for elderly care, in 1The fifth International Workshop on the Interplay between Natural and Artificial Computation (IWINAC2013), 2013.

[3] https://github.com/irvs/ros_tms/wiki.

[4] F. Cavallo, M. Aquilano, M. Bonaccorsi, R. Limosani, A. Manzi, M. Carrozza, and P. Dario, On the design, development and experimentation of the astro assistive robot integrated in smart environments, in IEEE International Conference on Robotics and Automation (ICRA), 2013.

[5] F. Martinez, A. Carbone, and E. Pissaloux, Combining firstperson and third-person gaze for attention recognition, in IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013.

[6] K.M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, Fast unsupervised ego-action learning for first-person sports videos, in Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[7] I. Laptev, On space-time interest points, International Journal of Computer Vision, vol. 64, no. 2-3, pp. 107-123, 2005.

[8] M.S. Ryoo and L. Matthies, First-person activity recognition: What are they doing to me?, in Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[9] Y. Iwashita, A. Takamine, R. Kurazume, and M.S. Ryoo, First-person activity recognition from animal videos, in International Conference on Pattern Recognition, 2014.

[10] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, Visual categorization with bags of keypoints, in Workshop on statistical learning in computer vision, European Conference on Computer Vision. Prague, 2004, vol. 1, pp. 1-2.

[11] F. Perronnin, J. Sanchez, and T. Mensink, Improving the fisher kernel for large-scale image classification, in European Conference on Computer Visions, pp. 143-156. Springer, 2010.

[12] H. Jegou, M. Douze, C. Schmid, and P. Perez, Aggregating local descriptors into a compact image representation, in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 3304-3311.