Journal of
Sensor and Actuator Networks

MDPI

Article

Activity Recognition Using Gazed Text and Viewpoint Information for User Support Systems

Shun Chiba [†]**, Tomo Miyazaki, Yoshihiro Sugaya and Shinichiro Omachi** [*]

Graduate School of Engineering, Tohoku University, Aoba 6-6-05, Aramaki, Aoba-ku, Sendai 980-8579, Japan; chiba@iic.ecei.tohoku.ac.jp (S.C.); tomo@iic.ecei.tohoku.ac.jp (T.M.); sugaya@iic.ecei.tohoku.ac.jp (Y.S.)

* Correspondence: machi@ecei.tohoku.ac.jp

† Current address: Future Architect, Inc., 1-2-2 Osaki, Shinagawa-ku, Tokyo 141-0032, Japan.

Received: 30 June 2018; Accepted: 31 July 2018; Published: 2 August 2018

Abstract: The development of information technology has added many conveniences to our lives. On the other hand, however, we have to deal with various kinds of information, which can be a difficult task for elderly people or those who are not familiar with information devices. A technology to recognize each person's activity and providing appropriate support based on that activity could be useful for such people. In this paper, we propose a novel fine-grained activity recognition method for user support systems that focuses on identifying the text at which a user is gazing, based on the idea that the content of the text is related to the activity of the user. It is necessary to keep in mind that the meaning of the text depends on its location. To tackle this problem, we propose the simultaneous use of a wearable device and fixed camera. To obtain the global location of the text, we perform image matching using the local features of the images obtained by these two devices. Then, we generate a feature vector based on this information and the content of the text. To show the effectiveness of the proposed approach, we performed activity recognition experiments with six subjects in a laboratory environment.

Keywords: activity recognition; eye tracker; fisheye camera; viewpoint information

1. Introduction

Advanced information technology and various information devices have made the environment where we live remarkably convenient. On the other hand, society is becoming more complicated because we have to deal with various kinds of information. It might be difficult for elderly people or those who are not familiar with information devices to enjoy the merits of the advanced informationization of modern society. Consider a case where you are at a station and are traveling to a destination by train. If you are not good at handling information devices, the first thing you have to do is look at a route map and find the destination station. Then, you will need to find the route information from the current station to your destination: the first railway line where you take a train, transit station, railway line after transit, and so on. Next, you will need to find an appropriate ticket vending machine and purchase a ticket for that destination. Then, it is necessary to find the ticket gate where you catch the train. This procedure will continue until you arrive at your destination. However, such necessary information can be easily provided by information devices. If it was possible to recognize each person's fine-grained activity and provide appropriate support based on that activity, this could be a useful technology for smart cities, where all people, including the elderly, can enjoy the benefits of advanced information technology.

In this paper, we propose a novel fine-grained activity recognition method for user support systems that focuses on identifying the text at which a user is gazing. Text exists everywhere around us and provides various kinds of useful information. Our proposal is based on the idea that the content of the text is related to the activity of the user.

However, it is necessary to keep in mind that using only the content of gazed text is insufficient to estimate the activity. The meaning of the text depends on its location and situation. For example, if the gazed text is a number, you cannot judge whether the user is checking a price when shopping or looking at a sign indicating the distance to a destination. To tackle this problem, we propose the simultaneous use of a wearable device and fixed camera as sensors. These two are assumed to be Internet of things (IoT) devices and are connected to each other. Using these devices, we propose a method to simultaneously obtain the content of the text at which the user is gazing and its global location. Then, the activity of the user is recognized by a machine learning method using this information.

To achieve this, we use an eye tracker as a wearable device and a fisheye camera as a fixed camera. The eye tracker is used to measure the viewpoint of the user and acquire an image of the area around that viewpoint. Recognition of the gazed text of the user is possible by utilizing the eye tracker. The fisheye camera is capable of acquiring a 360° image of the area around the camera with a fisheye lens. The global location of the user's viewpoint is estimated by matching the images acquired with these devices. Finally, the user's activity is recognized using a feature vector calculated from the acquired features. The effect of the proposed approach was tested in the situation where a user buys a ticket at a ticket vending machine in a station.

The main contribution of this paper is twofold. First, we propose a general idea of an activity recognition algorithm utilizing content and location of the text at which a user is gazing. By using text information, more detailed activities can be recognized compared to the target activities of the existing methods described in the next subsection. Second, we construct a system with a wearable device and fisheye camera based on the proposed algorithm, and the effect of the system is experimentally shown. The results show the feasibility of the user support system mentioned above. To the best of our knowledge, this is the first attempt at using the content and global location of the text at which a user is gazing for activity recognition.

Related Work

Many studies on activity recognition have been carried out. One of the typical approaches is to recognize activities using fixed cameras, and this approach has a long history. Polana and Nelson defined activities to be temporally periodic motions possessing a compact spatial structure [1]. They used a periodicity measure for detecting an activity and classified the activity by a feature vector based on motion information. Yamashita et al. proposed a method for human body detection, posture estimation, and activity recognition using an image sequence acquired by a fixed camera [2]. Human body detection and posture estimation were performed using a single frame, and the activity was recognized by combining the information of several frames. Chen et al. used a panoramic camera located at the center of a living room to classify activities [3]. Moving subjects and TV switching were detected by background subtraction. It should also be noted that the method that uses a fixed camera can also be applied to recognize the activities of a group of people. The method proposed by Gárate et al. was used for the tracking and activity recognition of a moving group in a subway station [4]. This method had the advantages of robustness, the ability to process the data for a long video, and the ability to simultaneously recognize multiple events.

Using wearable sensors is another option for activity recognition. Ouchi and Doi proposed a method of utilizing the sound acquired by a microphone, in addition to data from an acceleration sensor [5]. In this method, a user's activity is first roughly classified into resting, walking, or performing an activity using an accelerometer. If it is classified as performing an activity, a more detailed work analysis is conducted using the sound. Because this method uses sound, it is not possible to classify work that has no distinctive sound, and it cannot perform accurate classification in places with loud noises. Zeng et al. proposed a method for recognizing activity by convolutional neural networks using mobile sensors [6]. In their method, the local dependency and scale invariant characteristics could be extracted. Pham used an acceleration sensor in a smart phone or wristwatch-type mobile device [7]. Real-time activity recognition was performed by data processing, segmentation, feature extraction, and classification. Xu et al. introduced the Hilbert–Huang transform to handle nonlinear and non-stationary signals [8]. They proposed a method for extracting multiple features to improve the effect of activity recognition. Liu et al. focused on housekeeping tasks and developed a wearable sensor-based system [9]. In their method, the activity level was also evaluated for each task. Rezaie and Ghassemian focused on the lifetime of sensor nodes considering the actual use situation [10]. Their approach nearly doubled the system lifetime. Twomey et al. conducted a survey of activity recognition methods that use accelerometers [11]. They selected six important aspects of human activity recognition and discussed these topics. In terms of devices, Wang et al. reviewed the wearable sensors for activity recognition [12].

Among the approaches using wearable sensors, the technique of using a wearable camera to recognize the environment and user's activity is called a first-person vision or an egocentric vision method, and has been attracting attention in recent years [13]. Yan et al. proposed a multitask clustering framework to classify daily activities [14]. They introduced two novel clustering algorithms to determine partitions that are coherent among related tasks. Abebe and Cavallaro proposed the use of a long short-term memory network to encode temporal information [15]. They derived a stacked spectrogram representation for global motion streams so that 2D convolutions could be used for learning and feature extraction. Noor and Uddin used the information of objects to increase the accuracy of activity recognition [16]. They showed that adding object information not only improved the accuracy but also increased the training speed. Nguyen et al. reviewed daily living activity recognition methods that used egocentric vision [17].

Examples of the target activities mentioned in the above references are summarized in Table 1.

Table 1. Example of target activities.

Method	Devices	Target Activities
Polana [1]	Fixed camera	walking, running, swinging, skiing, exercising, and jumping
Yamashita [2]	Fixed camera	walking, picking, bending, boxing, clapping, waving, jogging, running, and walking
Chen [3]	Fixed camera	standing, walking, sitting, falling, and watching television
Ouchi [5]	Wearable sensor	washing dishes, ironing, vacuuming, brushing teeth, drying hair, shaving, flushing the toilet, and talking
Zeng [6]	Wearable sensor	jogging, walking, ascending stairs, descending stairs, sitting, and standing
Pham [7]	Wearable sensor	running, walking, sitting, standing, jumping, kicking, going-up stairs, going down-stairs, laying, and unknown activities
Xu [8]	Wearable sensor	lying, sitting, standing, walking, running, cycling, nordic walking, watching television, computer work, driving a car, ascending stairs, descending stairs, vacuuming, ironing, folding laundry, house cleaning, playing soccer, and rope jumping
Liu [9]	Wearable sensor	hanging clothes, folding clothes, wiping furniture, sweeping floor, mopping floor, vacuuming floor, scrubbing floor, digging, filling, moving items (on the floor), moving items (upstairs), and moving items (downstairs)
Rezaie [10]	Wearable sensor	standing, sitting, lying down, brushing, eating, walking, and running
Twomey [11]	Wearable sensor	walking, ascending stairs, descending stairs, sitting, standing, lying down, working at computer, walking and talking, standing and talking, sleeping, etc.
Yan [14]	Wearable camera	reading a book, watching a video, copying text from screen to screen, writing sentences on paper, and browsing the internet
Abebe [15]	Wearable camera	going upstairs, running, walking, sitting/standing, and static
Noor [16]	Wearable camera	reaching, sprinkling, spreading, opening, closing, cutting, etc.

However, it is difficult to realize the above-mentioned user support system using these existing approaches. When using the approaches with fixed cameras and wearable sensors, it is difficult to recognize activities other than the motion of the whole body. For example, assuming that the user stands in front of a ticket vending machine at a station, the necessary information is quite different depending on whether the user watches the route map or instructions for the ticket vending machine. It is difficult to distinguish these activities using wearable acceleration sensors or fixed cameras because there is almost no movement of the body. As for the first-person vision approach, it is not easy to obtain the global location of the user, which is also important information for user support. These problems can be solved by using the content and location of the text at which the user is gazing. Textual information exists everywhere around us, and the possibility of looking at text is considered to be high not only when reading a book but also when users are engaged in other activities. The textual information provides a good clue to recognize the user's activity.

2. Materials and Methods

2.1. Proposed Method

The outline of the proposed method is shown in Figure 1. A wearable eye tracker and fixed fisheye camera are used as input devices. For convenience, the images obtained from the eye tracker and fisheye camera are called the eye-tracker image and fisheye image, respectively. Information about the user's viewpoint can also be obtained from the eye tracker. The text at which the user is gazing is detected using the eye-tracker image and viewpoint information. This text is regarded as the text of interest, and it is recognized by an optical character reader (OCR). On the other hand, the eye-tracker and fisheye images are matched to calculate the user's viewpoint in the fisheye image. Hence, the global location of the gazed text is detected. Then, the activity of the user is estimated using the information about the text and its location. Note that the fisheye camera has a drawback that the acquired image is distorted and its resolution is low, which makes image recognition difficult. Therefore, the fisheye image is only used to detect the user's viewpoint, and text recognitoin is performed using the eye-tracker image.

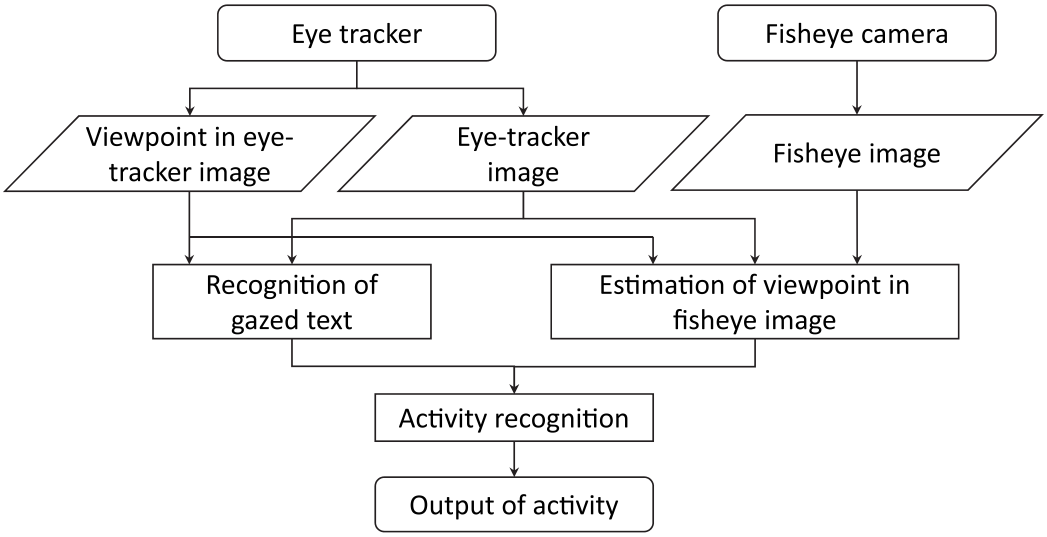

Eye tracker
Fisheye camera
Viewpoint in eye-tracker image
Eye-tracker image
Fisheye image
Recognition of gazed text
Estimation of viewpoint in fisheye image
Activity recognition
Output of activity

Figure 1. Outline of the proposed method.

2.1.1. Recognition of Gazed Text

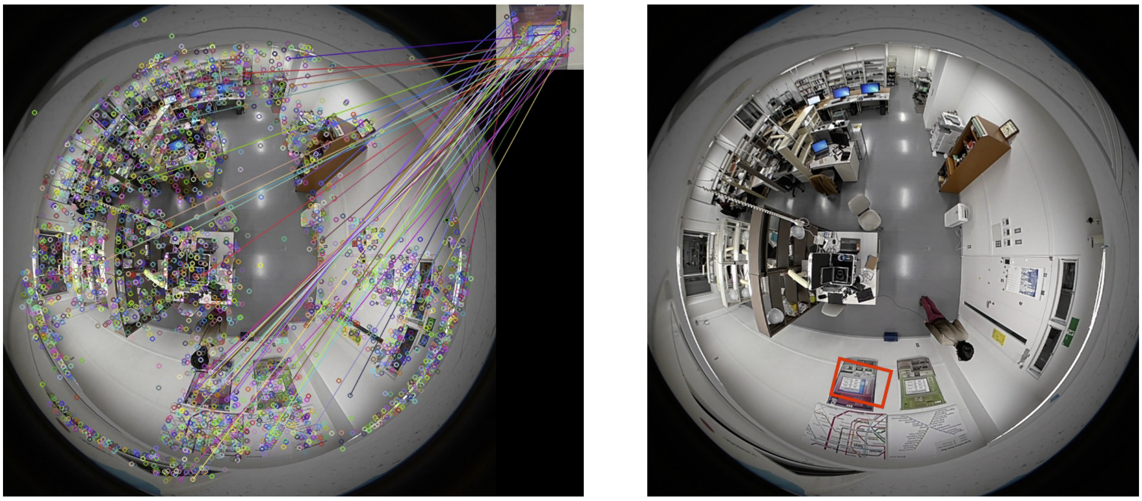
An image of the area around the viewpoint is segmented from the eye-tracker image and is binarized. Then, the text recognition is performed using the Tesseract-OCR [18]. When a text is detected in the image, the coordinates of a rectangle surrounding the recognized text area, the recognition result, and its reliability are output. The distance between the recognized text and the viewpoint position is calculated, and the nearest text is regarded as the gazed text. If the segmented image does not include any text, it is judged that the user is not gazing at any text.

Then, the recognized text is matched using a text database prepared in advance. This database consists of a set of texts and their categories. For the purpose of database construction and text matching, we used SimString [19]. The category of the text is determined through the matching with the database. If there is no matched text in the database, the recognized text is used as is and its category is determined to be *others*.

2.1.2. Estimation of Viewpoint in the Fisheye Image

In order to estimate the user's viewpoint in the fisheye image, image matching is performed between the fisheye image and the eye-tracker image. Because the distortion of the fisheye image is different from that of the eye-tracker image, we adopt two-stage matching. First, the eye-tracker image is used as is to roughly detect the region. Then, the eye-tracker image is converted according to the detected location, and it is used for precise detection. We use the speeded-up robust features (SURF) [20] as the feature for image matching.

In the first stage, we extract the keypoints of SURF from each image and perform feature matching. An example is shown in Figure 2a. The large image is the fisheye image, and the small image at the upper-right corner is the eye-tracker image. The matched keypoints are connected by lines. Each extracted keypoint has information about the rotation angle and scale. Using this information, the scale and difference in the rotation angles between the two images are adjusted. The position of the eye-tracker image in the fisheye image is estimated by searching for the position that minimizes the sum of the distances between matched feature points. An example of a roughly detected region is shown in Figure 2b.



(**a**) (**b**)

Figure 2. First stage of image matching. (**a**) feature point matching. The large image is the fisheye image, and the small image at the upper-right corner is the eye-tracker image. (**b**) roughly detected region. This region that corresponds to the eye-tracker image is indicated as a red rectangle in the fisheye image.

The accuracy of this matching is not very high because the distortion of the fisheye lens is not taken into account. Therefore, we convert the eye-tracker image so that the distortion is the same as that at the detected position in the fisheye image by calculating the corresponding points of these images [21]. Then, the image matching of the second stage is performed using the converted image. Figure 3 shows an example of the result of the second-stage image matching. It can be confirmed that the accuracy was much improved compared with the first stage matching (Figure 2a). Finally, we calculate the viewpoint position in the fisheye image using the positional relationship obtained by the image matching.

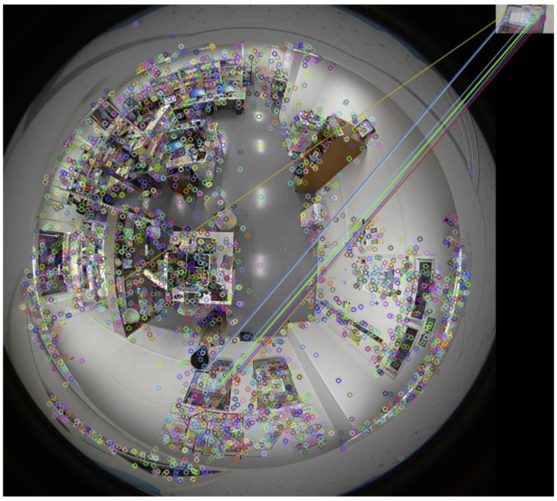


Figure 3. Second stage of image matching.

2.1.3. Activity Recognition

Activity recognition is performed using the acquired gazed text and viewpoint location in the fisheye image. The accuracy of the eye-tracker varies depending on the user or misrecognition of the text. Therefore, we adopt the random forest [22] method, which is less sensitive to noise, as the classification algorithm. In the training phase, subsamples are selected from the training data by random sampling to construct decision trees. These decision trees are used for classification.

We use six-dimensional feature vector (x, y, n, t, d, c) as the feature vector. The meaning of each element is listed in Table 2. The number of characters is used for distinguishing long and short texts, which normally represents guidance and place names, respectively. The average of character codes is used for distinguishing between numbers and alphabets. The distance between the viewpoint and the gazed text is used to judge whether or not the user is really gazing at the text.

Table 2. Elements of feature vector.

Element	Meaning
x	x-coordinate of the viewpoint location in the fisheye image
y	y-coordinate of the viewpoint location in the fisheye image
n	number of characters in the gazed text
t	average of character codes in the gazed text
d	distance between the viewpoint and the gazed text
c	category of the gazed text

2.2. *Experiment*

2.2.1. Equipment

As a wearable eye tracker, we used SMI eye tracking glasses (https://www.smivision.com/). This device is equipped with a camera for obtaining an infrared image of the eyes of the user, and a camera for obtaining the field of view of the user. This makes it possible to record the field of view and viewpoint of the user at the same time. Because the frame rates for the field of view and viewpoint are 24 fps and 30 fps, respectively, synchronization is required.

As a fixed fisheye camera, we used the Kodak PIXPRO SP360 4K (https://www.kodak.com/). This is an omnidirectional camera equipped with one fisheye lens. It is possible to acquire an image that covers 360° in the horizontal direction and 235° in the vertical direction.

2.2.2. Experimental Environment

We constructed an experimental environment to simulate a ticket vending machine in a station. An image of a ticket vending machine and route map was printed on paper and affixed to the wall to reproduce the vicinity of the ticket vending machine. The constructed environment is shown in Figure 4a. A portion of the text database used is listed in Table 3.

In order to reproduce the surveillance camera in the station, we installed the fisheye camera on the ceiling. It was installed in front of the ticket vending machine, about 2.5 m away from the wall. A fisheye image acquired by the fisheye camera is shown in Figure 4b.

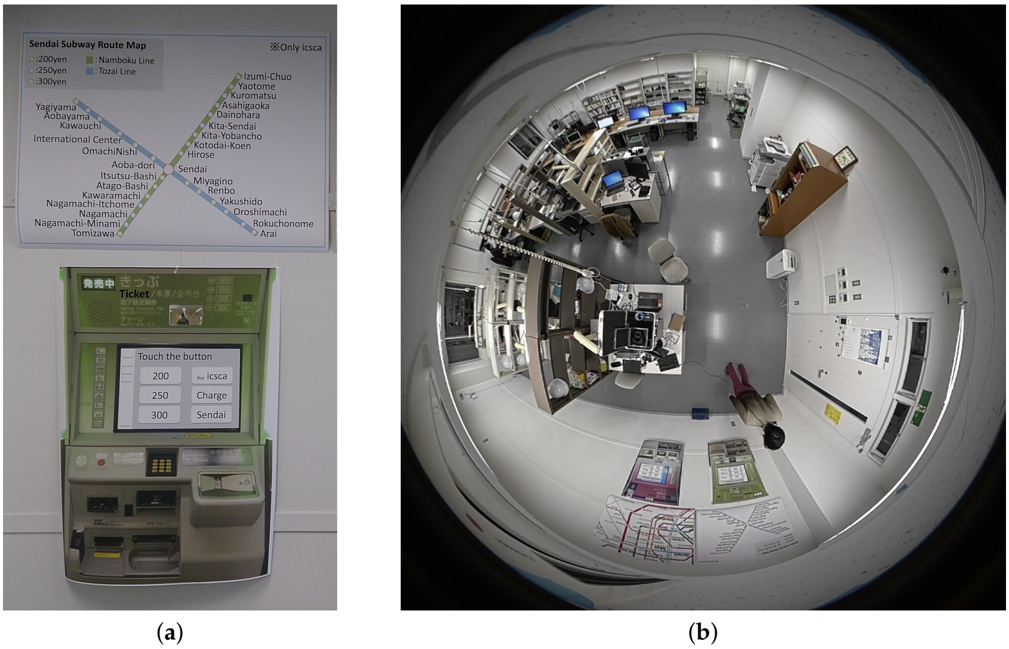


Figure 4. Experimental environment. (**a**) printed image of ticket vending machine and route map affixed to wall; (**b**) omnidirectional image acquired by fisheye camera.

Table 3. Portion of text database.

Category	Texts
Guidance	Touch the button, Tozai Line, Nanboku Line,...
Station name	Aobayama, Sendai, International Center,...
Price	200, 250, 300,...

2.2.3. Training Data

The training data were constructed as follows. First, a subject was asked to stand in the experimental environment. Then, he or she was asked to look at the ticket vending machine and route map by moving their viewpoint (see Figure 5). We simultaneously recorded the viewpoint information and video of the field of view. Frames were extracted by synchronizing these data, and feature vectors were created with the values described in Section 2.1.3. The length of the captured video was approximately 110 s, and the number of feature vectors was 2040.

A label describing the user's activity was manually assigned to each feature vector. Considering the gazed text and location of the text, the activity was classified into the following eight types:

- Looking at the route map to check the price,
- Looking at the route map to check the station name,
- Looking at the route map to look for guidance,

- Looking at the ticket vending machine to check the price,
- Looking at the ticket vending machine to check the station name,
- Looking at the ticket vending machine to look for guidance,
- Operating the ticket vending machine,
- Others.

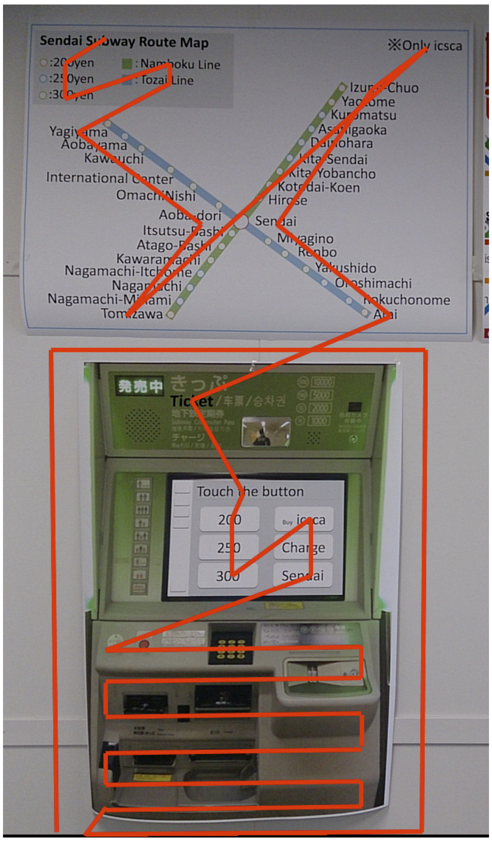


Figure 5. Viewpoints of training data.

3. Results and Discussion

3.1. Experimental Results

First, the accuracy of the proposed method was verified. We conducted the experiment with six subjects in the constructed experimental environment. The subjects were instructed to move their viewpoint to purchase a ticket for Aobayama Station while assuming that they were currently at Sendai Station. Because the distance to the ticket vending machine was not provided, each subject took the usual distance. The method used to create the experimental data was the same as that used to create the training data. The activity in each frame was recognized by the proposed method.

The number of video frames for each subject, number of correctly recognized frames, and recognition accuracy are listed in Table 4. The accuracy calculated with all the frames of all the subjects was 75.4%. Note that this is the accuracy for individual frames. Considering that recognition using several frames and recognition results in a time series can also be used when the method is actually used for user support, it is considered that activity recognition was possible by the proposed method under the experimental environment. There was a difference of more than 20 percentage points in the recognition accuracy, and it could be confirmed that there were variations among the subjects.

Table 4. Experimental results.

Subject	Number of Frames	Number of Correctly Recognized Frames	Accuracy
1	903	668	74.0%
2	620	403	65.0%
3	750	558	78.4%
4	717	462	64.4%
5	441	344	78.0%
6	1314	1141	86.9%
Total	4745	3576	75.4%

3.2. Discussion

Next, we analyze the causes of recognition failures and show improvement plans. The causes of recognition failures were roughly classified into two types. The first was the motion blur that occurred during movement of the viewpoint, as shown in Figure 6. There are two ways to move the viewpoint: moving only the eyes without moving the head, and rotating the head without moving the eyes. Motion blur frequently occurs in the experimental data acquired from subjects who frequently rotate their head. As a result, both the text recognition accuracy and image matching deteriorated. To overcome this problem, it will be useful to use an eye tracker with a higher frame rate.

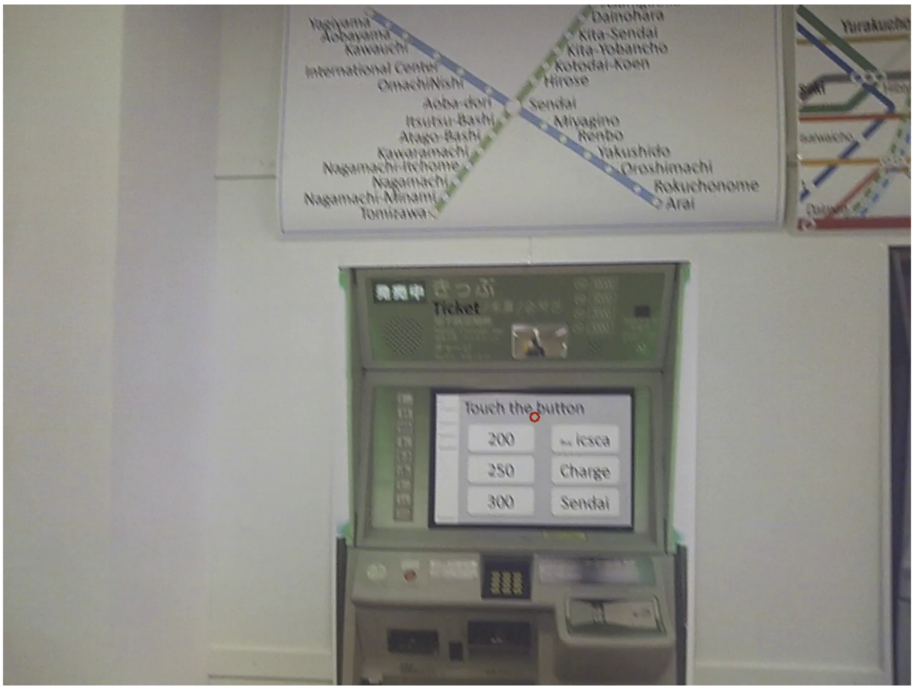


Figure 6. Example of failure caused by motion blur.

The second reason was the error of the viewpoint location detected by the eye tracker. If the gazed text was incorrectly detected, the activity of the user was not correctly recognized. An example of this is shown in Figure 7. Although the subject was looking at the guidance of the "Sendai Subway Route Map", the detected viewpoint was slightly shifted downward. As a result, the activity was incorrectly recognized as "Looking at the route map to check the price". The eye tracker used in this experiment has the characteristic that the viewpoint location to be detected tends to be shifted when looking at the edge of the field of view. Therefore, if the subject moves only their eyes without moving their head, the detected location of the viewpoint is greatly shifted. To solve this problem, using multiple texts included in a wide area around the viewpoint will be effective.

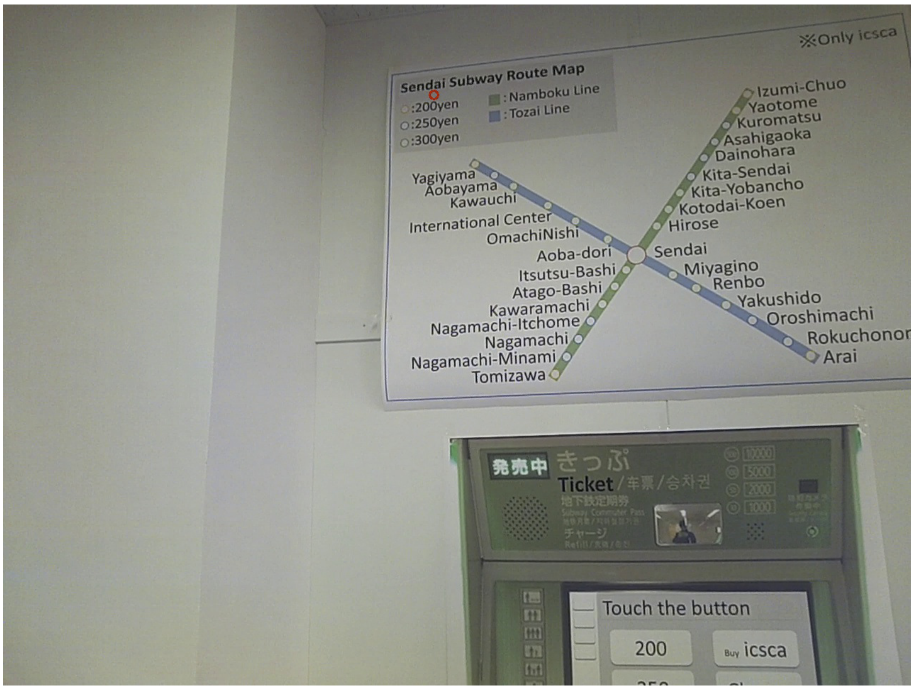


Figure 7. Example of failure caused by misdetection of viewpoint. Although the subject was looking at the guidance of the "Sendai Subway Route Map", the detected viewpoint was slightly shifted downward.

4. Conclusions

We proposed a fine-grained activity recognition method that uses a wearable eye tracker and fixed fisheye camera. The information about the text at which the user is gazing is utilized based on the idea that the content of the text is related to the activity. The proposed activity recognition method consists of three processes: the gazed text recognition, estimation of the viewpoint in the fisheye image, and classification of the activity using the feature vector. To obtain the global location of the text, we performed image matching of the images obtained by these two devices. We demonstrated that activity recognition was possible under the experimental environment created by simulating the vicinity of a ticket vending machine in a station.

Although only an experiment in this limited environment was conducted at this time, we believe that the proposed approach can be applied for various purposes. In order to recognize activities in various situations, it will be necessary to further select and add features. Then, the method can be applied to recognize the activities in stores and libraries, in addition to stations. It is also important to construct a concrete user support system with the proposed approach.

Author Contributions: Funding acquisition, S.O.; Methodology, S.C.; Project administration, S.O.; Software, S.C.; Supervision, T.M. and Y.S.; Validation, T.M. and Y.S.; Writing—Original draft, S.C.; Writing—Review & editing, S.O.

Funding: This work was partially supported by JSPS KAKENHI Grant Numbers JP16H02841, JP16K00259.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Polana, R.; Nelson, R. Recognizing activities. In Proceedings of the 12th International Conference on Pattern Recognition, Jerusalem, Israel, 9–13 October 1994; Volume 1, pp. 815–818.
2. Yamashita, T.; Yamauchi, Y.; Fujiyoshi, H. A single framework for action recognition based on boosted randomized trees. *IPSJ Trans. Comput. Vision Appl.* **2011**, *3*, 160–171. [CrossRef]

3. Chen, O.T.-C.; Tsai, C.-H.; Manh, H.H.; Lai, W.-C. Activity recognition using a panoramic camera for homecare. In Proceedings of the 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, Lecce, Italy, 29 August–1 September 2017; pp. 1–6.
4. Gárate, C.; Zaidenberg, S.; Badie, J.; Brémond, F. Group tracking and behavior recognition in long video surveillance sequences. In Proceedings of the 2014 International Conference on Computer Vision Theory and Applications, Lisbon, Portugal, 5–8 January 2014.
5. Ouchi, K.; Doi, M. Smartphone-based monitoring system for activities of daily living for elderly people and their relatives etc. In Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication, Zurich, Switzerland, 8–12 September 2013; pp. 103–106.
6. Zeng, M.; Nguyen, L.T.; Yu, B.; Mengshoel, O.J.; Zhu, J.; Wu, P.; Zhang, J. Convolutional neural networks for human activity recognition using mobile sensors. In Proceedings of the 2014 6th International Conference on Mobile Computing, Applications and Services, Austin, TX, USA, 6–7 November 2014; pp. 197–205.
7. Pham, C. MobiRAR: Real-time human activity recognition using mobile devices. In Proceedings of the 2015 Seventh International Conference on Knowledge and Systems Engineering, Ho Chi Minh City, Vietnam, 8–10 October 2015; pp. 144–149.
8. Xu, H.; Liu, J.; Hu, H.; Zhang, Y. Wearable sensor-based human activity recognition method with multi-features extracted from Hilbert-Huang transform. *Sensors* **2016**, *16*, 2048. [CrossRef] [PubMed]
9. Liu, K.-C.; Yen, C.-Y.; Chang, L.-H.; Hsieh, C.-Y.; Chan, C.-T. Wearable sensor-based activity recognition for housekeeping task. In Proceedings of the 2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks, Eindhoven, Netherlands, 9–12 May 2017; pp. 67–70.
10. Rezaie, H.; Ghassemian, M. An adaptive algorithm to improve energy efficiency in wearable activity recognition systems. *IEEE Sens. J.* **2017**, *17*, 5315–5323. [CrossRef]
11. Twomey, N.; Diethe, T.; Fafoutis, X.; Elsts, A.; McConville, R.; Flach, P.; Craddock, I. A comprehensive study of activity recognition using accelerometers. *Informatics* **2018**, *5*, 27. [CrossRef]
12. Wang, Y.; Cang, C.; Yu, H. A review of sensor selection, sensor devices and sensor deployment for wearable sensor-based human activity recognition systems. In Proceedings of the 10th International Conference on Software, Knowledge, Information Management & Applications, Chengdu, China, 15–17 December 2016; pp. 250–257.
13. Kanade, T.; Hebert, M. First-person vision. Proc. IEEE. **2012**, *100*, 2442–2453. [CrossRef]
14. Yan, Y.; Ricci, E.; Liu, G.; Sebe, N. Egocentric daily activity recognition via multitask clustering. *IEEE Trans. Image Process.* **2015**, *24*, 2984–2995. [CrossRef] [PubMed]
15. Abebe, G.; Cavallaro, A. A long short-term memory convolutional neural network for first-person vision activity recognition. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 1339–1346.
16. Noor, S.; Uddin, V. Using context from inside-out vision for improved activity recognition. *IET Comput. Vision* **2018**, *12*, 276–287. [CrossRef]
17. Nguyen, T.-H.-C.; Nebel, J.-C.; Florez-Revuelta, F. Recognition of activities of daily living with egocentric vision: A review. *Sensors* **2016**, *16*, 72. [CrossRef] [PubMed]
18. Smith, R. An overview of the Tesseract OCR engine. In Proceedings of the Ninth International Conference on Document Analysis and Recognition, Parana, Brazil, 23–26 September 2007; pp. 629–633.
19. Okazaki, N.; Tsujii, J. Simple and efficient algorithm for approximate dictionary matching. In Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, 23–27 August 2010; pp. 851–859.
20. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-up robust features (SURF). *Comput. Vision Image Underst.* **2008**, *110*, 346–359. [CrossRef]
21. Mori, T.; Tonomura, M.; Ohsumi, Y.; Goto, S.; Ikenaga, T. High quality image correction algorithm with cubic interpolation and its implementations of dedicated hardware engine for fish-eye lens. *J. Inst. Image Electron. Eng. Jpn.* **2007**, *36*, 680–687.
22. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]