Intention Inference from 2D Poses of Preliminary Action Using OpenPose

Ryuji Tanaka
Graduate School of Science and Engineering, Saga University
Saga, Japan
19704011 @edu.cc.saga-u.ac.jp

Chika Oshima
Faculty of Science and Engineering, Saga University
Saga, Japan
karin27@sa3.so-net.ne.jp

Koichi Nakayama
Graduate School of Science and Engineering, Saga University
Saga, Japan
knakayama@is.saga-u.ac.jp

ABSTRACT

Caregivers in nursing facilities are too busy to pay attention constantly to care receivers. Recently, some cameras have been set up in some nursing facilities. However, the caregivers cannot continuously monitor the care receivers on a display in the daytime. Now two-dimensional (2D) poses of many people in an image can be detected by OpenPose software; it can detect skeletons of humans by using a deep learning method. Therefore, we thought that if 2D poses of care receivers were detected by OpenPose immediately before a new action (preliminary action) and if the poses could be classified by a deep learning method, it might be possible to infer the care receivers' intentions. In this paper, we created a learning model that can discriminate the preliminary action based on coordinate data of keypoints detected by OpenPose from an image of a person. We examined whether a subject's action that was going to interrupt a conversation could be predicted or not. The result suggested that the learning model can discriminate the preliminary action of the subject by the coordinate data[1].

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence;** *Distributed artificial intelligence*, Intelligent agents • **Human-centered computing** → **Human computer interaction (HCI)**; *HCI theory*, concepts and models

KEYWORDS

Conversation, Interruption

ACM Reference format:

R. Tanaka, C. Oshima, and K. Nakayama. 2019. SIG Proceedings Paper in word Format. In *Proceedings of ACM GECCO conference, Prague, Czech Republic, July 2019 (GECCO'19)*, 4 pages.
DOI: 10.1145/3319619.3326886

GECCO'19, July 13-17, 2019, Prague, Czech Republic

DOI: 10.1145/3319619.3326886

1 INTRODUCTION

Many caregivers in nursing facilities want to consider each care receiver's feelings and identify the care receiver's irregular state before an accident. However, the caregivers cannot continuously watch each care receiver because there are too few caregivers in Japan. Some care receivers often go outside the nursing facilities. Others cannot tell the caregivers that they want to go to the toilet, go back to their beds, not do a recreation activity.

Some nursing facilities set cameras in an entrance hall, a corridor, and/or in some rooms. Especially at night, one caregiver alone can monitor the states of care receivers in some spaces simultaneously through a display. However, in the daytime, it is difficult to continuously perceive and monitor via the display. Furthermore, it is difficult for the caregivers to observe and predict what the care receivers are going to do.

Tamaki proposed a method that senses actions that indicate the desire to speak to the other participants of a Web conference [1]. Tamaki called this type of action "preliminary action." He extracted four actions as preliminary actions. The first is that a person moves his/her hands to and/or around his/her face. The second is that the person inclines his/her head to one side. The third is that the person nods. The fourth is that the person gives feedback using positive sounds, laughter, and/or interjections that express agreement, and so on.

In this way, if the caregivers can recognize a preliminary action of the care receivers, the caregivers can plan how they will deal with the care receivers. Furthermore, the care receivers will be happy because the caregivers recognize the care receivers' feelings and desires without using explicit words.

The important thing is that the caregivers become able to recognize the care receivers' states without continuously monitoring the display. Now two-dimensional (2D) poses of many people in an image can be detected. OpenPose [2][3] detects human bodies, hand, faces and feet in 135 "keypoints (characteristic points)," even in a single image. Therefore, we suppose that if the coordinates that express a person's pose can be

used to classify it into some categories of preliminary actions, caregivers can receive a notification only when the care receiver is going to do something that needs help.

In this paper, as the first step of the research, we recorded actions of healthy university students who were asked to interrupt a conversation of other people. Their actions (2D poses) were detected by OpenPose. We examined whether their actions could be classified under the preliminary actions or not by using machine learning.

2 EXPERIMENT

2.1 Data Collection

Nine healthy male university students participated in an experiment. They were divided into three groups. As shown in Fig.1, in each group, two of them (speakers) were asked to have a conversation for 20 minutes on some themes provided in advance by an experimenter. The other (interrupter) was asked to interrupt the conversation. Twenty objects were prepared and set to one side of each speaker. The interrupter had to ask either speaker to take an object according to orders displayed on a smartphone in front of him. Some examples of the order are listed below. In this situation, both Mr. Smith and Mr. Brown are the speakers.

- You (interrupter) ask Mr. Smith to pass an eraser on the table to Mr. Brown.
- You ask Mr. Brown to pass a pen on the table to Mr. Smith.

The actions of the interrupters were recorded by a video camera (Sony Handycam HDR-CX480). We acquired moving images of 20 minutes' duration.

2.2 Data Process

At first, the first author of this paper divided the images into the preliminary actions for interrupting the conversation and the others (non-preliminary actions). The actions in the three seconds before interrupting the conversation were considered to be preliminary actions.

As shown in Fig.2, coordinates of the keypoints of the interrupter's actions were acquired per 1/30 seconds (one frame) by OpenPose. As shown in Fig.3, we call a "set" the coordinates of three frames per second. Then, the set was sifted per one frame 30 times. Namely, 30 sets of the coordinates were prepared per one preliminary action. Because there were 41 preliminary actions in 20 minutes' conversation, there were 1230 sets for the preliminary actions.

Following the same process, 1230 sets for non-preliminary actions were created.

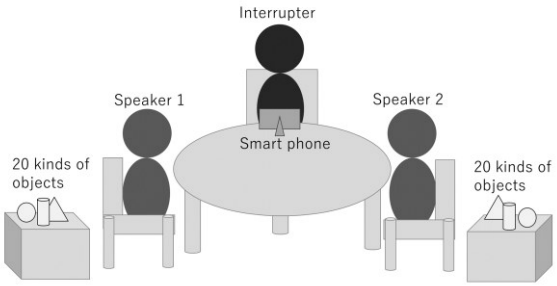

Figure 1: The interrupter was asked to interrupt the conversation.

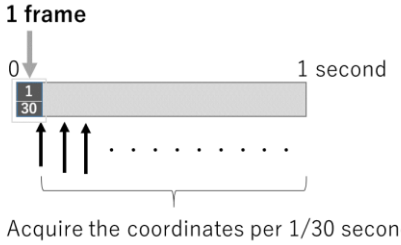

Figure 2: Coordinates of the keypoints were acquired per one frame.

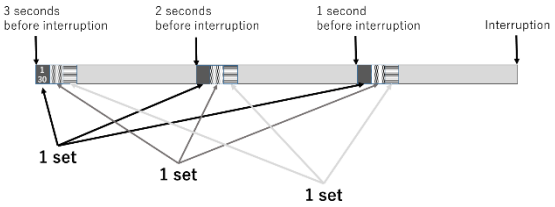

Figure 3: 30 sets of the coordinates were prepared per one preliminary action.

2.3 Analysis

In this paper, we used the coordinate data of one interrupter for analysis. The coordinate data were divided into training data, validation data, and testing data for evaluating a model of machine learning. The data for 17 minutes of the 20 minutes of data were used for training and validation (holdout method) in proportion of nine to one. The rest of the data were used as the testing data. The test was performed by the k-fold cross-validation method.

We used 2214 training data of the coordinates that signified preliminary actions (1107 data) and other actions (1107 data) for training. In the training, 256 data were selected at random from all the training data until all data had been learned. The system learned whether the data should be considered preliminary actions or other actions. Then, the result of the learning was evaluated by

using the validation data. One epoch means that all data had been learned and evaluated. We repeated this process till 2000 epochs. Finally, we examined whether the test data could be classified under the preliminary actions or not.

Table 1 shows the development environment.

Table 1: Development environment.

Classification	Specific
OS	Ubuntu 16.04
GPU environment	GTX 1080 CUDA 8.0 cuDNN 6.0.21
Library	Tensorflow Keras
Execution environment	Jupyter notebook
Development language	Python

2.4 Result

Figure 4 shows the result of one of the validation test using 0-6 and 9-20 minutes for the learning and validation. The ordinate axis shows the accuracy rate. The abscissa shows the number of epochs. We can see that the accuracy rate rose as the number of epochs increased.

Table 2 shows the result of discrimination in 6-9 minutes for the test data. The left column indicates the time of starting the actual preliminary action. The right column indicates the time that the system judged the data to be the preliminary action. The values for recall, precision, and the F-measure were 0.56, 0.71, and 0.63, respectively.

In the same way, the results of discrimination using data for the intervals at the minutes 0-3, 3-6, 9-12, 12-15, 15-18, and 17-20 for the test are shown in Table 3. The F-measures varied according to the time used for the test. The highest value was for the case of using data from minutes 6-9 for the test.

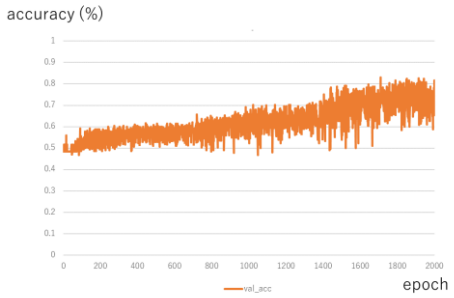

Figure 4: The result of the validation test for the data from minutes 6-9.

3 DISCUSSION

The F-measure suggested that the preliminary action can be discriminated only by the coordinates of the actions using the machine learning. However, the F-measures varied according to the times used as test data. The moving images of the experiment suggest that the F-measure was high when the preliminary action of the interrupter was well-defined. For example, it was easy for the system to consider an action that the subject reseated a chair as the preliminary action. On the other hand, it was difficult for the system to consider the action that the subject had restless eyes before the interruption as the preliminary action because the Openpose cannot detect eyeballs.

In this experiment, we used participants (two speakers and one interrupter) who were close friends because we expected that they would demonstrate clear preliminary actions. However, in fact, most interrupters made modest gestures although they made well-defined actions in the beginning of the conversation. The interrupters did not hesitate about interrupting the conversation of the speakers because they were used to talking together. We think that we should have requested a person who was younger than the speakers to interrupt the conversation. Moreover, we think that the interrupter's action may become well-defined when the conversation is lively. We need to examine this matter more.

Table 2: Results for discrimination within 3 min. from the beginning.

Beginning of preliminary action (correct data)	Action considered preliminary action by the system
	6'03''
6'25''	
6'32''	
6'50''	6'50''
7'17''	7'17''
7'39''	
	7'42''
7'48''	7'48''
8'24''	8'24''
8'29	8'29''
8'40''	

Table 3: Results for recall, precision, and F-value per 3 min. interval.

Data for test (min.)	Recall	Precision	F-measure
0-3	0.50	0.43	0.46
3-6	0.57	0.44	0.50
6-9	0.56	0.71	0.63
9-12	0.43	0.43	0.43
12-15	0.50	0.50	0.50
15-18	0.25	0.29	0.27
17-20	0.57	0.36	0.44
Average	0.48	0.45	0.46

4 CONCLUSION

In this paper, we created a learning model that discriminates the interruption of a conversation from the preliminary action by a machine learning method. The coordinates of the characteristics points of people's actions were detected from images with OpenPose. The system learned whether each 2214 sets of the coordinates' data should be considered preliminary actions or other actions. Then, the experimental result suggested that the learning model can discriminate the preliminary action by the coordinate data.

In the near future, we will this experiment on care receivers with dementia.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 17H01950.

REFERENCES

[1] H. Tamaki, S. Higashino, M. Kobayashi, M. Ihara, and K. Okada. 2012. Method of Reducing Speech Contention in Distributed Conferences. *Journal of Information Processing*, 53,7(2012), 1797-1806. (in Japanese)

[2] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7291-7299.

[3] GitHub. OpenPose. https://github.com/CMU-Perceptual-Computing-Lab/openpose/releases