# Eating and Drinking Recognition via Integrated Information of Head Directions and Joint Positions in a Group

Naoto Ienaga, Yuko Ozasa and Hideo Saito
*Graduate School of Science and Technology, Keio University, Yokohama, Japan*
{*ienaga, saito*}*@hvrl.ics.keio.ac.jp, yuko.ozasa@keio.jp*

Keywords: Action Recognition, Long Short-term Memory, Information Fusion.

Abstract: Recent years have seen the introduction of service robots as waiters or waitresses in restaurants and cafes. In such venues, it is common for customers to visit in groups as well as for them to engage in conversation while eating and drinking. It is important for cyber serving staff to understand whether they are eating and drinking, or not, in order to wait on tables at appropriate times. In this paper, we present a method by which the robots can recognize eating and drinking actions performed by individuals in a group. Our approach uses the positions of joints in the human body as a feature and long short-term memory to achieve a recognition task on time-series data. We also used head directions in our method, as we assumed that it is effective for recognition in a group. The information garnered from head directions and joint positions is integrated via logistic regression and employed in recognition. The results show that this yielded the highest accuracy and effectiveness of the robots' tasks.

## 1 INTRODUCTION

Of late, service robots have been deployed in many fields and for a variety of reasons. For example, they have been used as cyber serving staffs in restaurants, cafes, and so on (Pieska et al., 2013; Qing-Xiao et al., 2010). Most of these robots do not appreciate the subtleties of a situation (that is, what customers are doing) and merely serve dishes in a mechanical manner. This is in contrast to their human counterparts, who are able to serve at appropriate times and behave in a suitable manner. For example, they never talk to customers when the latter are eating or drinking. In order to improve the performance of the robots, it is important to make them understand when customers are eating or drinking.

There are many methods by which such actions can be recognized (Amor et al., 2016; Chua et al., 2014; Iosifidis et al., 2012; Wang et al., 2014). Assessing the joint positions of the human body is a particularly powerful technique; Du et al. (Du et al., 2016) presented one such approach. What makes it effective is that we also use joint positions for eating and drinking recognition.

It is common for customers to frequent restaurants or cafes in groups. This means that a conversation will most likely be going on alongside the produce consumption. We may also make the following assumptions in such a scenario: (1) when one of the customers is speaking, the others may be looking at him or her; (2) one cannot speak while eating or drinking; (3) when some of the customers are eating or drinking, the others may be speaking. Considering these assumptions, head directions are also useful for eating and drinking recognition when it comes to groups.

In this paper, we propose a method to enable such recognition for robot servers. There have been no studies performed eating and drinking recognition in a group. Both joint positions and head directions are used, and integrated with logistic regression (LR) (Bishop, 2006). To handle the time sequence of actions, we use long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) as same as the method in (Du et al., 2016). In the experiment, we compared the results of using only joint positions, only head directions, and the integration of both. Since there have been no datasets suitable for our purpose, a dataset is constructed. The results show that recognition accuracy is the highest when we integrate both joint positions and head directions.

The main novelties of our work are summarized as follows:

- A new task setting; eating and drinking recognition as applied to each customer who come to restaurants or cafes in a group, and interact each other such as having a conversation.



Ienaga, N., Ozasa, Y. and Saito, H.
Eating and Drinking Recognition via Integrated Information of Head Directions and Joint Positions in a Group.
In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)*, pages 527-533
ISBN: 978-989-758-222-6

- Integrating joint positions and head directions via LR.

## 2 RELATED WORKS

Action recognition is one of the most widely studied tasks in the computer vision area. It has various applications, such as in video surveillance (Nayak et al., 2015), sports analysis (Swears et al., 2014), robot action (Koppula and Saxena, 2016), and so on.

Recently, skeletons have frequently been used for action recognition. For example, Amor et al. proposed a framework for analyzing human actions using the shape evolutions of skeletons (Amor et al., 2016). Ding et al., meanwhile, recognized human actions based on skeletons with a profile hidden markov models (Ding et al., 2015). In another study, Vemulapalli et al. represented skeletons in a new way that explicitly modeled the 3D geometric relationships between various body parts using rotations and translations in 3D space (Vemulapalli et al., 2014). Elsewhere, Wang et al. achieved high speed recognition by extracting discriminative features (Wang et al., 2014). Finally, Du et al. proposed an end-to-end hierarchical RNN model with LSTM neurons for skeleton-based action recognition (Du et al., 2016). We also use joint positions and LSTM.

Head directions are often used in speech analysis or communication modeling areas (Ishii et al., 2016; Nihei et al., 2014), but rarely in the computer vision field. In this study, the focus is on enabling service robots to recognize eating and drinking actions of customers who are having a conversation with their dining compatriots while eating or drinking, meaning that head directions are also useful for our purposes. For example, when talking to others, we look at their faces and therefore cannot eat or drink also because we can't talk with our mouth full. In addition, when we listening to others, we also look at them. Considering these assumptions, the analysis of head directions is appropriate for our study.

There have been a few studies focusing on eating and drinking recognition. For example, Iosifidis et al. recognized eating and drinking by using regions of the head and hands in a binary image (Iosifidis et al., 2012). They concatenated these images to produce a 3D volume, in which the third dimension referred to time. In our study, we handle time sequence using LSTM. In another study, Chua et al. made a feature of hand grasping postures and some rules to recognize drinking (Chua et al., 2014). They recognized only drinking, as we recognize both drinking and eating. Both of the abovementioned studies looked at individual actions, which we also do, but related them to the context of the group.

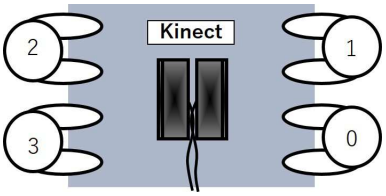


Figure 1: Environment setup in our work. The numbers denote ID labels of subjects *i*.

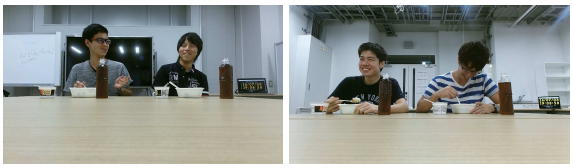

Figure 2: One frame of the dataset.

Our method assess both joint positions and head directions. To consider the differences in characteristics between these, we integrate the recognition results of both and use them for the recognition. This is called the score level fusion method (Ozasa et al., 2015). LR is widely used for such a method. While there have been some studies that have used the Bayesian network instead of LR, the recognition accuracy of the former is lower than that of the latter (Ozasa and Ariki, 2012). Therefore, in our study, we use LR.

## 3 TASK SETTING

As there have been no datasets that are suitable for our purposes, we construct a dataset. In order to reproduce a restaurant setting, we build our dataset in the environment shown in Figure 1.

There are four subjects, who are facing each other in two sets. Two dishes (a main dish and a dessert) and one drink are prepared for them. Two Microsoft Kinect v2s are located at the center of the table. To record the joint positions and head directions of the subjects of each frame, we put Kinect v2s while robots usually have embarked cameras.

There are no special requests, they are allowed to talk freely for about 30 minutes while eating and drinking. Figure 2 shows one frame of the dataset.

## 4 EATING AND DRINKING RECOGNITION IN A GROUP

In this section, we explain eating and drinking recog-

nition of a group based on joint positions, head directions, and their integration. The joint positions and head directions of each person are obtained by the Kinect v2s, as described in (Clark et al., 2015).

## 4.1 Eating and Drinking Recognition based on Joint Positions and Head Directions

As may be observed from Figure 2, we can only see the joints of the upper body. Even in such a situation, Kinect v2 can occasionally record 25 joint positions; however, those of the lower body could be inaccurate. Accordingly, we only use the 15 joint positions of the upper body (see Figure 3). The positions we used are a value of x-axis $u$ and a value of y-axis $v$ in a 2D image. The input signal of joint positions $\mathbf{X}_\mathrm{s}$ is:

$$\mathbf{X}_\mathrm{s} = \{\mathbf{x}_\mathrm{s}^0, \cdots, \mathbf{x}_\mathrm{s}^t, \cdots, \mathbf{x}_\mathrm{s}^\mathrm{T}\} \quad (1)$$

$$\mathbf{x}_\mathrm{s}^t = \{u_0^t, v_0^t, \cdots, u_j^t, v_j^t, \cdots, u_\mathrm{J}^t, v_\mathrm{J}^t\} \quad (2)$$

where $\mathbf{x}_\mathrm{s}^t$ is the input signal at frame $t(=0,\cdots,\mathrm{T})$, and $j$ denotes the label of joints ($j=0,\cdots,\mathrm{J}$). $\mathrm{J}=15$ as mentioned above. $\mathbf{x}_\mathrm{s}^t$ is 30 dimensions.

The signal is input to the architecture of our model. Figure 4 illustrates this architecture, including the LSTM block. It is clear from Figure 4 that the architecture is simply constructed of three layers; that is, the first fully connected layer, the LSTM block, and the second fully connected layer. Hereafter, the

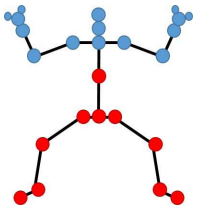

Figure 3: 25 joint positions that can be recorded by Kinect v2. Blue circles denote the upper body's joints.

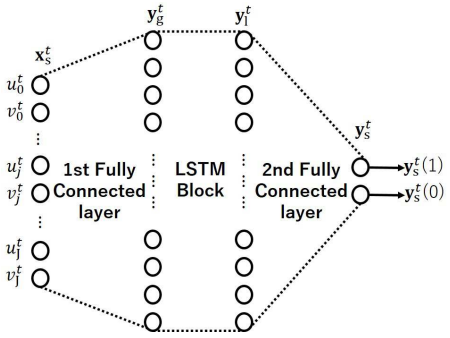


Figure 4: Our architecture, including LSTM block (please see also Figure 5). This is an example in the case that the input data is $\mathbf{x}_\mathrm{s}^t$. Hereafter, the architecture will be referred to as the "LSTM Arch".

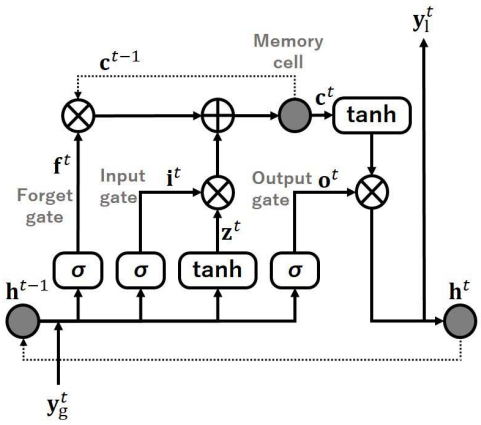


Figure 5: LSTM block with a single cell.

architecture will be referred to as the "LSTM Arch." First, $\mathbf{x}_\mathrm{s}^t$ passes through the first fully connected layer. $\mathbf{y}_\mathrm{g}^t$ is then calculated as follows:

$$\mathbf{y}_\mathrm{g}^t = \mathbf{W}_\mathrm{xg}\mathbf{x}_\mathrm{s}^t + \mathbf{b}_\mathrm{g}^t \quad (3)$$

where $\mathbf{W}_\mathrm{xg}$ denotes the connection weights from the input layer to the LSTM block, and $\mathbf{b}_\mathrm{g}^t$ is a bias vector.

LSTM architecture was proposed by (Hochreiter and Schmidhuber, 1997) to solve vanishing gradient and error blowing up problems. Figure 5 illustrates an LSTM block with a single cell. It contains one self-connected memory cell $\mathbf{c}$ and three multiplicative units, that is, the input gate $\mathbf{i}$, the forget gate $\mathbf{f}$, and the output gate $\mathbf{o}$. LSTM stores and accesses the long-range contextual information of a temporal sequence. The output of LSTM block $\mathbf{y}_\mathrm{l}^t$ is calculated from $\mathbf{y}_\mathrm{g}^t$ using the activations of the memory cell and three gates:

$$\mathbf{z}^t = \tanh(\mathbf{W}_\mathrm{gz}\mathbf{y}_\mathrm{g}^t + \mathbf{W}_\mathrm{hz}\mathbf{h}^{t-1} + \mathbf{b}_\mathrm{z}^t) \quad (4)$$

$$\mathbf{i}^t = \sigma(\mathbf{W}_\mathrm{gi}\mathbf{y}_\mathrm{g}^t + \mathbf{W}_\mathrm{hi}\mathbf{h}^{t-1} + \mathbf{b}_\mathrm{i}^t) \quad (5)$$

$$\mathbf{f}^t = \sigma(\mathbf{W}_\mathrm{gf}\mathbf{y}_\mathrm{g}^t + \mathbf{W}_\mathrm{hf}\mathbf{h}^{t-1} + \mathbf{b}_\mathrm{f}^t) \quad (6)$$

$$\mathbf{c}^t = \mathbf{z}^t\mathbf{i}^t + \mathbf{f}^t\mathbf{c}^{t-1} \quad (7)$$

$$\mathbf{o}^t = \sigma(\mathbf{W}_\mathrm{go}\mathbf{y}_\mathrm{g}^t + \mathbf{W}_\mathrm{ho}\mathbf{h}^{t-1} + \mathbf{b}_\mathrm{o}^t) \quad (8)$$

$$\mathbf{y}_\mathrm{l}^t = \mathbf{h}^t = \tanh(\mathbf{c}^t)\mathbf{o}^t \quad (9)$$

where $\sigma(\cdot)$ is the sigmoid function, and all the matrices $\mathbf{W}$ are the connection weights between the two units.

Finally, the second fully connected layer of LSTM Arch. calculates the final output:

$$\mathbf{y}_\mathrm{s}^t = \mathbf{W}_\mathrm{ls}\mathbf{y}_\mathrm{l}^t + \mathbf{b}_\mathrm{s}^t \quad (10)$$

$\mathbf{y}_\mathrm{s}^t(=\{\mathbf{y}_\mathrm{s}^t(1), \mathbf{y}_\mathrm{s}^t(0)\})$ is the output value corresponding to the two classes (eating/others or drinking/others). Therefore, the predicted label $a(=0 \text{ or } 1)$ is:

$$a = \arg\max_x \mathbf{y}_\mathrm{s}^t(x) \quad (11)$$

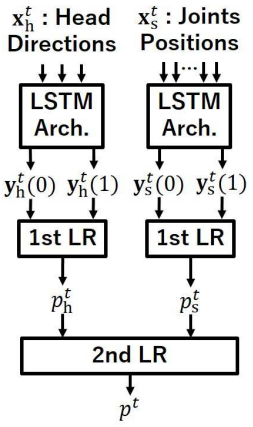


Figure 6: Integration of joint positions and head directions by logistic regression.

In the training phase, the $t$-th training sample is given as the pair of the input signal $\mathbf{x}_s^t$ and the teaching signal $d^t$. For eating recognition, the teaching signal $d^t = 1$ when the subject is eating and 0 otherwise. For drinking recognition, $d^t = 1$ when the subject is drinking and 0 otherwise.

We now turn our attention on to head directions. There are three such directions: roll, pitch, and yaw. The input signal of head directions $\mathbf{X}_h$ is:

$$\mathbf{X}_h = \{\mathbf{x}_h^0, \cdots, \mathbf{x}_h^t, \cdots, \mathbf{x}_h^T\} \tag{12}$$

$$\mathbf{x}_h^t = \{x_h^0, x_h^1, x_h^2\} \tag{13}$$

where $\mathbf{x}_h^t$ is the input signal at frame $t$, and $x_h^0$, $x_h^1$, $x_h^2$ denote roll, pitch, and yaw ($0 \leq \{x_h^0, x_h^1, x_h^2\} < 360$) respectively. $\mathbf{x}_h^t$ is 3 dimensions.

To calculate the predicted label $a$, we simply replace $\mathbf{x}_s^t$ with $\mathbf{x}_h^t$ in Equation (3). We then get $\mathbf{y}_h^t (= \{\mathbf{y}_h^t(1), \mathbf{y}_h^t(0)\})$, as in Equation (10).

## 4.2 Eating and Drinking Recognition based on Integration of Joint Positions and Head Directions

It is conceivable that recognition accuracy can be improved by integrating joint positions and head directions. However, since they have different properties, we should not simply concatenate $\mathbf{x}_s^t$ with $\mathbf{x}_h^t$. Our integration method is described in Figure 6. We first normalize the output values of LSTM Arch. with LR because the output values are not probability values, but simply results of multiplications and additions through the architecture. The output values are converted to the values between 0 and 1 with the following LR:

$$p_s^t = \frac{1}{1 + \exp\{-\alpha^{\mathrm{tr}} \mathbf{Y}_s^t\}} \tag{14}$$

where $(\mathbf{Y}_s^t)^{\mathrm{tr}} = (1, \mathbf{y}_s^t(1), \mathbf{y}_s^t(0))$, and $\alpha^{\mathrm{tr}} = (\alpha_0, \alpha_1, \alpha_2)$ are LR coefficients. To normalize $\mathbf{y}_h^t$, we use $(\mathbf{Y}_h^t)^{\mathrm{tr}} = (1, \mathbf{y}_h^t(1), \mathbf{y}_h^t(0))$ to get $p_h^t$.

Finally, we again use LR to normalize and integrate $p_s^t$ and $p_h^t$. The LR is as follows:

$$p^t = \frac{1}{1 + \exp\{-\alpha^{\mathrm{tr}} \mathbf{P}^t\}} \tag{15}$$

where $(\mathbf{P}^t)^{\mathrm{tr}} = (1, p_s^t, p_h^t)$. The predicted label $a$ is:

$$a = \begin{cases} 1 & (p^t \geq 0.5) \\ 0 & (p^t < 0.5) \end{cases} \tag{16}$$

# 5 EXPERIMENT

In the experiment, we compared joint positions, head directions, and integration of both to verify the effectiveness of head directions and the integration.

## 5.1 Dataset

As mentioned above, to conduct the experiment, we constructed a dataset since there have been none suitable for our purposes.

### 5.1.1 Annotation

We briefed the four subjects to engage in about 30 minutes of conversation while eating and drinking. There were no special requests for what they were to talk about. They were recorded by Kinect v2s, with their joint positions and head directions being obtained by Kinect v2 SDK, as described in (Clark et al., 2015). We discarded frames at which the Kinect v2s could not capture certain joint positions or head directions. The number of remaining frames was 33,336, each frame having $\mathbf{X}_s$ and $\mathbf{X}_h$ of the four subjects. All frames were manually annotated with labels for eating, drinking, and others. The number of frames for each label is shown in Table 1. *i* in Table 1 means the label of subjects. Comparing eating and drinking, the number of frames for the former was more than that of the latter, as eating takes much longer than drinking.

We recognized eating and drinking separately for each subject; accordingly, the annotations for eating, drinking, or others were done respectively. For eating recognition, the teaching signal $d^t = 1$ when a subject was eating, and $d^t = 0$ when a subject was not eating. The same annotations were applied for drinking recognition.

Table 1: The number of frames of each label.

| $i$ | Eating | Drinking | Others |
|---|---|---|---|
| 0 | 7644 | 2385 | 23307 |
| 1 | 11891 | 1030 | 20415 |
| 2 | 11812 | 1025 | 20499 |
| 3 | 10127 | 700 | 22509 |

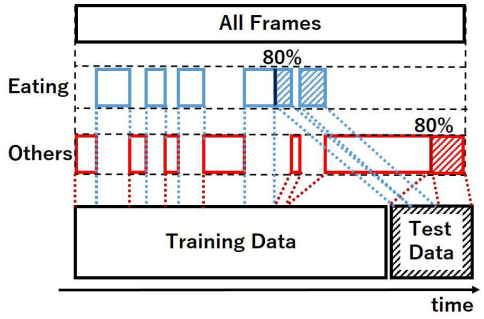


Figure 7: Data division method when joint positions or head directions were used. 80% of each class was selected for training, while the remaining were used for the test while sticking to the time sequence as much as possible.

#### 5.1.2 Data Division

We have to divide all frames into training data and test data. In this paper, we decided that 80% of all frames were for training and the rest were for test. However, all 80% of frames from the beginning should not have been used for training. Since there were no special requests, eating and drinking frames were concentrated until the middle. Therefore, using the above method of division, the test data included no or few eating/drinking frames. This was not adequate for testing eating and drinking recognition capability. To avoid this, we took 80% from each class for training and 20% from each class for the test, while sticking to the time sequence as much as possible. Figure 7 illustrates this strategy.

For eating and drinking recognition based on integration of joint positions and head directions, we have to train three times: LSTM Arch, first LR, and second LR as we explained in 4.2. 40% of frames were used for LSTM Arch training, 20% for first LR training, 20% for second LR training, and 20% for the test. The details are described in Figure 8. Figure 7 explains how to divide the data for the recognition based on joint positions and head directions, while Figure 8 explains for the recognition based on the integration of them. All frames were also broadly divided. 0-80% of all frames were for training, 0-40% were for LSTM Arch training, and 40-80% were for LSTM test. By using half of the results of LSTM Arch, first LR was trained, while second LR was trained by using half of the results of first LR. 80-100% of all frames would

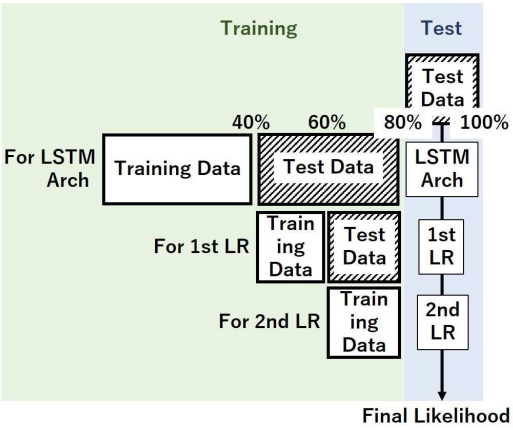


Figure 8: Data division method when joint positions and head directions are integrated.

show the final likelihood by passing through LSTM Arch, first LR, and second LR.

### 5.2 Implementation

We implemented our architecture, LSTM Arch., by using Chainer which is a standalone open source framework for deep learning models (Tokui et al., 2015). We used Adam (Kingma and Ba, 2014) for optimization. Before training, we performed standardization. We computed the mean and standard deviation from the training data, before standardizing the training and testing data by centering and scaling.

### 5.3 Hyperparameters

LSTM Arch. has some hyperparameters. Those we used are as follows. The number of units of the middle layer was 512 for $\mathbf{X}_\mathrm{s}$, and 32 for $\mathbf{X}_\mathrm{h}$ because there is a difference between the number of the units of the input layer of $\mathbf{X}_\mathrm{s}$ and that of $\mathbf{X}_\mathrm{h}$ (they are 30 and 3 respectively as described before). We conducted full batch training (the batch size was equal to the training data size) with 200 epochs.

### 5.4 Results

First, we will explain how we calculated the accuracy. Usually, accuracy $ac$ means a value using the following formula:

$$ac = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}} \qquad (17)$$

where TP, FP, FN, TN denote true positive, false positive, false negative, and true negative respectively. This evaluation index is inappropriate for us because there is a leaning between two classes as mentioned

Table 2: Experimental results (%).

(a) Accuracy of eating recognition.

| $i$ | $\mathbf{X}_\mathrm{s}$ | $\mathbf{X}_\mathrm{h}$ | $\mathbf{X}_\mathrm{s}+\mathbf{X}_\mathrm{h}$ |
|---|---|---|---|
| 0 | 59.0 | **76.0** | **76.0** |
| 1 | 57.0 | **82.4** | 81.6 |
| 2 | 80.4 | 78.6 | **81.9** |
| 3 | 56.5 | **57.0** | 55.9 |
| Avg. | 63.2 | 73.5 | **73.9** |

(b) Accuracy of drinking recognition.

| $i$ | $\mathbf{X}_\mathrm{s}$ | $\mathbf{X}_\mathrm{h}$ | $\mathbf{X}_\mathrm{s}+\mathbf{X}_\mathrm{h}$ |
|---|---|---|---|
| 0 | **59.1** | 50.0 | 48.7 |
| 1 | 61.3 | 50.0 | **70.0** |
| 2 | **52.4** | 50.0 | 51.0 |
| 3 | 51.1 | 50.0 | **56.3** |
| Avg. | 56.0 | 50.0 | **56.5** |

in 5.1.1. To evaluate class 1 (eating or drinking) and 0 fairly, we also should not use precision and recall. Therefore, we used the following evaluation index *ev*:

$$ev = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} + \frac{\mathrm{TN}}{\mathrm{FN}+\mathrm{TN}}\right) \tag{18}$$

We refer to *ev* simply as accuracy in this paper. Accuracies calculated by the formula (18) are described in Table 2. $i$ is the label of subjects. Their sitting positions are illustrated in Figure 1. We will discuss the result in the next section.

### 5.5 Discussion

We will investigate the accuracy of eating recognition shown in Table 2 (a). First, we consider the accuracy of each subject. When Eating, the subject with $i = 3$ had a very low accuracy. The subjects with $i = 0,1,3$ had a better accuracy regarding their head directions $\mathbf{X}_\mathrm{h}$ rather than joint positions $\mathbf{X}_\mathrm{s}$, specially when $i = 0$ and $i = 1$. For instance, the subject with $i = 0$ had 17 points higher when compared with its joint positions, while the subject with $i = 1$ had 25.4 points higher with the same comparison. However, $i = 2$ had a better accuracy when using its joint positions, although the values remained relatively close. The subjects with $i = 0,2$ had their maximum value when using the integration method. For $i = 1,3$, however, their maximum value was achieved when using their head directions. The reason is that the accuracies of joint positions are very low in their case. Next, we consider the average of accuracies of all subjects. On average, we achieved the best accuracy when using the integration method. This result shows that the integration of joint positions and head directions is effective for eating recognition.

We discuss drinking recognition shown in Table 2 (b). First we discuss the accuracy of each subject. The accuracy of each subject was quite low except for the subject $i = 1$, because the number of frames of drinking was quite smaller than that of eating, as shown in Table 1. When we compared between joint positions and head directions, we found that the accuracies of the method using head directions were lower (it couldn't recognize at all). The joint positions were more effective for drinking recognition since the heads had lower movement when they were drinking. For subjects $i = 0,2$, the accuracy was the highest when using the joint positions, whereas for subjects $i = 1,3$, the accuracy was the highest when using the integration method. The reason why the integration was not so effective is that the accuracies of the head directions would seem to be random, and they were quite low. Second, we discuss the average accuracy of all subjects. On average, we achieved the best accuracy when using the integration method. This result shows that the integration of joint positions and head directions is also effective for drinking recognition.

For both eating and drinking recognitions, the integration of joint positions and head directions achieved the best accuracy on average. Since the movement of each subject was different, there was variability among the accuracy of each subject. In our future work, we will present the method that considers the differences between the subjects. On this occasion, however, when we recognized the eating or drinking for the subject with $i = 0$, we used only his training data. In future, we will use data that does not include the current subject being tested to achieve eating and drinking recognition.

## 6 CONCLUSION

In this paper, we have proposed a new task: eating and drinking recognition as applied to each customer who come to restaurants or cafes in a group, and interact each other such as having a conversation. We use head directions to demonstrate that the integration of joint positions and head directions leads to the highest recognition accuracy on average, using logistic regression.

At present, our constructed dataset only has four subjects. In future research, we should increase the variety of the dataset in several aspects, such as the number of subjects, sitting places, different subjects, and so on. If the dataset is small, we must overcome certain particularities of subjects, such as handedness. In addition, we will consider the recognition of speaking and performing multiclass classification.

Our goal is to make cyber serving staff determine the appropriate times for waiting on tables. One of the times is when a customer in a group finishes the eating or drinking. We will predict the time by using our proposed method in future work.

## ACKNOWLEDGEMENTS

This work was partially supported by JST CREST "Intelligent Information Processing Systems Creating Co-Experience Knowledge and Wisdom with Human-Machine Harmonious Collaboration."

## REFERENCES

Amor, B. B., Su, J., and Srivastava, A. (2016). Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):1–13.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Chua, J.-L., Chang, Y. C., Jaward, M. H., Parkkinen, J., and Wong, K.-S. (2014). Vision-based hand grasping posture recognition in drinking activity. In *International Symposium on Intelligent Signal Processing and Communication Systems*, pages 185–190.

Clark, R. A., Pua, Y.-H., Oliveira, C. C., Bower, K. J., Thilarajah, S., McGaw, R., Hasanki, K., and Mentiplay, B. F. (2015). Reliability and concurrent validity of the microsoft xbox one kinect for assessment of standing balance and postural control. *Gait & posture*, 42(2):210–213.

Ding, W., Liu, K., Cheng, F., Shi, H., and Zhang, B. (2015). Skeleton-based human action recognition with profile hidden markov models. In *CCF Chinese Conference on Computer Vision*, pages 12–21.

Du, Y., Fu, Y., and Wang, L. (2016). Representation learning of temporal dynamics for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 25(7):3010–3022.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Iosifidis, A., Marami, E., Tefas, A., and Pitas, I. (2012). Eating and drinking activity recognition based on discriminant analysis of fuzzy distances and activity volumes. In *International Conference on Acoustics, Speech and Signal Processing*, pages 2201–2204.

Ishii, R., Otsuka, K., Kumano, S., and Yamato, J. (2016). Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. *ACM Transactions on Interactive Intelligent Systems*, 6(1):4:1–4:31.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, pages 1–9.

Koppula, H. S. and Saxena, A. (2016). Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29.

Nayak, N. M., Zhu, Y., and Chowdhury, A. K. R. (2015). Hierarchical graphical models for simultaneous tracking and recognition in wide-area scenes. *IEEE Transactions on Image Processing*, 24(7):2025–2036.

Nihei, F., Nakano, Y. I., Hayashi, Y., Huang, H.-H., and Okada, S. (2014). Predicting influential statements in group discussions using speech and head motion information. In *International Conference on Multimodal Interaction*, pages 136–143.

Ozasa, Y. and Ariki, Y. (2012). Object identification based on color and object names using multimodal information. *The journal of the Institute of Image Electronics Engineers : visual computing, devices and communications*, 45(1):105–111.

Ozasa, Y., Nakano, M., Ariki, Y., and Iwahashi, N. (2015). Discriminating unknown objects from known objects using image and speech information. *IEICE TRANSACTIONS on Information and Systems*, 98(3):704–711.

Pieska, S., Luimula, M., Jauhiainen, J., and Spiz, V. (2013). Social service robots in wellness and restaurant applications. *Journal of Communication and Computer*, 10(1):116–123.

Qing-Xiao, Y., Can, Y., Zhuang, F., and Yan-Zheng, Z. (2010). Research of the localization of restaurant service robot. *International Journal of Advanced Robotic Systems*, 7(3):227–238.

Swears, E., Hoogs, A., Ji, Q., and Boyer, K. (2014). Complex activity recognition using granger constrained dbn (gcdbn) in sports and surveillance video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 788–795.

Tokui, S., Oono, K., Hido, S., and Clayton, J. (2015). Chainer: a next-generation open source framework for deep learning. In *Workshop on Machine Learning Systems at Neural Information Processing Systems*.

Vemulapalli, R., Arrate, F., and Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a lie group. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595.

Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2014). Learning actionlet ensemble for 3d human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(5):914–927.