Behavioral Pattern Modeling of Human-Human Interaction for Teaching Restaurant Service Robots

Eun-Sol Kim, Kyoung-Woon On, Byoung-Tak Zhang

School of Computer Sci. & Eng., Seoul National University
{eskim, kwon, btzhang}@bi.snu.ac.kr

Abstract

In this paper, we consider the problem of making restaurant-serving robots to learn more about the human and to plan and act more interactively. To resolve this problem, it is inevitable to understand what situation the customer is in. In order to understand the customer's situation automatically, we suggest sensing the behavioral signal of the customer and using the data to predict the customer's situation. Here, we propose a machine learning algorithm for modeling the customer's behavioral pattern while having dinner. First of all, we collect the behavioral data from customer using two kinds of wearable devices, an eye tracker and a watch type EDA device, while having dinner. Furthermore we show a novel algorithm which can analyze the data efficiently and extract the individual behavioral patterns. The suggested model has a hierarchical structure: the bottom layer combines the multi-modal behavioral data based on causal structure of the data and extracts the feature vector. Using the extracted feature vectors, the upper layer predicts the customer's situations based on the temporal correlation between feature vectors. Experimental results show that the suggested model can analyze the behavioral data efficiently and predict the current situation of the customer.

Introduction

How can we improve the ability of robots interacting with humans more fluently? What kind of technique would be a breakthrough for developing human-likely behaving robots? Despite impressive progress in robotics and artificial intelligence, still most of the robots act unnaturally when interacting with a human. Specifically, even though many service robots can behave naturally in some fixed situations, the robots find it hard to fluently manage the unknown situations which occur due to humans accidentally. [Christensen et al., 2010] Here, we consider a problem of helping restaurant-serving robots to learn more about the human customer and plan and act more interactively like human clerks. The key idea to achieving this goal is observing how the human clerk interacts with the human customer and imitating the action of the clerk.

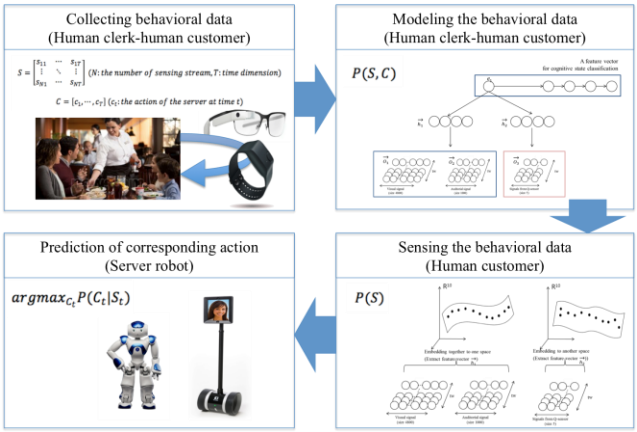


Figure 1: System architecture of the suggested method

In this paper, we suggest a novel idea of observing and imitating the interaction between the clerk and the customer through sensing the behavioral signal of the human and modeling the signal with a machine learning algorithm. Here, we focus on using the behavioral signal for imitating the action of the clerk and modeling of the human customer.

From the early 1900s, there have been wide cognitive researches aiming at modeling human cognitive processes using behavioral data. [Ballard et al., 1997; Newell 1987] Specifically, various wearable devices have been used for sensing the behavioral data recently. Current wearable devices help the experimenter move more comfortably and it is possible to wear more than two devices simultaneously. Moreover, not only experiments in laboratory with strict constraints but also unaffected and compositive experiments out of laboratory have become feasible.

Inspired by other cognitive researches, we use wearable devices to obtain the information about the human and apply the information to human-robot interaction problems. As the first step for resolving the problem, we focus on modeling the behavioral data from wearable devices.

We suggest a novel machine learning algorithm which can analyze behavioral data from multiple wearable devices efficiently and correctly. The key idea of the suggested method is to choose only a few kinds of data streams among the entire set of data streams from multiple wearable devices. [Zhang 2013; Zhang 2014]

For the experiments, we collected behavioral data in meal situations. To obtain behavioral data, two kinds of wearable devices are used: one is an eye tracker which collects the first-person video and audio signal. The other is a watch-type EDA sensor which collects the user's electrodermal activity, temperature and movement of wrist. With these five kinds of data streams from the wearable devices, we tried to model the cognitive process of the human. The cognitive process we aim to find is what stage the user is in, e.g. choosing the menu, eating, and looking for a server to ask something.

From the experimental results, we show that the cognitive processes can be modeled with behavioral data, in particular the suggested method dramatically improved the performance of predicting what the cognitive state the user is in.

Behavioral Dataset in Restaurant

To predict the situation the human customer is in, we collected behavioral data of 50 hours during a meal using 2 wearable devices, which are a watch-type EDA sensor and an eye tracker. (Figure 1)

First, the watch-type EDA sensor, the Q-sensor made by Affectiva, is embedded with multiple sensors which measure user's electrodermal activity, skin temperature and movement of wrist. An electrodermal activity (EDA), referred to as skin conductance, provide a sensitive and convenient measure of assessing alterations in sympathetic arousal associated with emotion, attention, and cognition. Also, a skin temperature and movement of wrist are important features for identifying user's activity.

The next thing, the eye tracker, Glass Smart IR made by Tobii, is embedded with a forward camera, a microphone and an infrared lamp for collecting first-person video, audio and gazing point data. The first-person video and audio data from real, dynamic environments is a novel factor for recognizing real-world perception of person. Also, gazing point data can help to analyze real-time attention, intention and interest of person.

In conclusion, 6 heterogeneous stream data is collected for this this research: electrodermal activity, skin temperature, movement of wrist (3-axis acceleration), first-person video, audio, and gazing point. Time frequency of each data is 16Hz, 32Hz, 32Hz, 30Hz, 30Hz, 30Hz respectively. Among these data stream, we used 5 streams (except the gazing point). And we consider the data streams from Q-sensor (electrodermal activity, skin temperature, movement of wrist) as a stream. To evaluate our model, we also labeled each instance with Greeting, Spot guidance, Natural conversation, Manu selection, Order, Food service, Meal, Staff call, Request to staff, Payments. Figure 1 shows the details of data.

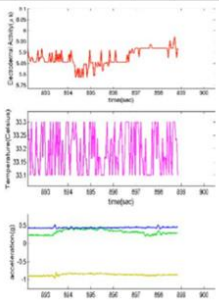
	Watch-type EDA sensor		Eye-tracker	
equipment				
Data Type	First-person video	30 Hz	Electrodermal activity	16 Hz
	First-person Audio	30 Hz	Skin temperature	32 Hz
	First-person gaze point	30 Hz	Movement of wrist (3-axis acceleration)	32 Hz
Example of actual data				
Label	{Greeting, Spot guidance, Natural conversation, Manu selection, Order, Food service, Meal, Staff call, Request to staff, Payments}			

Figure 2: Wearable sensor devices used in this research and the data specification

Model

In this paper, we suggest a machine learning model which has hierarchical structure with two layers. The lower layer extracts a feature vector from multiple wearable sensor data stream in accordance with the causal structure within the data. And the upper layer temporally classifies the feature vector from the lower layer into the cognitive process stage.

Lower layer: combine the feature vector

This is the second paragraph. It in formatted with the Text-indent style. This is example text. This is example text. It is 10 point The lower layer extracts the feature vector from multiple sensor data stream. The key idea of this layer is to select highly correlated data stream and to combine them into a feature vector.

Let define the first-person video data stream, auditorial data stream and Q-sensor data stream in time interval $[t, t+\Delta t]$ as $O^1_{t:t+\Delta t}, O^2_{t:t+\Delta t}, O^3_{t:t+\Delta t}$. From these data stream, the main purpose of the low layer is to combine these streams as a single feature vector $\vec{h}$.

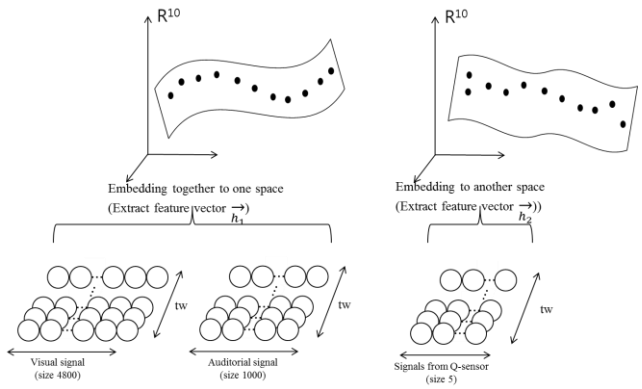


Figure 3: Structure of the lower layer: combining the multiple sensor stream

To combine the multiple data streams, we adopt the strategy of how humans integrate the sensory information: sensory cue integration. Humans combine sensory cues if they are highly correlated to reduce the uncertainty. While, if they are weakly correlated, humans handle the sensory cues separately [Koerding 2007]. Inspired by the literatures about the sensory cue integration, we designed the method for combining multiple wearable sensor streams.

In this research, we used two wearable sensor devices and collected 3 kinds of data: first-person video, speech, sensor value from Q-sensor. From these data, the lower layer decides which data streams are highly correlated in stochastic, and then combines the related data streams.

There are 5 ways combining 3 kinds of behavioral data. 1) C_1 : as 3 kinds of behavioral data are correlated together, combining all of the data together. 2) C_2, C_3, C_4 : as only 2 behavioral data are correlated, combining two data stream and handling the other data stream independently. 3) C_5 : as 3 kinds of behavioral data are not correlated, handling all of the data stream independently. The distribution over the h_1, h_2, h_3 can be represented with 5 terms of combining way as below equation.

$$\begin{aligned}&\mathrm{P}(\mathrm{h}_1, h_2, h_3)\\&= P(C_1)P(h_1, h_2, h_3|C_1) + P(C_2)P(h_1, h_2, h_3|C_2)\\&+ P(C_3)P(h_1, h_2, h_3|C_3) + P(C_4)P(h_1, h_2, h_3|C_4)\\&+ P(C_5)P(h_1, h_2, h_3|C_5)\end{aligned}$$

Probabilistic distribution of each term can be represented as follow [Koerding 2007].

$$\begin{aligned}&\mathrm{P}(\mathrm{h}_1, h_2, h_3|C_5)\\&= \int P(h_1, h_2, h_3|s_1, s_2, s_3)ds_1 ds_2, ds_3\\&= \int P(h_1|s_1)ds_1 \times \int P(h_2|s_2)ds_2 \times \int P(h_3|s_3)ds_3\end{aligned}$$

Here, the distribution of s when h is given could be modelled as multivariate Gaussian distribution. Then the distribution over h_1, h_2, h_3 is also derived as multivariate Gaussian distribution. The mean and the covariance of the distribution over h_1, h_2, h_3 can be derived with closed form solution as follow. [Koerding 2007] (Due to the limitation of space, we just derive the mean and variance of the C_1 .)

$$\begin{aligned}&\mu_{\mathrm{P}(\mathrm{h}_1, h_2, h_3|C_1)}\\&= \frac{1}{\Sigma_{P(h_1|s)}}\mu_{P(h_1|s)} + \frac{1}{\Sigma_{P(h_2|s)}}\mu_{P(h_2|s)} + \frac{1}{\Sigma_{P(h_3|s)}}\mu_{P(h_3|s)}\\&+ \frac{1}{\Sigma_{P(s)}}\mu_{P(s)}\end{aligned}$$

$$\frac{1}{\Sigma_{\mathrm{P}(\mathrm{h}_1, h_2, h_3|C_1)}} = \frac{1}{\Sigma_{\mathrm{P}(\mathrm{h}_1|C_1)}} + \frac{1}{\Sigma_{P(h_2|C_1)}} + \frac{1}{\Sigma_{P(h_3|C_1)}} + \frac{1}{\Sigma_{\mathrm{P(s)}}}$$

Upper layer: Classify the feature vector temporally

The upper layer classifies the extracted feature vector $\vec{h}$ into several catogories which represent the restaurant customer's situations. We defined 10 possible situations and used as the label value for classification. (Figure 1)
As the wearable data is entered sequentially, we use HMM model as upper layer model for temporal classification. (Figure 4)

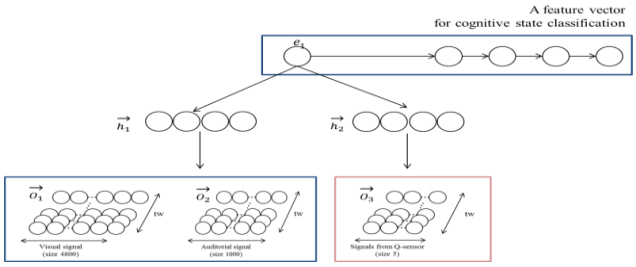


Figure 4. Structure of the upper layer: temporal classification with HMM.

Preliminary Experimental Results

As experiments, we designed a classification problem to verify the performance of the suggested model. First of all, we verify the performance of the suggested model with classification accuracy. As each frame has class label (which situation the customer is in), we can calculate the accuracy with the predicted result of HMM. With HMM, two kinds of classification problem are designed. First, we classify the behavioral data into 10 classes which describe the situation the customer is in (Figure 1). Second, we just predict whether the customer would ask help to the clerk or not.. As comparative experiment, we used Naïve Bayes algorithm and k-NN algorithm.

The preliminary results are summarized in Figure 5. From the figure, we can argue as follow. Even though the

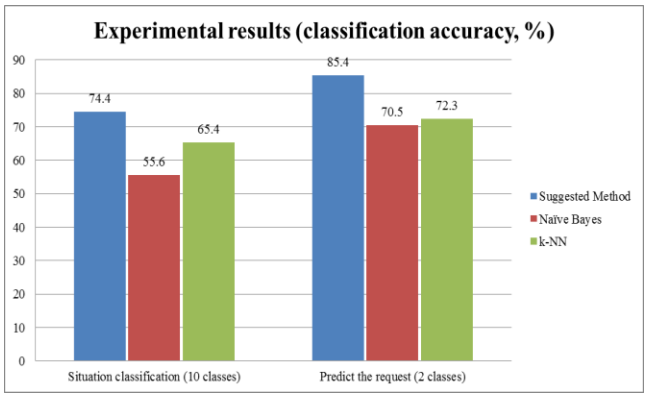


Figure 5. Experimental results. Two kinds of classification are performed.

wearable sensor data is very noisy and the HMM is quite simple model, the lower layer of the model which is inspired by human cue integration mechanism could extract informative feature vector. Also, we can see that the classification accuracy is meaningful compared to the result of Naïve Bayes and k-NN method.

Furthermore, the status board containing the input data stream and the real-time classification result is shown in Figure 6.

Discussion

This research suggests a system which applies the human behavioral to human-robot interaction problems. As the first step of the research, here, we suggest a bio-inspired machine learning algorithm which can analyze multi-stream data efficiently. As the preliminary experiment, we collected behavioral data using wearable devices while having dinner.

In addition to this preliminary research, to apply this model to the real restaurant server robot would be an interesting future direction.

References

Christensen, H., Geert-Jan M. K., and Jeremy W. 2010. Cognitive systems. Vol. 8. Springer Science & Business Media.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., and Rao, R. P. N. 1997. Deictic codes for the embodiment of cognition. Behavioral and Brain Sciences 20, 723-767.

Newell, A. 1987. Unified Theories of Cognition. Harvard University Press.

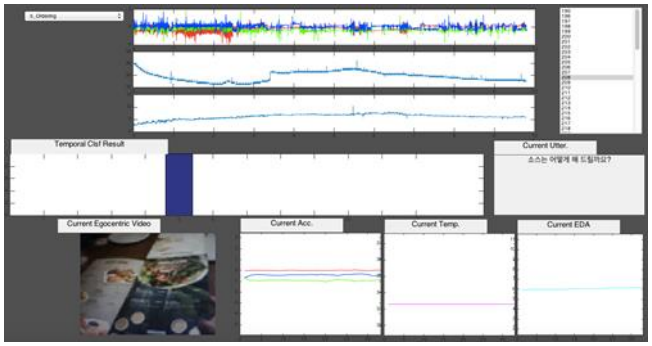


Figure 6. The status board shows the input data stream and the real-time classification results.

Zhang, B. -T, 2014. Ontogenesis of agency in machines: A multidisciplinary review, AAAI 2014 Fall Symposium on The Nature of Humans and Machines: A Multidisciplinary Discourse.

Zhang, B.-T. 2013. Information-Theoretic Objective Functions for Lifelong Learning, In AAAI 2013 Spring Symposium on Lifelong Machine Learning, pp. 62-69, Stanford University, AAAI Press.

Koerding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B. and Shams, L. 2007. Causal Inference in Multisensory Perception, PLoS One 2.9: e943.