KUSK Dataset: Toward a Direct Understanding of Recipe Text and Human Cooking Activity

Atsushi Hashimoto
Graduate School of Law
Kyoto Univ.
Yoshida Honmachi, Sakyo-ku
606-8501, Kyoto, Japan
ahasimoto@mm.media.kyoto-u.ac.jp

Shinsuke Mori
Academic Center for Computing and Media Studies
Kyoto Univ.
Yoshida Honmachi, Sakyo-ku
606-8501, Kyoto, Japan

Tetsuro Sasada
Academic Center for Computing and Media Studies
Kyoto Univ.
Yoshida Honmachi, Sakyo-ku
606-8501, Kyoto, Japan

Michihiko Minoh
Academic Center for Computing and Media Studies
Kyoto Univ.
Yoshida Honmachi, Sakyo-ku
606-8501, Kyoto, Japan

Yoko Yamakata
Graduate School of Informatics
Kyoto Univ.
Yoshida Honmachi, Sakyo-ku
606-8501, Kyoto, Japan

UbiComp '14, September 13 - 17 2014, Seattle, WA, USA

ACM 978-1-4503-3047-3/14/09. . . $15.00.
http://dx.doi.org/10.1145/2638728.2641338

Abstract
In this paper, we provide a multimodal dataset for understanding cooking activities. To build the dataset, we instructed the subjects to perform cooking according to instructional texts shown on a display one by one. The instructional texts were generated from flow graphs, which were automatically extracted from recipes sampled from a Web site. The main identity of this dataset is the correspondence between the steps automatically extracted from recipes, and real human activities. Typical uses of our dataset are to construct classifiers for understanding human activities in the kitchen, text generation through observing the activities, and so on.

Author Keywords
Dataset, Multimedia, Text Generation, Activity Recognition

ACM Classification Keywords
H.5.1 [Information interfaces and presentation (e.g., HCI)]: Multimedia Information Systems.

Introduction
Everyday food is directly related to our quality of life. In recent years, there is a set of studies which analyse the structure of cooking activity from recipes, or on the basis of observing human activity.

Those studies have treated recipe and human cooking activities independently; however they are related to each other in the nature. Firstly, assuming a cooking assist system, the system will assist users along with any authorized information. One typical example is recipes. Secondly, a recipe describes what a chef should do to make a dish. This is a description of human activity. In other words, it will be a human-readable explanation by computer if a system can automatically understand the activity from the observation, and generate a recipe-style texts.

Our grand goal is to make correspondence between human activity and the words in the recipe text directly, toward realizing the assist system and the explanation generation. As a first step, we prepared a public dataset, Kyoto University Smart Kitchen (KUSK) dataset [1]. In this paper, we introduce how the dataset is organized.

Related Works

There are already several dataset of cooking activities [5, 9]. CMU-MMAC is a dataset with cameras, microphones, radio frequency identifiers (RFIDs), and several wearable sensors such as motion capture (MoCap), and e-watch [5]. Currently, observation of 43 subjects and 5 recipes are openly available. Half of the data are labeled finely with motion and object labels with propositions.

There is a different dataset from an algorithm contest in ICPR2012[9]. In the dataset, cooking activities consist of five different recipes of egg cooking, and are observed by Kinect. Seven actors performed all of the recipes. Provided annotation data include eight different cooking motions with one background label.

[1] http://kusk.mm.media.kyoto-u.ac.jp

Different from CMU-MMAC, we observed cooking by various sensors embedded in a built-in kitchen, but not wearable sensors. Currently, we observed the cooking activities with 20 Japanese recipes, which are sampled from COOKPAD [2]. Hence, the difficulty of making the dish is not controlled.

Our annotation data are also unique; they correspond to instructional texts in the recipes. The texts were generated semi-automatically from the flow-graph representation of recipe [6]. The annotations were automatically collected while cooking through subject's operating actions to our recipe-displaying system CHIFFON [3]. Hence, the data are linked to the instructional texts in the recipe directly and subjectively by subjects.

Smart Kitchen Environment

Considering to realize the cooking assistive system, we decided on a policy of no-use of wearable sensors, but using those embedded in an environment. As a daily tool, the kitchen should be under a highest usability, and any preparation of wearing sensors is undesirable.

We set up a sensor-embedded kitchen following the above policy, and selected the sensors for understanding the progress of cooking along with recipes. The built-in kitchen in our laboratory is shown in Fig. 1, and all the installed sensors are listed in Table 1.

Optical Cameras

An optical video is considered as a most informative sensor because human can understand what happens near-completely through a video. We embedded three cameras (A, B, C) on the ceiling over the cooktop. Each camera respectively observes sink, work-area, and stove area of the cooktop. We also have one super-wide angle

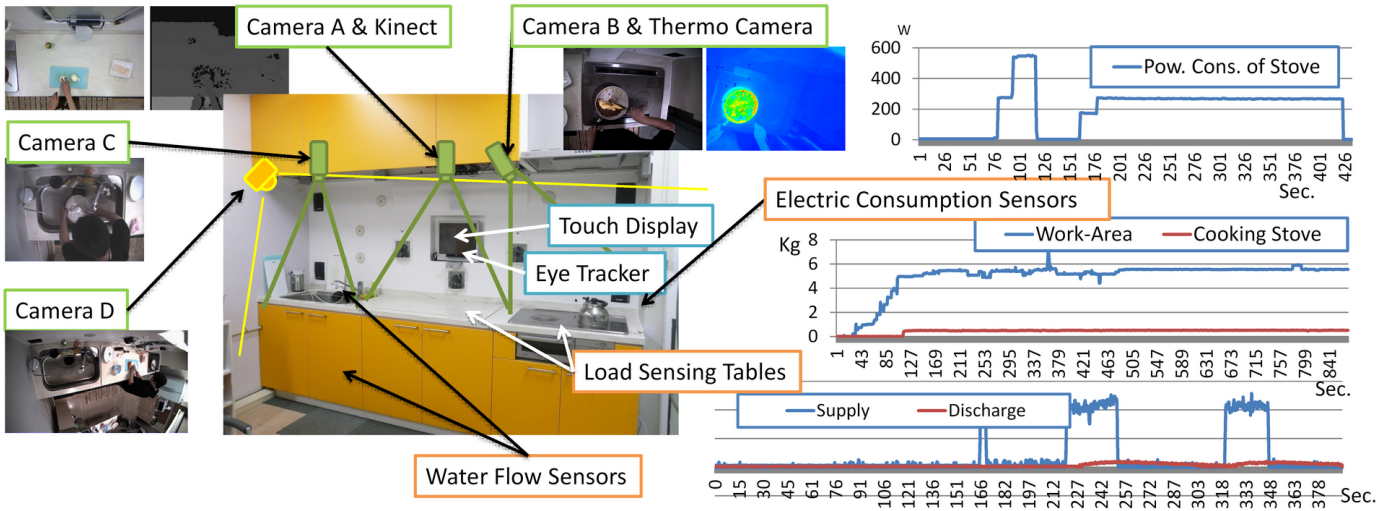

Figure 1: Sensors in the smart kitchen environment.

camera (D) on the corner of the kitchen to overview all activities in the kitchen.

Camera A, B and C capture narrower range of the cooktop than camera D. Thus, they are more suitable to observe objects. Camera D will facilitate motion recognition by the other cameras because of its different camera angle.

The cameras observe cooking activities in 30fps and each activity is roughly 45 minutes long. They are not strictly synchronized. To ensure a high speed of the storage, we prepared a RAID0 drive with four SSDs. In order to save storage space, we checked any kind of motion in each camera frame, and skip to store frames if no difference was found between the frames.

The precise algorithm is following. We checked only pixels sampled from each frame. A dense sampling with a regular interval of 8 pixels was used to sample the pixels. The value of each pixel was compared with the value observed in the past. We used the frame observed five frame ago to reinforce the difference caused by any motion.

When any of the sampled pixels had a larger difference than $2\sqrt{2}\sigma$, we regard it as a motion. Here, σ is the standard deviation of camera noise, and $\sqrt{2}\sigma$ is that on the difference of pixel values. Hence, $2\sqrt{2}\sigma$ is the percent point of 97.8%. This is a tight threshold, but missing detection can still happen about one time per 100 frames. Therefore, we saved five consecutive frames after detecting a motion. Hence, the algorithm fails to detect a motion only $5.15\times10^{-7}(=(1-0.978)^5\times100)\%$ theoretically. We roughly estimate the threshold at

Table 1: The list of sensors used in KUSK Dataset.

Sensor	Manufacturer	Model Number	Remarks Column
Optical cam. (A,B,C)	Point Grey Reseach	FL3-U3-32S2C-CS	Lense: Tamron M12VM412
Optical cam. (D)			Lense: Tamron 13FM22IR
RGBD	Microsoft	LPF00006	
Electric Cons.	Ubiquitous Corp.		
Water Flow (supply)	Keyence	FD-MZ5AT	Data Logger: HBM QuanumX MX440A
Water Flow (discharge)		FD-MZ10AY	
Load Sensor	HBM	PW6D	Data Logger. HBM QuantumX MX840A
Thermo Cam.	Artray	ARTCAM-320-THERMO-HYBRID	
Eye Tracker	Tobii	Tobii X-2 30 Compact Edition	

$2\sqrt{2}\sigma = 20$ for all four cameras.

RGB-D Cameras
RGB-D camera is often used in contemporary fine-grained activity recognition studies [7]. In our dataset, RGB-D camera observes the work-area, where a variety of activities are carried out.

Electric Power Consumption Sensor
Many electric appliances are used while cooking, and there is a system using sensor-embedded appliances [10]. The embedded sensors sense user's use of the appliance. This is helpful to understand the activity; however they are not available in the market. Instead, we set the commercial sensors whose original purpose is for visualizing power consumption in a smart home.

In the dataset, the power consumptions of cooking stove, microwave, and food processor is measured. The stove has two inductive heating (IH) spots, one electric heating and one small oven for fish grill. This is a standard cooking stove in Japan.

Water Flow Sensor
Water flows are an important factor of the cooking activity. While subjects wash something, both supply and discharge are active. Active supply without discharge indicates subject's putting water into any container. The last case of only active discharge may be dumping water. We set a fluid meter to both water supply and discharge.

We note that the maximum flow speed of discharged water is slower than that of supply because of the different pressure, and the limitation of pipe diameter of fluid meter.

Load Sensing Table
Load on the cooktop are competent evidence to detect human-object interaction [11]. We set a load sensing table [8] on the work area and stove area. This setting is similar to Chi's work [1].

Only the sink area has no load sensors. To measure load on a top, the top board must be supported only by load sensors. The sink has a discharge pipe, and it has a physical connection to the ground. Hence, it was difficult

to measure the load in the sink area.

Thermo Camera
We set thermo camera experimentally over the stove area. It observes the temperature of food stuff surface, or is useful to recognize the lid of a pot.

Eye Tracker
Eye trackers are placed at the touch display. It facilitates to analyse user's activity of referring recipes. We note that this sensor is invalid for a person with thick glasses.

Recipes

For KUSK dataset, we selected 20 recipes from the flow-graph corpus [6]. The corpus contains 200 recipes which were randomly extracted from COOKPAD [2]. The selection was made on the basis of the ingredients appeared in the recipes and those listed in our previous work [4]. Because a number of ingredients appear in the corpus, it is costly to collect enough number of training samples for recognizing all of them.

We saved this cost by choosing the recipes which include more ingredients listed in our previous work, and less others. This is done semiautomatically; the 200 recipes are first sorted automatically by the number of listed ingredient. Then the number of other ingredients is checked manually.

Generally, instructions in a recipe are described in the free format. One sentence often directs two or more complex tasks (i.e. "cut tomato and onion," or "cut carrot after peeling the skin"). Hence, we regenerated instructional texts for each vertex with the named entity tag of "Action by the chef" (Ac). The named entity is tagged automatically by the method proposed in [6].

The text is generated manually in following manner; human constructed a sentence from the words picked up automatically in the flow graph. Only minimum grammatical modification was applied. No semantic modifications were performed even when the sentence does not sense. This imitates output from full automatic natural language processing.

Data and Annotations

In the experiment, 25 different subjects performed cooking along with two of the 20 recipes, which were manually assigned. We explained to the subjects that the instructional texts are generated automatically and sometimes they do not sense. This explanation is done while showing the display of CHIFFON system (Fig. 2). The subjects are allowed to refer the original COOKPAD recipe by switching the display whenever they are confused by the instructional texts.

Figure 2: Information displayed on the touch display.

The subjects are also instructed to tap a checkbox on the touch display at every completion of the directed tasks.

All those operations on CHIFFON were logged with its time stamps. Hence each text is generated from a word with Ac entity tag, the data observed during two time stamps correspond to the word directly. This is annotation data of KUSK dataset.

The image collection of 23 ingredients from [4] is also published in KUSK datasets. These ingredients were used to select the 20 recipes. Please refer our Web site for the detail of any format and specifications of the dataset.

Conclusion

In this paper, we provided a new dataset for studies targetting to understand cooking activity. The annotation data are obtained in a subjective manner by subject's operation of referring the guidance text via CHIFFON system. Images of 23 ingredients are also published, which often appears in the selected 20 recipes. Any donation of annotation data helps us a lot. We will keep updating the dataset.

Acknowledgment

This work was supported by JSPS Grants-in-Aid for Scientific Research Grant Numbers 26280084, 24240030, and 26280039.

References

[1] Chi, P. Y., Chen, J. H., Chu, H. H., and Lo, J. L. Enabling calorie-aware cooking in a smart kitchen. *Lecture Notes in Computer Science 5033* (2008), 116.

[2] COOKPAD inc. COOKPAD - The best Japanese recipes to make your cooking fun! http://en.cookpad.com.

[3] Hashimoto, A., Inoue, J., Funatomi, T., and Minoh, M. How does userfs access to object make HCI smooth in recipe guidance? In *Proc. of HCII2014*, Springer (2014), 150–161.

[4] Hashimoto, A., Inoue, J., Nakamura, K., Funatomi, T., Ueda, M., Yamakata, Y., and Minoh, M. Recognizing ingredients at cutting process by integrating multimodal features. In *Proc. of CEA2012*, ACM (2012), 13–18.

[5] la Torre, F. D., Hodgins, J., Montano, J., Valcarcel, S., Forcada, R., and Macey, J. Guide to the carnegie mellon university multimodal activity (CMU-MMAC) database. In *Tech. report CMU-RI-TR-08-22*, Robotics Institute, Carnegie Mellon University (2009), 1–17.

[6] Mori, S., Maeta, H., Yamakata, Y., and Sasada, T. Flow graph corpus from recipe texts. In *Proc. of LREC'14* (May 2014), 2370–2377.

[7] Rohrbach, M., Amin, S., Andriluka, M., and Schiele, B. A database for fine grained activity detection of cooking activities. In *IEEE Conf. on CVPR 2012*, IEEE (2012), 1194–1201.

[8] Schmidt, A., Strohbach, M., Van Laerhoven, K., Friday, A., and Gellersen, H.-W. Context acquisition based on load sensing. In *Proc. of UbiComp* (2002), 333–350.

[9] Shimada, A., Kondo, K., Deguchi, D., Morin, G., and Stern, H. Kitchen scene context based gesture recognition: A contest in ICPR2012. In *Advances in Depth Image Analysis and Applications*. Springer, 2013, 168–185.

[10] Stander, M., Hadjakos, A., Lochschmidt, N., Klos, C., Renner, B., and Muhlhauser, M. A smart kitchen infrastructure. In *ISM 2012*, IEEE (2012), 96–99.

[11] Yasuoka, R., Hashimoto, A., Funatomi, T., and Minoh, M. Detecting start and end times of object-handlings on a table by fusion of camera and load sensors. In *Proc. of CEA'13* (2013), 51–56.