Slice&Dice: Recognizing Food Preparation Activities Using Embedded Accelerometers

Cuong Pham and Patrick Olivier

Culture Lab, School of Computing Science, Newcastle University

Abstract. Within the context of an endeavor to provide situated support for people with cognitive impairments in the kitchen, we developed and evaluated classifiers for recognizing 11 actions involved in food preparation. Data was collected from 20 lay subjects using four specially designed kitchen utensils incorporating embedded 3-axis accelerometers. Subjects were asked to prepare a mixed salad in our laboratory-based instrumented kitchen environment. Video of each subject's food preparation activities were independently annotated by three different coders. Several classifiers were trained and tested using these features. With an overall accuracy of 82.9% our investigation demonstrated that a broad set of food preparation actions can be reliably recognized using sensors embedded in kitchen utensils.

1 Introduction: Ambient Intelligence in the Kitchen

Ambient intelligence has seen significant progress since Weiser's original vision of ubiquitous computing [1], a world in which computers and sensors disappear as they are woven into the fabric of our surroundings. The technical basis and infrastructure to realize miniaturized computing and sensory devices, connected by wireless networks, is already apparent. However, it is not enough simply to demonstrate that miniaturization and networking are possible. If this technology is to find its way into our homes it must be made to fit into existing home environments and then be able to support the things people wish to do there.

Kitchens offer a unique challenge for the development of the situated services envisaged by Weiser, not only because there is a readily identifiable user group in people with dementia, whose lives would be transformed by effective situated support, but because kitchens are not typical sites for the deployment of digital technologies. Whilst the kitchen contains much existing technology in the form of appliances for cooking, washing and food preparation the level of integration of such devices is minimal and the very notion of the "appliance" emphasizes the stand alone character and well defined function of each device. Unlike many aspects of modern life, the kitchen is still a space where physical interaction with real objects (food and kitchen utensils) is valued and information access is furnished through traditional media such as cookbooks.

In particular, the development of cognitive prosthetics lies at the heart of our interest in activity recognition in the kitchen. Specifically, the clinical problem of prompting people in the early stages of dementia through multi-step tasks [2, 3].

M. Tscheligi et al. (Eds.): AmI 2009, LNCS 5859, pp. 34–43, 2009.

Carefully conducted interviews with people with dementia and their carers, and in depth observational studies have revealed both the nature of the support that people with dementia require in the kitchen, and just how important being able to prepare food and drink was to their sense of their own autonomy. The development of a system to support a person with dementia undertaking even quite simple food and drink preparation requires intelligent technologies that are still well beyond the state-of-the-art, in particular, the detection of a user's actions and intentions, and the provision of situated prompts when things go awry.

The literature contains many proposals as to how one might use ambient intelligence to support people with dementia complete daily tasks. Only a few of these have been implemented as research prototypes, and none so far are commercial products. There are significant problems to be solved before such technologies can become a commercial reality. An effective prompting system has to infer context: what activity the user is engaged in, what have they done so far, and what is the current state of the environment. For example, the COACH hand washing system [4] has to sense that someone is at the sink oriented in such a way that there is a high probability that they want to wash their hands. It has to track the steps in carrying out this activity and prompt only when necessary. It has to sense whether the tap is on or off and where the soap is. The COACH system is sensitive to errors in this process of sensing and inference. Our goal is to provide the sensing foundations for a system that can support people undertaking food preparation activities and we therefore concentrate on the detection and classification of activities in which cooking utensils (in this case knives and spoons) are involved.

2 Previous Work: Activity Recognition

A key element of intelligent situated support technology is context recognition, how a system can understand when and where a service should be provided and furnishing this service in a manner that is sensitive to the current location, activity and characteristics of the user. The notion of *context* is typically very broad and includes any information that characterizes a situation, but in particular, the activity that the user is engaged in. Representing and automatically recognizing activity remains a fundamental challenge and plays a vital role in a broad range of applications such as automated prompting, preventive health care systems, proactive service provision, and visual surveillance. Such applications are often proposed as being particularly valuable for assisting elderly and cognitively impaired people.

Previous work has often approached the problem of activity recognition through sensors worn on different parts of the users body to detect activities such as running, and walking. Although much of this prior work has yielded significant results, it is well understood that users are generally not comfortable wearing such sensors. Moreover, the actions performed in the kitchen context relating to food preparation (such as chopping, peeling, slicing and coring ingredients) are highly

dependent on the motions of the kitchen instruments themselves (i.e. kitchen utensils such as knives and spoons) which are rather distinct from the movements of the user's body. Wearable sensor research, and computer vision based approaches [8, 15] have generally explored sets of body-level activities such as lying, standing, sitting, walking, running and cycling. The majority of these studies [7, 9, 13] collected data under controlled laboratory conditions, with the minority [5, 6] using only semi-realistic conditions.

A small number of previous projects [8, 10, 11] have utilized embedded sensing in objects to recognize human activities. For example, by attaching RFID tags to everyday items such as sugar bowls, cups and kettles [12] and requiring users to wear a sensor in the form of a wrist worn RFID reader. Such systems aim to identify everyday kitchen task performance such as boiling water, making tea, making cereal or other activities of daily living such as making a phone call, having a drink or taking medicine. Such worn RFID sensor systems are notoriously noisy, but can provide an accurate picture of activity levels in a kitchen (characterized in terms of the number of object manipulations). However, fine grained activity support with RFID is not feasible; it might be possible to detect that a knife is in use, but not to distinguish between a user's chopping or peeling actions.

3 Detecting Food Preparation Activities

Our examination of a wide range of sensor-based activity recognition systems (see Table 1) shows that existing systems are not always evaluated independently of the subjects for which the underlying models have been trained. Even the best subject independent evaluations achieve accuracies for the detection of high level activities of around 80-85%. Furthermore, participatory design activities with users usually reveal two key requirements for activity recognition systems: that

Table 1. Summary of examples of previous work on human activity recognition

	Ind.	**Dep.**	**Example activities**	**No.**	**Sensors**
[5]	81%	94%	Walking, lying, cycling	21	Worn
[6]	84%	n/a	Walking, lying, cycling	20	Worn
[7]	73%	99%	Standing, walking, brushing	2	Worn
[8]	81%	n/a	Boiling water, making tea	n/a	Embedded
[9]	n/a	79-92%	Driving, sitting, ironing	n/a	Worn
[10]	25-89%	n/a	Watching TV, preparing dinner	2	Worn
[11]	n/a	88%	Toileting, washing	14	Worn
[12]	n/a	63-99%	Vacuuming, ironing, mopping	12	Worn
[13]	n/a	98%	Sitting, standing, lying	18	Worn
[14]	80%	98%	Drinking, phoning, writing	4	Embedded

the sensing infrastructure should not intrude on the conduct of the activity itself (i.e. sensors hidden from users) and that the cost and ease of deployment (and maintenance) of such systems should be minimal. Even the most unobtrusive worn sensor system would appear to be in conflict with these needs and are not a practical solution for cognitively impaired users for who many situated support systems are targeted.

The level of abstraction of activity that existing systems are capable of detecting (e.g. walking, sitting, drinking) is also out of kilter with the granularity of action with which a system aimed at supporting cognitively impaired users requires. Wherton and Monk's studies of people with dementia [3] demonstrated that so-called failures in task completion occurred when low-level actions were unexpectedly suspended or prolonged, and fine-grained prompts to complete a low-level action were performed by carers (e.g. putting a tea bag in a tea pot, or buttering a piece of toast). Even situated support systems that are not targeted at cognitively impaired users, require prompts (and therefore action detection) at the sub-task level, for example, prompting the next step in a recipe based on the detection of the completion of the previous step. Such boundaries between subtasks might be the break between chopping one ingredient and scraping it off the chopping board to make space for the next ingredient. We can therefore see that situated support systems for food preparation require low-level recognition of food preparation activities that is beyond the theoretical capability of a worn sensor system.

3.1 Slice&Dice: Instrumented Utensils

With a view to detecting low-level food preparation activities we have developed Slice&Dice, a set of custom made cooking utensils created using FDM rapid prototyping, in which we embedded accelerometers. Slice&Dice comprises three knives and a serving spoon (see Figure 1) into which we have embedded

Fig. 1. Modified Wii Remotes embedded in specially designed handles

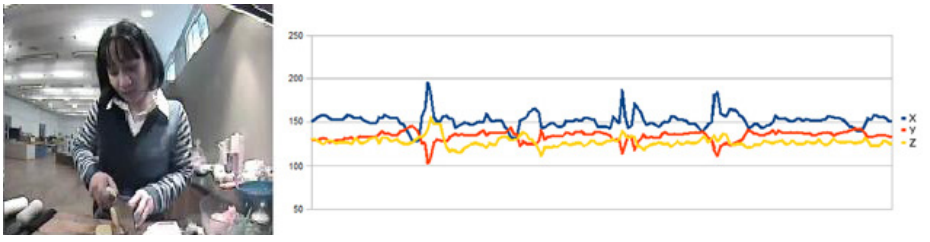

Fig. 2. X, Y and Z acceleration data for dicing with the big knife

the electronics of modified Wii Remotes [16]. An ADXL330 accelerometer, which is able to sense acceleration along three axis, is integrated in a Wii Remote. An ADXL330 is a thin, low power, 3-axis accelerometer with signal conditioned voltage outputs. The dynamic acceleration can be measured from motion, shock or vibration. An ADXL330 accelerometer can measure acceleration with a minimum full-scale range -3g to 3g. While we envisage that future versions of Slice&Dice might use the ADXL330 as a component of a smaller wireless sensor, the Wii Remote provides an excellent platform for developing and evaluating classifiers in utensils that can still retain a usable form factor.

Wii Remotes yield 3-axis values X, Y and Z. As the frequency for the embedded Wii Remotes was set to 40 Hz, this provided approximately 40 triples (samples) of acceleration data every second. The handles of the Slice&Dice utensils were designed so that they comfortably contained the Wii Remote board after its casing and the infrared camera were removed. Informal post-experiment discussions with users revealed that they were generally not aware of the embedded electronics inside the utensils while they were performing their food preparation actions.

3.2 Data Collection and Annotation

There are five IP cameras installed in the Ambient Kitchen [17], our laboratory-based instrumented kitchen environment (see Figure 3). Three of these cameras directly focus on the work surface where the food is prepared. The food ingredients used for the salad and sandwich preparation task that we set users included: potatoes, tomatoes, lettuce, carrots, onions, garlic, kiwi fruit, grapefruit, pepper, bread and butter. Twenty subjects without professional experience of food preparation were recruited from our institution. Subjects were asked to freely perform any actions relating to salad and sandwich preparation, with the ingredients provided, without any further direction from the experimenter, and subjects were not placed under any time constraint.

The time taken to complete the task varied widely, resulting in lengths of recorded sessions that varied from 7 to 16 minutes. To synchronize the videos with acceleration data, we recorded one time stamp for each sample written into the log files. In contrast to previous studies [5, 6], the subjects of our experiment were unencumbered by worn sensors, and were free to act as if they were in their home kitchens (to the extent that this is possible in a university research lab).

Subjects signed privacy waivers and the resulting accelerometer data and the annotated video (see next section) is freely available to other researchers. A set of 11 activity labels were decided upon based on an informal survey of language used in several hours of English language cooking videos found on YouTube, these were: *chopping*, *peeling*, *slicing*, *dicing*, *coring*, *spreading*, *eating*, *stirring scooping*, *scraping* and *shaving*. The recorded videos showed that all subjects performed a significant number of chopping, scooping, and peeling actions. Only small subsets of subjects performed eating (i.e. using the knife or spoon to eat ingredients), scraping (i.e. rather than peeling) or dicing (i.e. fine grained rapid chopping) actions.

The collected videos were independently annotated by 3 coders using the Anvil multimodal annotation tool [18]. Each annotator was provided with an informal description of the 11 activities for which we had labels, and were asked to independently annotate the video of each subject. After annotation, we created two data sets for training and testing the classifiers. Dataset A was the intersection of the 3 coded data sets where only labeled data for which all 3 annotators agreed was extracted. This corresponds to data where there is complete agreement between the annotators as to the action being performed. It should be noted that the boundaries between the 11 activity labels is often unclear. While dataset A is a subset of the data for which all the annotators agree, we also repeated our experiments on a second dataset (B) which is the complete data set coded by one annotator. Dataset B is therefore a larger single dataset corresponding to a single annotator's interpretation of the labels and the video data.

Although previous research on activity recognition used window sizes ranging from 4.2 seconds [5] to 6.7 seconds [6], our problem domain is quite different as our sensors are attached to the utensils themselves and we are well aware of the relatively short time window within which some actions might be best characterized. However, we still needed to determine the window size with which the best result could be achieved. Therefore, the collected data were grouped into windows sized: 0.8 seconds (i.e. 32 samples), 1.6 seconds (i.e. 64 samples), 3.2 seconds (i.e. 128 samples), 6.4 seconds (i.e. 256 samples) and 12.8 seconds (i.e. 512 samples). A 50% overlap between two consecutive windows was used. For each triple X, Y, Z, we also computed the pitch and roll of the Wii Remotes and made these available to our classifiers.

For each window, we computed *mean*, *standard deviation*, *energy* and *entropy* [19] as follows:

- *energy*: computed for values on the X-axis by: $energy(X) = \frac{\sum_{i=1}^{N} x_i^2}{N}$
- *mean*: computed by summing all values in an axis of a sliding window and divided by the window size.
- *standard deviation*: $\Delta_x = \sqrt{energy(X) - mean(X)^2}$
- *entropy*: for values on the X-axis, where N is the window size, $0\ log_2(0)$ is assumed to be 0; and $p(x_i) = Pr(X = x_i)$ the probability mass function of x_i: $entropy(X) = -\sum_{i=1}^{N} p(x_i) log_2 p(x_i)$

For each utensil, 12 features were computed for the X, Y, Z axes; 4 features were computed for pitch, and 4 features for roll. Therefore a total of 20 features are computed for one utensil within one window. As we used 4 utensils, a vector composed of all 80 features was used for training the classifiers.

4 Results

Feature vectors computed from 32, 64, 128, 256 and 512-sample windows were trained for the Bayesian Network, Decision Tree (C4.5) and Naïve Bayes classifiers found in Weka [20, 21]. The classification algorithms were evaluated on the two different data sets A and B. Since subject dependent evaluations have only very limited application in the contexts that we envisage we only evaluated our algorithms under the subject independent protocol. Under this protocol, we trained 19 subjects, tested the one remaining subject, and repeated the process.

Where n_i is the number of test instances of an activity a_i, $Acc(a_i)$ is the accuracy of an activity a_i, and M is the total number of test instances for all activities, the overall accuracy was computed as follows:

$$Overall\ Accuracy = \frac{\sum_i (n_i \times Acc(a_i))}{M} \tag{1}$$

The overall accuracy of classifiers evaluated on dataset A (data for which all three annotators agree) is given in Table 2 and this shows that the Decision Tree with window sizes of 128 (3.2 seconds) and 256 (6.4 seconds) exhibit the highest accuracies (i.e. over 80%). Table 3 shows a summary of the results for dataset B which has significantly more data (i.e. corresponding one annotator's mark-up of all the data), here a 128-sample windows demonstrated marginally better accuracy.

Table 2. Summary of results for three classifiers on dataset A

	Window size				
Algorithm	32	64	128	256	512
Decision Tree (C4.5)	77.7	78.7	80.3	82.9	80.2
Bayesian Net	70.8	72.5	79.7	78.9	69.1
Naïve Bayes	47.5	45.6	50.5	52.4	51.3

Table 3. Summary of results for three classifiers on dataset B

	Window size				
Algorithm	32	64	128	256	512
Decision Tree (C4.5)	77.0	76.8	80.2	77.5	80.1
Bayesian Net	67.5	70.2	73.6	71.3	74.5
Nave Bayes	61.3	61.8	62.2	73.5	72.7

5 Reflections

The aggregated results tell us little of the accuracy of the classifiers across the different actions and tables 4 and 5 reveal the performance of the Decision Tree classifier for different window sizes across the different activities. As one might expect, the classifier performs worst for actions which are intuitively less well defined and in themselves hard to label. Classifier performance for relatively unambiguous activities such as chopping, peeling, coring, stirring and scooping is above 80%, while classification of less distinct activities is significantly less accurate. This may in part be due to the low number of training instances for activities such as slicing, dicing and scraping. We also see significant improvements in accuracy for these activities in dataset B, where the number of

Table 4. Detailed results for the Decision Tree classifier on dataset A

	Window size									
	32		*64*		*128*		*256*		*512*	
Action	instances	%	instances	%	instances	%	instances	%	instances	%
chopping	2040	87.3	986	87.5	476	86.1	220	87.7	102	86.5
peeling	712	96.9	342	95.9	170	97.1	80	97.5	36	94.4
slicing	160	19.4	64	26.6	36	30.3	10	80.0	4	0.0
dicing	184	19.6	86	18.6	38	15.8	12	4.2	4	0.0
coring	354	73.4	170	77.7	80	76.3	36	80.6	10	60.0
spreading	224	46.0	126	44.4	56	57.1	26	53.9	10	40.0
eating	94	10.6	44	31.8	18	27.2	8	50.0	0	n/a
stirring	392	78.6	192	85.9	86	90.7	36	91.7	14	100.0
scooping	906	89.5	460	86.3	222	89.2	98	86.7	42	92.9
scraping	98	50.0	48	56.3	22	18.2	8	75	2	0.0
shaving	388	60.3	176	59.7	120	72.3	60	69.9	28	60.3

Table 5. Detailed results for Decision Tree classifier on dataset B

	Window size									
	32		*64*		*128*		*256*		*512*	
Action	instances	%	instances	%	instances	%	instances	%	instances	%
chopping	6184	88.0	3116	86.7	1510	88.1	726	86.6	328	90.9
peeling	472	72.5	230	75.7	86	67.4	50	70.0	16	75.0
slicing	788	31.6	378	27.5	150	35.3	82	24.4	36	19.4
dicing	162	3.7	78	14.1	38	15.3	16	12.5	4	12.5
coring	1246	94.5	612	94.8	300	92.7	130	86.9	64	82.8
spreading	2177	88.0	1034	90.5	498	90.4	260	86.9	116	82.8
eating	326	21.5	164	10.4	82	31.7	28	35.7	18	4.2
stirring	1122	74.7	558	70.8	190	77.4	134	70.9	64	81.3
scooping	796	37.3	390	39.5	186	51.1	76	47.2	32	81.3
scraping	716	62.2	364	64.3	172	65.2	82	65.8	32	56.3
shaving	848	75.9	416	79.6	200	90.5	88	86.4	42	92.9

Fig. 3. The Ambient Kitchen is pervasive computing prototyping environment

training samples is higher (this is particularly true for scraping). Notably, as one might expect, the activities themselves have different temporal scales (e.g. dicing vs. eating), and thus the window size for which a classifier is most accurate varies across the activities.

Slice&Dice is a first step in the development of a low-level activity recognition framework for fine-grained food preparation activities. As such it forms one component of the heterogenous sensor framework that we are developing in the Ambient Kitchen [17]. The Ambient Kitchen (Figure 3) also utilizes extensively deployed RFID readers (and tags associated with all moveable non-metallic objects), five IP cameras integrated into the walls of the kitchen, a mote-based wireless network of accelerometers attached to other kitchen objects, and a pressure sensitive floor. Projectors and speakers embedded in the environment facilitate the generation of spatially embedded cues. The current configuration uses blended projection from four small projectors placed under up-lighters in the main workbench. The projection is reflected by a mirror under the over-head cabinets onto the kitchen wall. Integrating spatially situated auditory cueing allows us to explore the various configurations by which auditory and visual cues can be realized. In this way we can explore the practical problems of creating technologies that are robust and acceptable in the home.

References

[1] Weiser, M.: The computer for the 21st century. Scientific American 265(3), 93–104 (1991)
[2] Wherton, J., Monk, A.: Designing cognitive supports for dementia. SIGACCESS Access. Comput. (86), 28–31 (2006)
[3] Wherton, J., Monk, A.: Technological opportunities for supporting people with dementia who are living at home. International Journal of Human-Computer Studies 66(8), 571–586 (2008)
[4] Mihailidis, A., Boger, J., Canido, M., Hoey, J.: The use of an intelligent prompting system for people with dementia. interactions 14(4), 34–37 (2007)
[5] Tapia, E.M., Intille, S.S., Haskell, W., Larson, K., Wright, J., King, A., Friedman, R.: Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In: Proceedings of 11th IEEE International Symposium on Wearable Computers, October 2007, pp. 37–40 (2007)

[6] Bao, L., Intille, S.S.: Activity recognition from user-annotated acceleration data. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 1–17. Springer, Heidelberg (2004)
[7] Ravi, N., Dandekar, N., Mysore, P., Littman, M.L.: Activity recognition from accelerometer data. In: Proceedings of the Seventeenth Conference on Innovative Applications of Artificial Intelligence (IAAI), pp. 1541–1546. AAAI Press, Menlo Park (2005)
[8] Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., Rehg, J.M.: A scalable approach to activity recognition based on object use. In: Proceedings of the International Conference on Computer Vision (ICCV), Rio de (2007)
[9] Huynh, T., Blanke, U., Schiele, B.: Scalable recognition of daily activities with wearable sensors. In: Hightower, J., Schiele, B., Strang, T. (eds.) LoCA 2007. LNCS, vol. 4718, pp. 50–67. Springer, Heidelberg (2007)
[10] Tapia, E.M., Intille, S.S., Larson, K.: Activity recognition in the home using simple and ubiquitous sensors. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 158–175. Springer, Heidelberg (2004)
[11] Philipose, M., Fishkin, K.P., Perkowitz, M., Patterson, D.J., Fox, D., Kautz, H., Hahnel, D.: Inferring activities from interactions with objects. IEEE Pervasive Computing 3(4), 50–57 (2004)
[12] Stikic, M., Huynh, T., Van Laerhoven, K., Schiele, B.: Adl recognition based on the combination of rfid and accelerometer sensing. In: Second International Conference on Pervasive Computing Technologies for Healthcare. PervasiveHealth 2008, 30 2008-February 1 2008, pp. 258–263 (2008)
[13] Kim, I., Im, S., Hong, E., Ahn, S.C., Kim, H.-G.: ADL classification using triaxial accelerometers and RFID. In: Proceedings of the International Workshop on Ubiquitous Convergence Technology (November 2007)
[14] Wang, S., Yang, J., Chen, N., Chen, X., Zhang, Q.: Human activity recognition with user-free accelerometers in the sensor networks. In: International Conference on Neural Networks and Brain (ICNN&B 2005), October 2005, vol. 2, pp. 1212–1217 (2005)
[15] Robertson, N., Reid, I.: A general method for human activity recognition in video. Computer Vision and Image Understanding 104(2-3), 232–248 (2006); Special Issue on Modeling People: Vision-based understanding of a person's shape, appearance, movement and behaviour
[16] Wii Remote: `http://en.wikipedia.org/wiki/Wii_Remote`
[17] Olivier, P., Xu, G., Monk, A., Hoey, J.: Ambient kitchen: designing situated services using a high fidelity prototyping environment. In: PETRA 2009: Proceedings of the 2nd International Conference on PErvsive Technologies Related to Assistive Environments, pp. 1–7. ACM, New York (2009)
[18] Kipp, M.: Anvil – a generic annotation tool for multimodal dialogue. In: Proceedings of In EUROSPEECH 2001, pp. 1367–1370 (2001)
[19] Shannon, C.E.: A mathematical theory of communication. SIGMOBILE Mob. Comput. Commun. Rev. 5(1), 3–55 (2001)
[20] Weka: `http://www.cs.waikato.ac.nz/ml/weka/`
[21] Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco (2005)