A Dynamic Time Warping Approach to Real-Time Activity Recognition for Food Preparation

Cuong Pham, Thomas Plötz, and Patrick Olivier

Culture Lab, School of Computing Science, Newcastle University, United Kingdom
{cuong.pham,t.ploetz,p.l.olivier}@ncl.ac.uk

Abstract. We present a dynamic time warping based activity recognition system for the analysis of low-level food preparation activities. Accelerometers embedded into kitchen utensils provide continuous sensor data streams while people are using them for cooking. The recognition framework analyzes frames of contiguous sensor readings in real-time with low latency. It thereby adapts to the idiosyncrasies of utensil use by automatically maintaining a template database. We demonstrate the effectiveness of the classification approach by a number of real-world practical experiments on a publically available dataset. The adaptive system shows superior performance compared to a static recognizer. Furthermore, we demonstrate the generalization capabilities of the system by gradually reducing the amount of training samples. The system achieves excellent classification results even if only a small number of training samples is available, which is especially relevant for real-world scenarios.

1 Introduction

Ambient Assisted Living (AAL) is a key domain of Ambient Intelligence (AmI), which focuses on the development of technology that supports humans during their activities of daily living (ADL). Applying such technology to people's private homes offers, for example, a realistic opportunity for age-related impaired people to live more independent and longer in their homes. Especially the kitchen plays an important role in people's life, as it is an indispensable place where many everyday life activities such as cooking and food preparation take place. Although the AmI research community has recently made a stride in real-world applications for health care at private homes, the technology driven support for cooking and food preparation activities is still relatively under-explored.

Technology support in the kitchen would include nutritional advice for, e.g., recipe selection and automatic guidance whilst cooking. Especially the latter is of major importance for people with age-related impairments, such as stroke and dementia. For these people it is often difficult to focus on the particular kitchen task and, hence, to complete the cooking. Monitoring kitchen activities and eventually providing situated advice would help substantially, e.g., for following a recipe and would thus give these people more independence in their private homes.

As a first step towards automatic kitchen assistants we developed an activity recognition system for kitchen environments that automatically recognizes relevant food preparation tasks [9]. Accelerometers were embedded into kitchen utensils and sensor

B. de Ruyter et al. (Eds.): AmI 2010, LNCS 6439, pp. 21–30, 2010.

data are recorded whilst the utensils are being used. Recognition was based on a decision tree classifier that processes fixed length statistical feature vectors extracted in a sliding window procedure. By means of a closed set evaluation we demonstrated the reliable recognition of ten common kitchen tasks, which represents the proof of concept for the overall approach.

In this paper we substantially improve our AR system for the kitchen by integrating dynamic time warping (DTW) based classifiers. The motivation is to adapt Activity Recognition towards the users' idiosyncrasies by means of an automatic maintenance of a proper template database, which is used for the actual recognition. DTW based classification aligns sequential data in a way that preserves potentially existing internal structures of the data, which is beneficial for the analysis of real-world time-series. Compared to [9], in this paper we apply the system to a more realistic evaluation tasks with an open inventory of kitchen activities to be recognized, i.e., integrating segmentation of continuous data streams. Furthermore, we investigate the effect smaller training sets have on the accuracy of the recognition system. Especially the latter is relevant for realistic tasks where the amount of labeled sample data is usually limited since manual annotation is tedious and labor intensive, hence costly. The whole system is implemented as a real-time recognition system, which is integrated into [9], our lab-based pervasive kitchen environment.

2 Related Work

Technology-augmented kitchens have been used to explore the application of AmI in the home. For example, the AwareKitchen was an intelligent environment equipped a number of sensors such as microphone, forces, accelerometer etc. to detect cutting food activity[1]. CounterIntelligence[2], an augmented reality kitchen, is able to support the users while they are performing cooking task by displaying instructive text. An application of RFID technology embedded in a kitchen counter was described in [6]. A number of other design proposals relate to the provision of situated advice on food and cooking (cf., e.g., [7]). Moreover, with the goal of helping old people to live more independent in the homes, the Ambient kitchen, a design of situated services using a high fidelity prototyping environment in which technologies such as wireless accelerometers, RFID, IP cameras, projectors are completely embedded into the environment, is proposed in [8].

A significant amount of previous work on activity recognition is based on sensors wore on user's body and the application of pattern recognition techniques. In [11], for example, a Dynamic Bayesian Network approach was described that aims at recognizing high-level household activities based on object use.[12] explored hidden Markov models (HMMs) for activity modeling and recognition. In [9], an AR framework for recognizing low-level typical food preparation activities using accelerometers and classical machine learning approaches has been presented.

Dynamic Time Warping (DTW) is a standard technique for the comparison of time series data [15]. The key idea is to align two sequences in an optimal way, i.e., minimizing "costs" for this "warping" (see Sec. 3.1). Numerous applications of DTW have been developed being as diverse as personalized gesture recognition [10], speech recognition [15], and hand printed signature verification [16].

3 Dynamic Time Warping Based Activity Recognition for Food Preparation

Aiming to support people in their activities of daily living, the focus of our research is on the development of a kitchen monitoring system that ultimately will provide automatic situated assistance for kitchen tasks like food preparation. The basis for such a system is reliable and efficient activity recognition. For this purpose we proposed to integrate accelerometers into kitchen utensils, which are used for food preparation [9]. Analyzing the data recorded by the sensor-equipped utensils allows to recognize typical food preparation activities in real-time, i.e., while the person is acting in the kitchen. Based on this, situated support for food preparation becomes possible.

Analyzing food preparation activities in more detail, it becomes clear that even the most fundamental activities exhibit substantial variance depending on the personal preferences of how to handle the food ingredients and the utensils. Consider, for example, the process of shaving a carrot using a knife. Some persons perform long, slow movements of the utensil along the carrot towards themselves, which is comparable to "carving" the vegetable. Others tend to perform short, fast cuts of the carrot's surface thereby using the knife more in a "chopping" way. Although both kinds of movements differ substantially, they represent the same kind of activity and an automatic recognition system needs to cope with it.

In order to deal with the aforementioned users' idiosyncrasies, we developed a fully automatic, real-time activity recognition system, which is outlined in Fig. 1. Accelerometer data are recorded while a person is working in the kitchen. As in our previous work, sensor equipped standard kitchen utensils integrating modified Wii remotes are used. For continuous sensor data streams (x,y,z) with a sampling frequency of 40Hz, then frames of 64 contiguous samples are extracted in a sliding window procedure (50 percent overlap; lower part of Fig. 1). Following some basic preprocessing and trivial movement detection (simple threshold based procedure), then the actual classification of the extracted frame regarding the activities of interest is performed (middle part of the figure). This recognition procedure is performed as a DTW-based template comparison with an automatically maintained template database. This database contains representative templates for the activities of interest together with activity specific thresholds for acceptance / rejection (upper right part). By means of the analysis of the DTW scores the template database is continuously adapted to represent the idiosyncrasies of the particular activities performed by different users. The output of the system consists of classification hypotheses for every extracted frame including possible rejection, which effectively means segmentation of continuous sensor data streams.

In the following we will first briefly summarize the theoretical foundations of Dynamic Time Warping before giving detailed descriptions of the key components of the overall recognition system.

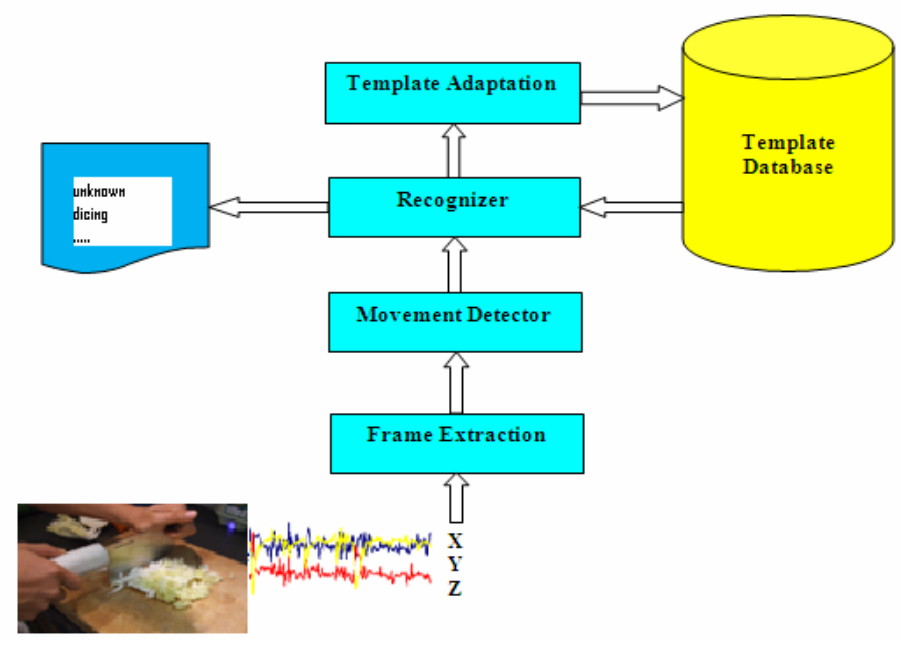

Fig. 1. Activity Recognition for Food Preparation Tasks – System Overview

3.1 Dynamic Time Warping – A Brief Overview

DTW (cf. [15]) is a constrained, non-linear pattern matching method based on dynamic programming to measure the dissimilarity between two time series. Let $O = o_1,o_2,..,o_m$ and $Y = y_1,y_2,..,y_n$ be two time series. DTW finds an optimal mapping from O to Y by reconstructing a *warp* path W that optimizes the maping of the two sequences: $W = w_1,w_2,\ldots,w_K$ where $\max\{m,n\} \leq K < m+n$, with K denoting the length of the warp path. The warp path constrained in the sense that it is anchored by the start and end points of both sequences. To map the in-between elements of both time series a step-wise distance minimization for every position is performed:

$$\delta(W) = \min \{ \sum_{k=1}^{K} \delta(w_{ki}, w_{kj}) \} \quad (1)$$

The actual distance calculation is usually (but not necessarily) based on the Euclidean distance $d(o_i,y_j)$ and dynamic programming [15]:

$$\delta(i,j) = d(o_i,y_j) + \min \{ \delta(i-1,j), \delta(i-1,j-1), \delta(i,j-1) \} \quad (2)$$

where i and j represent monotonically increasing indices of the time series O and Y. The resulting matching cost $\delta(W)$ is then usually normalized to the range of [0,1] to ensure comparability.

3.2 DTW-Based Recognition

By means of the accelerometers integrated into the kitchen utensils continuous three-dimensional (x,y,z) data streams are recorded. In a sliding window procedure 64 contiguous samples are summarized to frames. Adjacent frames overlap by 50 percent. The actual parameterization of the frame extraction process has been optimized in previous experiments (cf. [9]).

For every observation frame O[u], which is recorded for a particular utensil u, activity recognition is performed using the DTW-based algorithm as outlined in Fig. 2. Note that by means of a trivial threshold comparison only those frames are considered where the utensil was actually moving. After computing and sorting the DTW scores for all templates (lines 4:12 in Fig. 2), the set of the smallest min(K,n) scores is compared to the activity-specific thresholds (lines 13:19). If none of the DTW score s contained in the sorted list `cost` is smaller than the particular threshold the observation frame O is rejected, i.e., assigned to the unknown class. Heuristically we chose K=10, which provides reasonable results for acceptance / rejection on a cross validation set. The Threshold function (`Thresh,` line 14) retrieves the class-based threshold of the template Y[index[i]].

```
Input:    An observation frame O, utensil u, number K of
          sorted match scores to analyze for final result
Output:   Activity hypothesis
          //Extract templates
1:        CurrentTemplates= ExtractTemplateDB(u);
2:        n=the number of templates in CurrentTemplates;
3:        For i from 1 to n do
4:            Y[i]=CurrentTemplate.template[i];
5:            cost[i]= DTW(O, Y[i]);
6:            index[i] = Y[i].Id;
7:        End for
          // sort and maintain indices of the templates
8:        Sort(cost, index);
9:        For i from 1 to min{K,n} do
              //Acceptance
10:           If cost[i]<Thresh(Y[index[i]]) then
11:               activity(Y[index[i]])→activity_list;
12:               break;
13:           Else
              //Rejection
14:               unknown activity → activity_list;
15:           End if
16:       End for
17:       Return (activity_list);
```

Fig. 2. DTW-based activity recognition

The recognition procedure utilizes activity specific thresholds T_a. Applying Chow's rule [13] to our domain, an observation frame o is classified as being a particular activity a if:

$$\delta(o,t) = \min_{j=1..N} \{ \delta(o,t_j) \} < T_a \tag{3}$$

where t represents a template from the database, which represents the activity a. N is the overall number of templates in the database. A frame is classified as unknown, i.e., being rejected, if:

$$\delta(o,t) = \min_{j=1..N} \{ \delta(o,t_j) \} \geq \max_{i=1..n} \{T_i\} \tag{4}$$

where n denotes the number of classes of interest. The class-based thresholds were manually selected through a 5-fold cross validation procedure.

3.3 Template Adaptation

In order to adapt the overall recognition system towards users' idiosyncrasies, the template database is continuously being updated, i.e., templates are removed and added if necessary. The adaptation scheme used can be described as follows. Let f_k define the weighted histogram of recognition hypotheses for a particular time step k consisting of activity specific entries $f_k(a)$, where specifies the particular activity, and w denotes a (heuristically chosen) adaptation weight:

$$f_k(a) = \begin{cases} f_{k-1}(a) \cdot w & \text{if} \quad \delta(a,t) < \alpha \\ f_{k-1}(a) \cdot (1-w) & \text{otherwise} \end{cases} \tag{5}$$

α denotes an acceptance / rejection threshold, which is derived from the activity specific template thresholds: $\alpha = T_a/(1+T_a)$. All thresholds were optimized in a separate cross-validation procedure. Let $\gamma_k(a)$ be the cumulative number of recognitions of activity a at time k. The positive probability of activity a at time k $\rho_k(a)$ is computed as:

$$\rho_k(a) = \frac{f_k(a)}{\gamma_k(a)} / \sum_k \frac{f_k(a)}{\gamma_k(a)} \tag{6}$$

The negative probability of activity a at time k is hence defined as:

$$\varphi_k(a) = \frac{1-\rho_k(a)}{\sum_k 1-\rho_k(a)} \tag{7}$$

At time k, if an observation frame makes $\varphi_k(a) = \min \{ \varphi_i(a) \}$ for i=1..k, then this frame will be updated as a template into the *positive* list in the template database for later use (adaptation). The template which makes $\varphi_i(a) = \max\{ \varphi_i(a) \}$ for i=1..k is moved to the *negative* list in the template database at the same time. The negative list is used only when recognizer has recognized an unknown activity on the positive list, then recognizer does one more try on the negative list before returning the activity list.

4 Experimental Evaluation

In order to evaluate the applicability of the proposed DTW-based approach to activity recognition, we performed a number of practical experiments. Therefore, we used the (publically available) dataset from our previous work [9], which covers 20 persons pursuing typical food preparation tasks (salad and sandwich making) using our sensor-equipped utensils. No further constraints were given. Ten typical low-level activities were subject to recognition, namely *chopping*, *peeling*, *slicing*, *dicing*, *scraping, shaving*, *scooping*, *stirring*, *coring, spreading.* Additionally a considerable amount of sensor data, which does neither belong to one of the ten known activities nor to "idle" is included in the dataset. In total more than 6 hours of sensor data have been collected from four sensor equipped kitchen utensils (knifes and spoon).

Extending the proof-of-concept, which was given in [9], we aimed for realistic experiments using the recognition system in real-world scenarios. This implies that the recognition system was applied online, i.e., continuous data streams had to be segmented and classified (open lexicon with rejection) in real-time, i.e., with negligible latency (results given in Sec. 4.1). Furthermore, we are interested in the dependency of the recognition procedure on the number of annotated samples available for training. Since manual annotation is tedious and costly, it is somewhat unrealistic to rely on large training sets for setting up the recognition system. Consequently, we performed a second set of experiments where the number of training samples has been decreased step by step. Classification results are reported in section 4.2.

Recognition results are reported as frame-wise precision and recall values. The precision for some activity a was calculated by dividing the number of correctly classified frames by the total number of frames classified as being a (i.e. true positives/(true positives + false positives)). Recall was calculated accordingly as the ratio of the number of correctly classified frames of a and the total number of frames of a (true positives/total number of frames of A). Baseline results for the evaluation are given by the decision tree based system described in [9].

4.1 Results for Full Training Set

For the first set of experiments we used all available training data (see below) to estimate the recognition system. When the dataset was recorded, it was manually annotated by 3 independent subjects. The consensus of the three annotators serves as ground truth. Since we recorded complete cooking sessions the dataset for obvious reasons is dominated by idle "activities", i.e., frames recorded while the particular utensil of interest is not moved at all. A quick subject-independent test including all idle frames led to 96.36% overall accuracy as the recognition of "idle" is almost perfect (see Sec. 3: simple threshold comparison). To avoid this over-optimistic and not quite informative evaluation, we limited the set of "idle" frames to four per utensil, which were randomly selected per subject. This effectively truncates the dataset to 12,265 frames (of 64 samples each).

The evaluation was performed in a "leave-one-subject-out" manner, i.e., we trained the recognizer using the data from 20 subjects, and tested on those data recorded for the remaining subject. This process was repeated for all 21 subjects and results were averaged. The overall performance of the proposed DTW-based approach are presented in

Tab. 1. Additionally the results for the baseline system (Decision Tree C4.5) are given. It can be seen that the new approach clearly outperforms the baseline with overall precision rates of approx. 83% (CBT-DTW) vs. 77.9% (DT C4.5), and recall rates of 82.8% (CBT-DTW) vs. 76.7% (DT C4.5). All differences are statistically significant. Additionally Tab. 2 illustrates the aggregated confusion matrix for the subject-independent evaluation of CBT-DTW.

Table 1. One-subject-leave-out evaluation (all figures in percent)

Activity	CBT-DTW			Decision Tree C4.5		
	Precision	**Recall**	**False Positive**	**Precision**	**Recall**	**False Positive**
chopping	82.61	88.54	2.22	82.21	87.5	7.37
coring	77.02	81.94	0.21	74.02	77.7	4.12
dicing	51.16	54.63	0.25	24.87	18.7	4.25
peeling	72.76	80.63	0.53	88.7	95.9	3.91
scraping	80.09	81.1	0.12	56.8	56.3	3.37
shaving	72.79	82.73	0.28	55.11	59.7	2.91
slicing	70.31	70.73	1.21	33.47	26.6	4.95
spreading	71.06	86.57	0.77	54.33	44.4	2.32
scooping	97.92	94.55	0.78	91.2	86.3	2.6
stirring	84.77	86.98	0.08	81.63	85.92	1.26
idle	100	100	0	100	100	0
unknown	91	80.92	4.96	85.3	83.2	9.82
Overall	**83.02 ± 4.8**	**82.78 ± 5.5**	**2.61 ± 1.03**	**77.9 ± 8.7**	**76.7 ± 6.5**	**6.29 ± 2.1**

Table 2. Aggregated confusion matrix for one-subject-leave-out evaluation for CBT-DTW [%]

	a: chopp.	**b: coring**	**c: dicing**	**d: peel.**	**e: scrap.**	**f: shav.**	**g: slic.**	**h: spread.**	**i: scoop.**	**j: stir.**	**k: idle**	**l: unkn.**
A	88.4	0.14	5.5	0	0.3	0	4.56	0	0	0	0	1.09
B	0	81.9	0	6.4	3.1	0	0	38.9	0	0	0	4.72
C	36.6	0	54.6	0	1.2	0	5.2	0	0	0	0	2.48
D	0	4.0	0	80.6	5.8	3.7	0	0	0	0	0	5.9
E	2.3	1.06	0.1	4.05	81.1	2.6	0.96	0	0	0	0	7.81
F	2.0	0	0	8.43	1.21	82.7	0	0	0	0	0	5.62
G	15.0	1.71	7.63	0	1.37	0	70.7	0	0	0	0	3.53
H	0	0	0	0	5.76	1.68	0	86.6	0	0	0	5.99
I	0	0	0	0	0	0	0	0	94.6	2.08	0	3.38
J	0	0	0	0	0	0	0	0	8.07	87.0	0	4.95
K	0	0	0	0	0	0	0	0	0	0	100	0
l	3.28	0.92	0.03	2.3	2.96	0.58	1.4	3.51	3.18	0.95	0	80.9

4.2 Results for Reduced Training Sets

In the second set of experiments we analyzed the dependency of the classification accuracy on the number of samples available for training the recognizer. We therefore gradually reduced the amount of annotated training samples by randomly selecting frames to be excluded from training and evaluated the resulting recognizers on the remaining data.

Fig. 3 illustrates the classification accuracies of both the proposed CBT-DTW approach and the baseline system (DT C4.5). The number of frames used for model training is given at the x-axis, whereas the y-axis represents the classification accuracies. For the sake of clarity in the illustration the (discrete) figures are connected to continuous curves. It can be seen that the proposed DTW-based recognition approach greatly generalizes even if only small amounts of training data were available. For example, with only 5 labeled frames for training (per activity) the accuracy of CBT-DTW is still about 79%. The performance of the baseline system here drops to approximately 55%. Fig. 4 also illustrates the general superiority of the proposed approach compared to the baseline system.

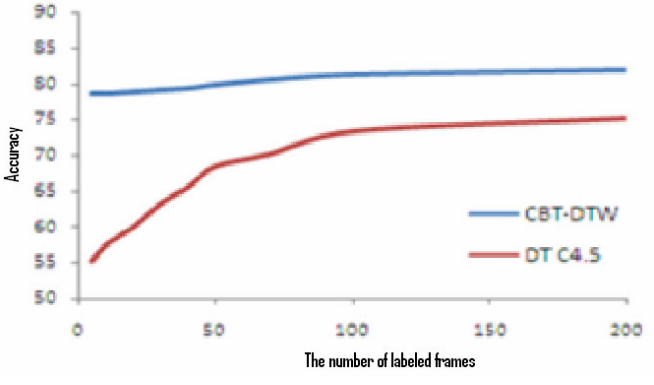

Fig. 3. Classification accuracies in dependence of the amount of training data

5 Summary

Automatically monitoring food preparation activities is a key for the development of kitchen assistance systems that, for example, provide situated cooking support for people with age-related impairments like dementia. For non-intrusive activity recognition we integrated accelerometers into kitchen utensils and analyze the sensor data recorded while people cook using these enhanced utensils.

In this paper we presented an activity recognition system that automatically adapts towards the idiosyncrasies of people using kitchen utensils. Based on a Dynamic Time Warping procedure a template matching system has been developed, which successfully segments and recognizes ten low-level kitchen activities by analyzing contiguous frames of sensor readings. The adaptation of the AR system towards variants of certain activities is pursued by an automatic maintenance procedure, which effectively updates the template database if necessary.

By means of an experimental evaluation on a large, realistic datasets that covers unconstrained food preparation we demonstrated the capabilities of the proposed approach. In a second set of experiments we gradually reduced the number of training samples. The proposed DTW-based recognition system shows superior generalization even if only a few samples are available for training.

References

1. Kranz, M., Schmidt, A., Rusu, B., Maldonado, A., Beetz, M., Hornler, B., Rigoll, G.: Sensing Technologies and the Player-Middleware for Context-Awareness in Kitchen Environments. In: Proc. 4th Int. Conf. on Networked Sensing Systems (2007)
2. Bonanni, L., Lee, C.H., Selker, T.: CounterIntelligence: Augmented Reality Kitchen. In: Proc. CHI 2005, pp. 2239–2245 (2005)
3. Bao, L., Intille, S.S.: Activity recognition from user-annotated acceleration data. In: Proc. Int. Conf. Pervasive Comp., pp. 1–17 (2004)
4. Ravi, N., Dandekar, N., Mysore, P., Littman, M.L.: Activity recognition from accelerometer data. In: Proc. Conf. Innovative Applications of Artificial Intelligence (IAAI), pp. 1541–1546 (2005)
5. Chang, K.-H., Liu, S.-Y., Chu, H.-H., Hsu, J., Chen, C., Lin, T.-Y., Huang, P.: Dietary–aware dining table: Observing dietary behaviors over tabletop surface. In: Proc. Int. Conf. Pervasive Comp., pp. 366–382 (2006)
6. Chi, P.-Y., Chen, J.-H., Chu, H.-H., Chen, B.-Y.: Enabling nutrition-aware cooking in a smart kitchen. In: Proc. CHI 2007 Extended Abstracts on Human Factors in Computing Systems, pp. 2333–2338 (2007)
7. Tran, Q.T., Calcaterra, G., Mynatt, E.D.: Cooks collage: Deja vu display for a home kitchen. In: Proc. HOIT 2005 Conf. on Home-Oriented Informatics and Telematics, pp. 15–32 (2005)
8. Olivier, P., Monk, A., Xu, G., Hoey, J.: Ambient Kitchen: Designing situated services using a high fidelity prototyping environment. In: Proc. Int. Conf. Pervasive Technologies Related to Assistive Environments (2009)
9. Pham, C., Olivier, P.: Slice&Dice: Recognizing Food Preparation Activities using Embedded Accelerometers. In: Proc. Europ. Conf. Ambient Intell., pp. 34–43 (2009)
10. Liu, J., Zhong, L., Wickramasuriya, J., Vasudevan, V.: uWave: Accelerometer-based personalized gesture recognition and its applications. In: Proc. Int. Conf. Pervasive Comp. and Comm., pp. 1–9 (2009)
11. Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., Rehg, J.: A scalable approach to Activity Recognition based on Object Use. In: Proc. Int. Conf. Comp. Vision (2009)
12. Wang, L., Gu, T., Tao, X., Lu, J.: Sensor-Based Human Activity Recognition in a Multi-user Scenario. In: Proc. Europ. Conf. Ambient Intell., pp. 78–87 (2009)
13. Chow, C.K.: An Optimal Error and Reject Tradeoff. IEEE Trans on Information Theory 16(1), 41–46 (1970)
14. Xi, X., Keogh, E., Shelton, C., Wei, L.: Fast Time Series Classification Using Numerosity Reduction. In: Proc. Int. Conf. Machine Learning (2006)
15. Myers, C.S., Rabiner, L.R.: A comparative study of several dynamic time-warping algorithms for connected word recognition. The Bell System Technical Journal 60(7), 1389–1409 (1981)
16. Jayadevan, R., Kolhe, S.R., Patil, P.M.: Dynamic Time Warping Based Static Hand Printed Signature Verification. Journal of Pattern Recognition Research 4(1), 52–65 (2009)