Recording a Complex, Multi Modal Activity Data Set for Context Recogntion

P. Lukowicz,G. Pirkl, D. Bannach, F. Wagner

Embedded Systems Lab, University of Passau, Germany

A. Calatroni, K. Förster, T. Holleczek, M. Rossi, D. Roggen, G. Troester

Wearable Computing Lab, ETH, Switzerland

J. Doppler, C. Holzmann, A. Riener, A. Ferscha

Institute Pervasive Computing, JKU Linz, Austria

R. Chavarriaga

Defitech Foundation Chair in Non-Invasive Brain-Machine Interface, EPFL Lausanne, Switzerland

Abstract

Publicly available data sets are increasingly becoming an important research tool in context recognition. However, due to the diversity and complexity of the domain it is difficult to provide standard recordings that cover the majority of possible applications and research questions. In this paper we describe a novel data set hat combines a number of properties, that, **in this combination**, are missing from existing data sets. This includes complex, overlapping and hierarchically decomposable activities, a large number of repetitions, significant number of different users and a highly multi modal sensor setup. The set contains around 25 hours of data from 12 subjects. On the low level there are around 30'000 individual annotated actions (e.g. picking up a knife, opening a drawer). On the highest level (e.g. getting up, breakfast preparation) we have around 200 context instances. Overall 72 sensors from 10 different modalities (different on body motion sensors, different sound sources, two cameras, video, object usage, device power consumption and location) were recorded.

1 Introduction

In most established fields related to pattern recognition and signal processing standard data sets exist, on which new algorithms can be evaluated and compared. Such data sets ensure that different approaches are compared in a fair and reproducible way. They also allow different groups to concentrate on method development rather then on repeating often considerable effort involved in data collection.

Recently publicly available data sets have also started emerging in the area of context recognition (see related work below). However, due to the diversity and complexity of the context recognition domain it is difficult to define a few "standard" task. Instead, there are many aspects that need to be considered in different applications.

1.1 Paper Contributions

In this paper we describe a large data set that has been collected as part of the OPPORTUNITY EU project and is currently being prepared for public release. The data set was recorded with the following goals in mind:

1. Complex, hierarchical, interleaved activity set.
2. Large number of properly labeled instances of activities on all hierarchy levels.
3. Complex, highly multi modal sensor setup that allows the effectiveness of different sensor combinations to be compared against each other.
4. Significant number of different users to allow the study of user dependent recognition.

The set contains around 25 hours of data from 12 subjects. On the low level there are around 30'000 individual actions (e.g. picking up a knife, opening a drawer). On the highest level (getting up, breakfast preparation) we have around 200 context instances. All of those were annotated during the recording and are currently being verified/re-annotated using the video stream. While the number of high level contexts is not unusual for this type of experiment, the number of annotated low level actions is far beyond what is available in other data sets. On the other hand, the availability of annotations for all low level activities is crucial for the development of complex, hierarchical recognition methods.

The experiment was carefully designed to provide realistic data. To this end the subjects were given loose high level instructions with respect to the activities and a good approximation of a real life environment was established. Nonetheless, this is clearly an artificial data set recorded in a laboratory setting. On the other hand, by choosing such a setting we were able to get a large number of repetitions of the same activity with the ability to annotate each individual instance. Both is difficult when recording in real life where people are free to do whatever they like and neither permanent observer presence nor detailed video recording are possible.

1.2 Related Work

PlaceLab data set The most popular data set available in pervasive / ubiquitous area is the so called PlaceLab data set (see [1]). Longtime data recordings with a rich multimodal sensor environment captures the behavior and activities of test subjects over days or weeks in a sensor equipped apartment. Environment sensors (like temperature, or humidity sensors) capture the environmental conditions of the living area. Sensors attached to objects allow to collect information about object interactions. In the beginning only 3 acceleration sensors capture on body posture and mode of locomotion, most information has been added in offline annotation sessions looking at the video stream or listening to audio recording. Only one data stream from each set of cameras and set of microphones have been recorded according to the current position of the person. The main goal of this data set is to provide a rich set of object interactions for behavior research and data for context algorithms. Neither specific and well defined gestures nor a high number of repetitions of gestures is the goal of this project. Capturing a single gesture with several sensor modalities also had a lower priority.

Kitchen data set Data recording in a kitchen environment has been performed by a group from TU Munich (see [2]). They focus on marker free motion capture of complex gestures. The data set provides video, motion capture, RFID reader and reed switch information. RFID reader and reed switches give timing information when the subject interacts with the kitchen environment. There have not been any on body sensors like acceleration or gyroscope sensors capturing body postures or modes of locomotion.

Activity Recognition in a homesetting Another data set has been presented in [3]. The authors recorded over a month the test subject's life. Digital or binary sensors (*idle* or *active*) like reed switches give information when the person interacts with furniture or objects of interest. Neither video, audio, modes of locomotion nor posture information have been recorded. The data set lacks in the missing number of sensor systems and number of test subjects.

2 The Scenario

As described in the introduction, the data set was intended to provide (1) a high number of instances of (2) different (3) multi level and (4) multi user activities recorded by (4) a high number of different sensor modalities.

A breakfast related scenario has been chosen as it has extensively been used in literature (for example in [4],[5],[6] or [7]). The tasks of the scenario are every day activities relevant for many applications. At the same time they involve complex hierarchies and overlaps of many divers actions (see below).

The experiment has been set up in a room (1) of the dimensions 8x5mx3m. The room has 3 doors, a kitchen section and a table in the center of the room. We divided the case study into two parts both providing a high number of atomic instances: The first part of the recording has been introduced to provide a high number of low level activities for training. The test subject sequentially has to go through a highly scripted sequence of simple actions (20 repetitions): (1) open and close the fridge (2 activities), (2) open and close the dishwasher (2 activities), (3) open and close 3 drawers (each at different heights, 2 activities each), (4) open and close door 1 (2 activities), (5) turn on and off the lights (2 activities), (6) open and close door 2 (2 activities), (7) clean table (1 activity), (8) drinking and standing, (2 activities), (9) drinking and sitting (2 activities). For each run we therefore record 21 different activities resulting in 420 instances per subject.

The activities have been chosen to be representative of the second, main part of the recording which was a semi realistic morning routine. The person at first gets up and goes out of the apartment for a walk. After coming back to the apartment, the breakfast is prepared. At first she prepares coffee, she fetches the sugar, spoon, milk and cup from their specific locations. After coffee preparation all dishes and food are fetched from the different locations and the subject sets the table. The bread is sliced and she puts some spread cheese and slices of peperoni on the bread. Water is poured in the water glass and after that she starts eating and drinking. After having finished she cleans up the table and puts the dishes in the dishwasher, the food is put back in the drawers and the fridge. She then turns off the lights, closes all doors and goes back to sleep.

The above includes overlapping activities like walking and moving of items at the same time or moving items and closing of doors. Especially when working with acceleration sensors such overlapping and simultanious occuring activities add complexity to the recognition task.

Figure 2 depicts the decomposition of activities at different temporal zoom levels. Level I are high level actions which are the abstract building blocks of the morning routine. The temporal sequence of these activities is static.

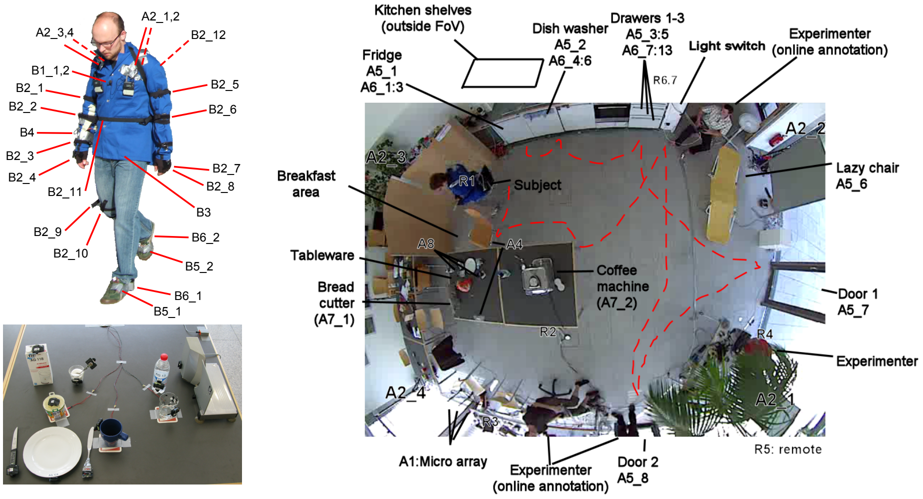

Figure 1: Left top: The configuration of on body sensors. Left bottom: Some of the objects instrumented with acceleration and gyroscope sensors. Right: The room in which the experiments were conducted including the location of sensors and activities. The red trails shows the path taken during the drill session.

If we pick out one of these high level activities and look at it more closely it can again be decomposed into lower level (but still complex) actions (symbolized as ellipses) on level II. The order of these actions is not fixed and differs from subject to subject. Zooming in on level III shows that the activities of level II are dominated by modes of locomotion (for example walking, standing, sitting) and by manipulative gestures (like moving, reaching, grasping or releasing). We want to point out that it is possible that manipulative gestures and modes of locomotion overlap. Logical, physical and spatial limitations distinguish and influence the order of these activities. *All of the above activity levels are exactly annotated* allowing complex multilevel reasoning to be performed on the data.

2.1 Sensors

As described in the introduction a key aim of the experiments was to provide a highly multimodal data set to allow different sensor types and combinations to be compared and dynamically exchanged in the recognition method. To this end we have used 72 sensors belonging to 10 different modalities distributed on the users' body, on selected objects and in the environment. The sensors were selected to be both complementary and redundant.

2.1.1 On-body Sensors

Sensors attached to body parts capture body postures, modes of locomotion, object interaction and environmental events. Magnetic field, acceleration and gyroscope (MARG) sensor combinations integrated in the so called motion jacket (see [8]) are attached to the subjects upper and lower arms and the back. They give a good estimation of the arm and torso posture. Gestures and object interactions are captured by specific arm positions and movements.

A magnetic coupling based sensor systems ([9]) estimates relative positions between the field transmitter and the receiver. Attached to the scapula the transmitter emits an oscillating magnetic field and the receiver attached to the wrist of the dominant arm captures the arm posture relatively to the scapula. The relative position information determines the arm position in a different way compared to the MARG systems.

Wireless microphones attached to the wrist and collar of the subject record environmental sound. Interactions for example with the coffee machine produce specific sounds recorded by the two on body microphones.

Additional acceleration and gyroscope sensor systems (Sun Spots and Inertiacube3) log modes of locomotion as they are attached to the shoes. Acceleration and gyroscope sensors are attached to the upper and lower legs, upper and lower arms of the subject to simulate sensor displacements. In addition an ECG sensor was also used during most of the recordings.

2.1.2 Object Sensors

Interaction with objects is known to be an important piece of information for activity recognition. We therefore attache acceleration and gyroscope sensors (see [10]) to a set of objects most relevant to the investigated actions. Specifically sensors were attached to (1) the breakfast knife, (2) the steak knife, (3) the spread cheese box, (4) the milk container, (5) the coffee mug, (6) the water glass, (7) the water bottle, (8) the sugar glass, and (9) the spoon. Two power sensors([11]) measure the current power consumption of the attached device. Note that since we were looking at a single user at a time scenario, motion signals from an object are an unambiguous indication of the user interacting

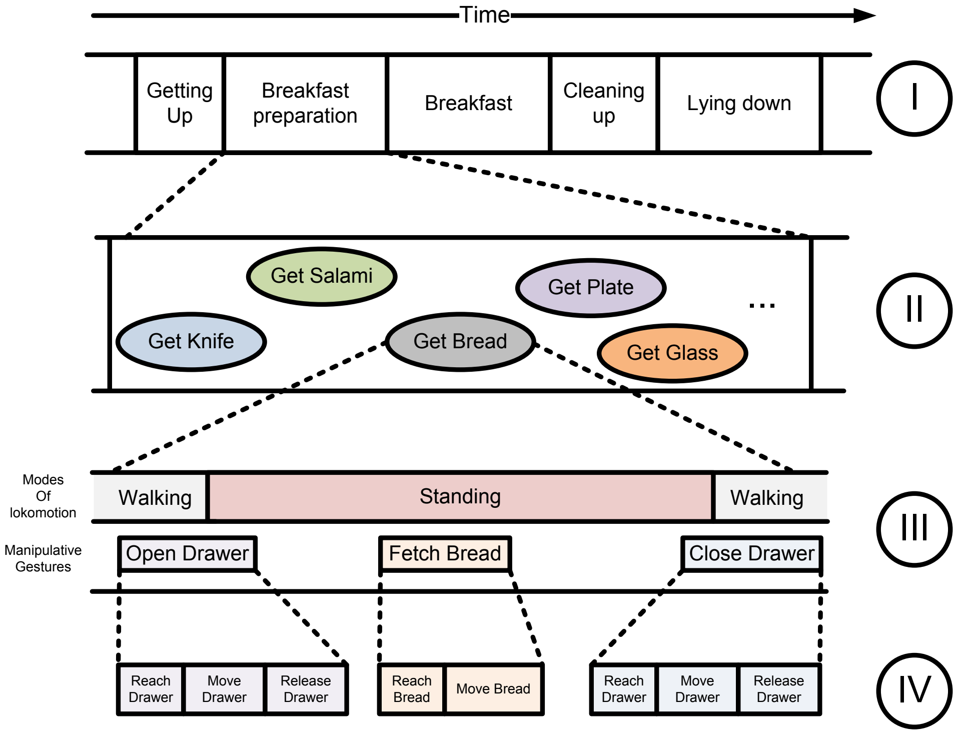

Figure 2: Temporal decomposition of activites. Level I is the highest activity level available in the setup. Level II zooms in into one high level activity, in this level the activities are not temporal ordered and depend on the execution sequence of the test subject. Logical, physiological and spatial limitations distinguish the order of activities in Level III. Here the activities are modes of locomotion and manipulative gestures. Level IV encapsulates the atomic gestures forming the manipulative gestures of level III.

with the corresponding object.

2.1.3 Environment Sensors

Sensors have also been integrated into the environment. First, we equipped the room with the Ubisense ultra wide band based location system. Two wide angle webcams made sure that all relevant actions are visible in video (attached to the ceiling and a side wall of the room). They can be used as additional means of localization (see e.g. [12]), for later labeling, or as vision based activity sensor. In addition there were and four microphones. The audio signals also allowa degree of localization plus the ability to recognize sound related actions (e.g. coffee machine).

Reed sensors and acceleration sensors attached to kitchen furniture log interactions of the person with the fridge, the dishwasher, three drawers and two doors. Vibrations caused by the person and bread slicer are captured by an acceleration system attached to the table and the chair.

We put 3 force resistive sensors on the table. The water glass, the coffee cup and the plate are put on top of the sensors. These sensors give information about whether there are objects on it and about the pressure (force) applied to the sensor. This force information can for example be used to roughly estimate the liquid level in the cup This force information can for example be used to roughly estimate the liquid level in the cups.

2.2 Experimental Protocol

Before we started the experiments we prepared the room and instrumented it with the sensors. The Ubisense ultra wide band localization system has been calibrated to see whether there ware interferences in the environment degrading the localization accuracy. Thus we measured 15 positions with 6 tags (exact coordinates measured with a laser meter with sub cm accuracy) at different heights. The accuracy of the localization system was found to be within specification (20cm to 30 cm).

Overall seven computers were used to capture different sets of sensor modalities. The computers ware ntp time synchronized to a local time server. Since for many sensors the accuracy of NTP synchronisation is not sufficient synchronization gestures were used in addition (clapping, and foot stamping)

During the runs there were several persons involved in labeling the activities. Each person labeled the synchronization gesture. One was responsible for labeling modes of locomotion (*standing*, *walking*, *sitting*), three other labeled different level of activities (high level activities like *Preparing breakfast*, mid level activities like *Slicing Bread* or low level activities like *Moving Bread*). The labels are currently being adjusted to more exactly fit the timing of the actions and remove false labels using the video feed.

Before the first run of the experiment the instructor explained the tasks and the sequence of the activities to the subject. The subjects were given high level instructions only (e.g. get up, walk around checking doors and looking into drawers, get yourself a coffee, make yourself a sandwich, clean up). A typical run took 15 to 25 minutes. We recorded 5 runs for each of the 12 subjects.

3 Data Examples

Due to the enormous amount of data fully describing the signals obtained during the experiment is beyond the scope of the paper. An example video showing the activities, annotations and some signals can be retrieved from "http://www.opportunity-project.eu". In this section we give a short discussion of two simple activities.

3.1 Sipping from the coffee cup

Sipping from the coffee cup has certain distinct properties: The person usually stands or sits (modes of locomotion), holds the cup in the hand and moves the hand near the mouth. After drinking, the cup is put back to the table. Thus, key modalities are sensor combinations which give information about body / arm posture, modes of locomotion and object interaction:

On body sensors :Several MARG units attached on the arm provide information about arm posture. Acceleration and gyroscope sensors measure the acceleration and rotation values plus the gravity ratios at different sensor axis. Relative position (distance, angle) information between chest and hand wrist in addition to wrist orientation is derived from the oscillating magnetic field system. MARG Sensors on the shoes capture the current mode of locomotion, together with upper body acceleration sensors and acceleration sensors attached to the knee.

Environmental sensors : Video and audio based localization determine the position of the person. Video stream can also be used to spot motions and the coffee cup. With 30cm accuracy the ultra wide band based localization system can also be used to distinguish between standing walking and sitting (since we attached the tags at the shoulders. One MARG unit attached to the chair detects interaction with the chair, another one attached to the table detects vibrations (e.g. from putting down the cup). Force resistive sensors measure when the person takes the cup and when she puts it back on the table giving a rough time interval when to spot gestures on on body sensor signals.

Object embedded sensors : An acceleration and gyroscope combination attached to the coffee cup captures movements and orientation changes while the person is interacting with the cup.

In figure 3 on body signals of the MARG system and the magnetic sensor are depicted. Both sensor modalities provide information about cleaning gesture (blue area) and about drinking (2 orange areas).

3.2 Taking Milk out of the fridge

An example where an environmental and object sensors make a contribution to the classification process is the high level event of taking a bottle of milk out of the fridge.

On body sensors : MARG units, acceleration and gyroscopes attached to the arm and to the upper part of the body capture body posture and gestures. The oscillating magnetic field sensor gives different (but also useful)posture information about gestures. The microphone attached to the lower arm can detect when the door is opened as this opening sound is very specific. Standing (mode of locomotion) is detected by Acceleration / gyroscope sensors and by the MARG units.

Environmental sensors : Video and ultra wide band localization capture proximity information (person is next to the fridge). Again video can be used to extract activity details. Reed switches attached to the fridge door capture opening and closing events giving a rough time frame. Acceleration and gyroscope information from sensors attached to the fridge door help to recognize this opening and closing events.

Object Embedded Sensors : Opening and closing of the fridge door is also captured by acceleration and gyroscope sensors at the water bottle and milk box.

Figure 3 depicts the signals which have been recorded by on body gyroscope sensors (upper plot), acceleration and gyroscops attached to the milk box when the box is taken out of the fridge. The elipses highlight the different gestures / activitis:

(I) The fridge is opened. The milk box is rotated as it is in a drawer in the fridge door. The on body sensor signals show a clear rotation signal.

(II) The fridge is closed. As the milk box is in the subject's hand, this activity is only captured by the on body sensor system.

(III) The person then carries the milk box from the fridge to the table.

(IV) The box is put on the table, a peak due to the impact of the box on the table is clearly captured by the gyroscope and the acceleration sensor attached to the box.

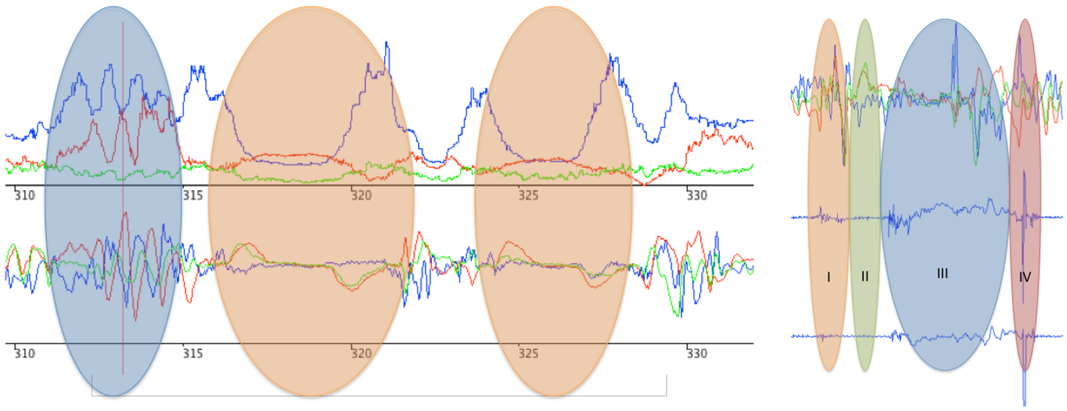

Figure 3: Left: The upper plot of this picture are data streams from the magnetic relative positioning sensor system (axes ratios), the lower plot depicts gyroscope information. A typical cleaning gesture is highlighted in the blue elipse, the orange elipses show drinking gestures. Right: Gyroscope information of the sensor attached to the right wrist is presented in the upper plot. The middle and the lower plot are linked to gyroscope and acceleration sensors being attached to the milk box. The first elipse I is the opening gestures, II highlights the the signals when the fridge is closed. III presents the movement of the box to the table and IV highlights the signal when the box is put on the table.

It can be seen that the presented examples show that actions performed during the recording are captured by different sensor modalities and there are always at least two systems contributing to the classification process.

References

[1] Intille, S., Larson, K., Tapia, E., Beaudin, J., Kaushik, P., Nawyn, J., Rockinson, R.: Using a live-in laboratory for ubiquitous computing research. (2006) 349–365

[2] Tenorth, M., Bandouch, J., Beetz, M.: The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In: IEEE Int. Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS). In conjunction with ICCV2009. (2009)

[3] van Kasteren, T., Noulas, A., Englebienne, G., Kröse, B.: Accurate activity recognition in a home setting. In: UbiComp '08: Proceedings of the 10th international conference on Ubiquitous computing, New York, NY, USA, ACM (2008) 1–9

[4] Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., Rehg, J.: A scalable approach to activity recognition based on object use. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. (Oct. 2007) 1–8

[5] Wilson, D.H., Atkeson, C.: Simultaneous tracking and activity recognition (star) using many anonymous, binary sensors. In: Pervasive Computing. (2005) 62–79

[6] Kranz, M., Schmidt, A., Maldonado, A., Rusu, R.B., Beetz, M., Hoernler, B., Rigoll, G.: Context-aware kitchen utilities. In: TEI '07: Proceedings of the 1st international conference on Tangible and embedded interaction, New York, NY, USA, ACM Press (2007) 213–214

[7] Tenorth, M., Bandouch, J., Beetz, M.: The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In: IEEE Int. Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS). In conjunction with ICCV2009. (2009)

[8] Stiefmeier, T., Roggen, D., Ogris, G., Lukowicz, P., Tröster, G.: Wearable activity tracking in car manufacturing. IEEE Pervasive Computing **7**(2) (2008) 42–50

[9] Pirkl, G., Stockinger, K., Kunze, K., Lukowicz, P.: Adapting magnetic resonant coupling based relative positioning technology for wearable activitiy recogniton. Twelfth IEEE International Symposium on Wearable Computers (ISWC 2008) (2007) 47 – 54

[10] Bächlin, M., Roggen, D., Tröster, G.: Context-aware platform for long-term life style management and medical signal analysis. In: Proceedings of the 2nd SENSATION International Conference. (0 2007)

[11] Bauer, G., Stockinger, K., Lukowicz, P.: Recognizing the use-mode of kitchen appliances from their current consumption. In: Smart Sensing and Context, EUROSSC09. (2009) 163–176

[12] Bauer, G., Lukowicz, P.: Developing a sub room level indoor location system for wide scale deployment in assisted living systems. In: Proc. 11th Int. Conf. on Computers Helping People with Special Needs, Springer LNCS (2008)