

A multimodal data collection of daily activities in a real instrumented apartment

A. Cappelletti, B. Lepri, N. Mana, F. Pianesi, and M. Zancanaro

FBK-irst

via Sommarive, 18 – Povo (Trento), Italy

E-mail: {cappelle,lepri,mana,pianesi,zancana}@fbk.eu

Abstract

This paper presents technical setup and methodology used for the data collection in progress within the NETCARITY project, as of middle of February, 2008. The goal of this work is to collect a large amount of high quality acoustic and visual data, concerning people doing common activities of daily living. The final expected structured and annotated database of activities will be helpful to develop systems that, starting from audio-visual cues, automatically analyze the daily behaviour of humans and recognize different kinds of daily living activities (distinguished in single activities, parallel activities, and single activities with some background noises).

1. Introduction

European society is strongly ageing. In 2005 people aged over 65 was 13% of the population [Czaja and Hiltz, 2005] and such figure is expected to increase. It has been estimated that by 2020 one out of four Europeans will be over 60 years old, and one out of five over 65 [Mikkonen et al, 2002]. Consequently there will be more and more aged people, in need of social, home and long-term care services.

Technology can play a crucial role in enhancing in the elderly people (and in their families and associated caring personnel) the feeling of confidence required for ageing-in-place, by assuring the basic support of everyday activities and health critical situations management.

On this way it is located the NETCARITY project¹, aiming to propose a new integrated paradigm to support independence and engagement in elderly people living alone at home. One of the final objectives of this project is the development of a light technological infrastructure to be integrated in the homes of old people at reduced costs. Such technologies should allow both assurance of basic support for everyday activities and detection of critical health situations.

As the project is targeting real everyday needs in real life contexts, one of its first steps is to carry out a data collection in order to obtain significant examples of activity of daily living (ADL) to be studied and modelled for the project purposes.

ADLs monitoring has become an important goal and a valued technological objective mainly for three reasons. Firstly, ADLs monitoring is important for healthcare [Katz, 1983]. Trained caregivers, clinicians and therapists spend much time measuring and tracking ADLs accomplishment in order to assess the functional status of a person and her/his degree of autonomy, detect problems

in performing those activities, and plan care interventions. However, current methods for recognizing and monitoring these activities consist of time and resource consuming manual tasks, relying on paid observers (e.g. a nurse, monitoring periodically an elderly patient) or on self-reporting (e.g. patients having to complete an activity report everyday). Automated aids that can improve the caregiver work practice would be of value.

Secondly, ADLs are common activities that people (not only old one) perform daily. Therefore, these activities become interesting also outside the elder-care field. In fact, a large variety of tasks, such as security monitoring or training, which are currently considered expensive human jobs, become amenable to automated support if the computer can recognize human activities.

Finally, home ADLs recognition is a highly interesting and challenging scientific task, aiming to the number of activities people get involved in, and the different ways they can be performed.

For all these reasons, jointly to the project objectives, we are currently working on a multimodal data collection, aiming to have a structured and annotated database of high-quality, realistic and synchronized acoustic and visual data of common daily activities performed in a home setting. This dataset may be very worth for researchers working on automatic activity recognition. Especially for people who use machine learning techniques and need large data corpora for training multimodal activity recognition algorithms.

At present there are already existing data collections coming from a number of smart homes in Europe and in USA, used to collect data on common daily activities. Among these:

- AwareHome project of the GeorgiaTech in Atlanta. In this project, the Georgia Tech Broadband Institute's Residential Laboratory is used as living laboratory for ubiquitous computing in home life [Kidd et al., 1999];
- Philips' HomeLab in Eindhoven, used as showroom and usability laboratory. Subjects live in the lab for several days while researchers may observe and study people in a naturalistic home environment in order to develop better products [Aarts and Eggen, 2002];

¹ NETCARITY (A Networked multisensor system for elderly people: health care, safety and security in home environment") is an Integrated Project, supported by the European Community under the Sixth Framework Programme (Information Society Technologies, Ambient Assisted Living, IST-2006-045508). For more details see <http://www.netcarity.org/>

- PlaceLab residential facility, maintained by the House_n research group at the MIT Department of Architecture. It is equipped with hundreds of sensing components and it is used as a multi-disciplinary observational facility for the scientific study of people and their interaction patterns with new technologies and home environments [Intille et al, 2006];

However, Philips' HomeLab and AwareHome project were not collecting multimodal data, useful for devising and testing activity recognition systems. On the contrary, the House_n research group was doing it, but it was not including audio and visual features.

There are other multimodal (audio-visual) corpora recently collected, not on common daily activities but on meetings, with people sat around a table. Among these, the MM4 corpus [McCowan et al., 2004] and the VACE corpus [Chen et al., 2005] include low-level cues of human behavior, such as speech, gesture, posture, and gaze, in order to interpret high level meeting events. Similar purposes have also been pursued by the AMI project, collecting a large multimodal corpus [Rienks et al., 2006].

We also collected two multimodal corpora on meeting scenarios, namely the "Mission Survival Corpus 1" [Pianesi et al., 2006], and the "Mission Survival Corpus 2" [Mana et al., 2007].

The paper is organized as follows: Section 2 presents the technical set-up used on the audio and video acquisition sides. Section 3 describes the architecture of the data acquisition system. Section 4 presents the recording procedure, while in Section 5 the expected final result is illustrated. Finally, Section 6 summarizes the present work, formulates some considerations and draws some future steps.

2. Technical Setup

To allow gathering of multimodal data in a real context we have instrumented two rooms of an apartment (specifically a living-room and a kitchen) with audio and video sensors (see Figure 1).

A third room is used to store computers and capture boards. All computers and webcam are connected via an Ethernet LAN. In addition, a wireless network let communication between control machine and PDA.

2.1 Audio sensors

Three groups of T-shape microphone arrays are installed into each room for a total of 24 audio sensors. An array (see dagger sign on Figure 2) is composed by 4 omni-directional microphones, mounted at 2 meters tall on wall. Microphones are connected to A/D converter that samples audio input with a frequency of 48 kHz and a resolution of 16bits. Converters are connected through optical cable to a 24 channels acquisition board, installed on the capturing workstation that provides also an internal synchronization clock to assure alignment between channels.



Figure 1: Camera view in the living room and kitchen

2.2 Video sensors

The two apartment rooms are covered by a total of three webcams (see oval sign on Figure 2) with Pan-Tilt-Zoom (PTZ) functionalities². They are mounted on the ceiling and offer a large (40° - 150°) field of view due to lens capability. Each camera has an IP address and dispatches "Motion-Jpeg" images over Ethernet Network at a variable frequency from 10 to 20 frames per second, depending on light conditions. To provide power supplying we use a Power Over Ethernet (POE) switch.

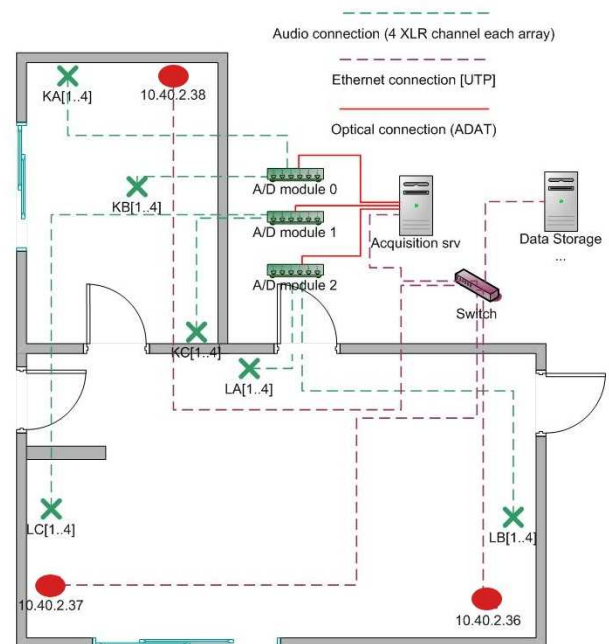


Figure 2: Sensors location in the apartment

² These features are not used at the moment.

Image resolution is 640x480 pixels, while jpeg compression level is set to 100 to reduce artifacts on captured images.

3. Data Acquisition System

The architecture of an acquisition system has a significant impact on the quality of collected data, as well as the data nature plays a crucial role into the architectural choice. Data we are collecting are audio and video image streams that are time-based information.

One of the main problems when multimodal data are collected is to guaranty time alignment between streams. That could not be trivial in real environment, especially when capturing systems are distributed, or timers are not precise. Some standards, such as MPEG7 [Martinez 2004], allow to handle multimedia data but we decided to develop our specific protocol to make infrastructure light and easy adaptable to requirements but absolute time remains the core indexer for all the data.

The procedure we are using follows the approach to acquire all streams synchronized referring to a unique time clock with enough resolution. For this reason we developed an ad hoc software application composed by libraries that access to each acquisition hardware. An operator, located in the third room, can initialize and manage all experiment through a GUI by controlling acquisition of streams, sending instructions to user, and making annotations. UI runs on the main machine, which uses a high resolution clock as reference timer³. Saving process is a I/O bound and can freeze capturing threads that lead to lack of data. All collected data are dispatched on distributed application over a local network. However, given network resources are limited, it is important to provide enough bandwidth to avoid saturation or data

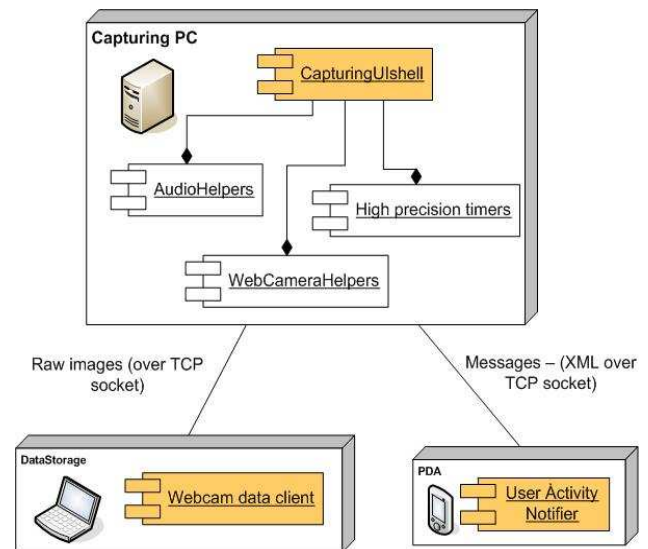


Figure 3: Deployment of acquisition system

loss.

Figure 3 shows the deployment of infrastructure. To handle experiments the operator uses a *CapturingUIShell*, (see Figure 4) running on the capturing PC.

From this shell the operator starts and ends experiments (see 1 on the figure). A counter indicates the experiment duration (2). All activities that a subject will perform during a session are listed in a randomized order and then numbered (3). Each one is marked with a different colour (4) according to its specific category (orange for the “single activities”, light blue for the “noised activities” and green for the “parallel” ones – see Section 5.1).

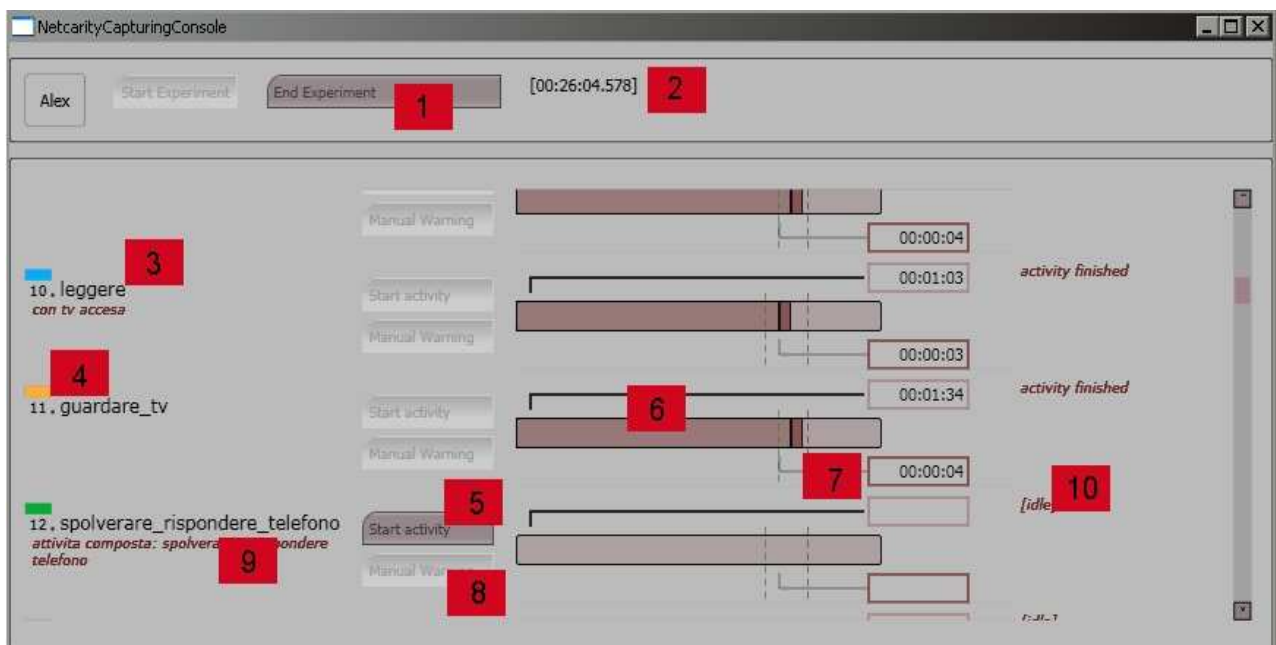


Figure 4: Capturing UI shell

³ See <http://msdn.microsoft.com/msdnmag/issues/04/03/HighResolutionTimer/>.

The operator announces to the subject the next activity to be performed by pressing the “Start activity” button (5). A progress bar keeps track of the effective time of the activity (6) and its closure time (7), i.e. the time that the subject takes to end the action after having received the automatic audio warning message in correspondence of the expected duration (see Table 1 and Table 2). If the subject exceeds the upper bound of closure time, the operator can manually warn the closure again (8). Also other properties as activity description (9) and status (running, finished or idle - 10) are shown on the interface. GUI links *AudioHelpers* and *WebCameraHelpers* libraries that are responsible of stream acquisition. They fetch data from dedicated hardware (audio-boards) or directly from sensor (webcams). *HighPrecisionTimer* provides a unique time service that all the system adopts to mark events. Given video from webcams does not have a fixed data rate, we annotate the timestamp of each frame. In this way we know when a specific information has been captured. The same idea is applied also for labels and event markers: when something happens, we register both information of time and content as a single item.

On the contrary, given audio data has a predefined sample frequency, we need to annotate just begin/end times.

All streams are then asynchronously saved on a different machine. Raw images are packed with their timestamp and sent via TCP connection to a remote *DataStorage* node, where clients are responsible to save incoming stream on disks. Later all distributed data are merged in a single repository by the operator.

Subjects are equipped with a PDA that runs an application, connected to the *CapturingUIShell*, which instructs them on the sequence of activities to do. The operator sends activity instructions to the mobile application. When subject reacts, a response is returned back to manager that records time, activity label, status of user and other surrounding information.

The whole infrastructure runs on Windows XP and it is built by NET framework. This offers capability to handle different acquisition devices, to allow interoperability with low level drivers and future sensor upgrading.

4. Recording Procedure

4.1 Instructions, task description and preparatory steps

Subjects involved in the data collection are mostly people of the administrative staff, but we are going to involve also old people, already collaborating within the project for the user study.

Subjects are firstly instructed on what to do during the session. They are informed that they are audio and video recorded while doing a sequence of common daily home activities, randomly generated from a limited set of activities, repeated more times (see Section 5.1). In doing these activities they are free to move between kitchen and living-room and to choose where and how to perform the task. Given the same activities are asked more times, subjects are invited to possibly perform them each time in

a different way (e.g. in case of “reading” activity: once reading a book on the couch and a second time sat at a table or, alternatively, reading a book, rather than a magazine or a newspaper, etc).

Subjects are also informed that during some activities a second person (called “actor”) enters in the scene, doing something else (e.g. the actor is answering to a phone call while the subject is watching TV). Subject and actor do not interact. The actor role is to be a sort of “noise generator” (see third action category in Section 5.1).

After having received instructions, subjects are asked to sign a consent form to make the collected audio and video data usable for research purposes.

Before starting a session recording, a visual representation (or model) of each session participant (subject and actor), as well as the recording ambient without any person (background), have to be acquired. These models will be supplied to the person tracker (Lanz, 2006) that will be used later in order to detect and track subject motions during their activities.

4.2 Recording procedure

As seen in Section 3, a software application is used in order to guide subjects in executing a sequence of activities and manage all incoming data: through a GUI the experimenter sends predefined warning audio messages to users via wireless network during experiments. Subjects have a PDA, where the client is running, and wear standard headphones to avoid vocal commands could be captured by microphones. For each activity subjects mark beginning and end of that activity on client in according with received audio warnings. In particular, each activity has a pre-defined duration (60 or 90 minutes). When subjects receive the message announcing the next activity to be done and they are ready to start it, they touch the PDA screen and in correspondence the beginning of the activity is marked. After a certain time (corresponding to the pre-defined duration), subjects receive via headphones a warning message inviting them to finish that activity. Only when they have ended the activity, subjects touch again the PDA screen (and in this way the activity end is marked). Activities' metadata and timestamps are then annotated on server.

5. Final Result: A Structured and Semi-automatic Annotated Database

5.1 Collected Activities

We are collecting some daily activities that people usually perform when at home. We have grouped these activities in three main categories: (1) single activities; (2) parallel activities (i.e. performed concurrently); and (3) single activities performed with some background noises.

The first category includes six basic activities. In particular: (a) phone answering; (b) cleaning-dusting; (c) TV watching; (d) ironing; (e) reading; (f) eating-drinking. A detailed description of these basic activities is depicted in Table 1.

Activity	Description	Location	Subject position	Activity duration (sec)
eating-drinking	eat a snack (chips/biscuits/fruit/ yogurt) and/or drink some water (taking the water from a bottle or a carafe)	kitchen or living room	stood or sat on a chair/armchair /couch	90
reading	read a book/newspaper/magazine	kitchen or living room	stood or sat on a chair/armchair /couch	60
ironing	iron a handkerchief or a napkin on a table or an ironing board	living room	stood	90
TV watching	do zapping or watch a TV program	living room	stood or sat on chair/armchair /couch	90
cleaning-dusting	clean or dust, by using a dust mop and a squirt gun or a feather duster	kitchen or living room	stood	60
phone answering	answer to a call on the mobile phone	kitchen or living room	stood or sat on chair/armchair /couch	90

Table 1: List of single activities

Activity durations are fixed (60 or 90 minutes). As seen in Section 3 and Section 4.2, after that time subjects receive an audio warning message inviting them to end the activity in progress. Given that subjects usually do not suddenly interrupt the activity, the actual duration of the recorded activity is longer than the fixed one (as evident from square 7 in Figure 4). Furthermore, still from Table 1, it is evident that subjects are free to choose where to perform these activities (kitchen and living room) and how (stood, sat, walking around, etc...).

Our choice of these daily activities is mainly motivated by the following reasons: first of all, they are common activities that all people (also elderly one) make during their daily living at home. Secondly, the audio-visual cues extractable from the recorded data are significant features for the recognition of these activities (e.g. on the acoustic side, the sound of the phone ringing is crucial for recognizing the “answering to a phone call” activity; at the same way, on the visual side, the head orientation for the recognition of TV watching activity). However, detection and recognition of some activities may be also quite challenging because some activities are very similar from the viewpoint of the audio-visual features (e.g. how to distinguish “eating a snack” and “watching TV” activities, when in both cases subject is sat on the couch?). Finally, the selected activities allow to use both the available rooms in the apartment (kitchen and living room) and to make data more variable.

The second activity category includes three parallel activities performed concurrently by subjects. In particular, these activities are: a) cleaning-dusting and phone answering; b) ironing and TV watching; c) eating-drinking and TV watching.

The choice of focusing our attention also on “parallel activities” has been guided by the consideration that people often perform activities concurrently in their daily living. However, there are few works in activity recognition field devoted to model and recognize the co-temporal relationships among multiple activities performed by the same subject [Wu et al., 2007]. Therefore, from this point of view, this kind of collected data could be helpful.

Finally, the third activity category includes three different

activities performed by subjects with some background noises generated by a second person (as seen in Section 4.2). Specifically: a) reading with a TV watching activity as background noise; b) eating-drinking with a TV watching activity as background noise; and c) TV watching with a phone call as background noise.

This last set of activities may be very useful to test the robustness of multimodal activity recognition systems. A robust activity recognition system should be able to distinguish parallel activities (e.g. a subject is eating while he/she is watching TV) and single activities performed while there are some background noises in the apartment (e.g. a subject is eating while another subject is watching TV).

activity	description	location	subject position	activity duration (sec)
reading (bkg_noise=phone call)	basic activity + phone calling in background (subject is ignoring the call; another person is answering)	kitchen or living room	sat around a table or on a armchair	60
eating-drinking (bkg_noise=TV watching)	basic activity + TV noise in background (note: TV is ignored by subject)	kitchen or living room	sat around a table or on a armchair	90
TV watching (bkg_noise=phone call)	basic activity + phone calling in background (subject is ignoring the call; another person is answering)	living room	sat on chair/armchair	90
cleaning & phone answering	clean or dust and in the same time answer to a phone call	kitchen or living room	stood	60
ironing & TV watching	watch TV while ironing	living room	stood	90
eating & TV watching	watch TV while eating/drinking	living room	stood or sat on chair/armchair	90

Table 2: List of parallel activities and single activities with background noises

A detailed description of the three parallel activities and the three single activities with background noises is depicted in the Table 2.

In addition, we are going to collect also one hundred examples of selected acoustic events (e.g. entry phone ringing, door knocking, cooking alarm) and about fifty examples of complex activities as tea making and coffee making. These activities are performed by the subject following a fixed script: (a) the subject enters in the kitchen; (b) he/she puts some water in the teapot/Italian coffee pot; (c) he/she reads a newspapers or a magazine on the kitchen table while he/she is waiting for the teapot/Italian coffee pot whistle; (d) then he/she turns off the hotplate and puts the water/coffee in a cup; (e) finally, the subject gets out from the kitchen bringing the cup.

These instances of tea/coffee making may be useful as data-set for training and testing learning algorithms able to recognize subjects’ intentions and plans [Pollack et al, 2003].

5.2 Expected Outcome

At the end of the data collection we will have a multimodal structured database, having synchronized audio and video streams. As summarized in Table 3, this database will encompass activities performed by 50 subjects. For each subject and activity the database will have four instances/examples.

In short, the database will be set up by about an hour of recorded data for each subject. The total audio-video recordings will be longer than fifty hours.

Subjects	50
Examples per subject and activity	4
Recorded data per subject	~ 1 h – 1h 20'
Total estimated recordings	> 50 h

Table 3: Expected collected data

6. Conclusion and Future Work

In this paper we presented technical setup and methodology of the NETCARITY data collection.

The goal of this work is to collect a large amount of high quality data, concerning people doing common activities of daily living. As the data collection is in progress, we cannot provide any detailed descriptions of its content but we can formulate some preliminary considerations on technical issues.

To satisfy evolutions of multimodal feature extraction systems, rate and dimension of audio/video information must be at maximum possibilities nowadays hardware can provide. Given this constraint, the first consideration is that, as made evident by the experience we are doing, collecting such raw data requires a lot of resources.

Secondly, the capturing process is I/O bound: that means the bottlenecks are network infrastructure and access to data storage (all open/write/close operations). In particular, we have tested that such acquisition architecture produces 3MB/sec for audio channels and 20.63MB/sec for video streams. The nature of video image structure is enough to drastically reduce capability of I/O Bus. Saving process must store 60 (20 frames x 3 cameras) relative small images (170kByte) per second; this means that operative system must perform 60 “open/write/close” calls each seconds. Such operations are strongly time consuming and can lead quite fast to a lack of data or a freeze of the system. To avoid this bottleneck we are using a client workstation with SATA disks and a 1Gbit network connection.

On the other side, having been able to synchronized audio and video streams will let us to save time in doing any post-processing (otherwise necessary in order to cut and synchronize collected audio and video files). It requires time information must be trusted, in a sufficiently precise and fast way to retrieve.

Finally, the architecture of the acquisition system, where subjects mark directly start and end times of the activities, jointly to structured files (XML) including all information about the order of the performed activities and the corresponding times, let us to have a semi-automatic annotated database where we know which activity is carried on, when it starts and ends, and consequently which are the corresponding audio and video cues.

At the end of data collection, this annotated and

structured database will consist of audio and video recordings of 12 activities, repeated 4 times during each recording session, for each subject (50 subjects), for a total length of more than 50 hours.

The next step will be to extract audio and visual cues from the recorded data. Finally, we plan to devise learning algorithms, based on audio-visual features, able to classify single and parallel activities performed by a subject in home setting.

More in general, this database may be helpful for whoever want to develop systems that, starting from audio-visual cues, automatically analyze the daily behavior of the subjects and recognize different kinds of daily living activities (single activities, parallel activities, single activities with some background noises).

7. Acknowledgements

The data collection described in this paper is supported by the European Union within the NETCARITY Project, under contract number IST2005-045508. The authors would like to thank all the subjects that participated in the experiments, as well as all colleagues collaborating in carrying on the data collection.

8. References

- Aarts, E.H.L., and Eggen, B. (eds.) (2002) *Ambient Intelligence in HomeLab*. Eindhoven: Neroc.
- Chen, L., Rose, R.T., Parrill, F., Han, X., Tu, J., Huang, Z., Harper, M., Quek, F., McNeill, D., Tuttle, R., Huang, T (2005): VACE multimodal meeting corpus. Proc. of Multimodal Interaction and Related Machine Learning Algorithms.
- Czaja, S.J and Hiltz, S. R (2005).: Digital aids for an aging society. In *Communications of the ACM*, 48(10).
- Katz, S. (1983) *Assessing Self-Maintenance: Activities of Daily Living, Mobility, and Instrumental Activities of Daily Living*. *Journal of American Geriatrics Society*. vol. 31, no 12, pp.712-726.
- Kidd, C.D., Orr, R., Abowd, G.D., Atkeson, C.G, Essa, I.A., MacIntyre, B., Mynatt, E., Starner, T.E., Newstetter, W. (1999) *The Aware Home: A Living Laboratory for Ubiquitous Computing Research*. In the Proceedings of the Second International Workshop on Cooperative Buildings - CoBuild'99. Position paper.
- Intille, S.S., Larson, K., Munguia Tapia, E., Beaudin, J, Kaushik, P., Nawyn, J., and Rockinson, R. (2006) *Using a live-in laboratory for ubiquitous computing research*. In K.P. Fishkin, B. Schiele, P. Nixon, and A. Quigley (eds.) *Proceedings of PERSASIVE 2006*, vol. LNCS 3968,. Berlin Heidelberg: Springer-Verlag, pp. 349-365.
- Lanz, O.: *Approximate Bayesian Multibody Tracking*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, September 2006 (Vol. 28, No. 9), pp. 1436-1449.

- Mana, N., Lepri, B., Chippendale, P., Cappelletti, A., Pianesi, F., Svaizer, P., and Zancanaro, M. (2007) Multimodal Corpus of Multi-Party Meetings for Automatic Social Behavior Analysis and Personality Traits Detection. In Proceeding of Workshop on Tagging, Mining and Retrieval of Human-Related Activity Information, at ICMI07, International Conference on Multimodal Interfaces, Nagoya, Japan.
- Martínez, J M: MPEG-7 Overview (version 10) (2004). Coding of moving picture and audio. International Organization for standardization (ISO/IEC JTC1/SC29/WG11 N6828), Palma de Mallorca.
- McCowan, D, Gatica-Perez, S. , Bengio, Y., Moore, D., and Boulard, H. (2004): Towards Computer Understanding of Human Interactions. In: Ambient Intelligence, E. Aarts, R. Collier, E. van Loenen & B. de Ruyter (eds.), Lecture Notes in Computer Science, Springer-Verlag Heidelberg, pp. 235-251.
- Mikkonen M., Väyrynen S., Ikonen V., Heikkilä M.O. (2002): User and Concept Studies in Developing Mobile Communication Services for the Elderly. in Personal and Ubiquitous Computing, 6(2):113-124..
- Pianesi, F., Zancanaro, M., Lepri, B., Cappelletti, A. (in press): Multimodal Annotated Corpora of Consensus Decision Making Meetings. To appear in The Journal of Language Resources and Evaluation.
- Pollack, M.E., Brown, L., Colbry, D., McCarthy, C.E., Orosz, C., Peintner, B., Ramakrishnan, S., and Tsamardinos, I. (2003) Autominder: An Intelligent Cognitive Orthotic System for People with Memory Impairment. *Robotics and Autonomous Systems*, 44, pp. 273-282.
- Rienks, R., Zhang, D., Gatica-Perez, D., Post, W. (2006): Detection and Application of Influence Rankings in Small Group Meetings. In Proceedings of ICMI'06. Banff, CA.
- Wu, H, Lian, C, and Hsu, J.Y. (2007). Joint Recognition of Multiple Concurrent Activities using Factorial Conditional Random Fields. In C. Geib and D. Pynadath (eds.) AAAI Workshop on Plan, Activity, and Intent Recognition. Technical Report WS-07-09. The AAAI Press, Menlo Park, California..