Netcarity Multimodal Data Collection

Alessandro Cappelletti
FBK-Irst
Via Sommarive 18
Povo-Trento 38050, Italy
cappelle@fbk.eu

Bruno Lepri
FBK-Irst
Via Sommarive 18
Povo-Trento 38050, Italy
lepri@fbk.eu

Nadia Mana
FBK-Irst
Via Sommarive 18
Povo-Trento 38050, Italy
mana@fbk.eu

Fabio Pianesi
FBK-Irst
Via Sommarive 18
Povo-Trento 38050, Italy
pianesi@fbk.eu

Massimo Zancanaro
FBK-Irst
Via Sommarive 18
Povo-Trento 38050, Italy
zancana@fbk.eu

CHI 2009, April 4 – April 9, 2009, Boston, MA, USA
ACM 978-1-60558-247-4/08/04.

Abstract

In this paper we describe the multimodal data collection, carried out within the NETCARITY project. We shortly present the acquisition system, the recording procedure and the collected data. Finally, we draw some considerations, while trying to point out positive characterizing aspects, as well as met problems and pitfalls.

Keywords

Multimodal data, ADL, home behavior, audio and video recording, semi-automatic annotation.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction

In the last years, Activities of Daily Living (ADLs) monitoring [3] has become an important goal in relation to effort such as the Ambient Assisted Living and the Assisted Cognition. To this end, a few corpora have been developed, mostly exploiting sensors such as RFIDs, position detectors, accelerometers, on-object motion sensors (e.g. MITes) and the likes [1]. Audio-video databases of ADLs, on the other hand, are substantially missing.

In this paper we introduce a data collection effort that aimed at providing a large audio-video database of ADLs.

Collection and management of audio-visual signals is more challenging than data collections based on other type of sensors because of the huge amount of data generated, the greater demand for bandwidth during collection and the need of high accuracy in time synchronization. These problems are all exacerbated by the necessity of providing for as realistic data as possible, hence giving up to lab-setting in favor of closer to reality ones.

The purpose of the Netcarity [2] data collection is to build a corpus that captures at least some of the complexity found in real environment while a) maintaining some amount of control on the activities generated, and b) providing for a compact corpus that covers a relevant number of variations in relatively short time. The data were collected in a real home setting, where such 'normal' disturbances such as the daily natural change of lighting conditions and the external noises typical of apartments and houses occur. The subjects could perform the requested actions in a rather unconstrained way. On the other hand, the subjects were instructed when to start and end an action, this way yielding an automatic segmentation of the data stream, with only a limited impact on naturalness. Part of the corpus consists of actions performed in parallel, either by the subject or through the intervention of a cognate, which allow to capture part of the complexity of the real world in a structured way.

Data acquisition system and recording procedure

The living-room and the kitchen of an apartment were instrumented with 2 web-cameras in the living-room and one in the kitchen, plus 3 T-shape microphone arrays per room, each consisting of 4 omni-directional microphones, for a total of 24 audio sensors inside the apartment (see Figure 1).

In order to control the acquisition process and to maintain time synchronization, a software application, running on Windows XP and built upon the .NET framework and consisting of libraries that access and control each acquisition hardware (audio and video sensor), was developed. The application uses and makes available a unique time clock to stamp and synchronize the data from the different sensors.

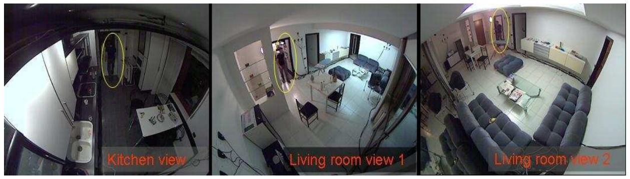

Figure 1: Views from the cameras in the apartment

A mobile application running on PDA was used to guide subjects in executing the activities by signaling them when it was time to close the current action (see below), the kind of activity they should engage in, etc. Moreover, the PDA application was used by the subjects themselves to mark the beginning and end of each activity, this way providing a reliable automatic segmentation and annotation of activities.

Besides the target subjects, a cognate was exploited to add controlled noise to some activities (see below). Subject and cognate did not interact.

The Dataset

We focused on daily activities that people perform when at home. They are grouped into three main categories: (1) basic activities: phone answering, cleaning-dusting, TV watching, ironing, reading, eating-drinking; (2) noisy activities, that is single activities as in (1) performed by the subject, X, in the presence of background noise due to the activity of the cognate, Y: X reading while Y is watching TV; X eating-drinking while Y is watching TV; and X is watching TV while Y is engaged in a phone call; and (3) parallel activities, i.e., two single activities performed simultaneously by the subject: cleaning-dusting + phone answering; ironing + TV watching; eating-drinking + TV watching.

The choice of these activities was motivated by the following reasons: a) the single activities are among the typical ones that people perform during their daily living at home; b) they are associated with audio visual cues that can be used for activity recognition (e.g. phone ringing is a cue for “answering to a phone call” and head orientation can be useful for “TV watching”); c) the perceptual similarities among some of our activities (e.g. “eating a snack” and “TV watching”, when both are performed while seating on the sofa) introduces ambiguities that are typical of ‘natural’ data; d) the selected activities can be naturally performed in both the rooms of the apartment (kitchen and living room) we have instrumented, this way providing for a higher data variability.

As to class 2 activities (“noised activities”), the presence of noise will be useful to test the robustness of multimodal activity recognition systems.

Finally, “parallel activities”, though very common in natural settings, introduce an additional level of complexity. Nonetheless, only few works have endeavored to model and recognize the co-temporal relationships among multiple activities performed by the same subject [4].

20 subjects (13 male and 7 female) participated in the study. Each subject performed the same activity 4 times; hence, the database contains 200 examples per each activity, for a total of more than one hour of recorded data per subject. The total audio-video recordings are about 23 hours and half (see Table 1).

Subjects were free to choose the location where to perform each activity (the kitchen or the living room) and aspects such as posture (standing, sitting), whether performing it while walking, etc. They had limitations on time, though: each activity was required not to exceed 60 or 90 seconds (see Table 1). The subjects received a message from the experimenters alerting them when it was about time for them to close the current activity, and what the next activity to perform was. The actual durations are reported in Table 1.

	Activities	Estimated Duration (sec)	Average Actual Duration (sec)	Collected Data
Basic	eating-drinking	90	96.663475	2h 8’ 53”
	reading	60	66.325825	1h 28’ 26”
	ironing	90	97.381838	2h 9’ 51”
	TV watching	90	96.558625	2h 8’ 45”
	cleaning/dusting	60	65.711175	1h 27’ 37”
	phone answering	90	101.1854	2h 14’ 55”
Noised	reading(background noise=TV)	60	66.42855	1h 28’ 34”
	eating-drinking(background noise=TV)	90	96.897675	2h 9’ 12”
	TV watching(background noise=PHONE)	90	97.825113	2h 10’ 26”
Parallel	cleaning&phoning	60	71.326988	1h 35’ 6”
	ironing&TV watching	90	99.801263	2h 13’ 4”
	eating&TV watching	90	97.865638	2h 10’ 29”
			Total =	**23h 25’ 18”**

Table 1: Collected data

Lessons Learnt and Considerations

In collecting audio-visual signals, a huge amount of data is generated. Consequently greater bandwidth and storage capability are needed.

The amount of data increases even more when high-quality (raw) data are targeted.

The quality of data collected in a real houses is obviously much worse that of lab data, due to: 1) variability of lighting conditions, which negatively impacts on frame rate and quality of the captured images; 2) noisier environment, because of typical external and internal noises in a house; 3) incomplete coverage of the room space by audio and video sensors.

Finally, having synchronized audio and video data is crucial to conduct research on multimodal fusion and to develop recognition systems using multimodal information. In our data collection the architecture of the acquisition system (where subjects mark start and end times of the activities) provides the advantage of a semi-automatic chunking of activities. This information is stored in structured XML files along with information concerning the activity type, and the expected and actual duration. Despite our many efforts, the strong video frame rate variability due to changing lighting conditions and unexpected problems in data buffering during I/O processes resulted in small (around 200-500 ms) though frequent, de-synchronizations of video data.

Acknowledgements

The data collection described in this paper has been supported by the EU NETCARITY Project.

References

[1] http://boxlab.wikispaces.com/Related+Work

[2] http://www.netcarity.org/

[3] Katz, S. Assessing Self-Maintenance: Activities of Daily Living, Mobility, and Instrumental Activities of Daily Living. Journal of American Geriatrics Society (1983), Vol 31, no 12, pp. 712-726

[4] Wu, H., Lian, C., and Hsu, J.Y. 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 (2007). The AAAI Press, Menlo Park, California.