Collecting Datasets from Ambient Intelligence Environments

Piero Zappi, University of California San Diego, USA
Clemens Lombriser, ETH Zürich, Switzerland
Luca Benini, University of Bologna, Italy
Gerhard Tröster, ETH Zürich, Switzerland

ABSTRACT

This paper describes a methodology and lessons learned from collecting datasets in Ambient Intelligence Environments. The authors present considerations on how to setup an experiment and discuss decisions taken at different planning steps, ranging from the selection of human activities over sensor choices to issues of the recording software. The experiment design and execution is illustrated through a dataset involving 150 recording sessions with 28 sensors worn on the subject body and embedded into tools and the environment. The paper also describes a number of unforeseen problems that affected the experiment and useful considerations that help other researchers recording their own ambient intelligence datasets.

Keywords: Ambient Intelligence, Data Collection, Data Generation, Data Mining, Data Sharing, Dataset, Wireless Sensor Networks

INTRODUCTION

Ambient Intelligence (AmI) describes a paradigm where smart, electronic environments are sensitive and responsive to the presence of people and their activities (Ramos et al., 2008). AmI systems use sensors invisibly embedded into objects of daily use and environments of everyday life. People moving in such settings engage many computational devices and systems simultaneously even if they are not aware of their presence (Jaimes & Sebe, 2007). The pervasive sensing enables context-aware computing by providing software that is able to adapt to aspects of the situations in which it operates (Li et al., 2009). A major challenge is the detection of human factors in context information, which can be grouped in three categories (Schmidt et al., 1999): information about the user (habits, emotional state, etc.), his/her social environment (social interaction, group dynamics, etc.), and his/her tasks (spontaneous activity, general goals, etc.).

DOI: 10.4018/jaci.2010040103
This work focuses on the recognition of human activities, which allows for activity-based computing (Davies et al., 2008), and is seen here as a sense and classify problem. Common approaches can be divided into those relying on video tracking systems (Mitra et al., 2007) and others relying on multimodal sensor networks that include body-worn sensors, smart sensors and sensors embedded in furniture (Ward et al., 2006; Jeong et al., 2008; Amft, 2007). In this paper, we focus on the second group. In such settings, data is collected from a large number of sensors and is fused using machine learning algorithms to recognize people activity. Typical algorithms include decision tree classifiers, Bayesian networks, linear discriminant classifiers, neural networks, Hidden Markov Models, voting techniques and many others (Duda et al., 2000). These algorithms belong to the class of supervised classifiers which learn by example. In a training phase, they require a large set of sample instances of all classes. From these instances, they optimize their model parameters to reflect the classes. The recognition performance can then be tested using a validation set, a second set of class instances, which have not been seen used in training. Complex classification algorithms or complicated signal patterns often require large datasets for reliable training and testing.

Establishing large datasets from AmI environments is costly and time-consuming, but crucial for the research community (Ponce et al., 2006). Besides deploying and maintaining sensors, data needs to be collected over weeks, months, or even years. All data needs to be annotated by humans, such that activities of interest are marked in the sensor streams. Aware of these limits, many researchers share the datasets they have obtained from their instrumented environments (Intille, 2009; BoxLab, 2009). The objective is to accelerate the creation of novel applications in the fields of human-computer interaction, healthcare and ubiquitous computing. But what are the specific problems encountered when collecting datasets? What should researchers be aware of when they plan experiments? We present the methodology and we have learned from recording various complex datasets of everyday activities. For every step we present, we describe how we have solved it during a joint project of the University of Bologna and ETH Zürich. The experiment consists of 28 sensors implementing 5 sensing modalities and recording 64 atomic activities within 8 scenarios of everyday life. The dataset is freely available for research purposes. The article is organized as follows; the next section presents an overview of freely available datasets. We then describe our methodology and the individual steps we have followed to produce our dataset. The article concludes with a number of lessons learned.

RELATED WORK

The UCI Machine Learning Repository (Asuncion & Newman, 2007) is a well-known database for machine learning. Created in 1987, it includes at date of writing 187 datasets for machine learning algorithms. Among others, it also contains datasets related to natural human computer interaction. The Semeion Handwritten Digit dataset includes 1593 handwritten digits from 80 persons. The CMU Face Image dataset includes 640 black and white face images of people taken with varying poses, expressions, and size. The Parkinson dataset is composed of biomedical voice measurements from 31 people, 23 with Parkinson’s disease.

Such a database does not yet exist for activity recognition. However several instrumented home are developing activity recognition datasets. Some examples are the Smart Medical Home Research Laboratory (University of Rochester, 2001), the PlaceLab (Intille, 2006), the AwareHome (Georgia Institute of Technology, 2000), or the SmartHome (Duke University, 2003).

The PlaceLab Intensive Activity 1 and 2 (PLIA1, PLIA2) (Munguia et al., 2006) and PlaceLab Couple 1 (PLCouple1) (Logan et al., 2007) are three freely available datasets collected within the MIT PlaceLab database. PLIA1 and 2 include four hours of continuous household activities (cooking, cleaning, etc.).
Related Content

Statistical Discriminability Estimation for Pattern Classification Based on Neural Incremental Attribute Learning
[www.irma-international.org/chapter/statistical-discriminability-estimation-for-pattern-classification-based-on-neural-incremental-attribute-learning/173414/](www.irma-international.org/chapter/statistical-discriminability-estimation-for-pattern-classification-based-on-neural-incremental-attribute-learning/173414/)

Construction of Domain Ontologies: Sourcing the World Wide Web
[www.irma-international.org/article/construction-domain-ontologies/54064/](www.irma-international.org/article/construction-domain-ontologies/54064/)

Advanced Cellular Neural Networks Image Processing
[www.irma-international.org/chapter/advanced-cellular-neural-networks-image/10224/](www.irma-international.org/chapter/advanced-cellular-neural-networks-image/10224/)

Interactions and Applications of Fuzzy, Rough, and Soft Set in Data Mining

Efficient Identification of Structural Relationships for XML Queries using Secure Labeling Schemes