Gathering Datasets for Activity Identification

Lorcan Coyle, Juan Ye, Susan McKeever, Stephen Knox, Matthew Stabeler, Simon Dobson, and Paddy Nixon
CASL, School of Computer Science and Informatics
University College Dublin
Ireland
lorcan.coyle@ucd.ie

ABSTRACT

The area of activity identification is maturing well in the HCI and ubiquitous computing fields. However, although algorithm development is proceedings well, without publicly available datasets on which to compare results it is difficult to consolidate the disparate work being done. This problem exists because realistic datasets describing human activity are difficult and expensive to gather and because there are significant barriers to releasing the data once gathered. We review positive recent development with the release of two high-quality datasets. From our experiences using these datasets we list some recommendations for the gathering and release of future datasets. Finally, we propose a strategy of our own for gathering a new dataset from these recommendations.

INTRODUCTION

The machine learning community has for years benefited from access to shared datasets on which researchers can compare their evaluation results. The UCI Machine Learning Repository[1] hosts 174 datasets [1], and in certain fields particular datasets are well understood and act as benchmarks for comparing algorithms, e.g., Fisher's *Iris* dataset [2] is popular in pattern recognition research. It is clear that the dearth of datasets in the HCI and ubiquitous computing fields is holding back the comparison of algorithm development for important tasks like activity identification. While it is common practice to test algorithms on personal datasets and publish results, these results will always be questionable if the datasets are not also made publicly available for broad scrutiny. Furthermore, the unwillingness to release datasets places an unnecessary cost to the community when it seeks to build on earlier achievements.

Releasing datasets is a hard task to undertake. There are often ethical constraints forbidding the release of key data. There is a large effort required to make datasets available in a way that is transparent to other researchers; often the tools used to manipulate the datasets need to be packaged along with the datasets themselves. Fortunately there is a broadening recognition of the need to releasing datasets publicly. We describe our experiences with two publicly available datasets released recently: Logan et al.'s PlaceLab dataset [4] and van Kasteren et al.'s activity dataset [6], as well as our experiences with smaller toy datasets. We then abstract some guidelines for gathering other datasets and discuss our goals for developing a new public dataset for activity identification using reasoning techniques that include Bayesian networks, Dempster-Shafer evidence theory and case-based reasoning. Although our data gathering exercise is based in the office domain we believe this exercise will translate well to home environments.

[1]The UCI ML Repository home page is here: `http://archive.ics.uci.edu/ml/index.html`

REVIEW OF EXISTING DATASETS

There are a number of publicly available datasets for the office and home environments. These include Logan et al.'s PlaceLab dataset, which covers 104 hours of annotated data collected from a person living in an instrumented house [4]; van Kasteren et al.'s activity dataset, which covers 28 days of annotated sensor data from the home [6]; and Wren et al.'s motion detector dataset, which captures a year of infrared motion detector data in an office setting [7]. We have recently used the PlaceLab and van Kasteren's activity datasets to evaluate techniques for managing and reasoning with uncertainty and activity identification (this work has been submitted elsewhere and under review). This section gives a critical analysis of both the PlaceLab and van Kasteren's datasets from the perspective of our experiences.

The PlaceLab Dataset

The PlaceLab is a living laboratory in Cambridge, Massachusetts, which is designed to be a highly flexible and multidisciplinary observational research facility for the scientific study of people and their interaction patterns with new technologies and home environments[4]. The PlaceLab house consists of a living room, a dining room, a kitchen, an office, a bedroom, a bathroom and a powder room. It contains over nine hundred sensors, including wireless infra-red motion sensors, which detect motion in the regions of the laboratory, including switch sensors, environmental sensors, and RFID sensors and wrist-mounted RFID readers, which measure whether an object is being used by a user or not.

Data from the PlaceLab were gathered over a period of 15 days, during which a married couple lived in the PlaceLab. The Placelab was instrumented with the audio-visual recording infrastructure that was used to record activities of the subjects except for private activities (such as bathing). 104 hours of the video was annotated by a third party, which precisely described the activities of the male subject of the study. Only the male's activities were annotated as only the male carried an RFID bracelet, which was necessary to cap-

ture interactions with tagged household artifacts.

The PlaceLab dataset is rich and contains well annotated activities with plenty of overlapping sensor data and as such should be ideal for testing activity identification algorithms. However, it has a number of limitations (these are also included in [4]): the most serious of these is that it is impossible to distinguish between the sensor readings that are reacting to the female subject and those that are reacting to the male. Given this limitation, we suggest that it would be sensible to annotate the dataset with the times where only the male subject is in the home to make it possible to use these times to train the system to react to his activities alone. In reality, it is impossible to tell when both subjects are in the house and when only one is. Algorithms thus have to contend with noisy data coming from ghost readings from the female subject while trying to learn and predict the activities of the male. Without the ability to distinguish between which user is triggering many of the sensor readings, datasets with multiple participants are not as useful as they could be for activity recognition, unless the goal is to recognize group activities — in fact the reason the results presented by Logan et al's were so reasonable was that often the couple did things together in the same location [4].

Van Kasteren et al.'s Activity Dataset

Van Kasteren et al.'s activity dataset[2] captures the activities of a single user in a three-room apartment over a period of 28 day [6]. Data are gathered from 14 state-change sensors installed in doors, cupboards, the refrigerator and a toilet flush. Annotations were provided in real-time by the user himself using a Bluetooth headset. Although this dataset is not as comprehensive as the PlaceLab dataset it does have a number of advantages. Since the dataset captures the activities of only one user there is not the same problem of ghost sensor readings from activities that are not annotated. Also, since the annotations are done by the subject themselves in real-time the cost of annotating video post hoc is avoided.

Toy Datasets

Rather than using publicly available datasets it is possible to develop simple toy datasets that allow proof-of-concept demonstrations of algorithms to be presented. In fact, we have built a number of these datasets ourselves. McKeever et al. put together a five-day dataset using three sensors, which we used to demonstrate an uncertainty model for context [5]. A similar dataset, covering 5 different days and backed by diary annotation was used to demonstrate Ye et al.'s earlier work on Situation Lattices [9, 10]. The sensors used gave us three types of information: the first is precise location data from our in-house Ubisense system, whereby a user wears a locator tag. The second is a computer activity sensor is installed on the user's desktop computer that flags when mouse or keyboard is used. The third sensor polls the user's web calendar to determine their schedule of meetings versus unscheduled time. The user maintained a manual diary in order to annotate their activities during the period. These annotations were based on a small number of pre-determined activities of "busy", "break" or "meeting", where the user was always assumed to be in one and only one of these activities.

These toy datasets were useful for proof of concept. They also thought us valuable lessons, which we can employ on future dataset collection. The synchronization of each sensor system, and the diary itself must be carefully maintained. It is difficult to accept diary data generated by a user as being fully accurate as they go about their business - especially when one of the activities to be recorded is "busy". From these lessons we believe that the video-based annotations used in the PlaceLab dataset, and to a lesser extent the real-time spoken annotations gathered by van Kasteren are more trustworthy in this respect.

When speaking about the sensor data we gathered, the reliability of data from tag-based location systems is always questionable given users' propensity to leave their tag behind or the possibility that a tag may break. While gathering our datasets we had to throw away a day's worth of data because a tag was found to be broken. We also found the calendar to be somewhat unreliable as users did not always attend meetings marked in their calendars, nor did they take lunch at prescheduled times. Despite these limitations, and in contrast to our initial expectations, we found that even using these three simple sensors was enough to provide useful toy datasets.

GATHERING DATASETS FOR ACTIVITY IDENTIFICATION

When making recommendations about gathering new datasets for activity identification we must first acknowledge the PlaceLab dataset as the gold standard for developing new datasets. Van Kasteren et al.'s dataset offers significant benefits too as it releases not only the dataset itself but also the software used for the annotation process. We take inspiration from these datasets as well as from our own limited experiences with toy datasets when outlining our guidelines for gathering new datasets:

- The focus of attention must be on the annotation process - this is where the real value of any dataset is. This means that it is necessary to decide in advance which data is to be annotates and what the annotations will be used for.
- It should be possible to re-annotate a dataset again if the purpose of the evaluation changes. This might mean re-examining the video data and adding additional annotation streams in parallel to the original annotation. This requires the capability of representing and releasing multiple annotation streams for each dataset.
- It should be decided in advance whether to goal is to capturing a single user's activities or multiple people's activities. If the goal is to only capturing a single user's activities but the sensors may react to additional users it is necessary to internalize the externalities as possible, and ensure that all sensor readings are related to the user we are interested in and not to ghost users. If there are sensors react to other users, e.g., guests, an effort should be

[2]Van Kasteren et al.'s dataset is available for download here: `http://staff.science.uva.nl/~tlmkaste/research/software.php`

made to annotate those activities.

- To focus of annotation should be to capture the activities that are most interesting. This means balancing the annotation effort towards interesting activities that might not actually happen often rather than over-annotating common but uninteresting activities. In the home field it is typical to see datasets that have only a small quantity of annotations devoted to interesting activities (e.g., hygiene activities, which are interesting from a health perspective).
- When there is any doubt as to the correctness of a part of the dataset it must be possible to either remove it or flag it as being of dubious quality. This flagging should be included in the dataset as another annotation stream.

In order to make it easier to reuse a dataset the following considerations should be made:

- In order to reuse a dataset, there should be an unambiguous explanation of the data fields in each sensor file, clear listing of sensor reading filenames and mapping of annotations to users, if the dataset is tracking multiple users.
- When gathering new datasets, precision and accuracy should be quantified for each sensor and this data should be included this information in the dataset.

Although we gathered fine-grained location data for our toy datasets (i.e., x, y, z coordinates at a cm scale) this was not the form that the data was actually used. We abstracted location data and clipped the sensor readings into symbolic locations at a room granularity. We used similar mappings with time, raising it from a fine per-second granularity to a broader symbolic time of day. However, when we published our toy dataset these mappings were not also released. Without these mappings between sensor data and the actual context data used in our experiments it is difficult to reproduce our results. When experimental results are published along with a dataset, these mappings must also be exposed.

Although more sensors are better when gathering a dataset, there is a problem for third-party developers when there is no awareness of the interplay between sensors, or the level of uncertainty associated with them. In order to smooth the learning curve, it would be useful if the developers of datasets were able to highlight subsets of the sensor set and annotation space where sensor data matches well with expected accuracies. By publishing ideal results on these subsets of the dataset it is easier to bootstrap new users of the dataset. From a practical perspective, the ideal solution is to include as many sensors in the data gathering effort but to mark subsets of the dataset as being more accurate or useful for bootstrapping third-party researchers.

PROPOSED CASL ACTIVITY RECOGNITION DATASET

We are currently building a number of overlapping sensor networks across our research centre, the Complex & Adaptive Systems Laboratory (CASL[3]). CASL is a 2500 square meter research facility containing more than 175 researchers from a number of disciplines, including mathematics, engineering, and computational science. CASL occupies a five floor building and currently we are gathering Bluetooth and Ubisense data from the communal coffee/lunch area on the fifth floor and offices area on the third floor.

From our experiences with publicly-available datasets, and from the guidelines listed above we are gathering a new activity recognition dataset for public release. We are introducing a number of additional sensors to our research space beyond the Ubisense, computer activity monitor, and calendar sensors used in our earlier toy datasets. While many of these sensors are already functioning and data gathering is underway, the more important and difficult task of annotation has not yet commenced. Here we outline our existing capture strategies, then we outline additional sensors we will integrate, and finally we outline our annotation strategy. Our goal is to capture users' activity in our office environment, to test our sensor systems, and to use our experiences to gather more natural, home and office datasets.

Sensor Infrastructure

We will gather sensor data from various sources within the CASL building. Location data will be gathered at a precise granularity using *Ubisense* and at room-level granularity using *Bluetooth*. Using a number of Bluetooth spotters we are logging simple information about the location of all Bluetooth devices in CASL. The software that gathers this data is part of Basadaeir[4], which acts as an API exposing the sensor data gathered in CASL. Basadaeir's website currently exposes location data gathered from mobile phones to provide an in-out board application. We are also capturing calendar data from a number of users, which captures the times they are scheduled to attend meetings and take lunch.

Participants will be asked to keep their Bluetooth-enabled phone with them at all times and to wear an active Ubisense tag. We will also track user's schedules with a *google calendar* sensor and *instant messaging statuses* will be recorded. We have also placed *pressure pads* under the carpet in boundary areas between rooms and under desks to capture the times people enter and leave those locations. Since the pressure pads have no concept of identity they offer a challenge in a multi-user environment.

A separate data gathering exercise is going on in CASL that follows on from collaborative work done by Lavelle et al. [3], which gathers the Bluetooth and WiFi IDs and signal strengths of all devices seen by a number of smart phone users. In the future it should be possible to match up the CASL dataset, which captures the sensor readings gathered by the environment (e.g., Bluetooth spotters), with the data gathered by each user's smart phones looking out at the environment. This is made possible because all data is time-stamped, and because many of the users are participants in both studies. This collaboration would add value to both data gathering exercises.

Annotation Strategy

[3]The CASL website is here: `http://casl.ucd.ie`

[4]Basadaeir's webpage is here: `http://basadaeir.ucd.ie`

Initially we will focus on annotating a narrow set of activities — whether a user is in a meeting, on a break, busy working at their computer, or working at their desk but not working. This makes the annotation task more straightforward but lessens the potential value of the annotations. We will keep all annotations and data stored but leave the door open to increasing the value of the dataset by making it possible to re-annotate the dataset in the future. For our initial experiments we will extend our length of time covered from our earliest one-week single-user toy datasets to a longer 4-6 week dataset covering the activities of 5 users. Depending on the outcome of this study we will increase the scale of the experiment further.

Our annotations will be gathered from both user-generated diary data and from annotated video data. We believe that real-time self-reporting using diaries (as done by van Kasteren et al.) offer the fullest and most correct description of what the user was actually doing at any time. However, to overcome the limitations of self-reporting, we intend to replicate the annotation by examining video data. We will use video to capture the activities of users in the cubicle spaces and the common areas. These will give enough information to annotate most activities that will be needed to capture each of the activities taking place in those spaces.

In order to improve the speed of annotation we have placed pressure pads under the carpet at desks and at the entrance to the common areas. These sensors give highly accurate data about whether a person is sitting or standing in that exact location and we will use these to delineate the boundary conditions, e.g., counting people in and out of the common area, checking when somebody leaves or arrives at their desk. By synchronizing diaries against these times we will get the benefit of faster annotation by examining just the video around those times (although possibly at the loss of some accuracy). Furthermore, by coupling pressure sensors placed under the carpet with webcams, we will record data only when there is activity in a particular area. This will allow us to avoid analyzing hours of video where there is nothing happening.

To make it easier to share and reuse our datasets and annotations we will mark up the sensor data and annotations using the best practices in pervasive computing ontology design [8]. By exposing the annotation API it will be easier for third-party annotations to be shared across researchers using the same underlying dataset.

CONCLUSION

Although it is true that the HCI and ubiquitous computing fields are sorely lacking in publicly available datasets there is much to be optimistic about. There are a number of high-quality datasets available today and already there is positive research coming out of the sharing of these datasets. As the community sees the benefit of both the release and use of shared datasets these benefits will multiply. We have mentioned our experiences with working with two datasets in the area and have learned much from them already. We hope that with our next steps we will add to the number of publicly available datasets and we hope that there will be a heavy emphasis in the field for new publications to to expose their datasets.

ACKNOWLEDGEMENTS

This work is partially supported by Science Foundation Ireland under grant numbers 05/RFP/CMS0062 “Towards a semantics of pervasive computing”, 04/RPI/1544 “Secure and predictable pervasive computing”, and Enterprise Ireland under grant number CFTD 2005 INF 217a, ”Platform for User-Centred Design and Evaluation of Context-Aware Services”.

REFERENCES

1. A. Asuncion and D. Newman. UCI machine learning repository, 2007.
2. R. A. Fisher. The use of multiple measurements in taxonomic problems. *Eugenics*, 7:109–122, 1936.
3. B. Lavelle, D. Byrne, C. Gurrin, A. F. Smeaton, and G. Jones. Bluetooth familiarity: Methods of calculation,applications and limitations. In *MIRW 2007 - Mobile Interaction with the Real World, Workshop at the MobileHCI07: 9th International Conference on Human Computer Interaction with Mobile Devices and Services*, 2007.
4. B. Logan, J. Healey, M. Philipose, E. M. Tapia, and S. S. Intille. A long-term evaluation of sensing modalities for activity recognition. In *Ubicomp*, pages 483–500, 2007.
5. S. McKeever, J. Ye, L. Coyle, and S. Dobson. A multilayered uncertainty model for context aware systems. In *Late Breaking Results - Adjunct Proceedings of the 6th International Conference on Pervasive Computing*, pages 1–4, Sydney, Australia, 2008. OCG.
6. T. van Kasteren, A. Noulas, G. Englebienne, and B. Kröse. Accurate activity recognition in a home setting. In *UbiComp ’08: Proceedings of the 10th international conference on Ubiquitous computing*, pages 1–9, New York, NY, USA, 2008. ACM.
7. C. R. Wren, Y. A. Ivanov, D. Leigh, and J. Westhues. The merl motion detector dataset. In *MD ’07: Proceedings of the 2007 workshop on Massive datasets*, pages 10–14, New York, NY, USA, 2007. ACM.
8. J. Ye, L. Coyle, S. Dobson, and P. Nixon. Ontology-based models in pervasive computing systems. *The Knowledge Engineering Review*, 22:315–347, Dec. 2007.
9. J. Ye, L. Coyle, S. Dobson, and P. Nixon. Representing and manipulating situation hierarchies using situation lattices. *Revue d’Intelligence Artificielle*, 22:647–667, 2008.
10. J. Ye, S. McKeever, L. Coyle, S. Neely, and S. Dobson. Resolving uncertainty in context integration and abstraction. In *ICPS ’08: Proceedings of the 5th international conference on Pervasive services*, pages 131–140, Sorrento, Italy, 2008. ACM.