

Towards a Consistent Methodology for Evaluating Activity Recognition Model Performance

Tim van Kasteren, Gwenn Englebienne and Ben Kröse

Intelligent Systems Lab Amsterdam
University of Amsterdam
Science Park 107,1098 XG, Amsterdam
The Netherlands

Email: T.L.M.vanKasteren@uva.nl

WWW home page: <http://www.science.uva.nl/~tlmkaste/>

Abstract—Activity recognition has become a very active field of research over the past few years. Many high quality papers are published, but the community is lacking a consistent methodology for evaluating model performance. Furthermore, the term activity recognition is too commonly used, making it difficult to compare work. To address these issues, we propose to release a benchmark dataset which can be used to compare the performance of newly proposed models. And suggest a survey paper is written to clearly categorize the different approaches to activity recognition. This will make it easier to compare results from different papers and will help the community grow further.

I. WHAT IS THE COMMUNITY DOING WELL?

Over the past few years activity recognition has become a very active topic of research. Activity recognition is done using wireless sensor networks, cameras, wearables, RFID and GPS. It is performed in home and office environments, and in outside settings. The term activity recognition, therefore, covers a large body of current pervasive computing research.

Because activity recognition has become such an active field there is a lot of competition, resulting in many high quality publications. The pervasive computing community has many interesting and creative ideas of how activity recognition can be applied. Special issues on activity recognition are common in journals, and so are special sessions in conferences.

Activity recognition models are based on state of the art machine learning models. Newly introduced models presented at machine learning conferences quickly find their application in activity recognition. And models are adapted to address specific issues relevant to activity recognition.

II. WHERE DOES THE COMMUNITY NEED TO IMPROVE?

The key to progress in science are experiments and the biggest weakness of the activity recognition community lies in the presentation of experimental results. Currently many results are presented using relatively small datasets, often recorded by the researchers themselves. Due to privacy issues it is not always possible to release a dataset to the public for further testing. This makes comparison with other work using a different dataset difficult, because there is no clear definition of activities.

In work where public datasets are used, the set of activities used in the experiments often changes from paper to paper. Furthermore, it is not clear what a good measure for evaluation is, because this depends on the application in which activity recognition is used. Every paper therefore introduces its own set of evaluation criteria.

These issues require researchers to provide their own baselines for comparing their model performance. Some papers present highly specified models specifically engineered for activity recognition. Performance of these models is presented without any further comparison to their simpler counterparts. This makes it difficult to understand why the model is performing well.

Another issue is the frequent use of the term activity recognition. Activity recognition is a popular term and has been used to describe recognition of human activities of daily living (ADL) in a home setting [5], recognition of human behavior in an outdoor setting [1] and recognition of the role of a robot in a game environment [4]. Although the use of the term activity recognition in each of these cases is perfectly warranted, it is clear that these different recognition tasks have different characteristics. Presenting them all using the same keyword is likely to lead to confusion.

III. MY BEST RECOMMENDATIONS

To address these issues we would ideally like to see the following elements in the presented results:

The type of activity recognition that is performed should be clearly categorized. For each dataset that is used the elements in table I should be described.

| Category | Possible values |
|--------------|--|
| Type of data | Real world, Laboratory, Simulated, ... |
| Environment | Home, Office, Outside, ... |
| People | Single person, Multiple people, ... |
| Subject | 50 year old male, ... |
| Sensors | Sensor network, Cameras, Wearables, GPS, RFID, ... |

TABLE I
LIST OF CATEGORIES AND THEIR POSSIBLE VALUES OF SETTINGS
ACTIVITY RECOGNITION CAN BE PERFORMED IN.

| True | Inferred | | | FN |
|------|-----------------|-----------------|-----------------|---------|
| | 1 | 2 | 3 | |
| 1 | TP_1 | ϵ_{12} | ϵ_{13} | FN_1 |
| 2 | ϵ_{21} | TP_2 | ϵ_{23} | FN_2 |
| 3 | ϵ_{31} | ϵ_{32} | TP_3 | FN_3 |
| FP | FP_1 | FP_2 | FP_3 | $Total$ |

TABLE II

CONFUSION MATRIX SHOWING THE TRUE POSITIVES (TP), FALSE NEGATIVES (FN) AND FALSE POSITIVES (FP) FOR EACH CLASS. THE ϵ_{ij} TERMS SHOW THE ERROR BETWEEN TRUE CLASS i AND INFERRED CLASS j . FN IS THE SUM OF THE ERROR TERMS IN THE SAME ROW, FP THE SUM OF THE ERROR TERMS IN THE SAME COLUMN.

Sensor data often needs to be preprocessed to be used with the proposed model. This typically includes a feature extraction step and a discretization step. The paper should include an explanation for the choice of features and a clear description of how the features were calculated. Furthermore, a description of the discretization process should be given, together with the granularity at which data was discretized.

The measures used for evaluating the model performance can depend strongly on the application the authors have in mind. In general we propose to use a class average precision and recall. We are dealing with a multi-class classification problem and therefore define the notions of true positive (TP), false negatives (FN) and false positives (FP) for each class separately as shown in a confusion matrix in Table II. The class average precision and recall can be calculated using

$$\text{Precision} = \sum_{i=1}^N \pi_i \frac{TP_i}{TP_i + FP_i} \quad (1)$$

$$\text{Recall} = \sum_{i=1}^N \pi_i \frac{TP_i}{TP_i + FN_i} \quad (2)$$

(3)

where π_i is the weight of the class, typically set to $\pi_i = 1/N$, and N is the total number of classes. In some applications some classes are more important than other classes and it is possible to give these more important classes a higher weight in the evaluation measure. In particular, in activity recognition there often is an ‘other’ or ‘idle’ activity in which no annotation is available or the person is simply doing nothing. This activity often takes up a large amount of time slices but is usually not a very important activity to recognize. It might therefore be useful to weigh this activity less. Ideally the researchers include the confusion matrix for at least one of their experiments using absolute time slice numbers (instead of percentages). This will clearly show which activities the model is able to recognize and which it has trouble with.

With respect to the experimental setup, it is of great importance on which datasets the experiments are performed. Next to comparing the proposed model on any of their own recorded data, the researchers should compare the performance of their model on a publicly available dataset.

Furthermore, it is important to which model their proposed model is compared to highlight the strengths and weaknesses

of the proposed model. A good baseline model for activity recognition is the hidden Markov model (HMM). Its strengths and weaknesses are well understood [3], yet it generally gives a good performance in activity recognition [2], [5].

IV. NEXT STEPS

To achieve this ideal form of research methodology the community will need to provide a benchmark dataset for evaluation. A number of large real world datasets, described in a single paper, will provide a consistent benchmark of activity recognition performance. For each dataset there should be a thorough analysis of the challenges for recognizing activities in the data. The paper should describe a number of evaluation criteria, which provide a clear insight into the performance of a model. For a number of commonly used models these evaluation criteria should be provided in the paper to allow consistent comparison. Examples of such models which could function as baseline models are the hidden Markov model (HMM), naive Bayes model and conditional random fields. The code for these models and for the evaluation criteria should be made public together with the paper. Future papers using the benchmark should at least include the performance of their model compared to the benchmark dataset using the proposed criteria. Additional experiments can further highlight the performance of the model.

Another important step that needs to be taken is a survey of activity recognition work over the past years. A survey will provide a clear categorization of ongoing research of activity recognition and will help the community become more structured allowing better comparison of results. Clear terminology can be suggested in the survey paper to set apart the different fields.

V. OUR WORK

In our work we use temporal probabilistic models to recognize activities of daily living (ADL) from sensor data. Our focus lies on the development and understanding of models for activity recognition and on creating effective methods for learning model parameters. We have primarily used wireless sensor networks to observe the inhabitants in a home. Our



Fig. 1. Wireless network node to which sensors can be attached.

wireless sensor network consists of several wireless network nodes (Fig. 1) to which sensors can be attached. Examples of sensors used include reed switches to measure open-close states of doors and cupboards; pressure mats to measure sitting on a couch or lying in bed; mercury contacts for movement of objects (e.g. drawers); passive infrared (PIR) to detect motion in a specific area; float sensors to measure the toilet being flushed. Sensor output is binary and represented in a feature space which is used by the model to recognize the activities performed.

In the past four years we have recorded several real world datasets in home settings, each consisting of several weeks of data. We plan to release these datasets to the public together with a paper providing a benchmark for the activity recognition community, as described above. Our datasets were all recorded using a wireless sensor network in homes with a single occupant. We encourage other researchers who have recorded datasets of different configurations to contact us, to include their data in the benchmark dataset.

ACKNOWLEDGMENT

This work is part of the Context Awareness in Residence for Elders (CARE) project and the ‘Zorg(en) voor Morgen’ project. The CARE project is funded by the Centre for Intelligent Observation Systems (CIOS) which is a collaboration between UvA and TNO. The ‘Zorg(en) voor Morgen’ project is funded through the Pieken in de Delta-program by the Ministry of Economic Affairs and the cities of Utrecht and Lelystad and the provinces of Utrecht, Noord-Holland and Flevoland.

REFERENCES

- [1] Lin Liao, Dieter Fox, and Henry Kautz. Extracting places and activities from gps traces using hierarchical conditional random fields. *The International Journal of Robotics Research*, Vol. 26:No. 1, 119–134, 2007.
- [2] Donald J. Patterson, Dieter Fox, Henry A. Kautz, and Matthai Philipose. Fine-grained activity recognition by aggregating abstract object usage. In *ISWC*, pages 44–51. IEEE Computer Society, 2005.
- [3] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [4] Douglas L. Vail, Manuela M. Veloso, and John D. Lafferty. Conditional random fields for activity recognition. In *International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, 2007.
- [5] Tim van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben 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.