# Standing on the Shoulders of Other Researchers A Position Statement

Ulf Blanke, Diane Larlus, Kristof Van Laerhoven, Bernt Schiele
Computer Science Department
TU Darmstadt
Email: {blanke, larlus, kristof, schiele} @ cs.tu-darmstadt.de

***Abstract*—Activity Recognition has made significant progress in the past years. We strongly believe however that we could make far greater progress if we build more systematically on each other's work. Comparing the activity recognition community with other more mature communities (e.g., those of computer vision and speech recognition) there appear to be two key-ingredients that are missing in ours. First, the more mature communities have established a set of *well-defined or accepted research problems*, and second, the communities have a tradition to *compare their algorithms on established and shared benchmark datasets*. Establishing both of these ingredients and evolving them over time in a more explicit manner should enable us to progress our field more rapidly.**

***Index Terms*—Activity Recognition, Evaluation, Code and Database Sharing**

## I. Where does the community need to improve?

In this paper we argue that our community of activity recognition has to improve on two fronts.

**1. Research Problems** Develop and evolve well-defined and accepted research questions that we believe are essential to make progress in activity recognition.

**2. Evaluate, Analyze, and Share** In order to make progress in activity recognition we have to understand and analyze thoroughly the strengths *and* weaknesses of different approaches. Therefore we need to a) share datasets and establish benchmarks to enable direct comparison and b) enable reproducibility of algorithms and results so that others can profit from our work and build upon each other's work.

The first front, namely the definition and maturation of well-defined research problems seems obvious but is – in our view – one of the weaknesses of our area. In many communities such well-defined problems can be tackled (let's take again the example of computer vision, in which object class recognition or optical flow estimation exist as challenges). In our community however we often take our subjective ideas about activity recognition, motivate why we think this is an important problem, and then record our own – typically non-shared – datasets to evaluate our algorithms. While this is fine at an early stage of a community we strongly believe that we have to rethink this practice and establish attractive research problems that are relevant to pursue and consequently are dedicated to work on. It is important to note that these research problems will and have to evolve over time. One of the reasons but not the only one is the progress we are making on previous research problems. However, these well-defined problems are absolutely essential to enable comparison as well as to analyze and understand our progress.

The second front is equally important and again a weak spot of our field. As already mentioned most of us analyze their great new algorithm on a new dataset making it hard to understand the progress that was made. Instead, we should develop (or even enforce) the practice that all new algorithms are compared to previous ones, either on common datasets or using the code shared from previous algorithms. As many of our algorithms originate from machine learning research, it is often inappropriately taken for granted that a trendy algorithm there, translates to superior performance in activity recognition. It is also worth noting that it is not enough to simply state performance numbers in such comparative studies. Instead one has to analyze and discuss why which algorithm performs differently. While this is again standard practice in other research areas this type of analysis and scientific knowledge generation is nearly completely absent in our field.

## II. Our best recommendations

The discussion of the previous section is based on the following cyclic approach to research. Each cycle comprises four steps:

1) Start with a **clear problem definition**,
2) Evaluate the **State-of-the-Art**
3) Synthesize, propose, and implement a (typically novel) **Problem Solution**
4) **Analysis** of the proposed solution on real-world data

These cycles require both points mentioned in the previous section. Without a set of well-defined problems we cannot start with a clear problem definition and (equally important) cannot evaluate the state-of-the-art. In current activity recognition research it is often not clear how a particular approach might perform on the chosen problem as respective papers often do not formulate the problem definition clearly enough. This in turn is essential to develop a problem solution that is typically synthesized from previous research and often contains novel aspects. These novel aspects again rely on a better understanding of the respective algorithms' strengths and weaknesses. Probably the most important part of the cycle however is the analysis of the proposed solution where most of the novel scientific knowledge is created. In this last step the

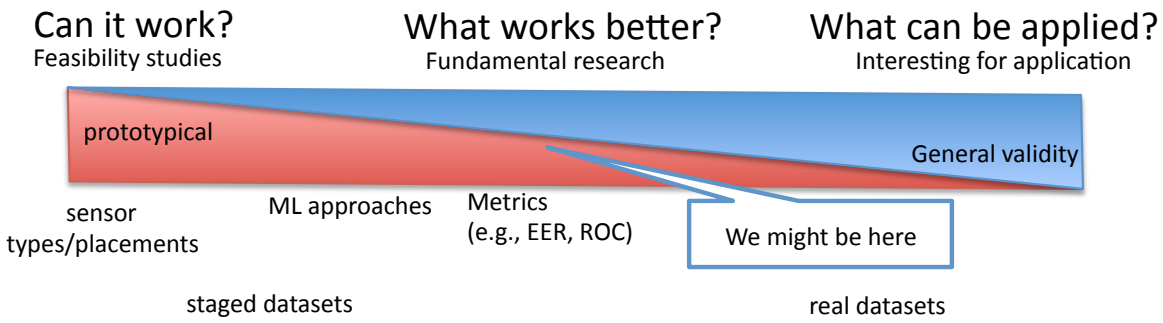


Fig. 1. From prototypical research to its application (or past to future)

performance of the algorithm is analyzed with respect to the clear problem definition allowing to understand the strengths as well as the weaknesses of the novel proposal.

In Germany we have the following saying: "if you can formulate a problem, you know the solution". The last step of the cycle therefore should result in a new and clear problem formulation that is then addressed in the next research cycle.

## III. What is the community doing well?

The community came already a long way since its very first papers, developing *implicitly* several good research methodologies. Figure 1 tries to give a rough overview and can be seen as timeline from left to right over the last decades.

**Evaluation metrics.** Whereas in the past, basic feasibility stood in foreground, e.g., in [1], today the use of expressive metrics and standard experimental designs (e.g., precision/recall, confusion matrices, EER) are common and meanwhile must-sees in papers.

**Experimental design.** The way especially classifiers are evaluated has also improved significantly over the years, with cross-validation being a standard requirement as well. Similarly, statistical significance, and size of the sample database are increasingly used in support of conclusive experiments.

**Realism.** While research needs to cover the complete spectrum, datasets were initially often motivated *application-wise* on a very small scale, leaving a large gap between the real-world application and the simplified dataset. This would rely heavily on the imagination of the reader to bridge the gap between what *was* studied, and what it actually *could* be applied to. The use of representative test subjects (instead of research students), the number of application-situated datasets, and the variety within the recorded data, are all noticeably increasing.

The points mentioned in this section are however not enough to guarantee that work will optimally build on that of others in the most efficient way. The next section will address what implicit methodologies other fields have in place.

## IV. What can we learn from other communities?

There are several additional aspects that have proven to work well in the more mature fields which were mentioned earlier. Here follow a few key observations that we believe are responsible for their current level of maturity.

**Commonly accepted datasets.** Regarding the usage of datasets, the vision community usually starts with complex datasets with respect to current state of the art research. Different authors iteratively approach these, improving performance progressively, until core problems of the datasets are solved. In the data mining community standard datasets are also used very rigorously. One of their main sources can be found here [3]. This is very different in our community, where a single dataset is often specifically recorded for the usage of a single specific approach.

**Comparative studies on multiple datasets.** In matured communities such as the machine learning (e.g., NIPS conference) or computer vision (e.g., CVPR, ICCV conferences) communities, the culture to create directly comparable results among the authors is strong. Evaluating algorithms on different datasets – common in other communities – is rarely done in our community. The benefits are obvious: not only does it allow generalizing conclusions, but also ensures more insight into the performance of an approach. Besides dataset-related issues, the experimental design plays a central role in comparative studies. If not done carefully, invalid conclusion can quickly be drawn. In [4], the author discusses several typical pitfalls of experimental comparisons.

**Explicit Challenges.** Projects in form of challenges are common [5], [6], [2], [7] and show evidence for this culture. Here evaluation criteria are defined a priori and have to be strictly met by all participants. In the following we will outline interesting aspects that characterize specific challenges.

- An interesting idea is done in PASCAL Challenge [6]. Here, the goal is to classify detect or segment visual objects. The number of classes is usually fixed (to 20). The participants are provided with a training- and a validation-set. Interestingly, test-data is provided without any labels. Instead of evaluating themselves, the participants send their predictions to the organizers. In turn, the participant receives the quantitative results. Clearly, outsourcing the evaluation step can support bias-free comparability between researchers. Besides a "third-party" evaluation, PASCAL organizes meetings, where usually PhD students

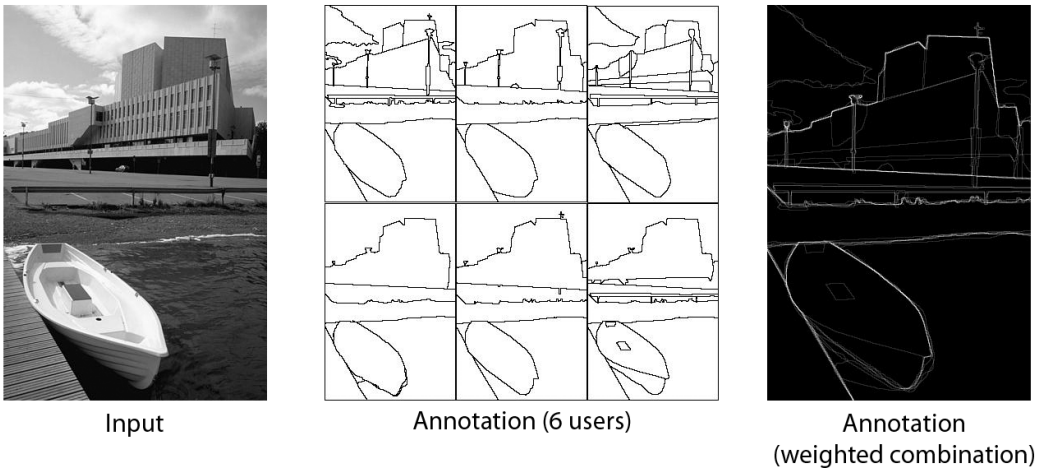


Fig. 2. Example annotation of the The Berkeley Segmentation Dataset and Benchmark [2]. Given a single image, multiple users annotate contours. The different annotation instances are then combined to a single annotation file.

gather to label datasets.

However, methods like this challenge might not be bulletproof and can be exploited to a certain extend (having often only few data available in activity recognition, labels could be guessed or the evaluation procedure can be abused – then approaches can be optimized on the test set).

What makes this challenge noteworthy for us is that the dataset provided through the challenge is not only used for the challenge itself. Researchers also refer to it as reference dataset for results provided in their papers, which are not necessarily within the scope of the challenge.

- The goal of BSDS [2] is to provide an empirical basis for research on image segmentation and boundary detection. Data, labels and benchmarking code is provided in a public repository. Additionally, benchmark results are gathered to show the cooperative scientific progress.

  This project is specifically interesting for us, as the problem of segmentation is perceived subjectively and differs between users, complicating the agreement on what has to be solved. To overcome this problem it is addressed as follows: A single image is annotated by multiple users, resulting in slightly different annotation instances (see Fig. 2). Each instance is regarded as valid and is combined to a final "soft" annotation file. Recall and precision are then calculated for different levels of the annotation. Keeping the statistics of annotation can help to evaluate across different applications, as different users have different perception or interests.

**Reproducibility.** Researchers in the more mature fields are watching each other's results closely and tend to demand reproducible experiments from authors who improve on previous benchmarks. Wrong use of a dataset or its parameters ("magic numbers" that only appear in experiment scripts) can quickly lead to false optimizations and deceptive conclusions. An interesting occurrence in this regard is [8], where an author in the data mining community was publicly singled out for wrong use of a dataset – regardless whether this truly was justified, this meticulous checking of others' results is missing in our community.

We outlined a few aspects of other communities that should be pursued within our community as well. Needless to say, the size of the community is an important factor which makes, e.g., collective databases such as [9] possible. While we are not at scale to allow such *implicit* development, we have to *explicitly* bring things forward.

## V. Next steps

Therefore, the field of activity recognition needs the following expansions.

**Clear problem definition.** For commonly accepted datasets to be established, a prime requirement is the need of a problem that is attractive enough to gain the interest of a larger community. This goes for both the application scenario that inspires the problem and the approaches to solve it.

**Improved evaluation methodology.** Once clear problems are defined, benchmarks in the form of valid datasets and the way they should be evaluated can be agreed upon within the field. More complex and realistic datasets are still desirable for the community to work on incrementally, hereby creating momentum for a challenge that cannot be solved at the first shot. This demands also the acceptance of initial state-of-the-art results that may be relatively low at the initial attempts, leaving space for successive improvements.

**Organize a competitive framework.** Although challenges have worked very well in other fields, the organization of a challenge is perhaps the hardest step to implement. For this to be successful, sufficient funding, clearly defined problems, a large interest from industry or the common public, and community-based momentum needs to be gained. We believe the timing for this is not yet right, but can nevertheless be kept as a future possibility.

**Demand reproducible results.** Sharing is the key to support this. Currently, motivation to deploy (clean) source code is low, as well as reimplementing previous work from others. As a result this often leaves non-recyclable code which goes to waste upon publication of the article. We think publication of datasets and *precise* description of approaches and evaluation techniques is the way to go. New approaches can then be directly compared with previous results on specific datasets. Ideally, authors should publish datasets and accompanying code straight after the publication of their paper.

## VI. Our work

Our work almost exclusively focuses on activity recognition from wearable motion sensors. Our field of interest ranges from efficient approximation of sensor signals [10], [11] and long term data collection through spotting gesture-like activities [12], [13] to approaches to infer high-level composed activities [14], [15], [16], [17]. Within this scope we are strongly motivated to exchange ideas about experimental methodologies of researchers with similar research aims and how we can improve reproducibility and comparability of our work within our community.

## Acknowledgements

The authors gratefully acknowledge the scholarship provided by the DFG research training group "Topology of Technology" and the partial funding through the DFG project "Methods for Activity Spotting With On-Body Sensor".

## References

[1] A. R. Golding and N. Lesh, "Indoor navigation using a diverse set of cheap, wearable sensors," in *ISWC*, 1999, pp. 29–36.

[2] The berkeley segmentation dataset and benchmark. [Online]. Available: http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/

[3] Ucr time series classification/clustering. [Online]. Available: http://www.cs.ucr.edu/ eamonn/time_series_data/

[4] S. L. Salzberg, "On comparing classifiers: Pitfalls to avoid and a recommended approach," in *Data Mining and Knowledge Discovery*, 1997.

[5] Contest on semantic description of human activities. [Online]. Available: http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html

[6] Pascal. [Online]. Available: http://pascallin.ecs.soton.ac.uk/challenges/VOC/

[7] Imageclef. [Online]. Available: http://www.imageclef.org/

[8] Jason chen's papers on making time series clustering meaningful are deceptive. [Online]. Available: http://www.cs.ucr.edu/ eamonn/JasonMeaningless.pdf

[9] Labelme. [Online]. Available: http://people.csail.mit.edu/torralba/research/LabelMe

[10] K. Van Laerhoven, E. Berlin, and B. Schiele, "Enabling efficient time series analysis for wearable activity data," in *Proceedings of the 8th International Conference on Machine Learning and Applications (ICMLA 2009)*, IEEE Press. Miami Beach, FL, USA: IEEE Press, 2009, pp. 392–397.

[11] K. Van Laerhoven and E. Berlin, "When else did this happen? efficient subsequence representation and matching for wearable activity data," in *Proceedings of the 13th International Symposium on Wearable Computers (ISWC 2009)*, IEEE Press. IEEE Press, 2009, pp. 69–77.

[12] A. Zinnen, C. Wojek, and B. Schiele, "Multi activity recognition based on bodymodel-derived primitives," in *LoCA*, 2009, pp. 1–18.

[13] A. Zinnen, U. B. Blanke, and B. Schiele, "An analysis of sensor-oriented vs. model-based activity recognition. (," in *13th Int. Symposium on Wearable Computing*, 2009.

[14] M. Stikic, K. Van Laerhoven, and B. Schiele, "Exploring semi-supervised and active learning for activity recognition," in *Proceedings of the 12th International Symposium on Wearable Computers (ISWC 2008)*, IEEE Press. Pittsburgh, USA: IEEE Press, September 2008, pp. 81–90.

[15] K. Van Laerhoven, D. Kilian, and B. Schiele, "Using rhythm awareness in long-term activity recognition," in *Proceedings of the 12th International Symposium on Wearable Computers (ISWC 2008)*, IEEE Press. IEEE Press, 2008, pp. 63–68.

[16] U. Blanke and B. Schiele, "Daily routine recognition through activity spotting," in *4rd International Symposium on Location- and Context-Awareness (LoCA)*, 2009.

[17] T. Huynh, U. Blanke, and B. Schiele, "Scalable recognition of daily activities with wearable sensors," in *3rd International Symposium on Location- and Context-Awareness (LoCA)*. Springer, 2007.