# Automatic Correction of Annotation Boundaries in Activity Datasets by Class Separation Maximization

**Reuben Kirkham**
Culture Lab
Newcastle University, UK
r.kirkham@newcastle.ac.uk

**Sebastian Mellor**
Culture Lab
Newcastle University, UK
s.j.i.mellor@newcastle.ac.uk

**Aftab Khan**
Culture Lab
Newcastle University, UK
Aftab.Khan@newcastle.ac.uk

**Daniel Roggen**
Culture Lab
Newcastle University, UK
Daniel.Roggen@newcastle.ac.uk

**Sourav Bhattacharya**
Helsinki Institute for
Information Technology HIIT
University of Helsinki, Finland
sourav.bhattacharya@cs.helsinki.fi

**Thomas Plötz**
Culture Lab
Newcastle University, UK
Thomas.Ploetz@newcastle.ac.uk

**Nils Hammerla**
Culture Lab
Newcastle University, UK
nils.hammerla@newcastle.ac.uk


*UbiComp'13 Adjunct*, September 8–12, 2013, Zurich, Switzerland.

ACM 978-1-4503-2215-7/13/09...$15.00.

http://dx.doi.org/10.1145/2494091.2495988

## Abstract
It is challenging to precisely identify the boundary of activities in order to annotate the activity datasets required to train activity recognition systems. This is the case for experts, as well as non-experts who may be recruited for crowd-sourcing paradigms to reduce the annotation effort or speed up the process by distributing the task over multiple annotators. We present a method to automatically adjust annotation boundaries, presuming a correct annotation label, but imprecise boundaries, otherwise known as "label jitter". The approach maximizes the Fukunaga Class-Separability, applied to time series. Evaluations on a standard benchmark dataset showed statistically significant improvements from the initial jittery annotations.

## Author Keywords
Annotation errors; Class Separability; Crowdsourcing; Human activity recognition

## ACM Classification Keywords
I.5.m [Pattern Recognition]: Miscellaneous.

## Introduction
Activity recognition typically requires a ground-truth of pre-annotated sensor data. These annotations comprise the *start* and *end* point of the activity together with the

*activity class*, usually identified from video footage synchronized with the acquired sensor data. Annotating datasets is challenging and a time consuming process. The challenges are to identify precisely the boundary and nature of activities, as they may be hard to define objectively well. This leads to so called "label-jitter", where the effective activity start and end points may not align perfectly with the annotation, or annotation errors, where the annotation is incorrect (although its boundaries may make sense). As the pattern recognition techniques used for activity recognition are trained on the annotated dataset, higher annotation jitter or errors tends to reduce activity recognition performance. An approach to study annotator behavior via exploring datasets with multiple annotators has been presented in [2]. Compensating annotation jitter or error by relying on more trained annotators would most likely reduce the number of available annotators and increase annotation cost. At the same time, lay annotators appear to be good at broadly determining between the general classes of behaviour although they would not desire to spend a great deal of additional time determining where the precise boundaries lie, as shown in [11]. Therefore crowd-sourcing paradigms have recently been investigated to speed up the annotation process [7]. One way to then cope with higher annotation jitter or errors of multiple lay annotators is to use a standard measure for disagreement (like Kappa scores [3]) in order to obtain correlation between various annotators. Algorithmic methods were proposed to handle errors in the activity class label in [1] (but not in the beginning and end time of the activity). The resilience to label jitter may also be improved in the actual pattern recognition step [6].

In our work, we explicitly address the problem of correcting annotation boundaries algorithmically. We rely on the Fukunaga Class Separability [4] to evaluate the quality of the annotations, which we modified to effectively deal with time-series data. This measure is a non-parametric and computationally efficient approach for measuring the challenge of a given classification problem. Our approach adjusts the annotation boundaries as a way to maximize the class separability using a simple search algorithm.

Through doing so, we find that annotations can be automatically improved as to yield a significantly greater classification performance, although not to the extent of recovering them in totality. Additionally we suggest potential improvements to our method that might be explored in order to improve both its reliability and performance.

## Simulating Errors in Annotations

We assume that the errors made by lay annotators will be normally distributed, and on the whole substantial. The normality assumption is because it is unlikely that there would be sufficient annotations from a given annotator as to identify a general error. In a practical scenario, we presume that the annotations will be of a fixed length – in other words through a constraint in the annotation software – and thereby the only error would be the location of that annotation.

We additionally presume that there would be a small amount of corrected data available to begin with. As such, our model will randomly introduce jitter through a normal distribution (with mean 0, and varying standard deviation (hereon s.t.d) X) a random percentage Y of the annotations and then try to correct the locations of these annotations and demonstrating improvement in performance. Whilst there are other forms of errors – such

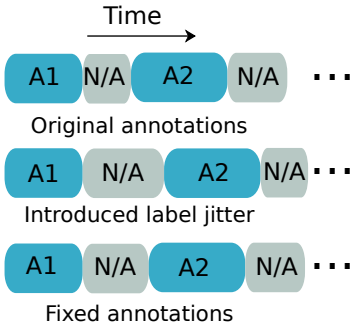


**Figure 1:** The above image provides an example of how the annotations might be adjusted in the three states of the annotation correction process, with the top being the original ground truth annotations, the middle after randomly introduced label jitter/shifting and the bottom after correction through class separability.

as a jitter in all of the annotations – we deliberately constrain this paper to one specific type of error.

Note that in this work, we rely upon each annotation being separated by a null class of events that we are not interested in. We then only adjust the meaningful annotations, as illustrated in Figure 1. We consider this to be realistic for most application scenarios, because there will almost always be a set of activities, which are not relevant to the problem at hand.

## Using Class Separability to Reduce Annotation Errors

Class separability is a parameter free means aimed at determining how challenging a machine learning problem could be. Whilst there are numerous approaches towards doing this, we use the Fukunaga Class Separability [4]. This means computing both the between and within scatter matrices of Linear Discriminant Analysis (where $c$ is the number of classes, the number of instances of the $i$th class is denoted $n_i$, and $\mathbf{m}_i$, $\mathbf{m}$ are the mean vectors for the ith class and the entire dataset respectively):

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \quad (1)$$

$$\mathbf{S}_W = \sum_{i=1}^{c} \left[ \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \mathbf{m}_i)(\mathbf{x}_{ij} - \mathbf{m}_i)^T \right] \quad (2)$$

We then computed the separation using the trace: $Separation = \mathrm{Tr}(\mathbf{S}_B/\mathbf{S}_W)$. To use this in a meaningful fashion for time-series data, we used the 23-dimension statistical feature representation using a sliding window procedure detailed in [8], with the mean, standard deviation, energy and entropy being computed for each of [x,y,z,pitch,roll] with the final three features being the correlation coefficients between channels.

In order to determine whether or not an annotation needs to be shifted, we compute the class separability over the whole dataset for each possible jitter using a grid-search, in sequential order. The result that generates the highest class separability is selected as the adjusted annotation.

There are several intrinsic advantages to this approach compared with simply refining classification performance on a given classifier. The class separability measure we used is similar to LDA (Linear Discriminant Analysis) and it does not require solving for eigen values. The overall computational costs to compute the class separability is linear in the number of instances and quadratic in the number of features. However, in subsequent evaluations, some of the elements in Equation 1 and 2 can be retained. Since it does not rely on pair-wise distances, it is more efficient to compute than classification approaches such as k-NN.

Additionally, the class separability is also parameter free, being a measure on how difficult a given classification problem is, rather than attempting to perform a classification in itself. Class separability as formulated in this paper also assumes no priors, and as a consequence, is not generally dependent on the properties of the data.

## Results

In order to evaluate our method, we conducted some initial experiments on the OPPORTUNITY dataset [9], taking a single subject and the acclerometer in the IMU on the right wrist. We used 180 annotations over a 15 minute time period, with a window length of 0.5 seconds and a shift of 0.25 seconds (as selected in [9] ) involving

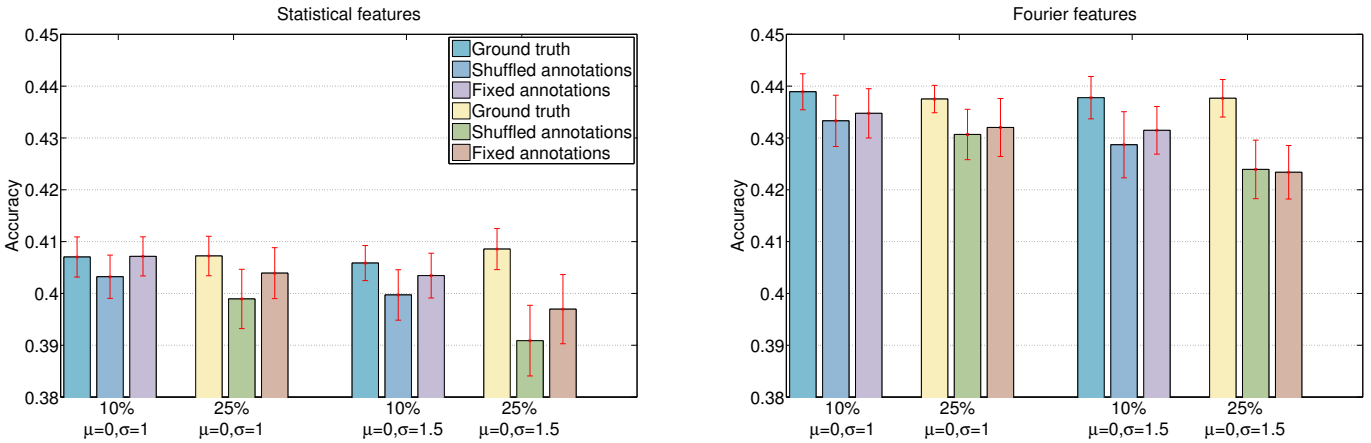


**Figure 2:** The k-NN classification accuracy results with standard error over 20 trials using statistical and Fourier features

18 distinct classes of data. This was intended to simulate the scenario of data being annotated in a single session.

Beginning with an existing ground-truth verified by multiple annotators (thereby ensuring that there is no label jitter in the original ground-truth), we randomly shifted (i.e., introduced label-jitter) 25% and 10% of the annotations using normal distributions with $\mu = 0, \sigma = 1$, and $\mu = 0, \sigma = 1.5$. This had the effect of substantially reducing the performance in all four circumstances. We then applied the foregoing methodology, with a grid of $[-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1]$ in respect of each boundary (left and right), making a total number of 81 possbilities. We did not use a larger grid because our aim was only to correct relatively minor errors, rather than the more ambitious and challenging goal of correcting major errors too, and the spacing was sufficient to ensure that. Because of the sliding window proceedure, we determined that increasing the spacing between annoations would be fruitless exercise, with the spacing specifically chosen to reflect this.

The accuracy results for 20 trials are documented in Figure 2 using k-nearest neighbours $(k = 3)$ with statistical and Fourier features. The change in class separability induced by the algorithm is also documented in Figure 3. We then performed t-tests, finding that the significance level for the improvement (see Table 1) implying statistical signifance of the improvement at the standard thresholds. The 'best' performance in that figure refers to the case where the optimal choice of annotations is selected (i.e., if an improvement is not detected, then we presume that the original annotations are reverted to).

## Discussion and Conclusions

The preliminary work above shows a promising approach towards improving annotations, automatically. One issue is that the class separability was overoptimised with

| | | Statistical | Fourier | Statistical | Fourier | Statistical | Fourier |
|---|---|---|---|---|---|---|---|
| Shifted annotations % | $\mu$,$\sigma$ | Best (t-test) | Best (t-test) | Actual (t-test) | Actual (t-test) | No. of Improvements | No. of Improvements |
| 10% | 0,1.5 | 3.789e-4 | 8.3715e-4 | 0.011 | 0.421 | 15/20 | 11/20 |
| 10% | 0,1 | 7.725e-4 | 0.001 | 0.008 | 0.322 | 13/20 | 10/20 |
| 25% | 0,1.5 | 5.242e-5 | 0.0057 | 0.659e-4 | 0.738 | 15/20 | 9/20 |
| 25% | 0,1 | 0.0012 | 3.27e-4 | 0.042 | 0.03 | 15/20 | 13/20 |

**Table 1:** Results showing the best, actual and the number of improved annotations resulting in superior performance

substantially improved results than the ground truth, as illustrated in Figure 3. Thus, future work should consider refinements in the search process, perhaps by pre-ordering the annotations using a leave one out (l.o.o.) methodology. Moreover, the class separability could potentially be complemented by other methods, for instance the histogram of the energy.

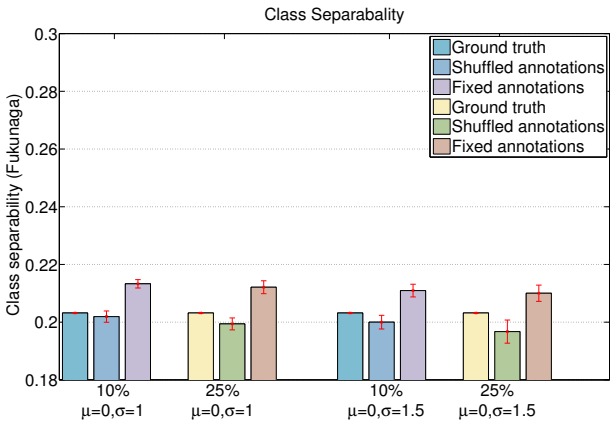


**Figure 3:** Class separability accuracy results with standard error over 20 trials

Another area of investigation is to characterize the nature of annotation errors done by experts and non-experts. This ties in to human factors, as well as HCI aspect, such as the nature of the user interface of the annotation tool that may influence the nature of annotation errors. We applied this method to the OPPORTUNITY dataset which is a standard benchmark dataset of naturalistic human behaviors used in many activity recognition papers. However, future work will apply this technique to additional datasets.

We will also explore whether using the ECDF [5] – which does not rely upon the normality assumption – instead of statistical features, as a time series compliant approach to representing the data for the purposes of the class separability. Additionally, we would seek to explore less common approaches, such as alternative class separability metrics (e.g., [12, 10]) in order to understand whether they may offer greater improvements in performance.

In summary, this approach offers a simple mean to search for improved annotation boundaries. If the approach fail to find better boundaries (i.e., increase classification accuracy), the experimenter is free to rely on the initial annotations. As such, our method is an annotation

pre-processing method, which can be selectively enabled when shown to be beneficial.

## Acknowledgments

This work was partly supported by the RCUK Digital Economy Theme [grant number EP/G066019/1 – SIDE: Social Inclusion through the Digital Economy], the EPSRC [grant number EP/I000755/1 – TEDDI: Transforming Energy Demand through Digital innovation] and the EPSRC DTG. S. Bhattacharya received funding from the Future Internet Graduate School (FIGS) and the Foundation of Nokia Corporation.

## References

[1] Bachrach, Y., Minka, T., Guiver, J., and Graepel, T. How To Grade a Test Without Knowing the Answers A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing. In *Proc. Int. Conf. on Machine Learning* (Edinburgh, UK, 2012).

[2] Bhardwaj, V., Passonneau, R. J., Salleb-Aouissi, A., and Ide, N. Anveshan: a framework for analysis of multiple annotators' labeling behavior. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV, Association for Computational Linguistics (Stroudsburg, PA, USA, 2010), 47–55.

[3] Carletta, J. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.* (1996), 249–254.

[4] Fukunaga, K. *Introduction to Statistical Pattern Recognition*, 2nd ed. 1990.

[5] Hammerla, N., Kirkham, R., Andras, P., and Plötz, T. On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution. *Proceedings ISWC* (2013).

[6] Nguyen-Dinh, L.-V., Roggen, D., Calatroni, A., and Tröster, G. Improving online gesture recognition with template matching methods in accelerometer data. In *Proc 12th Int Conf on Intelligent Systems Design and Applications* (2012), 831–836.

[7] Nguyen-Dinh, L.-V., Waldburger, C., Roggen, D., and Tröster, G. Tagging human activities in video by crowdsourcing. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval* (2013), 263–270.

[8] Plötz, T., Moynihan, P., Pham, C., and Olivier, P. Activity recognition and healthier food preparation. In *Activity Recognition in Pervasive Intelligent Environments*. Atlantis Press, 2011, 313–329.

[9] Roggen, D., et al. Collecting complex activity datasets in highly rich networked sensor environments. In *Proc. 7th Int Conf on Networked Sensing Systems* (June 2010), 233 –240.

[10] Sotoca, J., Sánchez, J., and Mollineda, R. A review of data complexity measures and their applicability to pattern classification problems. *Actas del III Taller Nacional de Mineria de Datos y Aprendizaje.–TAMIDA* (2005), 77–83.

[11] Vondrick, C., and Ramanan, D. Video Annotation and Tracking with Active Learning. In *Neural Information Processing Systems (NIPS)* (2011).

[12] Wang, L. Feature selection with kernel class separability. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (2008), 1534–1546.