Annotation Performance for multi-channel time series HAR Dataset in Logistics

Christopher Reining*, Fernando Moya Rueda†, Friedrich Niemann*, Gernot A. Fink† and Michael ten Hompel*
*Chair of Materials Handling and Warehousing
Email: christopher.reining@tu-dortmund.de
†Pattern Recognition in Embedded Systems Group
Email:fernando.moya@tu-dortmund.de
TU Dortmund University, August-Schmidt-Straße 4, 44227 Dortmund
Telephone: +49 (0)231 / 755-3228

***Abstract*—This contribution proposes an approach for annotating human actions and their coarse semantic descriptions for multichannel time-series. For this purpose, a new dataset that consists of Optical Motion Capturing and IMU time-series data for industrial deployment is created and annotated by 6 individuals. The expenditure of time for labelling, both classes and semantic attributes, and the annotation consistency are examined. The initial annotations are revised by a single domain expert to measure its effect on the overall between-individual consistency. Consistency measurements by means of Cohen's κ are analysed. The results give insights on the effort for dataset creation in the field of Human Activity Recognition for industrial application. The Cohen's κ for consistency assessment was moderate and substantial for the initial annotation, and it increased slightly after revision.**

***Index Terms*—Human Activity Recognition, Labelling, Annotation, Human Reliability, Logistics, Warehousing**

I. Introduction

Manual annotation of multi-channel times series data for Human Activity Recognition (HAR) is a time-consuming and expensive task [1]. The effort scales with the amount of data to annotate. Due to the intra- and inter-class variability of human motion, a large amount of observations from different subjects is necessary for methods of supervised learning [2]. Manual labelling by different annotators is prone to inconsistencies, also referred to as annotation noise [3]. Even the same annotator may label differently when repeating the process. Additional repetitions and corrections are needed to enhance the quality of the dataset, which would further increase the effort. Also, quantifying the human performance with respect to the annotation consistency provides useful information when deploying a classifier.

This annotation effort is further increased when not only activities, but also semantic descriptions are labelled. Recently, attribute representations have been proven successfully for solving HAR [4]–[6]. They show advantages where data are highly unbalanced or training and testing-data sets are disjoint. However, data sets with such representations do not exist yet.

The work on this publication was supported by Deutsche Forschungsgemeinschaft (DFG) in the context of the project Fi799/10-2, HO2463/14-2 "Transfer Learning for Human Activity Recognition in Logistics"

The scope of this contribution is to propose a method for annotating human activities along with their attribute representation, and to quantify both the annotation effort and its consistency when manually labelling multi-channel time series datasets. Therefore, an empirical investigation is conducted based on an optical Motion Capturing (oMoCap) dataset of human activities in warehousing. This oMoCap-dataset is synchronized with Inertial Measurement Unit (IMU) data streams for deployment in a real warehouse.

The remainder of this contribution is structured as follows. In Section II, the novelty value and the research goal of this contribution are outlined. The data set and the applied method for data annotation and its revision is presented in Section III. Next, the results regarding effort assessment and consistency assessment are evaluated in Section IV. A discussion is presented in Section V.

II. Related Work

As pointed out in a recent survey, publications dealing with novel approaches of multichannel time-series HAR rarely mention the annotation effort or its consistency [7]. In most cases, the dataset creation effort is not discussed. In only a few instances, the annotation method is described vaguely. There are only few quantitative statements about effort and consistency of annotations for HAR known to the authors. In [8], it took 26 min to annotate 1 min of order-picking activities from a video that was synchronized with an IMU data stream. All data originated from a real warehouse. Cross and repetition tests revealed a consistency of 74% and 90%, respectively. In [6], warehousing activities were recorded with an oMoCap-System in a laboratory environment. Activities were recorded in modular units and not within a scenario consisting of several process steps. On the one hand, annotation effort was rather low—it took 2.5 min per recorded minute, as some handling activities included approaching a stack of boxes, so annotating *Walking* was necessary. On the other hand, the classification performance was rather poor for some activities and attributes. Annotating the prominent OPPORTUNITY dataset took $7-10$ hours per 30 min of video [9]. The annotators were asked to label the activities in a human $3D$-Model of this dataset.

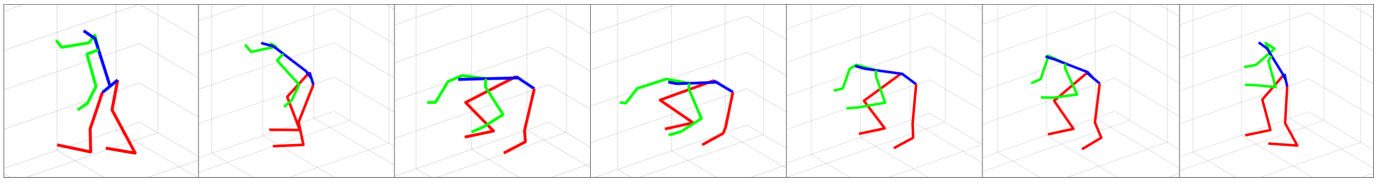

Fig. 1. Visualization of the MoCap skeleton performing the activity *Handling (downwards).*

Provided with a list of 11 activities, they reached an average accuracy of 56% [10].

Recently, semantic descriptions of activities have been used for solving HAR [6]. These, so called, attributes are useful when datasets are highly unbalanced, training and testing are disjoints, and for zero-shot learning problems. Random and expert-given attribute representations show a similar and superior performance than directly classifying human actions. However, these representations are created post-annotation, relating each activity class with a set of defined attributes; that is, they are not directly annotated.

Apart from HAR, the annotation of training data plays a major role in other domains of pattern recognition. For instance, [11] reports that the manual annotation of $27hours$ of video for car tracking took about eight person months. Annotating the SUN RGB-D dataset took $2,051$ hours for $10,335$ RGB-D images [12].

In [13], a description on the types of annotation and a novel method of model based annotation is presented, which can allow reasoning behind hidden properties and causal relations in the annotation. The consistency of the annotations set by different annotators was measured with Cohen's κ [14], and the Krippendorff's α.

Summarizing, the presented state of the art reveals that considering the data set creation effort and consistency is crucial for proper deployment of supervised learning. Dataset creation may require more effort than the deployment of a classifier. For HAR from multi-channel time series data using semantic attributes, an empirical study to assess effort and consistency has not yet been carried out.

III. Method

In the DFG-Project 'Transfer Learning for Human Activity Recognition in Logistics', a dataset of human warehousing activities is created in the 'Innovationlab Hybrid Services in Logistics' at the TU Dortmund University [15]. Data are captured synchronously using oMoCap, IMUs and RGB-D streams. This contribution focuses solely on the oMoCap data as, for real applications, identities of individuals must not be recorded, but it enables annotation by visualizing a skeleton. The software provides global poses of body segments and joints. A pose is a combination of position and angular values in [X,Y,Z] coordinates. All data are recorded with 200 fps. Short recording units of $120s$ are deployed for helping with the synchronization of the IMU recordings through a specific subject's movement as trigger signal. This results in a total of $24,000$ frames per recording.

A list of activities and attributes was defined by domain experts. The semantic meaning and examples of proper annotations are explained in annotation guidelines. There are eight activity classes, namely *Standing*, *Walking*, *Moving Cart*, *Handling (upwards)*, *Handling (centred)*, *Handling (downwards)*, *Synchronization* and *None*. The *Synchronization* class is used for synchronizing the other data sources with oMoCap time series. *None* identifies frames that shall not be used for training because of heavy noise or errors in the readings.

17 semantic attributes are separated into four major groups:

i Leg Motion: Gait Cycle, Step, Standing Still
ii Upper Body Motion: Upwards, Centred, Downwards, No Intentional Motion, Torso Rotation
iii Handedness: Right, Left, No Arms
iv Item Pose: Handy Unit, Bulky Unit, Utility-Auxiliary (e.g., a knife or adhesive tapes), Cart, Computer, No Item

A Python-Tool has been created for annotation. It loads oMoCap-data for the skeleton visualization to make the activities apparent, as seen in Figure 1. Each annotator was tasked to set the initial and final frame of a window that corresponds with the starting and ending of an activity. He or she then proceeds to pick one activity and, if the class chosen is not *None*, at least one attribute per group. There is no fixed assignment of an attribute representation incorporated in the tool. Thus, the annotators are free to choose combinations of activities and attributes they find appropriate. The only two exceptions are the *Synchronization* and *None* classes. The former is assigned a fixed representation while the latter one is assigned an $18th$ attribute of the same name. By the end of this process the initial annotation is created.

The human annotators are given a documentation sheet. First, they are asked to estimate their experience with respect to labelling data in general, and for HAR specifically. The sheet provides to the annotators a list of recordings they are expected to process. The beginning and the end of the annotation time can be entered by clicking a button. It is also possible to enter break times so that the total time spent on annotating is calculated. Remarks by the annotator can be entered in a separate field. The log's information may be helpful to trace errors found in the annotation tool or the guidelines. Furthermore, it is important to know if long excerpts of a recording cannot be annotated properly, resulting in the *None* class. This is because using the *None* class is expected to drastically decrease the expenditure.

Since the initial annotation is prone to errors, it is followed by an iterative revision process. The attribute representation helps to revise the annotated activity classes in a simple

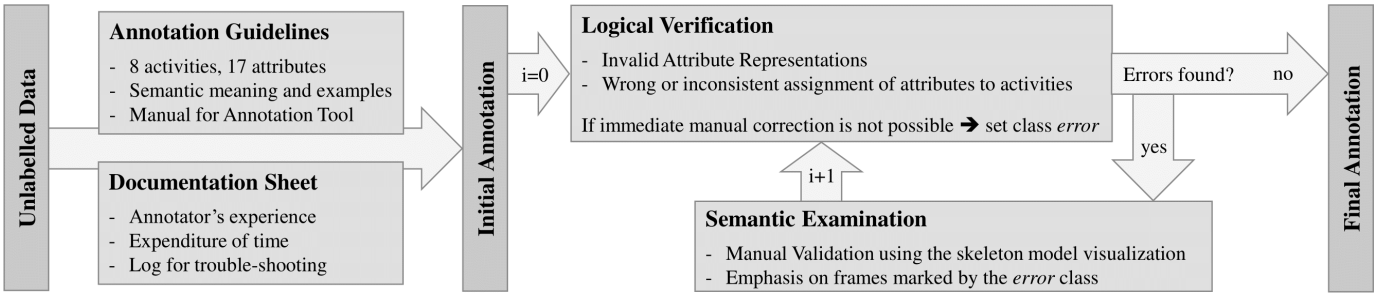

Fig. 2. Method for Annotation and Revision

manner. For that, the first step of each iteration is the logical verification that focuses on two kinds of errors. First, an attribute representation can be invalid for reasons such as missing or conflicting attributes. Second, the same attribute representation must not be assigned to two different classes on any occasion. Each logical error where there is no obvious solution apparent without considering the visualization, is assigned an auxiliary attribute called *Error*. The second step is the semantic examination. The visualization and labels are examined by a human. This step is guided by the $19th$ attribute called *Error*, which is highlighted in the tool's interface as well as the comments from the annotators listed in the documentation sheet. Logical and semantic errors are fixed manually. Since the alteration of the labels may add new logical errors and inconsistencies, the process starts anew with the logical verification. This iterations proceed until no more logical flaws are detected and the experts examining the annotated data agree with the labels. The entire method is illustrated in Figure 2.

Since all data for this publication originates from the same recording session laboratory set-up, there are five identifiers necessary to make the naming of each recording unambiguous:

- S - Subject: The Individual who has been recorded.
- R - Recording: The number of the recording unit of 120 s each within the recording session
- A - Annotator: The individual annotator
- N - Annotation Run: Count of the number of times that one annotator has annotated the same recording
- I - Revision: Count of revision runs (logical and semantic)

In total, the dataset will contain $R = 30$ recordings of $120s$ from 14 subjects (S) in three different scenarios, including 8 activities and 18 semantic attributes.

IV. Evaluation

We evaluated the annotation by means of the annotation time and consistency. The effort assessment quantifies the time spent on annotating a subset of the aforementioned dataset. The human annotation consistency examines the question - to what degree do the individuals agree on the labels. Six annotators (A) were asked to participate in an annotation study by labelling 30 oMoCap-recordings from nine subjects (S).

Assessing the documentation sheet revealed that $A4$ has little experience in labelling but none with respect to HAR, while $A6$ has a lot of experience at labelling in general as well as in HAR in specific. The remaining annotators did not have any experience with respect to dataset creation for supervised learning. $A1$, $A3$, $A4$ and $A6$ were scientists involved in creating the laboratory set-up, setting up the sensors and creating both the annotation guidelines and the tool. $A2$ and $A5$ joined the project when support for annotation became necessary.

A. Effort Assessment

The following questions regarding the expenditure for manual annotation will be answered by the method:

- What is the mean time and its standard deviation for annotating an oMoCap-recording of $120s$ / $24,000$ frames using both activity classes and semantic attributes?
- How much does the annotation time differ among the annotators? Is there a link to their prior experience?
- Does the annotation time per recording decrease over the course of several runs, hinting towards a learning effect?

Once the annotation is done and the documentation sheet is filled out by each annotator in accordance to the method described in section III, the aforementioned questions can be addressed.

When assessing the time spent on annotation, the *None* class has to be considered specifically. Long windows of this class may be necessary due to errors in the sensors readings, but they heavily influence the time spent on annotating.

Table I shows the number of recordings processed by each annotator, the total and mean time spent per recording. The amount of recordings each annotator processed varies as they were not given the same amount of recordings. Few assigned recordings were impossible to process and thus scraped. At this point, each recording was labelled once by an annotator.

$A6$, the annotator with the most experience, annotated the fastest. $A4$, the annotator with little experience, was annotating at a comparable rate to $A1$ and $A3$, who were also aware of the annotation pipeline beforehand. In contrast, $A2$ and $A5$ took more effort to annotate their recordings as they joined the project when the annotation process began.

The variance in annotation time is due to the properties of each recording. Apart from unusable data, which is dealt with by using the *None* class, there are frequent transitions between activities. For example, a recording that mainly

TABLE I

TOTAL NUMBER OF ANNOTATED RECORDINGS PER ANNOTATOR, TOTAL TIME AND MEAN TIME PER RECORDING. TIME IS GIVEN IN [HH:MM].

$N = 1$	Effort Assessment		
	Rec.	Total time[h]	Time per Rec.[h]
A1	38	$49:18$	$01:17 \pm 00:24$
A2	37	$72:06$	$01:56 \pm 00:34$
A3	36	$48:55$	$01:21 \pm 00:30$
A4	28	$37:19$	$01:19 \pm 00:37$
A5	30	$84:18$	$02:48 \pm 01:19$
A6	38	$23:11$	$00:36 \pm 00:17$
All	207	$303:27$	$01:25 \pm 00:57$

consists of the subject walking with the cart can be annotated faster than multi-stage handling processes. Since all annotators processed recordings that did not overlap, the mean values cannot be generalized entirely at this point. It is possible that the annotation time per recording would have deviated if they had been distributed differently. Still, 207 recordings and 303 hours spent on manual annotation provide an initial insight.

In Figure 3, the time spent by $A1$ for annotating 30 recordings from the same subject is illustrated. While the first annotation took the longest, a training effect cannot be observed. The same applies to the remaining annotators.

Among all annotators, a major influence on the annotation time per recording are the proportions of *Walking*, *Moving Cart* and *Standing* activities. This is because their attribute representations are simple to determine and their execution spans over a long period of time compared to *Handling* activities. Another reason is the problem with the sensor readings. For example, the drop in $R24$ and $R29$, see Figure 3, is due to problems with sensor readings that made most of the data impossible to interpret. Thus, the *None*-Class dominated.

B. Human annotation consistency

Considering the high frame rate and human mistakes, a frame-wise agreement on the labels seems unrealistic when annotating the same recording twice—neither when the labels are set by two different individuals nor by the same person. Thus, assessing the human annotation consistency must consider both between and within-subjects consistency of annotation. This contribution proposes to compute the Cohen's κ for this purpose [13], [14]. In contrast to performance metrics such as accuracy or the F1-measure, Cohen's κ emphasises agreement and it does not consider one annotation as the ground truth. Cohen's κ is defined as $\kappa_{A_a,A_b} = \frac{Pr(a)_{A_a,A_b} - Pr(e)_{A_a,A_b}}{1 - Pr(e)_{A_a,A_b}}$, where $Pr(a)$ represents the actual observed agreement among two annotation runs A by the same or two different individuals; this is referred to as the accuracy, the number samples that both annotators agree divided by the total number of samples. $Pr(e)$ is the expected chance agreement, and it used to take into consideration that two annotators may have guessed the same label by chance. It is defined as $Pr(e)_{A_a,A_b} = \frac{1}{M^2} \sum_C A_a^c \dot{A}_b^c$, where M is the number of samples, and C the activity class or attribute, A_a^c the number of times annotator a predicted class c.

To acquire the necessary data, 4 out of the 6 annotators—following the results regarding the annotation time—, are asked to annotate 6 recordings (R) of 3 individual subjects (S); that means 2 recordings or 4 min per subject, and 96 min of recordings in total. The annotators performed the labelling twice per recording ($N = 1, N = 2$).

Based on this, Cohen's κ metric is computed pair-wise for both, the between-individual cross-test as well as the within-individual repetition test. This is done for all classes and semantic attributes. First, the between-consistency is computed for the initial annotation ($N = 1$) of each of the 4 annotators for all 6 recordings. Next, the annotator $A1$ is asked to revise all 24 annotations ($I = 1$) according to the pipeline described in section III.

First, evaluating Cohen's κ for each class and attribute helps to reveal ambiguities concerning the rules for proper annotation according to the guidelines. Second, comparing the between-consistency of revision $I = 0$ and $I = 1$ may reveal a boost in consistency when labels are set by different individuals, but revised by a single annotator. Third, one can compare the between-consistency for $I = 1$ among all annotators with the within-consistency of $A1$, the person who revised the annotations. If the values come close to each other, this hints that it is not necessary that the same annotator sets all labels himself to ensure a high consistency. Instead, an initial annotation by a crowd of annotators that is finally revised by the same person may be sufficient. Fourth, the influence of the time spent on annotation on the between-consistency is of interest. For example, one may ask, whether a person who annotates rather fast agrees with a slower annotator. Also, the effort for revision may vary depending on the time consumed by initial annotation.

In table II, the within-consistency Cohen's κ was computed for the initial annotation of two recordings for 3 subjects. The given values are based on a total of $6 \cdot 2min = 12min$ recording data per annotator. At this point, solely the classes are taken into account.

TABLE II

WITHIN-CONSISTENCY κ FOR INITIAL ACTIVITY CLASS ANNOTATION OF $R1$ AND $R2$ ($2 \cdot 120sec = 240sec$) FROM SUBJECTS $7, 8, 9$ ($I = 0$).

$N = 1$ / $N = 2$	Within-consistency κ			
	Subject			
A	7	8	9	Mean
A1	0.75	0.75	0.81	0.77 ± 0.04
A2	0.64	0.64	0.73	0.67 ± 0.05
A3	0.71	0.94	0.76	0.80 ± 0.12
A6	0.83	0.82	0.67	0.78 ± 0.09

In general, the Cohen's κ values lie within a small interval and hint towards a strong within-consistency—for moderate agreement $0.41 \leq \kappa \leq 0.60$, and for substantial agreement $0.61 \leq \kappa \leq 0.80$ [14]. Interestingly, $A2$ spent the most time on annotation, but achieves the lowest within-consistency score. The highest value of $\kappa = 0.94$ was achieved by $A3$ for the two recordings of $S8$. This is due to the high amount *None*-class labels set during both annotation runs. $A3$ was the only

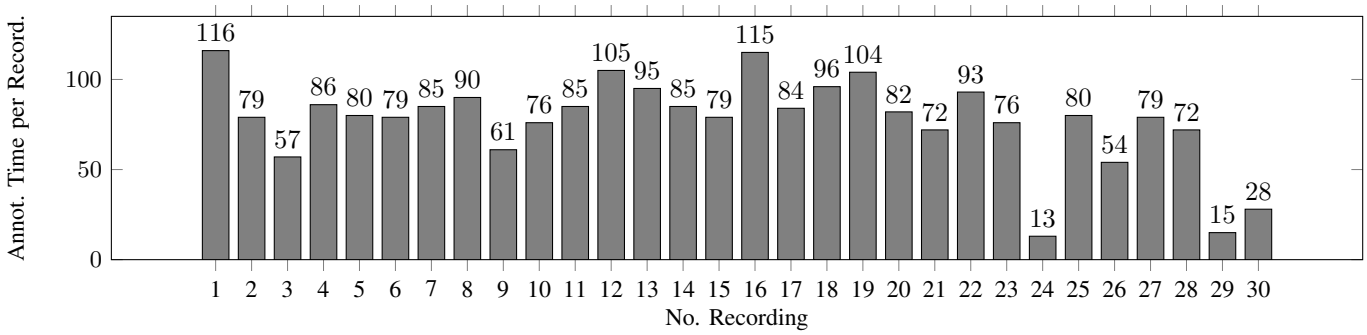

Fig. 3. Time $[min]$ spent by $A1$ for annotating a total of 30 recordings from the same subject.

annotator considering the visualization as incomprehensible due to noise.

Table III ($I = 0$) illustrates the between-consistency based on the same recordings. Solely the first annotation run is taken into account ($N = 1$). The non-revised values ($I = 0$) are shown first in each cell. In general, the values tend to be slightly lower than for the within-consistency hinting towards a "weak" to "moderate" agreement [14]. The low κ-values of $0.19 - 0.22$ of $A3$ for $S8$ are due to the high proportion of *None*-class labels. The agreement between $A6$ and the other annotators did not suffer from the quickly executed annotation.

The revision of all annotations was conducted by $A1$ and revealed the following issues. The revision of the labels set by $A6$ consumed more time than the other revisions. $A6$ considered activity classes of longer duration, mixing different attributes in a unique attribute representation. The revision of the labels set by $A1$ himself was the fastest. $A1$ found no systematic violations of the annotation guidelines by the other annotators with respect to the activity class labels.

Table III ($I = 1$) gives the between-consistency values after all recordings were revised—second value in each cell. Between-consistencies of regards $A2$ and $A3$ improved after the correction of *None* class. Nevertheless, there is no considerable improvements with respect to the other annotators. As the semantic examination in the following revision iterations ($I + 1$) solely focused on the error class, there were minimal changes of the annotations after the revisions.

Table IV shows the between-consistency κ of the classes in the initial annotation $I = 0$ among annotator pairs in more detail. Synchronization class shows the best consistency as it was the only activity class with fixed guidelines. Class None contains either zeroes or no values as $A3$ was the only one annotating such class. In general, activity classes *Walking*, *Moving Cart* and *Handling (downwards)* are substantially reliable. Activity *Handling (upwards)* is moderate reliable. The activities *Standing* and *Handling (center)* show less consistency among the annotators. This is as these classes tend to be mixed in the annotation.

Table IV also presents the between-consistency κ after the revision $I = 1$. Revision improved the consistency of activity classes *Handling (upwards)* and *Handling (center)*, and set *Standing* class as moderately reliable.

The evaluation of the between-consistency per attribute for the initial annotation $I = 0$ is given in table V. The κ-values vary widely within a span from 0.01 to 0.92. Seemingly, there was no common understanding among the annotators about how to use *Torso Rotation* attribute. Leaving out the *Torso Rotation*, the lowest value of 0.3 is provided by the *Centered* attribute. It refers to handling activity in a centered pose. This attribute is the most used of all which heavily influences the chance agreement and thus the κ-value.

In general, the attributes referring to Item Poses yield the highest agreement. Annotators are consistent with the use of the attributes *Utility-Auxiliary* and *Computer*. These attributes were not annotated, not even by error as they do not belong to the annotated scenarios. The overall best value was achieved by the *Handy Unit* attribute, which refers to a pose that is taken when handling a *Bulky Unit* such as a small load carrier.

During the revision process, several findings were made with respect to ambiguous utilization of attribute labels. They support the findings of table V ($I = 0$), meaning that those attributes with a low agreement required the most revision effort. Apart from that, there have been smaller violations of the annotation guidelines, in particular by $A6$. Nevertheless, there were no systematic violations among all annotators except for attribute *Torso Rotation*.

The evaluation of the between-consistency per attribute for the revised annotation is given in table V ($I = 1$). After revision, the consistency of attributes improved in general. However, this improvement is rather small. In addition, the attributes *Step* and *Torso Rotation* showed moderate and weak consistency.

V. Discussion and Conclusion

This contribution proposed and evaluated an approach for labelling multichannel time-series for synchronized MoCap and IMU datasets for HAR including annotation of semantic attributes. The results of this contribution originate from an excerpt from an extensive recording sessions in a laboratory set-up based on real warehousing scenarios. Activity class labels and semantic attributes were annotated using the MoCap skeleton visualization. Two aspects were evaluated in specific—the expenditure of time and the consistency of the annotations.

TABLE III

BETWEEN-CONSISTENCY κ FOR $I = 0/I = 1$ ACTIVITY CLASS ANNOTATION OF RECORDINGS 1 AND 2 ($240s$) FROM SUBJECTS $7, 8, 9$ ($I = 0$). SUBSTANTIAL AGREEMENTS ARE HIGHLIGHTED IN BOLD.

$N = 1$	Between-consistency κ			
	Subject $I = 0/I = 1$			
A	7	8	9	Mean
A1-A2	**0.72/0.74**	0.54/0.53	0.58/**0.64**	**0.61 ± 0.09 / 0.64 ± 0.11**
A1-A3	**0.71/0.67**	0.22/0.60	**0.65**/0.59	0.53 ± 0.26/**0.62 ± 0.04**
A1-A6	**0.75/0.81**	**0.74/0.76**	0.59/**0.77**	**0.69 ± 0.09/0.78 ± 0.03**
A2-A3	**0.66**/0.60	0.25/**0.62**	**0.62/0.73**	0.51 ± 0.23 / **0.65 ± 0.07**
A2-A6	**0.64/0.67**	**0.61**/0.55	0.53/**0.75**	0.59 ± 0.06 / **0.65 ± 0.10**
A3-A6	**0.62/0.68**	0.21/**0.66**	0.49/**0.72**	0.44 ± 0.21 / **0.69 ± 0.03**

TABLE IV

BETWEEN-CONSISTENCY κ FROM THE CLASSES IN THE INITIAL AND REVISED ANNOTATION ($I = 0/I = 1$). SUBSTANTIAL AGREEMENTS ARE HIGHLIGHTED IN BOLD.

	Stand.	Walk.	Mov. Cart	Handl. up.	Handl. cen.	Handl. down.	Synchron.	None
A1-A2	0.32/0.53	**0.65/0.74**	**0.77/0.81**	**0.65/0.71**	0.56/**0.63**	**0.73/0.67**	**0.69/0.82**	-
A1-A3	0.24/0.56	**0.74/0.76**	**0.79/0.82**	**0.61/0.76**	0.41/0.56	**0.62**/0.56	**0.96/0.97**	0
A1-A6	**0.69/0.73**	**0.71/0.83**	**0.77/0.87**	**0.65/0.75**	**0.64/0.76**	**0.76/0.86**	**0.98/0.97**	-
A2-A3	0.48/0.59	**0.85/0.80**	**0.83/0.82**	0.39/**0.69**	0.44/0.60	**0.70/0.64**	**0.70/0.82**	0
A2-A6	0.32/0.52	**0.64/0.76**	**0.77/0.80**	**0.62/0.64**	**0.65/0.61**	**0.82/0.73**	**0.69/0.81**	-
A3-A6	0.19/0.54	**0.63/0.83**	**0.75/0.85**	0.46/**0.78**	0.42/0.60	**0.70/0.64**	**0.96/0.97**	0

TABLE V

BETWEEN-CONSISTENCY κ FROM THE ATTRIBUTE REPRESENTATION FOR ANNOTATION ($I = 0$ / $I = 1$). SUBSTANTIAL AGREEMENTS ARE HIGHLIGHTED IN BOLD.

$I = 0/I = 1$	Between-consistency κ					
	Annotator Pairs					
Attribute	A1-A2	A1-A3	A1-A6	A2-A3	A2-A6	A3-A6
Gait Cycle	**0.69/0.68**	**0.64/0.75**	**0.67/0.70**	**0.83/0.76**	**0.69/0.74**	**0.69/0.84**
Step	0.53/0.54	0.47/0.37	0.41/0.45	0.57/0.5	0.47/0.52	0.45/0.51
Standing Still	**0.69/0.71**	**0.67/0.61**	**0.62/0.64**	**0.65/0.71**	**0.64/0.66**	0.50/0.56
Upwards	**0.67/0.79**	**0.77/0.70**	**0.75/0.83**	0.45/**0.72**	**0.63/0.72**	0.57/**0.80**
Centered	0.56/0.58	0.53/0.43	0.54/**0.69**	0.43/0.58	0.50/**0.62**	0.29/**0.62**
Downwards	**0.70**/0.55	0.56/0.60	**0.76/0.81**	**0.70**/0.57	**0.80/0.72**	**0.68/0.70**
No Int. Motion	0.53/**0.64**	**0.62**/0.57	**0.71/0.72**	**0.62/0.71**	**0.63/0.69**	0.52/**0.67**
Torso Rotation	0.28/0.41	0.17/0.18	0.40/0.48	0.01/0.09	0.16/0.28	0.11/0.17
Right	**0.67/0.70**	**0.71**/0.50	**0.72/0.71**	0.50/**0.80**	**0.68/0.74**	0.55/**0.72**
Left	**0.67/0.72**	**0.70/0.64**	**0.66/0.68**	**0.68/0.75**	**0.68/0.74**	**0.63/0.67**
No Arms	**0.66/0.80**	**0.83/0.74**	**0.83/0.78**	**0.79/0.90**	**0.66/0.81**	**0.65/0.82**
Handy Unit	**0.77/0.77**	**0.72/0.72**	**0.74/0.72**	**0.93/0.90**	**0.91/0.90**	**0.92/0.88**
Bulky Unit	**0.81/0.80**	**0.77/0.66**	**0.76/0.74**	**0.70/0.89**	**0.87/0.85**	**0.70/0.83**
Utility-Aux.	-	-	-	-	-	-
Cart	**0.82/0.82**	**0.80/0.74**	**0.83/0.82**	**0.82/0.89**	**0.87/0.89**	**0.79/0.87**
Computer	-	-	-	-	-	-
No Item	**0.82/0.84**	**0.87/0.80**	**0.86/0.82**	**0.77/0.92**	**0.83/0.85**	**0.72/0.85**

While the annotation time differed among 6 human annotators, it took on average $85min$ per $2min$ of recorded data. The results confirm previous findings in this field, but they are new with respect to the impact of annotating attributes.

Within and Between-consistency based on the Cohen's κ showed that annotators are moderately consistent when labelling recordings in the first trial. Within-consistency is higher than between-consistency as minor disagreement among annotators are present. After an annotation revision by single-domain expert, the consistency for activity classes and attributes improved. However, this improvement is rather small in comparison to the revision effort. This shows that more strict and clear guidelines for the first annotation shall be considered.

For future work, the authors plan to publish the entire dataset that consists of more recordings and more scenarios than incorporated in this contribution. The data will be labelled by a crowd of annotators and revised by few domain experts as this approach has been proven to give a high consistency while allowing for a fast labelling process.

REFERENCES

[1] K. Yordanova, A. Paiement, M. Schröder, E. Tonkin, P. Woznowski, C. M. Olsson, J. Rafferty, and T. Sztyler, "Challenges in Annotation of useR Data for UbiquitOUs Systems: Results from the 1st ARDUOUS Workshop," *arXiv:1803.05843 [cs]*, Mar. 2018.

[2] F. Ordóñez and D. Roggen, "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition," *Sensors*, vol. 16, no. 1, p. 115, Jan. 2016.

[3] L.-V. Nguyen-Dinh, A. Calatroni, and G. Tröster, "Robust Online Gesture Recognition with Crowdsourced Annotations," in *Gesture Recognition*, S. Escalera, I. Guyon, and V. Athitsos, Eds. Cham: Springer International Publishing, 2017, pp. 503–537.

[4] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-Based Classification for Zero-Shot Visual Object Categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, Mar. 2014.

[5] E. Rusakov, L. Rothacker, H. Mo, and G. A. Fink, "A Probabilistic Retrieval Model for Word Spotting Based on Direct Attribute Prediction," in *Proc. Int. Conf. on Frontiers in Handwriting Recognition*, Niagara Falls, USA, 2018.

[6] C. Reining, M. Schlangen, L. Hissmann, M. ten Hompel, F. Moya, and G. A. Fink, "Attribute Representation for Human Activity Recognition of Manual Order Picking Activities," in *Proceedings of the 5th International Workshop on Sensor-Based Activity Recognition and Interaction - iWOAR '18*. Berlin, Germany: ACM Press, 2018, pp. 1–10.

[7] C. Reining, F. Niemann, F. Moya Rueda, G. A. Fink, and M. ten Hompel, "Human Activity Recognition for Production and Logistics—A Systematic Literature Review," *Information*, vol. 10, no. 8, p. 245, Jul. 2019.

[8] S. Feldhorst, S. Aniol, and M. ten Hompel, "Human Activity Recognition in der Kommissionierung – Charakterisierung des Kommissionierprozesses als Ausgangsbasis für die Methodenentwicklung," *Logistics Journal : Proceedings*, vol. 2016, no. 10, Oct. 2016.

[9] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. d. R. Millàn, "Collecting complex activity datasets in highly rich networked sensor environments," in *2010 Seventh International Conference on Networked Sensing Systems (INSS)*, Jun. 2010, pp. 233–240.

[10] M. Ciliberto, D. Roggen, and F. J. O. Morales, "Exploring human activity annotation using a privacy preserving 3D model," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct - UbiComp '16*. Heidelberg, Germany: ACM Press, 2016, pp. 803–812.

[11] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently Scaling up Crowdsourced Video Annotation," *Int J Comput Vis*, vol. 101, no. 1, pp. 184–204, Jan. 2013.

[12] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, Jun. 2015, pp. 567–576.

[13] K. Yordanova and F. Krüger, "Creating and Exploring Semantic Annotation for Behaviour Analysis," *Sensors (Basel)*, vol. 18, no. 9, Aug. 2018.

[14] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochem Med*, pp. 276–282, 2012.

[15] A. K. R. Venkatapathy, H. Bayhan, F. Zeidler, and M. ten Hompel, "Human machine synergies in intra-logistics: Creating a hybrid network for research and technologies," in *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sep. 2017, pp. 1065–1068.