Attribute Representation for Human Activity Recognition of Manual Order Picking Activities

Christopher Reining
Michelle Schlangen
Leon Hissmann
Michael ten Hompel
christopher.reining@tu-dortmund.de
michelle.schlangen@tu-dortmund.de
leon.hissmann@tu-dortmund.de
michael.tenhompel@tu-dortmund.de
Chair of Materials Handling and Warehousing, Technical University of Dortmund
Dortmund, Germany

Fernando Moya
Gernot A. Fink
fernando.moya@tu-dortmund.de
gernot.fink@tu-dortmund.de
Pattern Recognition in Embedded Systems Group,
Technical University of Dortmund
Dortmund, Germany

ABSTRACT

Semantic descriptions or attribute representations have been used successfully for object and scene recognition, and for word-spotting. However, these representations have not been explored deeply on human activity recognition (HAR). Particularly, in the manual order picking process, attribute representations are beneficial for dealing with the versatility of activities in the process. This paper compares the performance of deep architectures trained using different attribute representations for HAR. Besides, it evaluates their quality from the perspective of practical application.

CCS CONCEPTS

• **Human-centered computing → Laboratory experiments**; Empirical studies in HCI; • **Computing methodologies → Neural networks**; *Batch learning*; Online learning settings;

KEYWORDS

Human Activity Recognition (HAR), Deep Learning, Attribute-based Representation, Order Picking, Motion Capturing

ACM Reference Format:
Christopher Reining, Michelle Schlangen, Leon Hissmann, Michael ten Hompel, Fernando Moya, and Gernot A. Fink. 2018. Attribute Representation for Human Activity Recognition of Manual Order Picking Activities. In *5th international Workshop on Sensor-based Activity Recognition and Interaction (iWOAR '18), September 20–21, 2018, Berlin, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3266157.3266214

1 INTRODUCTION

Order picking is the process of taking and collecting articles in a specified quantity to fulfill costumer demands. Order picking efficiency is crucial for the success of an entire supply chain [8, 14, p.13-30]. The associated manual processes need to be quantitatively determinable to allow for their proper planning, assessment and optimization. Recently, Human Activity Recognition (HAR) was used for analyzing this process. HAR associates activity labels to segments of multichannel time-series from sensors' measurements. However, this approach can hardly deal with the versatility of human activities in order picking. Human activities share a considerable amount of patterns. For example, an employee can carry a box while walking or simply walk. Different items such as boxes or articles can be picked with the left hand, the right hand or both. Adding further classes for each variant of similar activities results in immense annotation effort [5]. Beyond that, the desired definition of activities can differ with regards to the use case of warehousing. It needs to be adaptable and enable a posterior alteration of the activity definition even after data recording and annotation is concluded. An adaptable activity definition demands a high-level concept to semantically describe human activities in order picking.

Deep convolutional neural networks (CNNs) have been used successfully to analyze and to recognize human activities [7, 9, 15–17]. A CNN for HAR processes time-series

iWOAR '18, September 20–21, 2018, Berlin, Germany
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6487-4/18/09...$15.00
https://doi.org/10.1145/3266157.3266214

by using convolution and downsampling operations for extracting relevant features from raw measurements, e.g., from Inertial Measurement Units (IMUs) [7].

In addition, [2, 3, 12, 17] suggest that attribute-based representations of human activities are beneficial for HAR . Each activity is represented by a set of attributes that semantically and coarsely describe it. These attributes are for instance simple human movements and poses like moving the left foot or right foot, forward, and upright, which can be considered as "walking". Attributes serve as an intermediate layer between sensor measurements and activity labels. The definition of activities is not fixed and can be adjusted with regards to the application demands.

Attribute representations are suitable for cases where data is unbalanced, i.e., the number of sequences differs strongly per class. Besides, they are appropriate for zero-shot learning, which extends supervised learning to classify unseen classes at training. It allows for recognizing unseen activities by using familiar descriptions that are shared with known activities [10, 11, 24]. For example, activities like "walking" or "running" could share attributes related to the feet-movement; even though, they differ on the speed.

In [17], CNNs were deployed for computing attributes of human activities. As attribute annotations do not exist, these were found using an evolutionary algorithm starting from a random representation. However, they do not hold specifically any semantic descriptions of human activities. This is a major issue for deploying them on applications, e.g., human activities in order picking. Besides, they cannot be used to describe unseen activities. Attribute representations that have been provided by a domain expert from warehousing may be semantically clear. Thus, they might be usable to combine for recognizing new activities. However, the performance of representations provided by domain experts is unknown yet. There is no set of semantic attributes to describe order picking activities available at this point.

The goal of this paper is to compare different annotated attribute-representations and their performance in HAR for order picking activities. To address this issue, an order picking dataset from a motion capturing system along with competing attribute representations is utilized.

The remainder of this paper is structured as follows. Section 2 provides an insight on related work and underlines the novelty value of this contribution. Next, an order picking scenario as well as the seen and unseen activities are presented in section 3. In section 4, the proposed method to define and to compare attribute representations using deep learning is outlined. The quantitative and qualitative evaluation of the results follows in section 5. This contribution concludes with a discussion in section 6.

2 RELATED WORK

Deep convolutional neural networks (CNN) and recurrent neural networks (RNNs) have been shown to succeed in recognizing human activities on multichannel time-series, e.g., from IMUs [7, 9, 15, 17]. CNNs conveniently combine the learning of features and the classifier in an end-to-end manner directly from raw data. The authors in [16] proposed a CNN with convolutional layers, which are applied along the temporal axis and over all the sensor measurements from IMU's data. In [21], these convolutions were applied to individual sensor measurements where filters were shared among the sensors. In [15], the authors combined convolutional layers and recurrent units for HAR on activities of daily living (ADL). In [9], three deep architectures, namely a CNN, a long-short term memory (LSTM) network, and a bidirectional LSTM network, were applied to classify human-locomotion activities. The authors of [17] proposed an attribute-based representation for HAR. They utilized an evolutionary algorithm for deriving a suitable attribute representation for HAR starting from a random one. By using this representation, a comparable or even better performance contrasted to similar networks was achieved.

In the context of order picking, the authors in [6] processed raw measurements from IMUs for analyzing human movements. IMUs were attached to three workers during field experiments in two operating warehouses. Each of these units obtain measures from three different sensors, an accelerometer, a gyroscope and a magnetometer. The authors followed a standard approach in pattern recognition. They segmented sequences by means of a sliding-window approach, computed a set of handcrafted statistical-features, and trained a classifier. Specifically, they used a Support Vector Machine (SVM), a Bayes and a Random Forest as classifiers. The authors in [7] proposed a CNN for solving HAR in the order picking process. This architecture processes time-series segments through a stack of temporal-convolutions, pooling operations and fully-connected layers computing pseudo-probabilities of human actions. In contrast to previous architectures, this CNN contains parallel branches, one per IMU. Each of these branches is composed of two or three temporal-convolution layers and max-pooling operations processing segments per IMU. This architecture, called IMU-CNN, showed the state-of-the-art performance.

Attribute representations have been successfully used for image classification, scene recognition and word spotting [1, 3, 10–12, 18]. The authors in [10, 11] targeted attribute-based classification for detecting and recognizing objects in images based on their semantic descriptions, specifically on objects that are not used for training. The authors in [12] used high-level semantic concepts to create more descriptive models for human activity recognition, therefore evaluating video

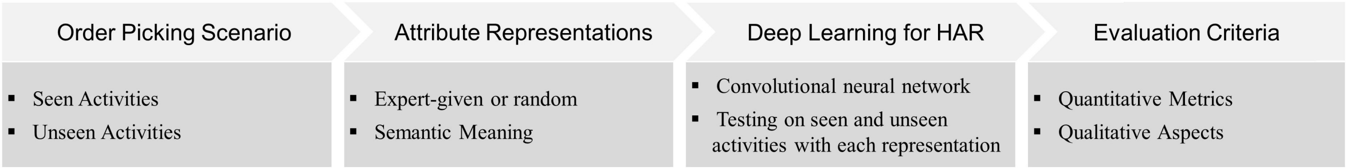


Figure 1: Four-step pipeline for evaluation of attribute-based activity representations in order picking scenarios

footage from daily or exercise related human activities. In this context, the suitability for zero-shot learning approaches was shown. A (human) activity recognition system called *NuActiv* has been introduced in [3]. Applying a two-layer zero-shot learning algorithm, two datasets were examined. The datasets include exercise as well as daily life activities from more than 20 subjects, where sensors were attached to wrists, hips and arms of the subjects.

Human motion capturing in general is a topic of interest in the field of industrial processes. The Carnegie Mellon University Motion Capture Database [4] is a free motion capture data base often used in other contributions, i.e., [13]. It provides data captured and processed using the Vicon System. Alongside data about activities of daily living or sports, it as well contains data about industrial work, for example, activities during building constructions [19]. However, the authors describe a method to monitor the human posture in industrial environment. Therefore, a 3D camera as well as IMUs were used simultaneously. Several tailoring operations were performed by eight workers in the stretch of two days.

3 ORDER PICKING SCENARIO

To evaluate competing attribute representations, an order picking scenario has been created, see Figure 2. Its definition is the first step of the pipeline as illustrated in Figure 1.

The scenario is a common picker-to-stock process. This means that the order picker moves to collect the products necessary for one order. The empty boxes in this scenario have the dimensions L 400 mm x W 300 mm x H 220 mm. They are provided via a conveyor at a height of 380 mm and manually placed on a cart's top deck at a height of 800 mm. The employee drives the cart to two shelves where the stored articles are put into the boxes. The articles are exemplary portrayed by ballast sacks with a weight of 500 g. The process ends when the filled boxes are placed on a second conveyor.

The scenario's process exists in two variants. They differ in the characteristics of the human motion, e.g., the handedness. Both variants of the scenario are illustrated in Figure 2. The first variant on the on the left solely includes process steps that consist of seen activities that are trained individually beforehand. The second variant on the right includes slightly altered process steps and thus different manual activities. For these activities no training data will be provided but they will be used for testing. The concerned process steps are highlighted in blue.

The unseen activities of the second variant are categorized into three groups in regards to their semantic link with the seen activities of the first variant. The idea is to train a method of deep learning using solely the seen activities and test it on both the seen and unseen activities. Figure 3 visualizes the expected role of the attribute-based representation as a semantic link between the seen and unseen activities. This is helpful to narrow down the practical benefits and shortcomings of competing attribute representations in regards to specific groups of activities and attributes.

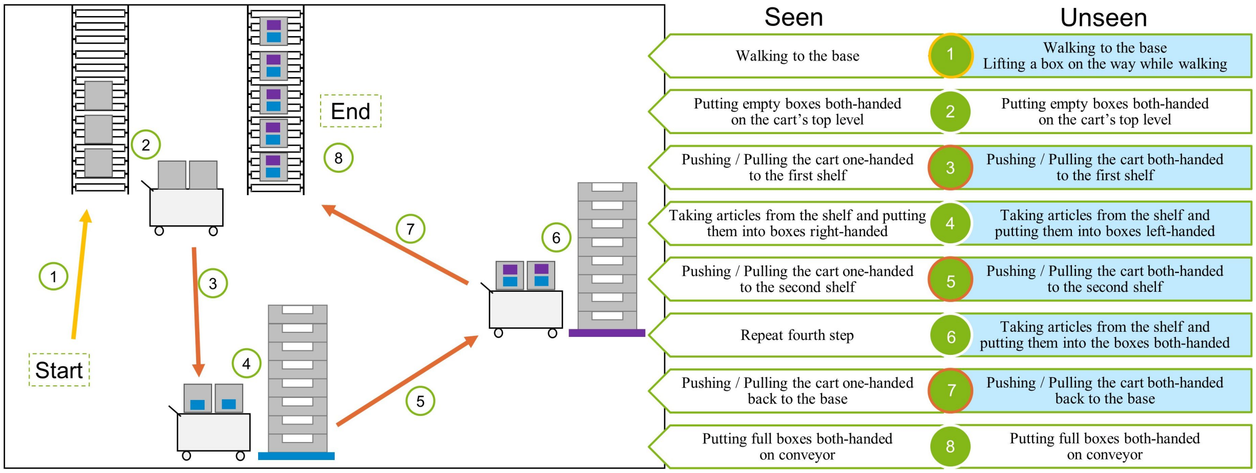
Group I: Lifting while walking. There exist recordings in which the participants walk and stand. These two activities are recorded with the participant having his hands free and while holding a box. Reaching forward and lifting the box with both hands are recorded as well. But during the latter activities, the participants were standing. It may be possible to recognize the activities of reaching forward and lifting when they are performed while walking. This is feasible as the sensor patterns of the upper body are expected to be very alike. The walking motion of the lower body is present in the walking activities.

Group II: Both-handed cart handling. The activities of pushing and pulling a cart are trained exclusively in their single-handed variant. With this data available, pushing and pulling a cart may be recognizable when performed with both hands.

Group III: Left-handed and both-handed article handling. Similar to Case II, it has been recorded how participants handle boxes with both hands and handle articles with the right hand only. The shared properties of these activities can be used to describe left-handed and both-handed article handling activities.

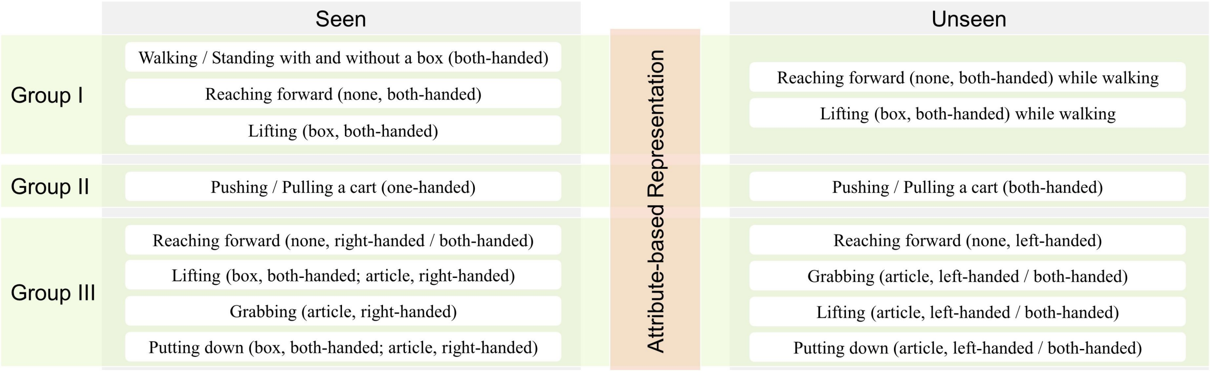
4 METHOD

The goal of the proposed method is to compare competing attribute representations in regards to their performance in HAR for seen and unseen activities. As illustrated in Figure 1, the method consists of defining attribute-based representations of activities present in a given scenario, deploying deep learning and applying predefined evaluation criteria.



Seen		Unseen
Walking to the base	1	Walking to the base Lifting a box on the way while walking
Putting empty boxes both-handed on the cart's top level	2	Putting empty boxes both-handed on the cart's top level
Pushing / Pulling the cart one-handed to the first shelf	3	Pushing / Pulling the cart both-handed to the first shelf
Taking articles from the shelf and putting them into boxes right-handed	4	Taking articles from the shelf and putting them into boxes left-handed
Pushing / Pulling the cart one-handed to the second shelf	5	Pushing / Pulling the cart both-handed to the second shelf
Repeat fourth step	6	Taking articles from the shelf and putting them into the boxes both-handed
Pushing / Pulling the cart one-handed back to the base	7	Pushing / Pulling the cart both-handed back to the base
Putting full boxes both-handed on conveyor	8	Putting full boxes both-handed on conveyor

Figure 2: Process Steps of the Order Picking Scenario. On the left is the scenario's variant with process steps that consist of seen activities. The list on the right contains slightly altered process steps with unseen activities that are highlighted in blue.



	Seen	Unseen
Group I	Walking / Standing with and without a box (both-handed) Reaching forward (none, both-handed) Lifting (box, both-handed)	Reaching forward (none, both-handed) while walking Lifting (box, both-handed) while walking
Group II	Pushing / Pulling a cart (one-handed)	Pushing / Pulling a cart (both-handed)
Group III	Reaching forward (none, right-handed / both-handed) Lifting (box, both-handed; article, right-handed) Grabbing (article, right-handed) Putting down (box, both-handed; article, right-handed)	Reaching forward (none, left-handed) Grabbing (article, left-handed / both-handed) Lifting (article, left-handed / both-handed) Putting down (article, left-handed / both-handed)

Figure 3: Three groups of unseen activities and the attribute-based representation as a semantic link to seen activities

Attribute Representations

Attributes are high-level semantic descriptions of classes, objects or scenes that are used for different recognition tasks [1, 10, 11, 18]. In HAR, collections of verbs and objects have been deployed for describing semantically and coarsely human actions in images and videos [22, 23]. For instance, attributes for HAR could be considered as simple particular movements of human body parts, e.g., *feet moving* or postures, which might define the action of *walking*. An advantage of using attribute representations is that human actions share simple movements. For example, *feet moving* could represent the activities *walking* and *running*, differing in speed. This sharing is suitable for tasks where datasets are unbalanced or testing sets contain unseen activities at training [10].

In HAR, a function $f : X \longrightarrow Y$ is learned. This function maps a sequence $x \in X$ to its respective activity class $y \in Y$. An attribute representation A can be seen as additional mapping between the sequences X and their classes Y, i.e., $f : X \longrightarrow A \longrightarrow Y$. This additional layer allows to share high-level concepts among classes.

An attribute representation $A \in \mathbb{B}_{[K,M]}$ contains K number of a binary vectors with M number of attributes. A binary vector a is defined uniquely per class $k_i \forall i = 1, 2, ..., K$ and it contains values of "1" when a certain attribute is present; otherwise the value is "0".

Deep Learning for HAR

For recognizing human activities using attribute representations, the CNN architecture, proposed in [7, 17], is used. This architecture uses temporal-convolution operations for extracting temporal-local dependencies of sequential inputs. By stacking temporal- convolutions and -pooling operations, CNNs extracts more complex features, being also robust against translations and noise. In general, a CNN for HAR classifies its input into classes using a softmax function

[7, 9, 15]. The proposed architecture, however, maps an input sequence into an attribute representation $\tilde{a} \in A$, replacing the softmax function by a sigmoid activation function, see Equation 1. Its output corresponds to pseudo-probabilities indicating if an attribute $\tilde{a}_i$ is present or not in the representation.

$$\tilde{a} = sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

In contrast to [7, 15, 17], input sequences are not comprised of IMUs' measurements. Sequences correspond to measurements of 3D global poses from a certain number of human body-segments, which are provided by the motion capturing system. For each segment, the motion capturing system records six different measurements, which are considered as a channel, similar to IMU's sensors. Thus, there are D channels in total. By segmenting sequences using a sliding window approach with window size of T and step of s, the input's size is $[T, D]$. Following [7, 17], channels are normalized to zero mean and unit variance. The architecture contains parallel branches, which are composed of convolution and max-pooling operations and a fully-connected layer. Each of these branches processes sequences from each of the human body segments. They are composed of two blocks of two stacked convolutional layers and a subsequent max-pooling layer. A convolutional layer has $C = 64$ filters of size $[5 \times 1]$. They perform convolutions along the time axis. Max-pooling operations find the maximum of $P = 2$ values along the time axis with a stride of 1. The network is trained using the binary-cross entropy loss. For predicting the activity class of a testing sequence, a nearest neighbor approach is utilized by computing the cosine distance between the computed $\tilde{a}$ and $a \in A$.

Evaluation criteria

The performance of competing attribute representations is evaluated in two steps. First, standard metrics for HAR are computed. Second, a qualitative aspects are presented.

Quantitative Metrics. The classification accuracy and the weighted $F1$ ($wF1$) are computed. The $wF1$ can be considered as the weighted average of the precision and recall using the proportion of activity classes in the testing set.

$$\omega F_1 = \sum_i 2 \cdot \frac{n_i}{N} \cdot \frac{\text{precision}_i \cdot \text{recall}_i}{\text{precision}_i + \text{recall}_i}, \tag{2}$$

where n_i is the number of sequences per activity and N is the number of sequences in the dataset. In contrast to the classification accuracy, the weighted F1 is more appropriate under highly unbalanced datasets.

Qualitative Aspects. Even though a specific attribute representation yields good quantitative results, it may be unfavourable for practical application. If the attributes' definitions are incomprehensible, their transfer to unseen activities will be impeded. A representation of high quality holds a low amount of attributes that are semantically easy to understand by a human, coherent and thus easy to transfer to unseen activities. The definition of attribute-based representation for new activities should require as few new attributes as possible. The evaluation of these aspects is performed by warehousing experts.

5 EVALUATION

This section provides an insight on the utilized data set and implementation details of the deep learning. Following the pipeline as illustrated in Figure 1, the achieved results are presented according to the evaluation criteria.

MoCap Dataset

A physical model of the order picking scenario has been built in the "Innovationlab Hybrid Services in Logistics" at the TU Dortmund University [20]. The MoCap dataset consists of recordings conducted in the given scenario, see Figure 2. While recordings of some activities already existed from previous work, the majority of activities has been recorded specifically for this contribution and they have been added to the MoCap dataset. The data is composed of multichannel time-series recorded by a motion capture system. This system is based on photogrammetry methods for measuring object poses on $2D$ and $3D$ spaces using a set of cameras. In total, it contains 38 cameras. The data recording in the controlled environment of the InnovationLab is illustrated in Figure 4. The recording frame rate is 300 fps. The MoCap provides global poses of 22 body segments. A pose is a combination of position and angular values in $[X, Y, Z]$. This leads a total of $D = 132$ channels. The dataset contains recordings from $K = 27$ classes.

The recording of the seen activities is not performed a single sequence within the entire scenario. Instead, they are recorded successively. Each activity is recorded with up to to 8 participants. Both versions of the scenario's processes, the seen and unseen activities, are performed by 4 participants. For participants $(1, 2, 3, 6)$ seen and unseen activities have been recorded. The remaining four participants solely performed seen activities. The recordings are annotated manually. The global pose sequences are normalized with respect to the lower back segment.

Recording from participants $(1, 2, 7, 8)$, $(4, 6)$ and $(3, 5)$ are used as the training, validation and testing sets respectively. The testing set is divided in two sets, for the seen and unseen activities. Sequences are extracted by means of a sliding-window approach with window size of $T = 200$ or $660ms$

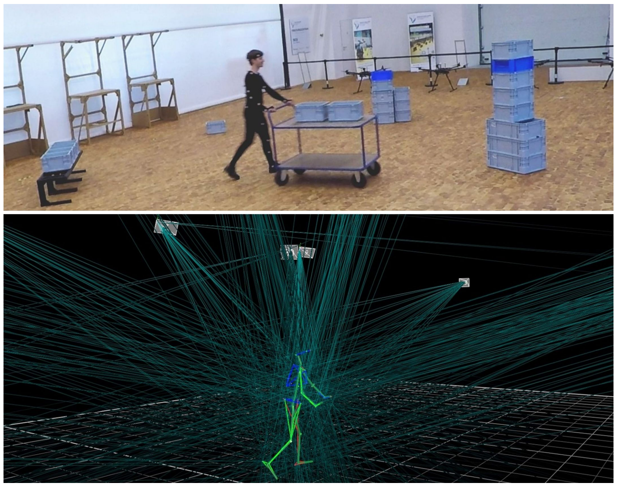


Figure 4: Data Recording in the InnovationLab. The image above shows the physical set-up in the InnovationLab, the lower image shows the Motion Capture Output.

and step of $s = 10$ or $33ms$. Sequences are assigned the most frequent activity label. Table 1 shows the number of sequences per set and the proportion of classes.

Implementation Details

Three CNNs are trained using three different attribute representations. CNNs are trained on the training set and deployed on the validation set for determining suitable hyperparameters for training. A final training is carried out on the training and validation sets. The CNNs are deployed on the testing seen and testing unseen sets. In addition, a standard CNN using a softmax layer for recognizing seen activities is also trained to compare the performance when using attribute representations.

As human activities definitions vary with respect to human motion, or to practical applications, a universal attribute representation for order picking does not exist. Three different attribute representations, two expert-given and a random one, are used for training a CNN for HAR on the MoCap dataset. The expert is a warehousing specialist who has practical experience with real-life order picking systems.

A_1, expert-given, 17 Attributes, see Table 2. Ten attributes describe various arm motions like *left arm reaching forward*, *arms pull* or *right arm stretched out*. The more general attributes *right Arm* and *left Arm* are used to have a feature shared by a majority of activities. Upper body movements and leg motions are characterized by two attributes each. Certain poses, which are taken while utilizing items, have three attributes, e.g. *pose box*. Table 2 shows the attribute representation A_1 for the $K = 27$ activity classes, as described in section 3.

A_2, expert-given, 27 Attributes. The second representation is more extensive than the first one. The attributes describe directional movements for individual body segments, e.g., *right hand moving forward*, *right hand moving up*, *left knee moving forward* and *head moving down*. In this aspect, the second representation is more detailed than the first, as $A1$ does not consider the movement of the knee, the elbow or the head. On the contrary, individual poses to express that a specific item is involved in the activity, such as a box-pose or a cart-pose, are not present in $A2$. In $A2$, there are attributes to identify special poses, for instance *right elbow angled*, *right forearm behind body* or *left arm raised*, but they do not imply a specific item.

The attributes of A_1 and A_2 are oriented towards body-part movements. For example, the action *Walking (none)* implies a movement of the right and left leg or foot. Both representations were created having such striking characteristics of the activities in mind.

A_3, randomly generated, 54 Attributes. This representation is generated randomly following the conclusion of [17], where random attributes present comparable performance for Locomotion and Gestures datasets. 54 attributes is double the amount of activities.

The presented CNN is implemented in the Caffe framework. CNN's parameters are learned by minimizing a loss function using stochastic gradient descent with RMSProp rule. This loss function depends on the classifier: softmax loss function, and the binary-cross entropy. The training hyperparameters are RMS decay of 0.95, base learning rate of 10^{-5}, batch size of 128 and the number of epochs of 2. Learning rate is decreased after 1 epoch by y factor of $\gamma = 0.1$. Channels are normalized to a range of $[0, 1]$. Dropout of $p = 50\%$ is applied to the first and second fully-connected layers. The network is initialized using an orthogonal initialization.

Results

The three CNNs, one per attribute representation $[A_1, A_2, A_3]$, and a CNN using a softmax layer are evaluated on the testing sets.

Quantitative Results. Table 5 shows the accuracy and $\omega F1$ of the four CNNs using the three attribute representations and a softmax classifier on the testing-seen set. In general, using an attribute representation is beneficial for HAR in comparison with a softmax classifier, following the conclusion in [17]. Even using a random representation, the performance is near to the softmax. Deep architectures learn features and classify sequence segments when targets are represented by a set of attributes. Besides, sharing attributes among activities is advantageous as information from frequent activities can be used for predicting infrequent activities.

Set	Proportion [%]																N_o sequences
	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9	k_{10}	k_{11}	k_{12}	k_{13}	k_{15}	k_{16}	k_{17}	
Training	14.5	13.8	15.3	15.7	1.4	1.7	3.8	1.8	1.2	3.4	0.5	1.5	6.3	6.3	6.3	6.3	288005
Validation	15.8	14.5	14.6	14.9	0.4	1.1	0.7	3.5	1.0	4.1	0.4	2.3	6.7	6.7	6.7	6.7	133934
Testing seen	14.0	13.7	16.1	15.8	3.7	2.6	3.8	5.1	0.9	1.7	0.5	0.7	5.4	5.4	5.4	5.4	165906

Table 1: Proportion of segmented sequences per activity class on the three sets.

	Activity	right arm	left arm	right arm reach forw.	left arm reach forw.	right arm reaching back	left arm reach back	right arm stretched out	left arm stretched out	arms push	arms pull	upper body move forw.	upper body move back	legs walk	legs stand	pose box	pose article	pose cart
		a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}
k_1	Walking *(none)*	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0
k_2	Walking *(box, both-handed)*	**1**	**1**	0	0	0	0	0	0	0	0	0	0	**1**	0	**1**	0	0
k_3	Standing *(none)*	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0
k_4	Standing *(box, both-handed)*	**1**	**1**	0	0	0	0	0	0	0	0	0	0	0	**1**	**1**	0	0
k_5	Reaching forward *(none,both-handed)*	**1**	**1**	**1**	**1**	0	0	0	0	0	0	**1**	0	0	**1**	0	0	0
k_6	Lifting *(box,both-handed)*	**1**	**1**	0	0	**1**	**1**	0	0	0	0	0	**1**	0	**1**	**1**	0	0
k_7	Putting down *(box,both-handed)*	**1**	**1**	**1**	**1**	0	0	0	0	0	0	**1**	0	0	**1**	**1**	0	0
k_8	Straightening up *(none)*	0	0	0	0	**1**	**1**	0	0	0	0	0	**1**	0	**1**	0	0	0
k_9	Reaching forward *(none,right-handed)*	**1**	0	**1**	0	0	0	0	0	0	0	**1**	0	0	**1**	0	0	0
k_{10}	Grabbing (article,right-handed)	**1**	0	0	0	0	0	**1**	0	0	0	0	0	0	**1**	0	**1**	0
k_{11}	Lifting (article,right-handed)	**1**	0	0	0	**1**	0	0	0	0	0	0	**1**	0	**1**	0	**1**	0
k_{12}	Putting down (article,right-handed)	**1**	0	**1**	0	0	0	0	0	0	0	**1**	0	0	**1**	0	**1**	0
k_{13}	Pushing *(cart,right-handed)*	**1**	0	0	0	0	0	0	0	**1**	0	0	0	**1**	0	0	0	**1**
k_{14}	Pushing *(cart,left-handed)*	0	**1**	0	0	0	0	0	**1**	**1**	0	0	0	**1**	0	0	0	**1**
k_{15}	Pulling *(cart,right-handed)*	**1**	0	0	0	0	0	0	0	0	**1**	0	0	**1**	0	0	0	**1**
k_{16}	Pulling *(cart,left-handed)*	0	**1**	0	0	0	0	0	**1**	0	**1**	0	0	**1**	0	0	0	**1**
k_{17}	Reaching forward Walking (none,both-handed)	**1**	**1**	**1**	**1**	0	0	0	0	0	0	**1**	0	**1**	0	0	0	0
k_{18}	Lifting & Walking *(box,both-handed)*	**1**	**1**	0	0	**1**	**1**	0	0	0	0	0	**1**	**1**	0	**1**	0	0
k_{19}	Pushing *(cart,both-handed)*	**1**	**1**	0	0	0	0	**1**	**1**	**1**	0	0	0	**1**	0	0	0	**1**
k_{20}	Pulling *(cart,both-handed)*	**1**	**1**	0	0	0	0	**1**	**1**	0	**1**	0	0	**1**	0	0	0	**1**
k_{21}	Reaching forward *(none,left-handed)*	0	**1**	0	**1**	0	0	0	0	0	0	**1**	0	0	**1**	0	0	0
k_{22}	Grabbing *(article,left-handed)*	0	**1**	0	0	0	0	0	**1**	0	0	0	0	0	**1**	0	**1**	0
k_{23}	Lifting *(article,left-handed)*	0	**1**	0	0	0	**1**	0	0	0	0	0	**1**	0	**1**	0	**1**	0
k_{24}	Putting down *(article,left-handed)*	0	**1**	0	**1**	0	0	0	0	0	0	**1**	0	0	**1**	0	**1**	0
k_{25}	Grabbing article,both-handed	**1**	**1**	0	0	0	0	**1**	**1**	0	0	0	0	0	**1**	0	**1**	0
k_{26}	Lifting article,both-handed	**1**	**1**	0	0	**1**	**1**	0	0	0	0	0	**1**	0	**1**	0	**1**	0
k_{27}	Putting down article,both-handed	**1**	**1**	**1**	**1**	0	0	0	0	0	0	**1**	0	0	**1**	0	**1**	0

Table 2: An expert-given attribute representation A_1. Classes k_1 to k_{16} correspond to the seen activities, followed by the three groups of unseen activities separated by horizontal double lines.

Attribute Rep.	Accuracy per activity class															
	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9	k_{10}	k_{11}	k_{12}	k_{13}	k_{15}	k_{16}	k_{17}
A_1	77.0	65.4	85.5	73.4	53.9	52.2	69.3	61.3	68.1	94.7	74.7	84.9	98.6	96.6	69.9	65.2
A_2	63.9	65.4	93.7	87.4	44.7	53.1	80.4	56.1	73.6	95.4	62.0	83.7	98.8	98.3	68.6	54.1
A_3	50.4	60.4	91.2	78.0	10.9	10.9	78.5	17.0	0.0	89.9	0.0	89.9	92.8	87.0	56.1	43.9

Table 3: Accuracy per class k_i on the testing-seen set.

Act. set	Metric	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}
Seen	Acc. [%]	97.2	93.4	82.7	72.5	61.0	53.1	96.6	98.0	96.3	67.5	80.5	60.3	97.4	85.7	87.5	94.9	99.1
Unseen	Acc. [%]	83.1	82.3	52.7	11.9	6.8	9.4	1.2	34.5	80.3	0.0	28.1	17.1	97.3	56.5	53.4	0.0	83.3

Table 4: Accuracy per attribute for the A_1 representation.

	Seen	
Attribute Rep	Acc. [%]	F1 [%]
A1	**75.11**	**75.74**
A2	74.46	73.76
A3	61.13	58.29
Softmax	64.21	62.18

Table 5: Accuracy and F1 of the CNNs for each of the three attribute representations and the CNN using a softmax layer on the testing-seen set.

Table 3 shows the accuracy per seen-class using the three attribute representations. Classes with small proportion in the dataset, i.e., $[k_5, ..., k_{10}]$ show poor accuracy using the random attribute representation.

An additional experiment is carried out on the testing-unseen set. This aims to evaluate the attribute prediction on sequences with activities that are not used for training. Table 4 shows the accuracy per attribute on testing-seen and -unseen sets using the A_1 representation. For the testing-seen set, the recognition of attributes in general yields high accuracies. The attributes $[a_5, a_6, a_{10}, a_{12}]$ present the lowest accuracies. On the testing-unseen set, the recognition of attributes show variable results, having high and low accurate attribute recognition. For both datasets, attributes $[a_{1,2,3}, a_9, a_{13,14,15}]$ yield the best performances. These results show that attributes, being mainly associated to the classes with the largest proportion in the testing sets, see Table 1, are better recognized. Similar results are found using the attribute representations A_2 and A_3.

Qualitative Results. On the one hand, A_1 is easy to understand for warehousing experts. Thus, they can easily adjust it according to application demands. This is because A_1 holds the lowest amount of attributes and their descriptions are easy to picture. On the other hand, the need to distinguish activities as precisely as possible is expected to lead to a high amount of attributes. For example, a new pose attribute is necessary for each new item. Once further aspects such as packaging processes are taken into account, this approach may come up against a limit of semantic comprehensibility. In a warehouse, the number of different items to interact with is virtually infinite. On the long run, a high level of granularity is necessary to maintain semantic comprehensibility.

Hence, A_2 uses more attributes than A_1 to describe the same amount of activities. This version already describes activities with a more detailed observation of each segment. The definition of further activities would therefore require a smaller amount of new attributes compared to A_1. This is also because there are no individual pose attributes for each item. However, in practical application it may be of interest to differentiate between plain walking, walking with a box, a picking list and so forth. Considering these differences, it is difficult as there are no attributes that explicitly state the respective items.

Even though the performance of the random representation A_3 can be improved, the lack of semantic meaning is inevitable [17]. A posterior alteration of the attribute definition is thus impossible for humans. An evolutionary algorithm would have to be deployed every time new activities are added or an activity's definition is changed. The resulting computational effort makes this approach unfeasible for practical application as an adaptable method.

6 DISCUSSION AND CONCLUSION

Three different representations, two expert-given and a random one, are compared by deploying deep architectures on multichannel time-series for order picking. These architectures process time-series computing attributes, which are used for predicting human actions. In addition, a comparison with a standard deep architecture for HAR using a softmax classifier is shown. In general, using attribute representations presents comparable or better performance than the standard CNN, even by using a random representation. Interestingly, representations with a lower quantity of attributes present a slightly better performance. Expert-given representations exhibit the best performances showing a semantic relation between the attributes and the activities. Deep architectures are able to compute attributes that belong to frequent activities in the training set.

During the annotation of the attribute representations, the semantic meaning of each attribute depends strongly on the expert. The semantic meaning is subjective and thus possibly ambiguous. There is no guideline for a consistent attribute definition known to the authors.

ACKNOWLEDGMENTS

The work on this publication has been supported by Deutsche Forschungsgemeinschaft (DFG) in the context of the research project Fi799/10-1, HO2403/14-1 ('Adaptive, Context-based Activity Recognition and Motion Classification to Analyze the Manual Order Picking Process').

REFERENCES

[1] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. 2014. Word spotting and recognition with embedded attributes. In *IEEE transactions on pattern analysis and machine intelligence (12)*, Vol. 36. IEEE, 2552–2566.

[2] Heng-Tze Cheng, Martin Griss, Paul Davis, Jianguo Li, and Di You. 2013. Towards Zero-shot Learning for Human Activity Recognition Using Semantic Attribute Sequence Model. In *Proc. of the 2013 ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing (UbiComp '13)*. ACM, New York, NY, USA, 355–358. https://doi.org/10.1145/2493432.2493511

[3] Heng-Tze Cheng, Feng-Tso Sun, Martin Griss, Paul Davis, Jianguo Li, and Di You. 2013. NuActiv: Recognizing Unseen New Activities Using Semantic Attribute-based Learning. In *Proc. of the 11th Annual Int. Conf. on Mobile Systems, Applications, and Services (MobiSys '13)*. ACM, New York, NY, USA, 361–374. https://doi.org/10.1145/2462456.2464438

[4] CMU. 2018. Carnegie Mellon University Graphics Lab Motion Capture Database. (June 2018). http://mocap.cs.cmu.edu/

[5] Sascha Feldhorst, Sandra Aniol, and Michael ten Hompel. 2016. Human Activity Recognition in der Kommissionierung – Charakterisierung des Kommissionierprozesses als Ausgangsbasis für die Methodenentwicklung. *Logistics Journal : Proc.* 2016, 10 (Oct. 2016). https://doi.org/10.2195/lj_Proc_feldhorst_de_201610_01

[6] Sascha Feldhorst, Mojtaba Masoudenijad, Michael ten Hompel, and Gernot A. Fink. 2016. Motion Classification for Analyzing the Order Picking Process using Mobile Sensors - General Concepts, Case Studies and Empirical Evaluation:. SCITEPRESS - Science and and Technology Publications, 706–713. http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0005828407060713

[7] Rene Grzeszick, Jan Marius Lenk, Fernando Moya Rueda, Gernot A. Fink, Sascha Feldhorst, and Michael ten Hompel. 2017. Deep Neural Network based Human Activity Recognition for the Order Picking Process. ACM Press, 1–6. http://dl.acm.org/citation.cfm?doid=3134230.3134231

[8] Jendrik Haase and Daniel Beimborn. 2017. Acceptance of Warehouse Picking Systems: A Literature Review. In *Proc. of the 2017 ACM SIGMIS Conf. on Computers and People Research (SIGMIS-CPR '17)*. ACM, New York, NY, USA, 53–60. https://doi.org/10.1145/3084381.3084409

[9] Nils Y. Hammerla, Shane Halloran, and Thomas Ploetz. 2016. Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables. *arXiv:1604.08880 [cs, stat]* (April 2016). http://arxiv.org/abs/1604.08880 arXiv: 1604.08880.

[10] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. IEEE, 951–958.

[11] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-Based Classification for Zero-Shot Visual Object Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 3 (March 2014), 453–465. http://ieeexplore.ieee.org/document/6571196/

[12] J. Liu, B. Kuipers, and S. Savarese. 2011. Recognizing human actions by attributes. In *CVPR 2011*. 3337–3344.

[13] Yang Liu, Lin Feng, Shenglan Liu, and Muxin Sun. 2018. Sensor Network Oriented Human Motion Segmentation With Motion Change Measurement. *IEEE Access* 6 (2018), 9281–9291. http://ieeexplore.ieee.org/document/8240911/

[14] Riccardo Manzini (Ed.). 2012. *Warehousing in the global supply chain: advanced models, tools and applications for storage systems.* Springer, London ; New York.

[15] Francisco Ordóñez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* 16, 1 (Jan. 2016), 115. http://www.mdpi.com/1424-8220/16/1/115

[16] Charissa Ronao and Sung-Bae Cho. 2015. Deep convolutional neural networks for human activity recognition with smartphone sensors. Springer, 46–53.

[17] Fernando Moya Rueda and Gernot A. Fink. 2018. Learning Attribute Representation for Human Activity Recognition. *arXiv:1802.00761 [cs]* (Feb. 2018). http://arxiv.org/abs/1802.00761 arXiv: 1802.00761.

[18] Gernot A. Fink Sebastian Sudholt. 2017. PHOCNet: A Deep Convolutional Neural Network for Word Spotting in Handwritten Documents. In *Proc. of the Int. Conf. on Document Analysis and Recognition.* https://arxiv.org/pdf/1604.00187.pdf Some info about attribute representations in the field of word spotting. Based on this paper, an attribute-based network for HAR could be implemented.

[19] Marco Tarabini, Marco Marinoni, Matteo Mascetti, Pietro Marzaroli, Francesco Corti, Hermes Giberti, Alberto Villa, and Paolo Mascagni. 2018. Monitoring the human posture in industrial environment: A feasibility study. IEEE, 1–6. http://ieeexplore.ieee.org/document/8336710/

[20] A. K. R. Venkatapathy, H. Bayhan, F. Zeidler, and M. ten Hompel. 2017. Human machine synergies in intra-logistics: Creating a hybrid network for research and technologies. In *2017 Federated Conf. on Computer Science and Information Systems (FedCSIS)*. 1065–1068. https://doi.org/10.15439/2017F253

[21] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiaoli Li, and Shonali Krishnaswamy. 2015. Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition. 3995–4001.

[22] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. 2011. Human action recognition by learning bases of action attributes and parts. IEEE, 1331–1338.

[23] Jingjing Zheng, Zhuolin Jiang, and Rama Chellappa. 2017. Submodular Attribute Selection for Visual Recognition. In *IEEE Trans. on Pattern Analysis and Machine Intelligence (Volume: 39, Issue: 11,) (11)*, Vol. 39. IEEE, 2242 – 2255. https://ieeexplore.ieee.org/abstract/document/

7776926/

[24] Maryam Ziaeefard and Robert Bergevin. 2015. Semantic human activity recognition: A literature review. *Pattern Recognition* 48, 8 (Aug. 2015), 2329–2345. http://www.sciencedirect.com/science/article/pii/S0031320315000953