Deep Neural Network based Human Activity Recognition for the Order Picking Process

Rene Grzeszick, Jan Marius Lenk, Fernando Moya Rueda, Gernot A. Fink
Department of Computer Science, TU Dortmund University
Dortmund, Germany
(rene.grzeszick,jan-marius.lenk,fernando.moya, gernot.fink)@tu-dortmund.de

Sascha Feldhorst, Michael ten Hompel
Fraunhofer IML / Department of Mechanical Engineering, TU Dortmund University
Dortmund, Germany
(sascha.feldhorst,michael.ten.hompel) @iml.fraunhofer.de

ABSTRACT

Although the fourth industrial revolution is already in pro-gress and advances have been made in automating factories, completely automated facilities are still far in the future. Human work is still an important factor in many factories and warehouses, especially in the field of logistics. Manual processes are, therefore, often subject to optimization efforts. In order to aid these optimization efforts, methods like human activity recognition (HAR) became of increasing interest in industrial settings. In this work a novel deep neural network architecture for HAR is introduced. A convolutional neural network (CNN), which employs temporal convolutions, is applied to the sequential data of multiple intertial measurement units (IMUs). The network is designed to separately handle different sensor values and IMUs, joining the information step-by-step within the architecture. An evaluation is performed using data from the order picking process recorded in two different warehouses. The influence of different design choices in the network architecture, as well as pre- and post-processing, will be evaluated. Crucial steps for learning a good classification network for the task of HAR in a complex industrial setting will be shown. Ultimately, it can be shown that traditional approaches based on statistical features as well as recent CNN architectures are outperformed.

ACM Classification Keywords

I.5.4 Pattern Recognition: Applications; Signal processing

Author Keywords

Activity Recognition; Deep Learning; Convolutional Neural Networks; Inertial Measurement Units.

iWOAR '17, September 21-22, 2017, Rostock, Germany

ACM ISBN 978-1-4503-5223-9/17/09…$15.00
https://doi.org/10.1145/3134230.3134231

INTRODUCTION

With the beginning of the fourth industrial revolution, many researchers expect Cyber-Physical Systems (CPS) to significantly change industrial processes, the way they are controlled and also their automation [15]. Nevertheless, human work will remain an important part of the industrial systems of the future. Completely automated factories and distribution centers are neither technologically nor economically reasonable. Especially in the field of logistics, where processes change rapidly and various tasks are performed by a worker, automation is difficult or simply not economically feasible. For instance, the order picking process, where a list of items is collected in a warehouse, is often done manually. In huge warehouses like the ones operated by Amazon or Zalando, hundreds of people are working simultaneously in order to deliver articles to customers in the shortest time possible. Thus, in high-wage countries in central Europe or north America, human work is a significant cost driver. It is therefore often subject of optimization efforts [11]. Many different technologies and approaches were developed in order to improve the efficiency of the manual order picking process [6]. However, for all optimization approaches a deep understanding of the manual steps within the process is a crucial requirement. In order to understand these manual processes, manual analysis (REFA) or time estimates (Methods-Time Measurement; MTM) are frequently used [8]. Recently, it was proposed to use methods of human activity recognition (HAR) for collecting and analyzing realistic data from manual processes [4]. In HAR, body-worn sensors record human activities and the data is subsequently analyzed. Frequently, intertial measurement units (IMUs) that contain an accelerometer, gyroscope and magnetometer are used. While most of these tasks deal with basic movements such as walking, jogging, driving or cycling, an industrial setting is much more challenging. The activities in industrial processes often consists of complex movements and require a high flexibility. For example, the picking or boxing of articles with different sizes, shapes and weights may show a great intra-class and especially inter-person variability. Similarly, different technical artifacts are handled by the workers. This variability is very challenging for HAR.

This work has been published at the Proc. Int. Workshop on Sensor-based Activity Recognition and Interaction. Please cite:
Deep Neural Network based Human Activity Recognition for the Order Picking Process, Rene Grzeszick, Jan Marius Lenk, Fernando Moya, Sasha Feldhorst, Michael ten Hompel, Gernot A. Fink, Proc. Int. Workshop on Sensor-based Activity Recognition and Interaction, 2017

Traditional approaches to HAR rely on statistical features in a sliding window approach [2, 10]. These are then processed by a classifier, e.g., an SVM, a Random Forest or a sequence based approach such as dynamic time warping (DTW) or a Hidden Markov Model (HMM). Typically, statistical features such as mean, median, min, max or the signal magnitude area are computed for each sensor signal (i.e. for accelerometer x,y and z separately). These features are often complemented by features computed on the derivative of the signal or in some cases by a correlation analysis between different signals [10]. The design or choice of statistical features is often a difficult process, due to the large intra-class and inter-person variabilities in HAR [7], but also due to the abstract nature of the sensor data. Nevertheless, a broad choice of these features tends to work comparably well, especially since learning features also poses a difficult task as the activity data is often scarce and also highly unbalanced [2]. In most tasks, there are a few dominant classes while several activities occur only infrequently.

A traditional approach has recently been applied to analyzing the order picking process in [4]. Three IMUs were used in order to analyze the activities within the order picking process. An evaluation of different classifiers in a sliding window setup has been performed using a set of handcrafted statistical features. The experiments were performed with data from two different warehouses and different workers, showing promising results.

More recently, approaches based on deep neural networks (DNNs) were also picked up in the field of HAR [12, 16]. These allow for learning features in conjunction with a classifier. In [12] and [16] CNN architectures that are applied in a sliding window framework are proposed. The CNNs use temporal convolutions which are applied over a fixed number of frames, i.e. 3 to 15 sensor values. While in [16] it is suggested to apply the convolutions for each sensor signal separately, in [12] the convolutions are applied over all sensor values simultaneously. The first approach handles the information of each sensor signal independently until the first fully connected layer. The fully connected layer then models the correlations between different sensor signal at feature level, i.e. movement in x and y direction. The second approach directly considers correlations between sensor signals at the first convolution. These CNN approaches outperform the traditional approaches for HAR. However, the traditional methods are not outperformed by a large margin as, for example, in the audio or most prominently the visual domain [13]. Note that both architectures are comparably shallow with up to three convolution layers, intermediate pooling layers and a single fully connected layer for classification.

A more extensive comparison of traditional methods and different neural networks is given in [5]. DNNs, CNNs and recurrent neural networks (RNNs) are compared on various tasks, ranging from every-day activities to more complex tasks. For the CNN the approach of [12] is followed, computing a convolution over all sensor values. For the RNNs, a long short-term memory (LSTM) as well as a bi-directional long short-term memory (B-LSTM) have been implemented. Both

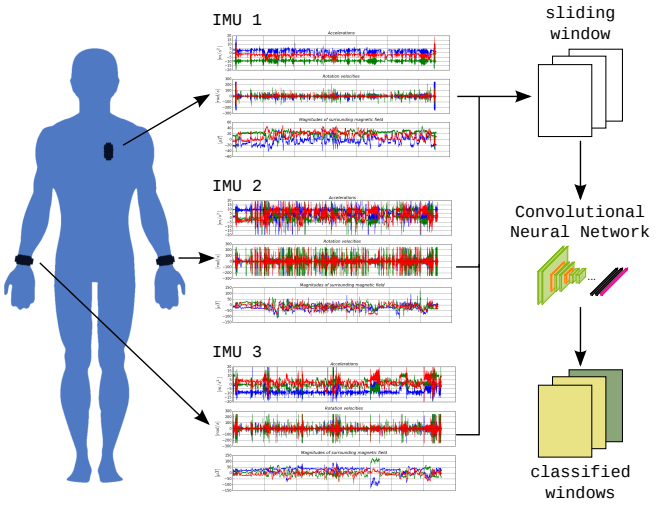

Figure 1. Overview of the proposed approach using an example of three IMUs worn by a worker at both wrists and the torso.

evaluate the signal in a frame-wise fashion. While the best performance has been achieved using the B-LSTM, most surprisingly a similar accuracy has been achieved by the CNN, which is applied in a sliding window. Furthermore, it is shown that the performance of the CNN is more robust to parameter changes, making the training of the network more easy.

Following up on the conclusions in [5], a CNN architecture for analyzing the order picking process is proposed. The contribution of this paper is two-fold: 1) The application of a deep neural network to the task of HAR in an order picking scenario is shown. 2) A novel network architecture that is IMU centered is introduced and compared to traditional methods based on statistical features as well as a recent CNN architecture.

The approach is evaluated on the dataset introduced in [4]. The effect of design choices in the network architecture, as well as pre-processing in the form of data augmentation and post-processing will be studied. It will be shown that the proposed CNN outperforms the traditional approach presented in [4] and that the novel IMU based architecture is able to improve the results compared to recent CNN architectures as, for example, discussed in [12, 5].

METHOD

The presented approach builds on a sensor setup where multiple IMUs are worn by a worker. An overview based on an example with three IMUs is given in Fig. 1. Each IMU records data from multiple sensors, i.e., accelerometer, gyroscope and magnetometer in x, y and z direction at a given Hz rate. The resulting sensor data is, therefore, a time series with multiple dimensions (3×9 values for the sensors described above). The sensors are synchronized at the beginning of every recording and then a sliding window is applied to the sensor data. Each sliding window contains multiple frames and is processed by a CNN in order to recognize different human activities. The output is a labeled sequence of human activities.

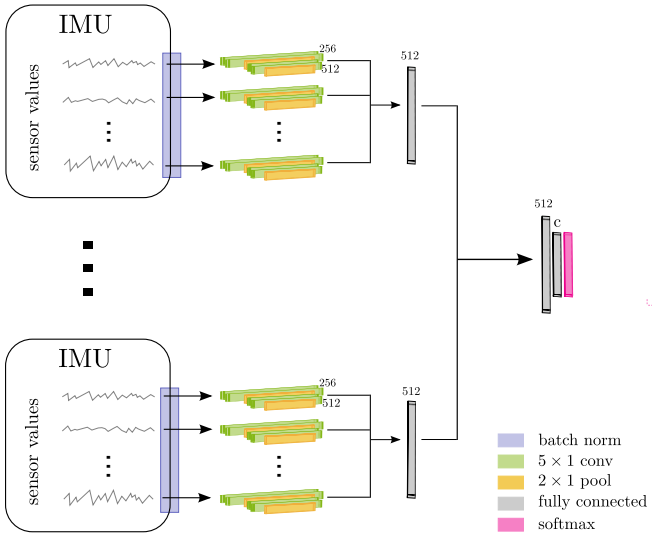

Figure 2. Overview of the proposed IMU-CNN architecture. First, the data of each IMU is processed independently and then the information is joined in a fully connected layer. In the last layer of size c (= number of classes), a softmax classification is performed (*best viewed in color*).

Data Augmentation and Class Imbalance

For training the CNN, all possible windows are extracted from a set of training sequences. Thus, a sliding window of size w is moved forward one frame at a time. As a result for each event in a training sequence multiple windows representing the activity are extracted. Although the information in these is highly redundant, the small frame-shift allows to generate a large number of samples, which is important for training a CNN.

As the sensor data is not invariant to most augmentations, such as mirroring, offsets or mixing two classes, two data augmentation strategies are proposed. First, Gaussian noise has been added to the sampling values, which simulates inaccuracies in the sensor's sampling. This is also a standard approach for data augmentation [9, 14]. Second, the sensor values within a given window are randomly re-sampled. New values between each two samples are computed. A random number $r \in [0, 1]$ is drawn and the point at time r between the two values is approximated using interpolation. Given the fast sampling rate in relation to often relatively long activities of multiple seconds, this augmentation preserves the coarse structure of the data, but a random time jitter in the sensor's sampling process is simulated.

During the data augmentation process, the imbalance issue is tackled by creating a larger number of augmented samples for the under represented classes. The samples are re-balanced such that each class has at least b percent of the largest number of samples per class in the training set.

IMU CNN

The proposed architecture builds on the idea of processing the data of multiple IMUs separately. An overview of the network is given in Fig. 2. Having multiple parallel processing blocks within the network, it roughly follows the idea of wider rather than deeper networks [17]. Here, each parallel block has a logical meaning as it represents the data of a single IMU. In theory, this abstraction should also allow for more robustness against the IMUs being slightly asynchronous.

To cope with different types of signals, i.e., accelerometer, gyroscope and magnetomer, that have different value ranges, the sensor values are processed by a batch normalization. The data is thus normalized to have zero mean and unit variance. The normalized data of each of the IMUs is processed by multiple convolutions and subsequent pooling operations. Following the approach of [16], each channel is separately processed by a temporal convolution. The temporal convolutions of each IMU share their weights. Instead of increasing the size of the temporal convolutions, two 5×1 convolutions are stacked and followed by a subsequent 2×1 max pooling. Two of these blocks are combined so that the network has a total size of four convolutional layers. Instead of scaling the network deeper, these layers are processed in parallel for each IMU, increasing the networks descriptiveness. From the resulting feature maps, an intermediate representation is computed for each of the IMUs by a fully connected layer. The intermediate representations are then concatenated by a subsequent fully connected layer that has a global view on the data and one fully connected layer of size c, representing a score for each of the classes. Dropout is applied to all fully connected layers, except the classification layer. As only one activity is considered to be present at each point of time, a softmax is used for deriving pseudo-probabilities from the c class scores. Thus, a softmax loss is used for training the network.

EVALUATION

The evaluation is performed on the dataset introduced in [4]. The proposed algorithm is evaluated with respect to its classification accuracy in an order picking setup and compared to both, traditional and recent CNN approaches.

A baseline system is evaluated using statistical features as described in [4]. A set of features, namely *min, max, avgerage* and *standard deviation*, is computed within a short time window separately for each signal. All features are concatenated into one feature vector representing the time window. This feature vector is then used as input for a classifier in order to predict an activity class. Here, three different classifiers are compared. On the one hand, a generative model is evaluated in the form of a Bayes classifier. On the other hand two discriminative approaches, a Random Forest and a linear SVM, are also evaluated.

For comparison with the proposed IMU-CNN architecture, a CNN baseline architecture is also evaluated. Following the architecture design presented in [5], a CNN with a global temporal convolution over all sensor values is used. This architecture is referred to as the *activity* CNN. Here, the same structure of 5×1 convolutions and max pooling operations as for the proposed IMU-CNN are implemented, which are then followed by three fully connected layers. The sizes of the layers are also chosen according to the proposed IMU-CNN architecture which is larger than the design proposed in [5]. As an intermediate step between the proposed IMU-CNN architecture and the *activity* CNN, an IMU-CNN with a global

activity class	warehouse *A*	warehouse *B*
walking	21465	32904
searching	344	1776
picking	9776	33359
scanning	0	6473
info	4156	19602
carrying	1984	0
acknowledge	5792	0
Unknown	1388	264
flip	1900	2933

Table 1. Overview of the dataset and number of frames for each of the activity classes in the different parts of the dataset.

temporal convolution is evaluated. Thus, a convolution over all sensor values is computed for each IMU separately. This architecture is referred to as *IMU-global*.

Order Picking Dataset

The data consist of two sets. Each of these sets contains recordings from three persons in two different warehouses, denoted as *A* and *B*. While the proposed approach is applicable to an arbitrary number of IMUs, the recording in the order picking setup has been restricted to three IMUs. In contrast to many setups which have been recorded in a controlled environment [1, 3], the number of sensors has been restricted in order to not interfere with the actual work. As shown in the example in Fig. 1, these were located at both wrists and an additional sensor at the torso. The IMUs collect data at a rate of 100Hz, using data from the accelerometer, gyroscope and magnetometer. Thus, in total there are $3 \times 9 = 27$ sensor values.

A cross-validation is performed so that two persons are used for training and one for testing. In total, the recordings contain 10 min and 23.30 min of data for warehouse *A* and *B* respectively. There are seven foreground classes in the dataset: *walking, searching, picking* (i.e. taking an order from a shelf), *scanning, info* (i.e. interaction with a paper list or a handheld), *carrying* and *acknowledge* (i.e. signing on a paper list) as well as two background classes *unknown* and a *sensor flip* (which has been used for synchronization and marking the beginning and end of an order line). Some parts of the dataset have been annotated with NULL and are excluded from the analysis. Out of the original recordings there are $54,079$ frames which equals ≈ 9.01 min labeled data for warehouse *A* and $99,941$ frames which equals ≈ 16.39 min labeled data for warehouse *B*. An overview is given in Tab. 1. It can be seen that the data is highly imbalanced and that not all classes are used in both parts of the dataset. Warehouse *A* requires interaction with a paper list for guiding the workers. This activity is annotated as *info*. This part of the dataset also has an *acknowledge* activity, where the list is signed manually. Warehouse *B* on the other hand used a handheld device for guiding the worker such that the *info* activity is represented by interacting with the handheld device. The *acknowledge* activity is therefore replaced by *scanning*.

In [4], the evaluation has been performed purely in a classification setup where windows of varying sizes have been extracted from the data. Each window has been assigned the ground truth label that covers the largest number of frames within the window. The windows are then classified and the results have been reported for various classifiers and window sizes. For an optimal choice of classifiers and window sizes, classification rates of as high as 72.6% and 85.6% have been reported on set *A* and *B* respectively. Note that this setup could potentially suppress events which are much shorter than the window length.

The setup used in this work is more precise and in general capable of online processing. Choosing a fixed window size of $w = 1$ sec, the sliding window is moved forward one frame at a time ($10ms$). For training and evaluation, each window is assigned the label at its center. Thus, a slight context of the past and future is available, resulting in a processing delay of approx. $w/2$ sec.

Implementation details

All networks have been trained on the data from two persons. Additional data augmentation has been performed, as described in the method section. Gaussian noise with $\sigma = 0.01$ has been added to the data, as well as the proposed random re-sampling. The class imbalance has been reduced so that each class contains at least $b = 20\%$ of the largest classes number of samples.

For training the network, stochastic gradient descent with momentum has been used. The initial learning rate has been set to 10^{-5} with a momentum of 0.9. $2,000$ training iterations with a batch size of 50 have been performed, so that the network uses $100,000$ augmented samples for training. Having a very limited number of original samples, it could be observed that more iterations lead to overfitting. The learning rate is reduced by a factor of 10 after $1,000$ iterations.

Warehouse A

The results for warehouse *A* are shown in Tab. 2. The left side of the table shows the traditional approaches using statistical features. The right side shows the CNN approaches: the *activity* CNN from [5], the proposed *IMU* architecture and *IMU-global* for the intermediate architecture with a global temporal convolution per IMU in the first layer.

Three observations can be made from the data: First, the traditional approaches are outperformed by the CNNs. However, similar to several results in HAR, not always by a large margin. In contrast, the *activity* CNN baseline, shows a similar performance as statistical features in combination with a linear SVM. Most interestingly, the generative Bayes classifier outperforms all discriminative models on the third fold of the dataset. Second, it can be observed that the *IMU*-CNN is able to learn the data more accurately and ultimately obtains a better classification accuracy. It can be assumed that learning different feature representations for each sensor signal and later in the network for each of the IMUs is an advantage compared to a global representation. Third, when comparing the proposed *IMU*-CNN architecture to the *IMU-global* architecture, where the first convolution considers all channels at once, it can be seen that joining the information of the individual sensor values later in the CNN is beneficial.

	statistical features			CNNs		
	Bayes	Random Forest	SVM linear	activity	IMU-global	IMU
P1	64.8	64.3	66.6	63.9	67.7	**70.6**
P2	51.3	52.9	60.1	65.5	68.2	**70.5**
P3	**69.9**	63.5	64.1	60.1	65.0	66.7
Warehouse A	62.0 ± 7.8	60.2 ± 5.2	63.6 ± 2.6	63.1 ± 2.6	67.0 ± 1.4	$\mathbf{69.2 \pm 1.8}$

Table 2. Classification accuracy [%] for warehouse *A*. (left) statistical features. (right) CNN architectures.

	Bayes	Random Forest	SVM linear	activity	IMU-global	IMU
P1	58.0	49.5	39.7	35.9	48.8	**70.1**
P2	62.4	70.1	62.8	65.5	65.9	**71.3**
P3	**81.8**	79.0	77.2	74.9	73.5	80.3
Warehouse B	67.4 ± 10.3	66.2 ± 12.4	59.9 ± 15.4	58.8 ± 16.6	62.7 ± 10.3	$\mathbf{73.9 \pm 4.6}$

Table 3. Classification accuracy [%] for warehouse *B*. (left) statistical features. (right) CNN architectures.

Context	none	50	100	150	200	250
warehouse *A*	69.2 ± 1.8	70.2 ± 2.2	70.5 ± 2.5	70.8 ± 2.0	$\mathbf{71.0 \pm 2.0}$	71.0 ± 2.0
warehouse *B*	73.9 ± 4.6	74.8 ± 4.2	75.4 ± 3.9	75.9 ± 3.9	$\mathbf{76.0 \pm 4.0}$	75.5 ± 3.7

Table 4. Classification accuracy [%] of the IMU-CNN using different context sizes for post-processing the classification results.

Warehouse B

Similar results can be observed on the second part of the dataset using the recordings from warehouse *B*. The results are shown in Tab. 3. Again, the best results are obtained by the IMU-CNN architecture. However, the other CNN approaches that consider all sensor values in the first convolution are outperformed by the traditional approaches. Here, it can be observed that there is a huge inter person variance between person one and the other two. Therefore, most approaches show a poor performance when testing on this data. The generative model is able to show a better generalization capability here. This inter person variance also makes it difficult to learn meaningful filter operations for all sensor values at once. At a later stage in the network when more abstract features are learned to represent the single sensor values, the generalization is more easily possible.

Data Augmentation

Besides the influence of different design choices in the CNN architecture, it is interesting to study the effect of data augmentation. Therefore, the networks are also trained without any augmentation using each training sample once. It is known that class imbalance is often an issue in HAR [2] and neural networks do easily overfit without proper augmentation. However, the results in Tab. 5 show that the drop in performance without augmentation is relatively small. Especially on the second part of the dataset where more samples are available for training, there is almost no difference.

Training samples	Augm.	warehouse *A*	warehouse *B*
100,000	yes	$\mathbf{69.2 \pm 1.8}$	$\mathbf{73.9 \pm 4.6}$
33,000-82,000 (*)	no	65.9 ± 0.7	73.7 ± 6.1

(*) Number of training samples depends on the warehouse and train/test split.

Table 5. Classification accuracy [%] of the IMU-CNN using a varying number of training samples.

Post-processing

In the last experiment, the influence of post-processing is shown. Typically, sequential data that is processed in a sliding window approach exhibits noisy labels (cf. [16]). In these cases, a few single frames within the sequence are incorrectly classified. Therefore, a simple majority voting is applied to the labels in a pre-defined neighborhood, smoothing the classification results. If no majority is achieved, the label remains unchanged. The results are shown in Tab. 4. While this requires a contextual knowledge of up to 2.0 seconds, it is possible to improve the results by 1.8% and 2.1% for the data from warehouse *A* and *B* respectively.

CONCLUSION

In this work, a novel CNN architecture for HAR has been introduced. The architecture follows an IMU centered design. Convolutions are applied for each sensor value separately from which an intermediate representation per IMU is derived. Ultimately, the features from the different IMUs are joined in order to derive a global representation, which allows for classifying human activities. The CNN is integrated in a sliding window approach for recognizing different classes in sequential activity data.

The proposed approach has been evaluated for the process of order picking using realistic data from two different warehouses. It has been shown that applying convolutions per sensor value as well as the IMU centered approach are beneficial for learning a good classifier. Furthermore, the influence of data augmentation and post-processing the labeled sequence has been investigated. While data augmentation is able to improve the performance of a network, a good accuracy can still be reached when using the plain data without any augmentation. Post-processing the labeled sequence by incorporating up to two seconds of context improves the results. On both parts of the dataset, traditional approaches based on statistical features as well as recent CNN architectures have been outperformed by the proposed network.

ACKNOWLEDGMENT

This work has been supported by the German Research Foundation (DFG) within project Fi799/9-1 ('Partially Supervised Learning of Models for Visual Scene Recognition') and Fi799/10-1 ('Adaptive, Context-based Activity Recognition and Motion Classification to Analyse the Manual Order Picking Process').

REFERENCES

1. Marc Bächlin, Daniel Roggen, Gerhard Tröster, Meir Plotnik, Noit Inbar, Inbal Maidan, Talia Herman, Marina Brozgol, Eliya Shaviv, Nir Giladi, and others. 2009. Potentials of Enhanced Context Awareness in Wearable Assistants for Parkinson's Disease Patients with the Freezing of Gait Syndrome.. In *ISWC*. 123–130.
2. Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)* 46, 3 (2014), 33.
3. Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R Millán, and Daniel Roggen. 2013. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters* 34, 15 (2013), 2033–2042.
4. Sascha Feldhorst, Mojtaba Masoudinejad, Michael ten Hompel, and Gernot A. Fink. 2016. Motion Classification for Analyzing the Order Picking Process Using Mobile Sensors. In *Proc. Int. Conf. Pattern Recognition Applications and Methods (ICPRAM)*. Rome, Italy.
5. Nils Y Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables. In *Proc. Int. Joint Conference on Artificial Intelligence (IJCAI)*.
6. R. Koster, T. Le-Duc, and K. J. Roodbergen. 2006. *Design and control of warehouse order picking: a literature review*. ERIM report series research in management Business processes, logistics and information systems, Vol. 5. ERIM, Rotterdam.
7. Matthias Kreil, Bernhard Sick, and Paul Lukowicz. 2016. Coping with variability in motion based activity recognition. In *Proceedings of the 3rd International Workshop on Sensor-based Activity Recognition and Interaction*. ACM, 4.
8. M. Krengel, M. Schmauder, Thorsten Schmidt, and K. Turek. 2010. *Beschreibung der Dynamik manueller Operationen in logistischen Systemen: Schlussbericht*. Dresden. available German only.
9. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.
10. O. D. Lara and M. A. Labrador. 2013. A Survey on Human Activity Recognition Using Wearable Sensors. *IEEE Communications Surveys and Tutorials* 15, 3 (2013), 1192–1209.
11. H. Martin. 2009. *Transport- und Lagerlogistik: Planung, Struktur, Steuerung und Kosten von Systemen der Intralogistik* (7th ed.). Vieweg+Teubner Verlag / GWV Fachverlage GmbH Wiesbaden, Wiesbaden. available German only.
12. Charissa Ann Ronao and Sung-Bae Cho. 2015. Deep convolutional neural networks for human activity recognition with smartphone sensors. In *Proc. Int. Conference on Neural Information Processing*. Springer, 46–53.
13. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252.
14. K Simonyan and A Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1 (2014).
15. M. ten Hompel, B. Vogel-Heuser, and T. Bauernhansl (Eds.). 2014. *Industrie 4.0 in Produktion, Automatisierung und Logistik: Anwendung, Technologien, Migration*. Springer Vieweg, Wiesbaden. available German only.
16. 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.. In *IJCAI*. 3995–4001.
17. Sergey Zagoruyko and Nikos Komodakis. 2016. Wide Residual Networks. *arXiv preprint arXiv:1605.07146* (2016).