DOI: 10.2195/lj_Proc_grzeszick_en_201610_01
URN: urn:nbn:de:0009-14-44556

Camera-assisted Pick-by-feel

Wearable assistiertes Kommissionieren

***Rene Grzeszick*[1]**
***Sascha Feldhorst*[2]**
***Christian Mosblech*[3]**
***Michael ten Hompel*[2]**
***Gernot A. Fink*[1]**

[1] *Pattern Recognition in Embedded Systems Group, Dept. of Computer Science*
TU Dortmund University

[2] *Chair of Materials Handling and Warehousing, Dept. of Mechanical Engineering*
TU Dortmund University

[3] *prismat Gesellschaft für Softwaresysteme und Unternehmensberatung mbH*

In this contribution a novel system to support order pickers in warehouses is introduced. In contrast to existing solutions it utilizes the tactile perception in order to reduce the systems impact on the visual and auditive senses. Therefore, a smartwatch and a low-cost camera which are both worn by the picker are combined with activity and object recognition methods for surveying the picking process. The activity recognition is used in order to determine whether an object is picked. Then, barcode detection and a CNN (Convolutional Neural Network) based object recognition approach are employed for recognizing whether the correct item is chosen. Beside the conceptional work, implementation details and evaluation results under realistic conditions and on a publicly available dataset are presented.

[Keywords: Order Picking, Deep Learning, Image Classification, Image Retrieval, Activity Recognition]

In diesem Beitrag wird ein neuartiges System zur Unterstützung des manuellen Kommissionierprozesses vorgestellt. Im Gegensatz zu existierenden Systemen wird taktiles Feedback genutzt um den Einfluss auf die audio-visuellen Sinne zu reduzieren. Eine vom Kommissionierer getragene Smartwatch und eine kostengünstige Kamera werden kombiniert mit Methoden der Aktivitätserkennung und der visuellen Objekterkennung, die den Kommissionierprozess überwachen. Zuerst wird mit Hilfe der Aktivitätserkennung bestimmt ob ein Gut vom Kommissionierer gegriffen wird. Im Folgenden werden eine Barcode-Erkennung und eine visuelle Objektklassifikation durch ein tiefes Faltungsnetz genutzt um zu erkennen ob das korrekte Gut gegriffen wurde. Neben der konzeptionellen Ausarbeitung werden Umsetzungsdetails erläutert und abschließend wird eine Auswertung unter realistischen Bedingungen sowie auf einem öffentlich verfügbaren Datensatz vorgestellt.

[Schlüsselwörter: Kommissionierung, Deep Learning, Bild Klassifikation, Bild Retrieval]

1 INTRODUCTION

Recently, many researchers and experts expect Cyber Physical Systems (CPS) to significantly change the way that industrial processes are controlled and automated [20]. These embedded systems are bridging the gap between the field-level, higher-level systems (e.g. ERP-systems) and Internet-based cloud solutions by means of communication. Nevertheless, human work will still play an important role in industrial systems of the future, especially in the field of logistics. For instance, the order picking process, where a list of items is collected in a warehouse, is often done manually even in huge warehouses like the ones operated by Amazon or Zalando. Especially in high-wage countries, human work is a significant cost driver and, therefore, often subject of optimization efforts [11]. Many different technologies, algorithms and design principles were developed to improve the efficiency of the manual order picking process [8].

One common approach is to guide the order picker with technical artifacts to quickly find the next order line and to reduce amount of picking errors [19]. Common solutions utilize head phones (pick-by-voice), lights on the racks or boxes (pick-by-light, put-to-light). A recent alternative uses head-mounted displays with augmented reality features (pick-by-vision). These head-mounted displays enrich the pickers' field of vision with all required data for processing the next order line [17]. Aside from static information, this includes dynamic data which is dependent on the pickers' current position and viewing direction [16]. While pick-by-voice and pick-by-light systems are widely adopted in real-world systems, the pick-by-vision approach has not reached a significant market share yet. On the one hand the difficulty of matching a virtual reality interface with a complete warehouse has to be faced and on the other hand working with such an interface is rather uncommon and very exhausting for the worker. The key functions of these systems are to provide the worker with all relevant information (e.g. the place of the next pick or

DOI: 10.2195/lj_Proc_grzeszick_en_201610_01
URN: urn:nbn:de:0009-14-44556

Figure 1. Overview of the proposed system: A smartwatch is used as the main interaction interface, displaying information and giving feedback to the user. A smartphone is worn on the workers chest. It recognizes the current activity and, if necessary, sends images to a server application for further analysis.

the amount of items to pick), to let him confirm picks, to keep his hands free and as a result to improve the overall speed of the order picking process.

In this paper Camera-assisted Pick-by-feel is introduced, a novel system for guiding and supporting order pickers. The system is based on standard consumer electronics components to reduce the costs for equipping all order pickers. It is designed to influence the work flow as little as possible. Similar to pick-by-vision, it incorporates activity and object recognition methods to survey the picking process, but the augmented reality features are replaced by a tactile feedback and interaction interface using a smartwatch. The system has been implemented and evaluated in a realistic scenario showing the capabilities of the proposed approach[1].

2 METHOD

Camera-assisted Pick-by-feel consists of three main components: a smartwatch (*Picker Watch*), a smartphone (*Picker Phone*) and a server running an application for recognizing items (*Picker Server*). The smartwatch displays all information necessary for the next pick and gives tactile feedback with respect to the pick. The smartphone is worn at the picker's chest with the help of a special belt. It monitors the current activity, e. g. whether he is walking or possibly making a pick. If a pick is recognized a stream of images is sent to the server application for further analysis. The server application manages the list of orders and assigns them to the pickers. If a picker handles an item, the application uses barcode detection and object recognition in order to determine whether the correct item is picked. The system and its main components are displayed in Fig. 1 and explained in more detail in the following.

2.1 PICKER WATCH

The *Picker Watch* is the main interaction interface for the picker. It displays all information about the next order line, i.e. the name of the product, its location in the warehouse and the number of items to be picked. It gives a tactile feedback whether a pick has been correct or not (i.e. a short or long vibration pattern). In case of a correct pick, it automatically displays the next order line. Furthermore, it allows the user to skip the current order line, to mark a product as missing, or to manually confirm an order line.

2.2 PICKER PHONE

The *Picker Phone* is monitoring the pickers' activity as well as the picking process. Furthermore, it is the communication interface between the Picker Watch and the Picker Server. It is worn on the pickers' chest with the camera facing outwards. The phone is worn at a height similar to the one indicated in Fig. 1.

2.2.1 ACTIVITY RECOGNITION

An activity recognition process [1, 2] is running in the background which distinguishes between walking and a pick. Activity recognition has been applied to the analysis of various situations in everyday life (cf. [2]). Furthermore, a similar approach has been proposed for measuring logistic activities in warehouses [5]. The acceleration sensor, the gyroscope and the magnetic field sensor of the phone are used for determining the picking process. The output values in x, y and z direction are selected from the sensors. Note that the sampling rates of the three sensors can differ (i.e. in the experimental setup, 99.74Hz, 199.52Hz and 47.89Hz). Hence, a sliding window over a period of ω_t seconds is used for deriving the features. For recognizing human activities, like gestures

[1] The authors would like to thank the Project Group Camera-assisted Pick-by-feel: Malte Baumann, Matthias Fey, Sebastian Kurth, Christian Matuschek, Matthias Neuhaus, Sigo Rosenkranz, Florian Schulz, Daniel Schulze-Bisping, Birol Sevim, Uthenthira Sivapatham and Jannik Zappe. All of them made a huge contribution to put a prototype of the system into effect and demonstrating the systems' capabilities.

DOI: 10.2195/lj_Proc_grzeszick_en_201610_01
URN: urn:nbn:de:0009-14-44556

Figure 2. Example of the pre-processing for the barcode detection. Only the detected candidate regions are evaluated in order to speed-up the processing.

or movement, statistical measures are shown to work well [2]. The following measures are currently used for the activity recognition: minimum, maximum, mean, standard deviation and the norm. This yields a 45 dimensional feature vector which is the input for the classification. The classification result is assigned to the center of the sliding window and it is moved forward by ω_s seconds at a time. Thus, there is only a processing delay of $\frac{\omega_t}{2}$ seconds. In the experiments it will be shown that good results can be achieved with $\frac{\omega_t}{2} = 0.5$ seconds, which is enough for all practical applications.

2.2.2 Image Acquisition

Rather than streaming continuously, images from the camera are automatically sent to the server when a possible pick is recognized without any further interaction from the picker. This allows reducing bandwidth and server workload as well as the risk of possible mis-classifications by the subsequent recognition.

2.3 Picker Server

The *Picker Server* manages the orders and sends the next order line to a picker. If an image is sent to the server from a Picker Phone, it tries to recognize whether the correct item is chosen. Two simultaneous analysis steps are carried out: a barcode detection and an object recognition. An object clearly displaying a barcode is the most secure way of checking whether the correct pick has been made or not. In the other cases a CNN based approach for identifying the item shown on the image is proposed.

2.3.1 Barcode Detection

The goal of the barcode detection is to find and recognize EAN-13 barcodes within a given image I. There are various implementations [3, 7, 10], however, mostly in very controlled environments. Here, the images are more cluttered. It can neither be guaranteed that the barcode is shown in a readable manner on the object nor that it is visible at all. Following common approaches, a line scanning method is applied that binarizes the image and identifies a barcode by decoding the sequences of black and white bars (cf. [3]). However, scanning the complete image is a very time consuming process. In order to tackle this issue, similar to [7], some pre-processing steps are executed that determine good candidates for barcode regions and allow for speeding-up the process: First, based on the knowledge that barcodes are black and white lines that should produce nice gradients, a gradient image I_s is computed by convolving the original image with a Scharr Operator S: $I_s = I * S$ [18]. Then, a threshold t is applied in order to select only prominent edges and binarize the image:

$$I_t(x,y) = \begin{cases} 0 \; if \; I_s(x,y) < t \\ 1 \; otherwise \; . \end{cases}$$

Following the binarization, a closing and opening [18] is applied in order to produce homogeneous regions that contain the prominent gradients. A set of image patches $\boldsymbol{I_p}$ is recognized by a connected component analysis. An example of this process is shown in Fig. 2. Patches that are too small for reliably identifying barcodes are discarded.

Second, in order to improve the readability the regions $\boldsymbol{I_p}$ are sharpened. Therefore, each patch I_p^l is first smoothed by a Gaussian G:$\hat{\mathrm{I}}_p^l = I_p^l * G$ and then the smoothed patch is subtracted from the original one $\tilde{I}_p^l = I_p^l - \hat{\mathrm{I}}_p^l$ leaving only the prominent edge information. The line scanning algorithm is then applied to the extracted, sharpened patches $\tilde{\boldsymbol{I}}_{\boldsymbol{p}}$. The improvement that is achieved for detecting barcodes in cluttered scenes by the pre-processing steps is analyzed in the evaluation.

2.3.2 CNN Object Recognition

As the object must not necessarily show a barcode, the image is in parallel processed with object recognition methods. Given a set of products in stock and an arbitrary set of images of these items, the task can either be formulated as a classification or an image retrieval task. In the first case, all images of a given item comprise a class and the task can be solved by a multiclass classification of a query image. In the second case, the query image is used for retrieving the most similar item from all product images. The retrieval might be the more realistic definition as it is only important to find the specific object instance from the given angle. The multi-class approach on the other hand might be able to generalize better as each class contains the information of an item shown from multiple viewpoints.

DOI: 10.2195/lj_Proc_grzeszick_en_201610_01
URN: urn:nbn:de:0009-14-44556

Figure 3. Experimental setup simulating a worker picking items from a rack.

In recent years deep learning architectures improved the state-of-the-art in object recognition [4, 9], detection [21], segmentation [15] and retrieval [13]. Training a CNN typically requires a large number of training samples and is computationally expensive. There are two remedies for this issue: First, additional training samples can be generated using data augmentation techniques which then allow to train a network with only a few original example images [15]. Second, it is possible to pre-train a CNN on a large dataset (i.e. ImageNet) and then adapt the networks parameters by a second training phase using samples from a different domain (cf. [13]). This iterative process is often referred to as *fine-tuning*. Here, a version of Alexnet [9] is adapted to the task of order picking. The set of images that represents the products from the warehouse is used for fine-tuning the network. Either the output of the last layer can be used for classification or the activations of one of the networks' intermediate layers can be used as a feature representation for retrieval. In [4] it has been shown that especially the last pooling layers or the fully connected layers show promising results as features.

Given a feature representation, derived from a network layer f, a distance metric can be used for retrieving the most similar items. Typical distance measures are the Euclidean distance, cosine distance

$$d_j = 1 - \frac{f_j \cdot f_q}{||f_j||\,||f_q||} \;\forall j$$

or the Bray-Curtis dissimilarity

$$d_j = \frac{\sum_i |f_j^i - f_q^i|}{\sum_i f_j^i + f_q^i} \;\forall j$$

If the object in question is retrieved, the picked item is assumed to be the correct one. The evaluation will compare the retrieval based approach with the multi-class classification.

Table 1. Results of the four-fold cross validation for the activity recognition using recordings from the different participants.

Test	SVM	Bayes	Rand. Forest
P1	36.6%	81.0%	**82.8%**
P2	25.3%	**85.9%**	77.3%
P3	48.8%	78.2%	**82.0%**
P4	39.1%	74.9%	**78.3%**
Avg.	37.5 ± 8.4%	80.0 ± 4.0%	**80.1 ± 2.3%**

Table 2. Confusion matrix for P3. Results are shown as the number of windows per annotation.

True / Prediction	NULL	WS	WC	PICK	FLIP
NULL	92	2	0	19	39
WS	19	2658	74	105	4
WC	18	189	295	113	0
PICK	44	35	13	855	1
FLIP	7	0	0	0	205

3 Evaluation

An experimental setup has been created at the Logistics Campus at TU Dortmund University containing a rack with 31 products as shown in Fig. 3. A dataset of 408 images showing the products has been recorded (approx. 13 per product). 237 of these images show the products from different perspectives and are used for training while 171 show the products being held by a worker and are used for testing. The training images have been obtained by taking images of the products on the rack. Note that taking images in a controlled setup would also be possible in many practical applications and reduce the complexity of the image recognition task. Here, the setting is completely uncontrolled and, therefore, more complicated but also more easy to implement. Furthermore, a dataset of 20 min. of order picking activities has been recorded under realistic conditions. Different persons have been tasked with picking items of various shapes from different position in shelving racks in a warehouse-like scenario.

For the hardware components, a Pebble Watch, a Samsung Galaxy S4 and a standard PC have been used. Qualitative and quantitative results are reported in the following.

3.1 Activity Recognition

For evaluating the picking activity, four different persons have been tasked with a list of items to pick in a realistic setting. In the setup no guidance system like Pick-by-Voice has been used, the transportation of the picked products was done with a cart and the products were stored at different heights and positions inside the racks. All persons had the same list of items and repeated the task twice. Each run took around 2:5 min., yielding a total of 20 min. activity data. A four-fold cross-validation has been performed using the data of three persons for training and one for testing.

DOI: 10.2195/lj_Proc_grzeszick_en_201610_01
URN: urn:nbn:de:0009-14-44556

Table 3. Recognition rates of the barcode detection with and without the proposed pre-processing steps on the images from the dataset that actually show a barcode.

Method	Recognition of clearly visible barcodes	Recognition of all barcodes	Runtime
w/o pre-processing	78%	41%	> 2000ms
w/o sharpening	78%	41%	**260ms**
with sharpening	**89%**	**47%**	310ms

The annotation of the data comprises five different activity classes: *start stop motion* (FLIP), *walking straight* (WS), *walking curve* (WC), *picking* (PICK), *none* (NULL) with the class of interest being picking.

The windows size was set to $w_t = 1$s and the step width to $w_s = 0.036$s which results from the shortest time frame that is typically used for measuring logistic activities (i.e. Methods-Time Measurement; MTM [6]). Hence, the processing delay is reduced to 0.5s which is enough for recognizing the object in the workers hand in most practical scenarios.

The results for different classifiers and different persons in the test set are shown in Tab. 1. Note that the random forest shows the most stable performance over all four participants. SVMs using an RBF kernel seem to perform rather poorly on this time series analysis, which has also been observed in similar tasks dealing with the analysis of time series (cf. [12]). Having a more detailed look at the confusion matrix of one of the participants in Tab. 2 it is revealed that the picking process is recognized quite well. About 90.1% of the picks are correctly recognized. It is rather the case that the picking is over estimated by the classifier, which is less of an issue than missing a pick completely.

3.2 Barcode Detection

A subset of 85 images showing a barcode has been used for evaluating the proposed barcode detection method. From these 85 images, 45 clearly showed a barcode, whereas the remaining images are more blurry or the barcode is not clearly visible. The recognition rates are shown in Tab. 3. The first row shows the analysis of the complete image. The second row shows the result of the line scanning only being applied to the set of extracted image patches $\boldsymbol{I_p}$. Note that the runtime can be greatly improved by the proposed pre-processing without sacrificing accuracy. For the parameterization of the pre-processing the threshold t is set to 225 and S is a 3x3 Scharr Operator. In addition, it can be seen that the sharpening of the image patches $\boldsymbol{I_p}$ further improves the recognition rates. Here, a Gaussian G of size 17x17 is applied to the image patches in order to emphasize the prominent edges.

3.3 Order Picking Item Recognition

The 171 test images showing objects held by a worker have been identified using the proposed CNN based object recognition approaches. The multi-class classification approach using the CNNs output layer has been compared with the retrieval approach based on different network layers. For the retrieval, an image I_q showing a hand-held object is used for querying the set of train images that were recorded at the shelf. Note that it is not necessary to retrieve all images of a given product as different perspectives have been recorded. It is rather important to find one instance that matches the product held by the worker. Therefore, the accuracy of the best retrieval result is reported.

Feature representations that were derived from different intermediate layers of the Neural Network have been evaluated: the last pooling layer (denoted as *p5*) and both fully connected layers (denoted as *fc6* and *fc7*). Furthermore, as the orientation of hand-held objects is not clear and the set of training images is limited, a data augmentation scheme that tackles this issue has been evaluated. All objects in the training set have been rotated in 45 degree steps and each of these images has been mirrored yielding a 16 times larger training set. Although augmentations do not necessarily need to resemble a completely realistic object, as they only need to be preserving the original classes, the rotation of objects is a very realistic augmentation for handling goods in an order picking process. Hence, the network can be fine-tuned without seeing an image multiple times, but is rather observing slightly different images, reducing the risk of overfitting. Note that the augmented images are not only used for fine-tuning, but the feature representations are also used for the retrieval itself.

The results are shown in Fig. 4. Three observations can be made from the experiments: 1) Augmenting the training images provides more robustness toward the arbitrary orientation of the test data. 2) The last pooling layer and the first fully connected layer provide the most discriminative features, which is similar to the classification results shown in [4]. 3) The cosine distance shows the best results and outperforms the other distance metrics with an accuracy of 83.6% using the last pooling layer.

DOI: 10.2195/lj_Proc_grzeszick_en_201610_01
URN: urn:nbn:de:0009-14-44556

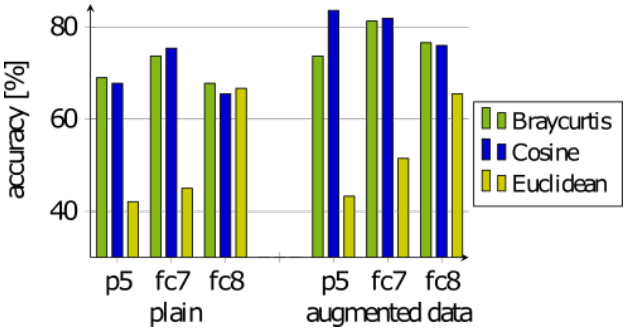

Figure 4. Comparison of different distance metrics for the image retrieval using different layers of the CNN as input features. The average accuracy over all query images is reported.

In the following, the retrieval results have then been compared to the multi-class classification where each product in stock comprises one of the classes. In addition, a Bag-of-Features image retrieval approach is shown as a baseline. It uses SIFT features whose location and orientation at points of interest are determined using the SIFT detector to account for rotations of the objects. The visual vocabulary has a size of $|V| = 1.500$ visual words. A cosine distance is used for determining the most similar object. The results are shown in Fig. 5. It can be observed that the image retrieval clearly outperforms the multi-class classification approach and that the features derived from the CNN outperform the Bag-of-Features approach. Although the CNN has been pre-trained with millions of images from ImageNet, this is not obvious as the collection of storage units is very specialized (cf. Fig. 3).

3.4 Small hand-held objects dataset

For comparison, the method has also been evaluated on the SHORT dataset for small hand-held object recognition [14]. The dataset is comprised of still images and video frames with annotated training and test images being available for 30 products. Although a little movement jitter might occur, the still images seem to be the more appropriate setting for the proposed application. Each training image has been recorded under pre-defined conditions from different angles and scales. However, the dataset is more complicated as all products are small and as a result heavily occluded in the test set. The authors therefore propose a Nearest Neighbor comparison of all local features (cf. [14]) which is however computationally expensive to evaluate and therefore hardly feasible for the application at hand. The accuracy of the different approaches is reported in Fig. 6. Similar trends as for the dataset of order picking items can be observed. Again the CNN features are very robust and useful for the application at hand and again the augmentation of the training images improves the recognition results. In contrast to the order picking items the performance of the classification comes very close to the retrieval accuracy. Here, the retrieval achieves an accuracy of 43.1% using the cosine distance on the first fully connected layer.

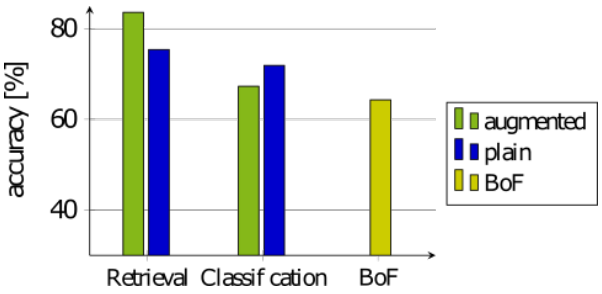

Figure 5. Comparison of the retrieval approach, a multi-class classification using a CNN and a Bag-of-Features baseline approach.

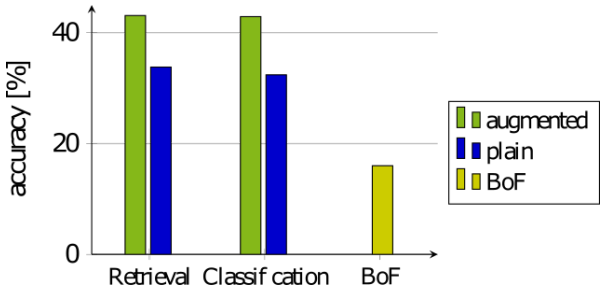

Figure 6. Recognition results on the SHORT dataset. Comparison of the retrieval approach, a multi-class classification using a CNN and a Bag-of-Features baseline approach.

3.5 Discussion

The results of proposed method are very promising. Especially, as in practice additional information can be exploited. It is known in advance which item should be picked from the rack. Therefore, the images retrieval only needs to verify that the pick is correct (i.e. the same item is retrieved as listed on the order line). Furthermore, multiple images can be sent to the server for analysis. It is not necessary to recognize the product on one image, but rather be able to make a reliable decision after a few images. While in practice the whole warehouse might contain several hundreds or thousands of items compared to the very small set of objects in the experimental setup, in practice the position of the pick and knowledge about its neighboring stock items helps limiting the set of candidate objects.

Qualitative experiments with a first prototypical system were conducted at the Logistics Campus. An evaluation protocol has been designed with a list of order lines and corresponding picks. Several correct items and a few incorrect items were chosen from a rack by a worker. It could be shown that after a few images have been analyzed quite stable recognitions can be made.

4 Conclusion

In this paper a new system design to support order pickers in warehouses has been presented. It builds on consumer electronics and uses activity and object

DOI: 10.2195/lj_Proc_grzeszick_en_201610_01
URN: urn:nbn:de:0009-14-44556

recognition methods for recognizing picks. Using a prior activity recognition for determining whether a worker is picking an object, images are sent to a server application for further analysis. A barcode detection and a CNN based object recognition approach are employed for recognizing whether the correct object has been picked. Here, a retrieval based approach that incorporates data augmentation showed very promising results. Tactile feedback about the pick is given to the user. The evaluation demonstrated the systems' capabilities as an alternative to the established order picking approaches.

LITERATURE

[1] L. Atallah and G.-Z. Yang. The use of pervasive sensing for behaviour profiling – a survey. Pervasive and Mobile Computing, 5(5):447–464, 2009.

[2] A. Bulling, U. Blanke, and B. Schiele. A tutorial on human activity recognition using body-worn inertial sensors. ACM Computing Surveys (CSUR), 46(3), 2014.

[3] D. Chai and F. Hock. Locating and decoding EAN-13 barcodes from images captured by digital cameras. In IEEE Int. Conf. on Information, Communications and Signal Processing, 2005.

[4] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In Int. Conf. on Machine Learning, 2014.

[5] S. Feldhorst and M. ten Hompel. Bewegungsklassifikation mithilfe mobiler sensoren zur analyse des kommissionierprozesses: Motion classification of the order picking process using mobile sensors. In Tagungsband 11. Fachkolloquium der WGTL. Duisburg, 2015.

[6] T. Gudehus and H. Kotzab. Comprehensive Logistics. Springer, Berlin, 2nd edition, 2012.

[7] M. Katona and L. G. Nyúl. A novel method for accurate and efficient barcode detection with morphological operations. In IEEE Int. Conf. on Signal Image Technology and Internet Based Systems (SITIS), 2012.

[8] R. D. Koster, T. Le-Duc, and K. J. Roodbergen. Design and control of warehouse order picking: a literature review. Volume 5 of ERIM report series research in management Business processes, logistics and information systems, 2006.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, 2012.

[10] D.-T. Lin, M.-C. Lin, and K.-Y. Huang. Real-time automatic recognition of omnidirectional multiple barcodes and DSP implementation. Machine Vision and Applications, 22(2), 2011.

[11] H. Martin. Transport- und Lagerlogistik: Planung, Struktur, Steuerung und Kosten von Systemen der Intralogistik. Vieweg+Teubner Verlag / GWV Fachverlage GmbH Wiesbaden, Wiesbaden, 7th edition, 2009.

[12] A. Plinge, R. Grzeszick, and G. A. Fink. A Bag-of-Features Approach to Acoustic Event Detection. In IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 2014.

[13] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2014.

[14] J. Rivera-Rubio, S. Idrees, I. Alexiou, L. Hadjilucas, A. Bharath, et al. Small hand-held object recognition test (short). In IEEE Winter Conference on Applications of Computer Vision (WACV), 2014.

[15] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention MICCAI, 2015.

[16] B. Schwerdtfeger. Pick-by-vision: Bringing HMD-based augmented reality into the warehouse. PhD thesis, Technische Universität München, 2010.

[17] B. Schwerdtfeger, R. Reif, W. A. Gunthner, G. Klinker, D. Hamacher, L. Schega, I. Bockelmann, F. Doil, and J. Tumler. Pick-by-vision: A first stress test. In IEEE Int. Symposium on Mixed and Augmented Reality (ISMAR), 2009

[18] R. Szeliski. Computer vision: algorithms and applications. Springer, 2010.

[19] M. ten Hompel, V. Sadowsky, and M. Beck. Kommissionierung: Materialflusssysteme 2 - Planung und Berechnung der Kommissionierung in der Logistik. VDI-Buch. Springer, Berlin, 2011.

[20] M. ten Hompel, B. Vogel-Heuser, and T. Bauernhansl, editors. Industrie 4.0 in Produktion, Automatisierung und Logistik: Anwendung, Technologien, Migration. SpringerLink. Springer Vieweg, Wiesbaden, 2014.

[21] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Partbased R-CNNs for fine-grained category detection. In European Conference on Computer Vision (ECCV), 2014.

Rene Grzeszick, M.Sc., PhD Student in the Pattern Recognition in Embedded Systems Group, TU Dortmund University. He received the bachelor's and master's degree in computer science from TU Dortmund University, Germany, in 2010 and 2012.

Address: Computer Science XII, TU Dortmund, Otto-Hahn Str. 16, 44227, Dortmund, Germany
Phone: +49 231 755-4642
Email: rene.grzeszick@tu-dortmund.de

Prof. Dr.-Ing. Gernot A. Fink, head of the Pattern Recognition in Embedded Systems Group, TU Dortmund University. He received his diploma in computer science from the University of Erlangen-Nuremberg, Germany, in 1991. From 1991 to 2005, he was with the Applied Computer Science Group at Bielefeld University, Germany, where he received his Ph.D. degree (Dr.-Ing.) in 1995 and his venia legendi (Habilitation) in 2002. Since 2005, he has been a professor TU Dortmund University. His research interests are machine perception, statistical pattern recognition, and document analysis**.**

Address: Computer Science XII, TU Dortmund, Otto-Hahn Str. 16, 44227, Dortmund, Germany
Phone: +49 231 755-6151
Email: gernot.fink@tu-dortmund.de

Dipl.-Inform. Sascha Feldhorst., PhD Student at the Chair of Materials Handling and Warehousing, TU Dortmund University. He received his diploma in computer science from TU Dortmund University in 2008.

Address: Chair of Materials Handling and Warehousing, TU Dortmund, Otto-Hahn Str. 16, 44227, Dortmund, Germany
Phone: +49 231 755-4073
Email: sascha.feldhorst@tu-dortmund.de

Dipl.-Inform. Christian Mosblech., PhD Student and former scientific researcher at the Chair of Materials Handling and Warehousing, TU Dortmund University. In 2007 he received his diploma in computer science from TU Dortmund University. Since 2016 he is a SAP Software Consultant at prismat GmbH.

Address: prismat Gesellschaft für Softwaresysteme und Unternehmensberatung mbH, Stockholmer Allee 30c, 44269, Dortmund, Germany
Phone: +49 231 567-6300
Email: christian.mosblech@prismat.de

Prof. Dr. Michael ten Hompel, head of the Chair of Materials Handling and Warehousing, TU Dortmund University. He is also head of the Fraunhofer Institute for Material Flow and Logistics.

Address: Chair of Materials Handling and Warehousing, TU Dortmund, Otto-Hahn Str. 16, 44227, Dortmund, Germany
Phone: +49 231 9743-600
Email: michael.tenHompel@tu-dortmund.de