Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 00 (2017) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

The 14th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2017)

Recognizing Grabbing Actions from Inertial and Video Sensor Data in a Warehouse Scenario

Alexander Diete[a,*], Timo Sztyler[a], Lydia Weiland[a], Heiner Stuckenschmidt[a]

[a] University of Mannheim, B6 26, 68159 Mannheim, Germany

Abstract

Modern industries are increasingly adapting to smart devices for aiding and improving their productivity and work flow. This includes logistics in warehouses where validation of correct items per order can be enhanced with mobile devices. Since handling incorrect orders is a big part of the costs of warehouse maintenance, reducing errors like missed or wrong items should be avoided. Thus, early identification of picking procedures and items picked is beneficial for reducing these errors. By using data glasses and a smartwatch we aim to reduce these errors while also enabling the picker to work hands-free. In this paper, we present an analysis of feature sets for classification of grabbing actions in the order picking process. For this purpose, we created a dataset containing inertial data and egocentric video from four participants performing picking tasks, modeled closely to a real-world warehouse environment. We extract features from the time and frequency domain for inertial data and color and descriptor features from the image data to learn grabbing actions. By using three different supervised learning approaches on inertial and video data, we are able to recognize grabbing actions in a picking scenario. We show that the combination of both video and inertial sensors yields a F-measure of **85.3%** for recognizing grabbing actions.

Keywords: machine learning, sensor fusion, action recognition

1. Introduction

In the field of modern warehouses a lot of attention is put on improving the process of order picking regarding accuracy and time to save on costs[1,2,3]. Order picking means the collection of items that make up an order for customers. Errors in this process are expensive because of the big organizational overhead of fixing an incorrect order. By using modern wearable technologies like data glasses and smartbands or -watches, the picker can be better aided and supported, thus minimizing the errors. Employees would immediately know they make an incorrect pick and could act accordingly early on. In addition, wearables could free up the workers hands and guide them to the correct item. This is especially useful for training new employees who have yet to learn each single step in the picking process. Solutions for improving the picking process can be grouped into two categories: 1.) The first category aims

[*] Corresponding author. Tel.: +49-621-181-2650.
E-mail address: alex@informatik.uni-mannheim.de

to equip the pickers with tools to speed up or even remove parts of their workload. This could be done by equipping pickers with voice control systems[4] or by giving the worker wearable devices that directly scan the item[5]. 2.) The second category augments the warehouse to reduce picking time and improve accuracy. An example could be the highlighting of shelves to be picked from while simultaneously showing the needed amount of the item[6]. Another example is the usage of RGBD-cameras to recognize item picking from a shelf[7].

Our work is within the first category, as it should be adaptable to different warehouses without a long installation process. In this work, we explore the usage of wearable devices for aiding the picking process. These devices include data glasses and a smartwatch that are worn by a picker. We focus on video and inertial data. In our case inertial data includes acceleration, gyration, and magnetic field. By considering both modalities at the same time we can deal with the shortcomings of each: video data may not capture the full motion of the arm while inertial data can be prone to wrongly identify arm movement as grabbing. We also put emphasis on finding the correct start of the action. This way we have the longest time to identify which item the picker is picking and can start the validation process early. For this purpose, we pose two research questions:

RQ1: Can inertial and video data be used to classify grabbing actions? Can we find the exact start of an action?

RQ2: What subset of features are best suitable for that task?

To answer these questions, we create a dataset for the picking scenario. It includes multiple participants performing different picking tasks in a simulated warehouse environment. We then analyze whether we can learn to distinguish grabbing actions from non-grabbing actions within this dataset.

The paper is structured as follows: In Section 2, we describe existing work in the field of multi sensors and feature selection in context of activity and action recognition. Afterwards, we describe our dataset in Section 3. Section 4 covers our methodology with a focus on the features we select for our experiments. These experiments are described in Section 5. Finally we conclude the results in Section 6 and give an outline for our future work.

2. Related Work

Modern warehouses often rely on RFID or QR codes to validate orders[1]. While these approaches are very precise, the validation happens at a late stage. By using wearables we aim to register the picking action earlier. This way the picker may know the location of the correct item early on which can be especially useful when training new employees. In this paper, we deal with action recognition on multi sensor data and the influence of different feature set on recognizing the action. We consider an action as an atomic subpart of an activity like a single step in a walking activity. On one hand, we look at work in the field of sensor fusion as we work with inertial and video data simultaneously. On the other hand, we look at related work in the field of activity recognition with a focus on feature selection as it is related to our approach of action recognition. Indeed, Kwapisz et al.[8] used acceleration data from a smartphone for activity recognition. By extracting features from short time intervals they are able to predict movement activities like walking, climbing stairs and jogging. Similarly, Preece et al.[9] did a feature analysis on accelerometer data for activity recognition. They consider sensors placed on different body parts to also recognize movement activities. A strong focus is put on comparing wavelet features to time and frequency features. Recently, San-Segundo et al.[10] used accelerometer features from smartphones for human activity segmentation. Their feature groups can be grouped in time based features and frequency based features to be classified with Hidden Markov Models. Neural networks for human activity recognition have been researched by Ordóñez et al.[11]. With a deep neural network, they are able to get high accuracy values on standard datasets. Indeed, they are able to show that by adding a new modality (e.g. adding gyroscope data to accelerometer data) to a network, new features can be extracted without any need for preprocessing. Many of the features considered in previous work are extracted from a long timespan. As we are considering actions instead of activities which span a much shorter time it has still to be shown if the same methods can be applied. Therefore, we evaluate the suitability of these and similar approaches for our grabbing scenario. Since deep learning needs a lot of labeled data for proper learning, it is not applicable in our scenario.

Analyzing only inertial data for activity recognition covers half of our analysis. We also want to consider the video sensor for our classification experiments. Combining different kind of sensors to create a multimodal dataset has been the focus of various previous studies[12,13,14]. Indeed, Torre et al.[12] published a dataset containing multiple recordings of participants cooking different recipes while recording inertial data and video data along with audio and motion capturing. On top of this work, researchers applied multimodal activity recognition experiments. Spriggs et al.[15] use both image and inertial features to recognize activities in the cooking domain (stirring, pouring etc.). By down sampling the inertial data to fit the frame rate of the video they classified frames with aligned inertial data as single entries. Therefore, this approach cannot make use of inertial features that are extracted from a window of inertial data. Recently, Song et al.[14] published an egocentric multimodal dataset recorded with data glasses which contains egocentric video and inertial sensor data. In their work, they also presented an approach for recognizing life-logging activities. By utilizing Fisher Kernels they combine video and sensor features and reach high accuracy values. In the context of our action recognition this approach may not suffice as it does not capture arm movement outside of the camera's frame.

3. Dataset

In this paper, we create a dataset by simulating order picking in a warehouse setting. In our previous work[16], we analyzed the impact of inertial data from a wrist worn sensor on action detection. As this dataset puts less focus on the egocentric video we create a new dataset that improves on that aspect. Observing a real-world picking process, lets us derive the following actions a picking process consists of, where we focus on the actions ”navigation” and ”item picking” and do not consider the preparational work, e.g., positioning of order boxes: Looking at the shelf number, then walking up to the shelf, finding the correct box, picking an item from the box, looking at the item to simulate scanning it, and finally dropping it off at the start. In a real world setting these actions may vary slightly, depending what type of picking technology is used. We record picking actions from four (three male and one female) participants each performing 20 picking actions in two different settings. The following four cases are performed and recorded:

Picking, with the arm activity fully in focus: In this scenario, the participants are focusing their view on the shelf while grabbing from a set of boxes. Half of the orders are from a shelf with boxes, the other half from an open shelf.

Picking, without arm activity in frame: Here the participants are asked to specifically not focus on the shelf and instead look at something else. Participants look at the smartphone they are provided to emulate reading from an order list. Such scenarios are also likely to occur in a real warehouse environment as experienced pickers often only glimpse at the shelf when working.

No activity, with the participants looking at the shelf and boxes: Participants are asked to walk to the shelf with the intent of picking an item but without actually performing the grabbing action. We add this scenario to include negative examples in our experiments.

No activity, with the participants looking at the shelf and moving their arm: This scenario serves a similar purpose as the previous one. But it adds arm movement (in the form of tacking out the smartphone from the pocket) as an additional action.

We record first person view and inertial data with data glasses and inertial data from a smartphone and a smartwatch. Additionally all scenarios are filmed from a third person perspective for improved labeling and easier validation of the actions. Figure 1 shows one participant with the devices and their on-body position. The tablet is used to record depth data which will be used in future work. All inertial data is recorded using a mobile application from previous work[17]. Each inertial sensor is recorded at a sampling rate of 50Hz. First person video is collected at a resolution of 1920x1080 pixel with 24 frames per second. The smartwatch is worn on the right wrist, while the connected smartphone is kept in the pocket of the participants. Our test environment consists of one shelf with multiple compartments. Each box or, in the case of open shelf picking, compartment has an unique QR code identifying the items.

For our experiments, we use a subset of the recorded data. Namely, we pick the acceleration data from the smartwatch and the egocentric video of the data glasses. This gives us better insights about the impact of each sensor towards the results. As the data is recorded with two devices we first have to synchronize it. For this purpose, we introduce an alignment motion at the beginning of each recording. This motion produces a distinctive curve in the plot of the

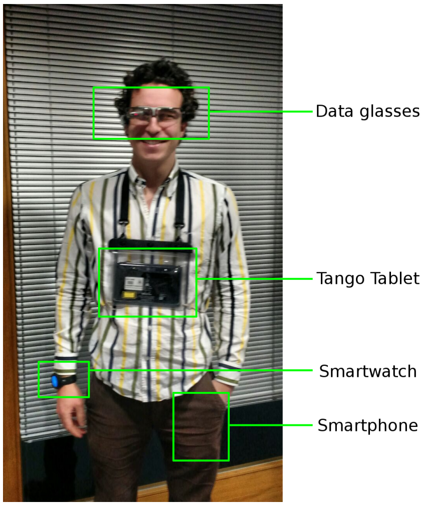

Fig. 1. Participant wearing all devices for data gathering.

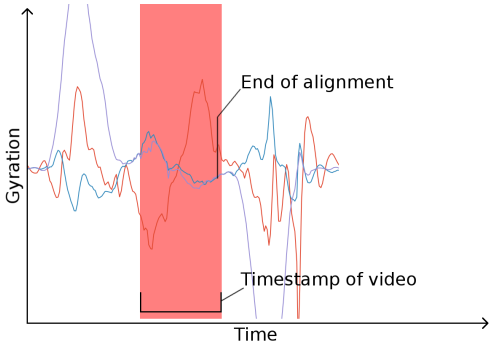

Fig. 2. Plot of the alignment motion of the smartwatch with an overlay of the adjusted timestamp of the egocentric video.

Table 1. Features extracted from different modalities. Inertial features are calculated on windows, image features on a per frame basis.

Inertial Features		**Image Features**	
Time	**Frequency**	**Color**	**Texture**
Mean, Variance, Correlation coefficient (Pearson), Gravity (pitch, roll), Standard Deviation, Median, Mean absolute deviation, Entropy (Shannon), Kurtosis, Interquartile Range (type R-5)	Energy (Fourier, Parseval), Entropy (Fourier), DC Mean	HSV-Histogram, Mean of each channel, Standard Deviation of each channel	Histogram of oriented Gradients

gyroscope data which we then use to calculate the time difference for each recording. We validate the difference by plotting inertial data of the watch and checking if the video timestamp is overlapping correctly (cf. Figure 2).
After recording, the data is annotated two-folds: the first person video and the third person video are both labeled with the BORIS software[18]. First person video annotation includes the exact end of the alignment action, the time span in which the hand is in frame while grabbing, and the timespan while an item is scanned. In the third person video we also label the end of the alignment action and the whole grabbing process if present in the scenario. We plan to publish the data*.

4. Methodology

Our essential idea for learning grabbing actions is to leverage the combination of extracted features from inertial and video data. We consider features in the frequency and the time domain for inertial data and color and image descriptor features for the video data. Figure 3 shows the process of feature extraction and merging. For the frames we extract histograms of the HSV color channels and histograms of oriented gradients (*HoG*[19]) (cf. Figure 3, Step 1.1, 1.2, and 1.3). The histograms of the HSV channel are extracted without binning, enabling us to bin the data later. We also add the mean and standard deviation of each channel. The HoG feature is extracted with 25 patches per frame as a trade of between amount of detail captured and feature size. All image features are extracted on a scaled down version of the original frame. In total this results in $(256 + 2) \cdot 3 + 25 \cdot 9 = 999$ features per frame.
Inertial features are extracted using a *sliding window* approach. This means, we consider a fixed timespan and calculate features on acceleration data within that span. Afterwards, the window is moved to the next point in time, in the end resulting in a set of windows (cf. Figure 3, Step 2.1, 2.2, and 2.3). Our features are calculated for a window size of 1000 milliseconds. This is a trade-off between too coarse window sizes for actions and windows without enough

[*] http://sensor.informatik.uni-mannheim.de

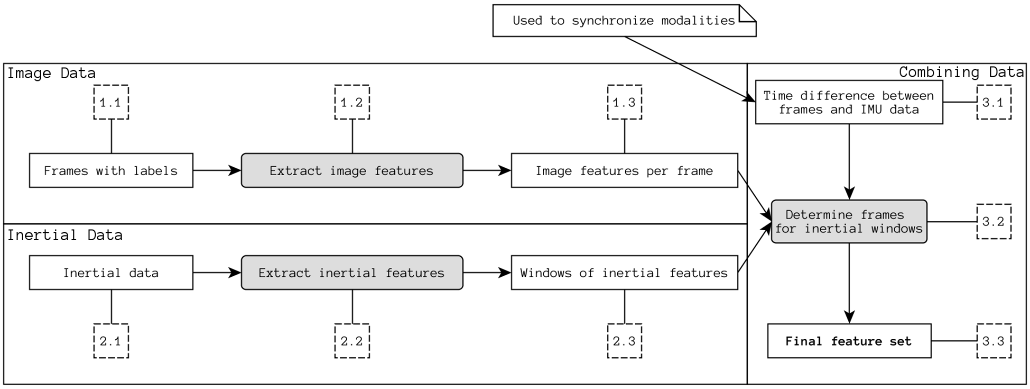
Used to synchronize modalities
Image Data
1.1
1.2
1.3
Frames with labels
Extract image features
Image features per frame
Inertial Data
Inertial data
Extract inertial features
Windows of inertial features
2.1
2.2
2.3
Combining Data
Time difference between frames and IMU data
3.1
Determine frames for inertial windows
3.2
Final feature set
3.3

Fig. 3. Process of feature extraction and combination.

information in them. Consecutive windows overlap, allowing us to determine the start of a grab more precisely. We choose an overlap of 70% resulting in 300 milliseconds between windows. Table 1 shows the features, we calculate from the acceleration data of the smartwatch. These can broadly be grouped into time based and frequency based features. Additionally, features can be grouped according to their properties, e.g., distribution, shape, and average. These feature groups are studied separately in our feature selection study of RQ2 (cf. Section 5). All inertial features are calculated on each of the three-way axis of the acceleration data yielding 42 features(14 different features * 3 axes) per window.

Since image features are calculated on a per frame basis and inertial features on windows, we have to combine them (Figure 3, Step 3.2). First, we have to align both feature sets with the alignment information we determine beforehand (Figure 3, Step 3.1). To merge the inertial and image features, we have to adapt the features extracted from the frames to fit the windows we calculated before. After we determined which windows a frame belongs to, we calculate the mean of each feature of all frames in every window, creating an *average frame*. As we store the labels of our dataset with the frames, we have to add that information to the windows. A window is thus labeled with the grabbing class if it contains at least one frame that also has this class. The combined windows are then stored per participant and scenario to enable different scenario combinations in our experiments (Figure 3, Step 3.2 and 3.3). In the following, we are going to use machine learning algorithms on the combined dataset.

5. Experiments

In the following, we present our experiments and their results in line with the research questions. First, we describe our experimental setup and subsequently conduct our experiments grouped by the research question.

5.1. Experimental setup

All experiments we conduct, are tested with three classification algorithms: Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Networks (ANN). These algorithms have shown to perform well in related problem domains[17,20,21]. Precision, Recall and F_1-Measure of the classifications are shown for each class separately with the measures for classifying the **grabbing** action being the focus in this work. Our dataset has a total of 8585 windows for non-grabbing actions and 1396 windows for grabbing actions. The classifiers use the following settings: A RF with a maximum of 100 trees and a depth of 10, a SVM-C with a polynomial kernel function, and a Multi-layer Perceptron with a maximal number of 500 iterations.

Table 2. **RQ1**: All features were used with a 5-fold cross validation with 100 runs on all data.

	SVM			RF			ANN		
Class	**Precision**	**Recall**	**F_1 ± SD**	**Precision**	**Recall**	**F_1 ± SD**	**Precision**	**Recall**	**F_1 ± SD**
None	0.977	0.974	0.976 ± 0.003	0.956	0.995	0.975 ± 0.002	0.962	0.956	0.958 ± 0.015
Grabbing	**0.845**	**0.862**	**0.853 ± 0.017**	**0.956**	**0.720**	**0.821 ± 0.019**	**0.775**	**0.761**	**0.751 ± 0.054**
Average	0.959	0.958	0.959 ± 0.005	0.956	0.956	0.953 ± 0.005	0.936	0.929	0.929 ± 0.019

Table 3. **RQ1**: Accuracy of all grabbing actions per participant (**P**) in the first 100%, 75%, 50%, 25% and 12.5% percent of each set of grabbing windows.

	SVM					RF					ANN				
P	**100%**	**75%**	**50%**	**25%**	**12.5%**	**100%**	**75%**	**50%**	**25%**	**12.5%**	**100%**	**75%**	**50%**	**25%**	**12.5%**
P1	0.851	0.883	0.845	0.744	0.574	0.640	0.643	0.589	0.473	0.314	0.681	0.706	0.662	0.564	0.405
P2	0.858	0.887	0.857	0.776	0.625	0.593	0.696	0.647	0.551	0.378	0.761	0.803	0.759	0.687	0.548
P3	0.875	0.900	0.869	0.792	0.607	0.797	0.880	0.892	0.793	0.586	0.803	0.839	0.798	0.658	0.450
P4	0.852	0.874	0.864	0.820	0.627	0.695	0.713	0.681	0.551	0.357	0.753	0.773	0.725	0.632	0.533

5.2. Experiments

To answer **RQ1**, we first apply the algorithms on the whole dataset with all features kept in place. We use 5-fold cross validation with stratified sampling for the evaluation. Each algorithm is run 100 times with different folds to check if the results are stable. The results are shown in Table 2. It can be seen that the RF yields a high precision at the cost of recall while the SVM balances these values out. The ANN yields slightly worse results than the other two algorithms and could be improved by increasing the number of max iterations. This trend continues in subsequent experiments throughout this work. It can be seen that the combination of both modalities is very promising for recognizing the grabbing action.

Still, we need to analyze how the classifiers perform within the timespan of a picking action. Our goal is to recognize a grabbing motion as early as possible, therefore we analyze how well the start of an action is recognized. For this purpose, we look at the accuracy of the prediction in the first 100%, 75%, 50%, 25% and 12.5% of all the windows of grabbing actions. Table 3 shows the result for our four participants. It can be seen that the results vary among the classifiers and participants. This is due to the fact that all the participants were grabbing at different speeds and also looked at the shelf at different angles. We can also see that the low Recall of the RF (c.f Table 2) is reflected in the accuracy of the grabbing windows. Generally, we have the highest accuracy in the first 75% of the grabbing windows. This can be attributed to the participants looking downwards at the end of a motion, not focusing on the shelf. Thus, relevant objects that are involved in the grabbing motion are not captured by the current camera frame(s) which makes it unfeasible to extract meaningful visual descriptors. Accuracy in the first 12.5% of the relevant windows drops to the lowest value. Since grabbing motions start when the arm moves towards the shelf, and participants are likely to not focus on the shelf yet, determining the correct start is hard. Therefore, we focus our next experiments on sub-featuresets to explore their influence on classification results.

Table 4. **RQ2:** Inertial features of all participants were used with a 5-fold cross validation and 100 runs.

		SVM			RF			ANN		
Features	**Class**	**Precision**	**Recall**	**F_1 ± SD**	**Precision**	**Recall**	**F_1 ± SD**	**Precision**	**Recall**	**F_1 ± SD**
Inertial	None	0.902	0.983	0.941 ± 0.002	0.923	0.978	0.950 ± 0.003	0.913	0.935	0.923 ± 0.010
	Grabbing	**0.765**	**0.342**	**0.472 ± 0.023**	**0.785**	**0.501**	**0.611 ± 0.025**	**0.549**	**0.448**	**0.478 ± 0.055**
	Average	0.883	0.893	0.875 ± 0.005	0.904	0.911	0.902 ± 0.006	0.862	0.867	0.861 ± 0.009
Image	None	0.949	0.994	0.971 ± 0.002	0.943	0.992	0.967 ± 0.002	0.957	0.959	0.957 ± 0.016
	Grabbing	**0.947**	**0.673**	**0.787 ± 0.018**	**0.992**	**0.629**	**0.750 ± 0.019**	**0.779**	**0.732**	**0.737 ± 0.061**
	Average	0.949	0.949	0.945 ± 0.004	0.941	0.942	0.937 ± 0.004	0.932	0.927	0.926 ± 0.020

To answer **RQ2**, we analyze the influence of different features on the recognition rate. First, we split up the image and inertial features and evaluate them separately (c.f. Table 4). For the inertial data it can be seen that among all algorithms precision and recall are dropping significantly. Our experiments indicate that image features have compa-

Table 5. **RQ2**: Different subsets analysis each with a 5-fold cross validation and 100 runs (only for the *Grabbing* class).

	SVM			RF			ANN		
Features	**Precision**	**Recall**	**F_1 ± SD**	**Precision**	**Recall**	**F_1 ± SD**	**Precision**	**Recall**	**F_1 ± SD**
Mean, SD, Var	0.697	0.099	0.173 ± 0.022	0.594	0.258	0.359 ± 0.022	0.640	0.222	0.328 ± 0.037
Gravity	0.625	0.264	0.369 ± 0.033	0.652	0.474	0.548 ± 0.021	0.639	0.344	0.445 ± 0.038
Time	0.739	0.302	0.429 ± 0.024	0.765	0.444	0.562 ± 0.022	0.701	0.530	0.599 ± 0.037
Frequency	0.647	0.077	0.134 ± 0.021	0.607	0.251	0.354 ± 0.025	0.476	0.291	0.338 ± 0.078
MAD, IQR, SD, Var	0.626	0.029	0.054 ± 0.014	0.506	0.134	0.211 ± 0.024	0.586	0.076	0.132 ± 0.041

rable results to experiments across all feature types. However, as recall drops in the experiments with SVM and RF, a more detailed study about the significance of visual features is required. We further analyze feature subgroups from the inertial data to find out if there are subsets of features that give us similar results to all inertial features. For this purpose, we create five feature subsets which can be seen in Table 5. Groups are created based on their domain, what they are representing, and on preliminary experiments. Table 5 shows the results of our feature subgroup analysis. We see that gravity by itself yields very good results. This is due to the fact that gravity consists of pitch and roll thus it contains the relative position of the smartwatch. With participants grabbing from the same shelves, the position of the smartwatch can be used to register the arms movement towards shelf height. Since shelves in warehouses are rarely located on different heights (to minimize unergonomic movement), gravity can be a good indicator for a grabbing action. Drawbacks in this approach are varying heights of people, and arm movements that are similar to a grabbing motion. While height variation can be compensated with a bigger dataset, similar arm movement has to be recognized by other features. All the features calculated from the time domain are also performing well. As gravity is part of the time domain features, the good performance may be attributed to it. Still, precision of all classification results improves when the whole domain is considered. The rest of our features perform worse, especially regarding the recall. It can therefore be seen that features from the time domain yield the best results for the task of grabbing recognition. This is due to the fact that our window size is smaller than the usual window size used for activity recognition. Since each participant performs the grabbing at different speeds and with different movements the acceleration data by itself may not be sufficient for recognizing the action. Adding gyroscope and magnetic field information may improve the results. With magnetic field data overfitting may be a problem as a classifier may learn a model based on the layout of a specific warehouse.

In addition, we also analyze the image features (cf. Table 4). Image features yield results close to using the combination of all features. Therefore, we analyze how the classifiers behaved in non-grabbing scenarios. We evaluate how often the algorithms classified non-grabbing windows as grabbing windows in the negative scenarios. We found out that on average **2.1%** of the windows in non-grabbing scenarios are labeled as grabbing actions.

Table 6. Accuracy of all grabbing actions per participant (**P**) in the first 100%, 75%, 50%, 25% and 12.5% percent of each set of grabbing windows for **inertial features**.

	SVM					RF					ANN				
P	**100%**	**75%**	**50%**	**25%**	**12.5%**	**100%**	**75%**	**50%**	**25%**	**12.5%**	**100%**	**75%**	**50%**	**25%**	**12.5%**
P1	0.383	0.433	0.323	0.170	0.214	0.488	0.510	0.423	0.236	0.198	0.507	0.511	0.486	0.423	0.380
P2	0.281	0.335	0.380	0.323	0.290	0.378	0.434	0.463	0.452	0.437	0.433	0.479	0.487	0.529	0.516
P3	0.473	0.513	0.522	0.530	0.440	0.620	0.672	0.696	0.655	0.581	0.572	0.597	0.591	0.597	0.613
P4	0.237	0.246	0.174	0.087	0.088	0.280	0.247	0.149	0.094	0.070	0.334	0.307	0.264	0.211	0.255

After the feature subgroup analysis we further evaluate the performance of the classifiers for the start of the action. For this purpose we again evaluate the accuracy of the algorithms for the first 100%, 75%, %25, and 12.5% of windows of all grabbing windows. Table 6 shows the results of this experiment. While the overall performance is in line with the feature experiments in Table 4, the performance for the different percentages differs greatly. It can bee seen that the accuracy varies stronger for the different participants when compared to the results in Table 3. This fact can be explained with arm movements having greater variation compared to the frames of the participants.

6. Conclusion

For **RQ1**, we are able to show that by merging features from image and inertial data grabbing actions can be recognized with an F-Measure of 85.3%. By combining the sensors, we are able to balance out the drawbacks of each. Image features register a grabbing action too late while inertial features are not reliable enough to distinguish arm movements. Finding the correct start of an action is still a task that needs further focus, as currently only 61% of the first 12.5% of grabbing windows are recognized. Improvements could be done by weighting the start of an action greater than the rest and therefore creating classifier focused on finding action starts. The feature analysis in **RQ2** shows that image features outperform inertial features. It also can be seen that for short actions inertial features from the time domain work better than features from the frequency domain. Future work will focus on two main topics: First, we want to explore the usage all the collected inertial data. Currently, only the inertial data from the smartwatch is analyzed in our approach. As the smartphone, connected to the watch, as well as the data glasses were recording inertial data, we could explore adding these to our current classification pipeline. Additionally, we could add gyroscope and magnetic field data to our features. The second topic we want to explore is a better merging of inertial and video data. Instead of calculating an *average frame* for each window we could find more elaborate methods to represent the image data within a window.

References

1. R. De Koster, T. Le-Duc, K. J. Roodbergen, Design and control of warehouse order picking: A literature review, European Journal of Operational Research 182 (2) (2007) 481–501.
2. L.-f. Hsieh, L. Tsai, The optimum design of a warehouse system on order picking efficiency, The International Journal of Advanced Manufacturing Technology 28 (5-6) (2006) 626–637.
3. T. Vaughan, The effect of warehouse cross aisles on order picking efficiency, International Journal of Production Research 37 (4) (1999) 881–897.
4. A. Miller, Order picking for the 21st century, Manufacturing & Logistics IT.
5. M. Wölfle, W. A. Günthner, Wearable RFID in order picking systems, in: Smart Objects: Systems, Technologies and Applications, Proceedings of RFID SysTech 2011 7th European Workshop on, VDE, 2011, pp. 1–6.
6. M. Funk, A. S. Shirazi, S. Mayer, L. Lischke, A. Schmidt, Pick from here!: an interactive mobile cart using in-situ projection for order picking, in: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2015, pp. 601–609.
7. X. Li, I. Y.-H. Chen, S. Thomas, B. A. MacDonald, Using kinect for monitoring warehouse order picking operations, in: Proceedings of Australasian Conference on Robotics and Automation, Vol. 15, 2012.
8. J. R. Kwapisz, G. M. Weiss, S. A. Moore, Activity recognition using cell phone accelerometers, ACM SigKDD Explorations Newsletter 12 (2) (2011) 74–82.
9. S. J. Preece, J. Y. Goulermas, L. P. Kenney, D. Howard, A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data, IEEE Transactions on Biomedical Engineering 56 (3) (2009) 871–879.
10. R. San-Segundo, J. M. Montero, R. Barra-Chicote, F. Fernández, J. M. Pardo, Feature extraction from smartphone inertial signals for human activity segmentation, Signal Processing 120 (2016) 359–372.
11. F. J. Ordóñez, D. Roggen, Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition, Sensors 16 (1) (2016) 115.
12. F. De la Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, P. Beltran, Guide to the carnegie mellon university multimodal activity (cmu-mmac) database, Robotics Institute (2008) 135.
13. C. Chen, R. Jafari, N. Kehtarnavaz, Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor, in: Image Processing (ICIP), 2015 IEEE International Conference on, IEEE, 2015, pp. 168–172.
14. S. Song, N.-M. Cheung, V. Chandrasekhar, B. Mandal, J. Liri, Egocentric activity recognition with multimodal fisher vector, in: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, IEEE, 2016, pp. 2717–2721.
15. E. H. Spriggs, F. De La Torre, M. Hebert, Temporal segmentation and activity classification from first-person sensing, in: Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Soc., IEEE, 2009, pp. 17–24.
16. A. Diete, T. Sztyler, H. Stuckenschmidt, A smart data annotation tool for multi-sensor activity recognition, in: 2017 IEEE International Conference on Pervasive Computing and Communications Workshops, IEEE Computer Soc., Piscataway, NJ, 2017, pp. 111–116.
17. T. Sztyler, H. Stuckenschmidt, On-body localization of wearable devices: an investigation of position-aware activity recognition, in: Pervasive Computing and Communications (PerCom), 2016 IEEE International Conference on, IEEE, 2016, pp. 1–9.
18. O. Friard, M. Gamba, Boris: a free, versatile open-source event-logging software for video/audio coding and live observations, Methods in Ecology and Evolution 7 (11) (2016) 1325–1330.
19. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 1, IEEE, 2005, pp. 886–893.
20. J. Yang, Toward physical activity diary: motion recognition using simple acceleration features with mobile phones, in: Proceedings of the 1st international workshop on Interactive multimedia for consumer electronics, ACM, 2009, pp. 1–10.
21. A. M. Khan, Y.-K. Lee, S. Lee, T.-S. Kim, Human activity recognition via an accelerometer-enabled-smartphone using kernel discriminant analysis, in: Future Information Technology (FutureTech), 2010 5th International Conference on, IEEE, 2010, pp. 1–6.