Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 110 (2017) 16–23

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.

Peer-review under responsibility of the Conference Program Chairs.

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

1877-0509
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.06.071

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