Exploring a Multi-Sensor Picking Process in the Future Warehouse

Alexander Diete
University of Mannheim
Mannheim, Germany
alex@informatik.uni-mannheim.de

Timo Sztyler
University of Mannheim
Mannheim, Germany
timo@informatik.uni-mannheim.de

Lydia Weiland
University of Mannheim
Mannheim, Germany
lydia@informatik.uni-mannheim.de

Heiner Stuckenschmidt
University of Mannheim
Mannheim, Germany
heiner@informatik.uni-mannheim.de

UbiComp/ISWC '16 Adjunct, September 12-16, 2016, Heidelberg, Germany
ACM 978-1-4503-4462-3/16/09.
http://dx.doi.org/10.1145/2968219.2968270

Abstract
Recognizing, validating, and optimizing activities of workers in logistics is increasingly aided by smart devices like glasses, gloves, and sensor enhanced wristbands. We present a system that augments picking processes with smart glasses and wristband that incorporates different types of sensors including ultrasonic, pressure, and inertial. We focus on low barriers for the adoption as well as the combination of video and inertial sensors. For that purpose, we create a new semi-supervised dataset to evaluate the feasibility of our approach. The system recognizes and monitors activities like grabbing and releasing of objects that are essential for order picking tasks.

Author Keywords
Augmented reality; pick-by-vision; sensor mining; depth video; activity recognition

ACM Classification Keywords
I.2.10 [Artificial Intelligence]: Vision and Scene Understanding; I.2.6 [Artificial Intelligence]: Learning; I.5 [Pattern Recognition]

Introduction
The development of wearable devices such as smart glasses features a variety of sensors and provides new opportunities for continuous monitoring and supporting of logistic pro-

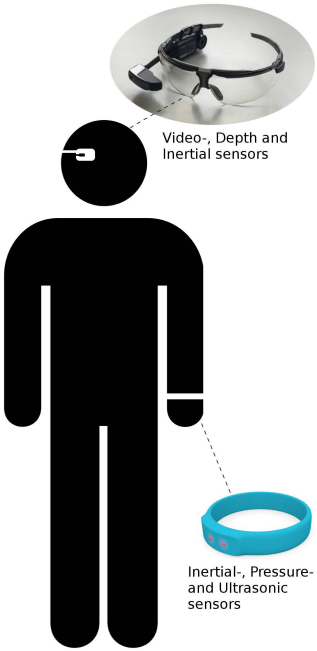

Figure 1: System setup with sensor positions

cesses. A common process is order picking where specific items must be picked that are typically part of a shipment or order. This includes navigating to the correct shelves, identifying the box that contains the required item, and scanning the corresponding barcode. The time exposure for most of these tasks can be reduced by relying on data that is already collected and stored in a data warehouse system. However, workers still have to log their activities, e.g., by using a handheld terminal that reduces their flexibility concerning movements as well as it requires additional time. Multiple solutions for this kind of problem have been proposed including pick-by-vision or pick-by-voice. In this paper, we focus on pick-by-vision by using smart glasses to analyze the field of vision of a worker. We aim to automatically recognize and scan the item of interest but also provide visual navigational information.

Compared to existing works, we also consider inertial sensors to recognize the performed activity (see Figure 1). Hence, on the one hand, the video sensor provides detailed information of the item of interest whereas the inertial sensors provide additional context information. Specifically the movement of a workers arm, its position, and its action is useful information that allows an aided picking system to better monitor and predict processes. Hence, the different kind of sensors and positions enable a precise monitoring of the different body regions. However, determining which sensor information has to be combined and interpreted depends on the performed activity. If a worker just stands idle in front of a shelf then acceleration sensor data probably do not cover useful information. In contrast, depth information may help to identify the box of interest in an image if several are visible. Therefore, we aim to investigate which sensor data is reliable to detect certain patterns during a picking process. We believe that this approach should be more robust than existing works as we compensate the weaknesses of the individual sensors by a multi-sensor setup.

Related work

In this paper, we investigate the detection of specific actions by relying on video and inertial sensors to support picking processes. Schwerdtfeger et al. conducted an extensive evaluation of their HMD-based industrial augmented reality system for picking processes [6], where augmented reality information is used for navigational support and user assistance. As the picking process itself is complex and sensor weaknesses need to be eliminated by complementary information, further methods from the field of image processing, such as hand/grabbing detection [2] and analyzing depth information for navigational purposes [1], are required. This also includes common methods of the activity recognition domain. In this context, recently, researchers investigated the recognition of common postures such as standing and walking in a real-world scenario relying on an acceleration sensor and considering several on-body position of the wearable device [8]. They stated that the distinction between dynamic (e.g., walking) and static (e.g., standing, sitting) activities is reliable. Considering our scenario, this enables to detect the point in time where other sensors could provide detailed information of the current activity, e.g., if the worker is navigating to or standing in front of a shelf. However, the possibility to distinct between similar activities depends on a sensors position, e.g. standing still versus standing while lifting the arm. Therefore detecting the movement of that body-part which is responsible for the difference between these actions is required. Consequently, a glove that is enriched with sensors would allow to derive more precise hand and hand gesture information [3, 4, 5]. However, the drawback of such an approach would be that gloves commonly reduce delicacy of feeling. Sikdar et al. addressed this problem by focusing on muscle activities [7]

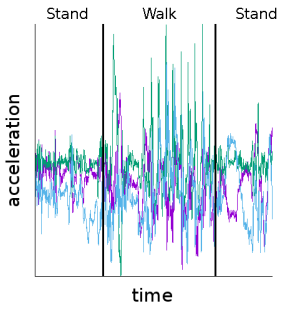

Figure 2: Inertial sensor data

Figure 3: Visual cues

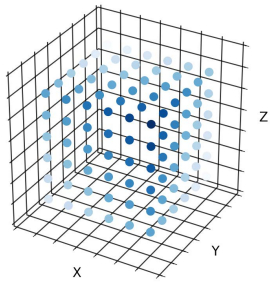

Figure 4: Depth information

with a wristband. Similar to Fraunhofer[1] we are using an RFID wristband, where we aim to register hand movement into and out of a box.

Methodology

We want to rely on common machine learning techniques in context of supervised learning to identify the performed action of a worker. Due to our scenario, we decided to create a new dataset which fulfills all of our requirements instead of using an existing one. For that purpose, we enhanced and used a mobile application which was initially developed for data collection for human activity recognition [8] where we added the capabilities to collect also video- and depth-information data. During the data collection, we recorded the sensor data of the wristband and smart glasses simultaneously. We introduced the subjects to stand idle for a minute so that we could easily synchronize the time depended data streams afterwards. To evaluate the feasibility of our approach, we decided to consider two different environments for data collection and evaluation. Hence, we build a testing environment locally for data collection and continuous evaluation but also use in addition the environment of a local logistics company.

We monitor and record a typical picking process as it occurs in logistic companies. This entails the navigation to the shelves, finding the correct shelf for the current item, picking up the item, and checking that the correct item was chosen. To label the recorded data afterwards we also use video cameras that record the whole process from the perspective of a third-person. Further, we attached a RFID scanner to the wristband to record which box with items was picked. In this process we only recognize the hand position inside or outside the box without information about grabbing motions. We consider the result of the scanner as a label for the picked box. Therefore, our dataset has to be considered as semi-supervised labeled.

Considering the devices, the smart glasses provide video and inertial sensors whereas the wristband also covers inertial but in addition pressure and ultrasonic sensors. The video stream enables to extract depth point clouds (see Figure 4) where the data from the inertial sensor (see Figure 2) can be used to recognize movement of the worker which in turn allows us to better understand the movements in the video recording. However, to decide which sensor data has to be interpreted or combined at which point in time is one of the problems which we are focusing. Further, by combining this sensor information we can eliminate the single disadvantages of each sensor on its own. Depth information is limited in distance and may not work for recognizing specific movements but can recognize if a certain box is in the process of being picked. Then, data from the wristband can be used to determine when a grabbing motion occurs. In case that these two events occur at the same time it is very likely that an item in a box was picked. Subsequently, the camera can be used to identify the specific product that was chosen. The depth sensor supports the interpretation of the video stream in case that several barcodes are visible. The specific methods for extracting information from video and depth data is out of scope for this paper.

Regarding the classification process, we consider machine learning techniques, i.e., neural network and random forest. The recorded sensor data is transformed into statistical features where commonly the time depended data streams are segmented into small windows. Video data will be evaluated separately by employing object recognition and tracking to find shelves and their labels. We want to rely on updatable models, i.e., online learning, to investigate the fea-

[1] http://www.iff.fraunhofer.de/en/business-units/material-handling-engineering/rfid-glove-object-identification.html

sibility regarding dynamic environments. In this context, we evaluate the influence of sensor combinations but also single sensor setups to determine their capability for aiding the process. Specifically, if the video camera and depth sensor are able to consistently identify shelves and boxes, if the wristband is sufficient to recognize grabbing motions and if the inertial sensors improve to overall accuracy of the process.

Conclusion & Future Work

We present an approach that supports workers in a warehouse, more precise, during a picking process. We create a labeled dataset to investigate the influence and advantages of the different kind of sensors in a multi-sensor setup where we aim to compensate the individual weaknesses. Further, we rely on common machine learning techniques to recognize the varying patterns that occur during the picking process. As a next step, we want to focus on the relation between sensor data and the activity patterns. Further, a productive system also requires a cross-device synchronization regarding time and data.

REFERENCES

1. J. Biswas and M.M. Veloso. 2012. Depth camera based indoor mobile robot localization and navigation. In *ICRA*. IEEE, 1697–1702. DOI: http://dx.doi.org/10.1109/ICRA.2012.6224766
2. A. Erol, G. Bebis, M. Nicolescu, R.D. Boyle, and X. Twombly. 2007. Vision-based Hand Pose Estimation: A Review. *Computer Vision and Image Understanding* 108, 1-2 (2007), 52–73. DOI: http://dx.doi.org/10.1016/j.cviu.2006.10.012
3. K.P. Fishkin, M. Philipose, and A. Rea. 2005. Hands-On RFID: Wireless Wearables for Detecting Use of Objects. *2012 16th International Symposium on Wearable Computers* 0 (2005), 38–43. DOI: http://dx.doi.org/10.1109/ISWC.2005.25
4. P. Hsiao, S. Yang, B. Lin, I. Lee, and W. Chou. 2015. Data glove embedded with 9-axis IMU and force sensing sensors for evaluation of hand function. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 4631–4634. DOI: http://dx.doi.org/10.1109/EMBC.2015.7319426
5. P. Plawiak, T. Sosnicki, M. Niedzwiecki, Z. Tabor, and K. Rzecki. 2016. Hand Body Language Gesture Recognition Based on Signals From Specialized Glove and Machine Learning Algorithms. *IEEE Trans. Industrial Informatics* 12, 3 (2016), 1104–1113. DOI: http://dx.doi.org/10.1109/TII.2016.2550528
6. B. Schwerdtfeger, R. Reif, W.A. Günthner, and G. Klinker. 2011. Pick-by-vision: there is something to pick at the end of the augmented tunnel. *Virtual Reality* 15, 2 (2011), 213–223. DOI: http://dx.doi.org/10.1007/s10055-011-0187-9
7. S. Sikdar, H. Rangwala, E.B. Eastlake, I.A. Hunt, A.J. Nelson, J. Devanathan, A. Shin, and J.J. Pancrazio. 2014. Novel method for predicting dexterous individual finger movements by imaging muscle activity using a wearable ultrasonic system. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on* 22, 1 (2014), 69–76. DOI: http://dx.doi.org/10.1109/TNSRE.2013.2274657
8. T. Sztyler and H. Stuckenschmidt. 2016. 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, 1–9. DOI: http://dx.doi.org/10.1109/PERCOM.2016.7456521