Development of a Holistic System for Activity Classification Based on Multimodal Sensor Data

Tony Rolletschke
Hochschule Mittweida
Technikumplatz 17
D-09648 Mittweida
rolletsc@hsmw.de

Rico Thomanek
Hochschule Mittweida
Technikumplatz 17
D-09648 Mittweida
rthomane@hsmw.de

Christian Roschke
Hochschule Mittweida
Technikumplatz 17
D-09648 Mittweida
roschke@hsmw.de

Marc Ritter
Hochschule Mittweida
Technikumplatz 17
D-09648 Mittweida
ritter@hsmw.de

As the world of portable computers has evolved, mobile activity and mobility monitoring has become one of the major trends of recent years using widely used technologies such as smartphone sensors and wearables. These technologies are the basis for a wide range of applications in the areas of health monitoring, fitness games and telematics systems in vehicles. In the resulting use cases, the focus is on recognizing, differentiating and qualitatively evaluating different types of movements. The key factor in this context is a high degree of recognition accuracy in almost real time. Due to the ongoing development of mobile devices and the associated increase in performance, it is now possible to use the interfaces provided in mobile operating systems for the use of deep learning technologies. Due to the high availability of the end devices, new context-sensitive applications can be created, which can adapt the program logic to the current environment of a user.

Wearable Sensors, Human Activity Recognition, Machine Learning, Ubiquitous Computing

1. INTRODUCTION

The paper includes the analysis, design and development of a system that allows real-time detection of activities using local and external mobile sensors and image capture systems. Due to the required computing power and reliable sensor technology, the required target platform is in the premium smartphone range. Using already existing, freely usable but also self-learned neural networks, the images captured by the camera are transferred in real time to relevant features, such as objects in an area. In this context, recently presented methods for the use of extended reality, recorded sensor data and real-time processing of camera images can be merged using machine learning techniques. The data extracted from this can be combined into a feature vector and enable the unique identification of individual activities. These feature vectors are then persistently stored in a database. Based on the detected metadata, the creation of a complex generic model is possible. This model enables generally valid context-related activity detection and therefore exceeds the value of traditional applications. This technology makes it possible to support people in complex work activities, avoiding errors and improving efficiency and quality during manufacturing (Hitachi, 2018).

The remaining part of the paper is organized as follows: Section 2 briefly describes the background conditions and previous achievements in this field of research. Section 3 explains the individual steps of the complete activity recognition procedure. Section 4 gives a summary of the project and an overview of the work planned in the future.

2. RELATED WORK

There are currently various scientific approaches to activity detection which, in addition to the usual sensor-based systems (Chen et al., 2017), also favour camera-based solutions (Jalal et al., 2017). Furthermore, Ordóñez et al. (2016) or Nunez et al. (2018) have shown that the use of CNNs and LSTMs achieves the highest accuracy in the detection of human activity and that the use of convolutional neural networks avoids the manual extraction of features.

The disadvantage of this is that the analysis of the recorded data usually does not take place directly on the end device, but on a server that detects the features and delivers the results back to the smartphone. Due to the outsourced feature recognition and the associated latency times for data transmission, real-time analysis cannot be

realized. Processing directly on the mobile device, on the other hand, allows the analysis of several images per second if the GPU is used for characteristic determination. Since mid-2017, Apple has created a powerful way for its mobile devices to use locally available neural networks within a feature detection app through the Core ML Framework. This eliminates the need to transfer data to a server and significantly improves processing performance (Apple ML, 2018).

The first own Core ML implementations have already delivered highly promising results in which all persons in the field of vision of the smartphone camera could be detected in real time. The newly introduced ARKit framework makes it possible to merge the data of the camera sensor with the data of the acceleration sensor using Visual Inertial Odometry (VIO) (Apple AR, 2018). This has already created a first possibility to detect the environment by means of detected feature points, which would make it possible to optimize the activity recognition in perspective (Eum et al., 2017). Since the framework is also supported by special processors (A9, A10 and A11), very fast processing times are possible.

3. SYSTEM ARCHITECTURE

Figure 2 shows the complete system, consisting of a prototypical iPhone app with smartwatch extension and an actioncam. The recorded sensor data and audiovisual data are then combined in a standardized exchange format. This allows the complete transfer and persistent storage of all data in a database. The efficient access mechanisms of the database enable a simple summary and an arbitrary combination of data to training and test data sets by means of search queries. A training data set generated in this way can be used to create a Core ML model. This Core ML model can then be integrated into an iOS app and enables real-time classification.

3.1 Mobile Application

To create a training data set, study participants will be equipped with 7-generation iPhones, Apple Watches Series 3 and GoPro Hero 5 cameras. A custom application for iOS and watchOS is installed on these devices, which was specially developed for recording motion data. This application records sensor data from iPhone and Apple Watch simultaneously in the background using the CoreMotion, CoreLocation and HealthKit frameworks. The sampling rate is 50 Hz. After the measurement has been completed, the data is transferred to the server as a JSON file and stored persistently in the database.

We record the raw data of accelerometer, gyroscope, magnetometer, altimeter, heart rate and location of the user. In addition, pre-processed data such as pedometers, device motion and simple motion events are also saved. External influences such as earth's gravitational force and, if necessary, noise are already partially eliminated by the interfaces contained in iOS. By this it is ensured that only the rotation rate and acceleration giving by the user is included in the further evaluation. The alignment of the device is recorded relative to a fixed reference frame. This makes it possible to track the rotation of the device around any axis in three-dimensional space. In addition, the occurring mechanical forces can be transformed from the device into a world coordinate system. This minimizes the dependency on the orientation of the device (Henpraserttae et al., 2011). For example, it does not matter whether the phone is placed with the connectors facing down or in the top of the pocket.

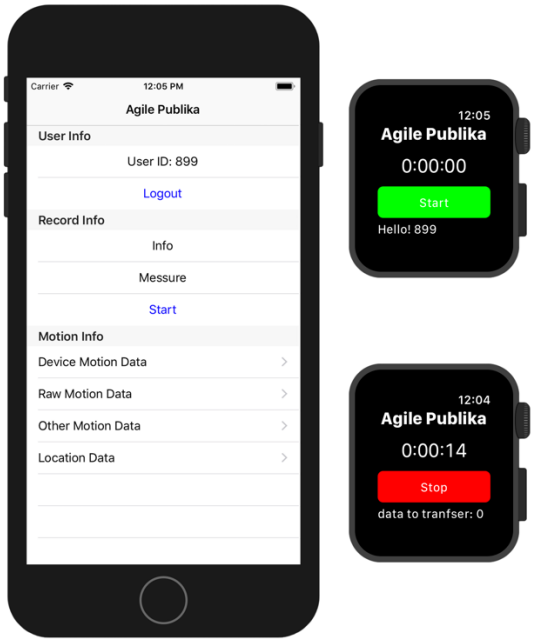

Figure 1 *– watchOS and iOS user interface*

However, the position of the telephone on the test persons bodies is problematic. In order to make the individual measurement series comparable, the possible positions of the devices are determined beforehand. Therefore, the phone should be carried in the front right trouser pocket and the watch on the dominant hand. This helps to record sensor data during activities where handedness (e.g. writing, opening doors) plays an overriding role.

The iOS App interface requires a user login. This will be entered once before the test person starts the measurement (see Figure 1). As a result, the measurement series are assigned to different persons and enable investigations regarding user-dependent and user-independent accuracies.

The watchOS application includes an interface to start and stop recording. In addition, it displays the previous measurement duration and further status information of the measurement (see Figure 1). This means that the user does not have to take the iPhone out of his or her pocket to control the recording and can focus entirely on performing the preset activities.

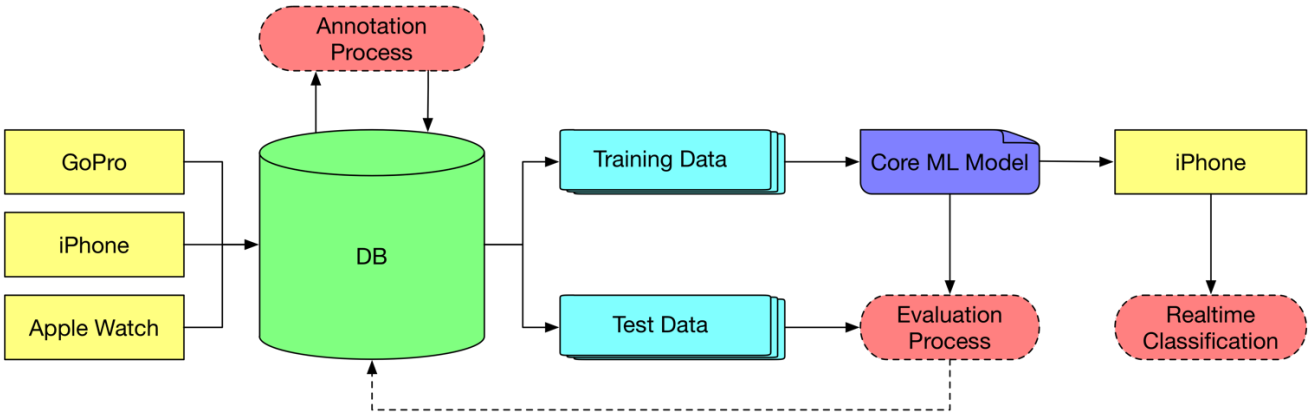

Figure 2: *schematic structure of the system*

3.2 Data recording

Machine learning achieves the best classification rates when the data basis is as heterogeneous and high-quality as possible. The higher the quality of a training data set, the better a machine learning process can recognize patterns in it.

In order to generate adequate sensor data of daily standard physical activities in appropriate quality, a study is carried out. This is used to collect data from several students. Each respondent performs a set of physical activities (walking upstairs, walking downstairs, sitting, standing, laying, stand to sit, sit to stand, sit to lie, lie to sit, stand to lie, lie to lie, lie to stand) for a fixed period of time. In addition, the participants collect further movement and video data from various public and private means of transport (walking, running, bicycle, car, bus, tram, train).

For each test, the test persons are provided with the necessary test equipment. In addition, help is given on the correct handling and operation of hardware and software.

3.3 Annotation

To be able to assign the collected sensor data to the correct classes, the video stream recorded simultaneously during sensor data acquisition is evaluated. A wide-angle camera is attached centrally to the test person's body with the aid of a chest strap. In this way, the environment of the test person is displayed over a wide area and a "first person" view is realized. This video material is then used for manual annotation of activities performed. For this purpose, a web-based software is being developed which offers a preview for recorded video and sensor data. Each activity class is assigned a key on the keyboard. Hotkeys for play, stop, rewind back and rewind forward are also set. This means that several hours of material can be annotated by hand with a minimum of effort.

In addition, further metadata such as prominent objects, location-dependent features and characteristic sound frequencies are extracted from the image and sound material with the aid of an automatic process. This supplemental meta information and the features of the sensor data are linked using time stamps. The idea behind this is that certain activities are related to certain objects and environments. For example, a meal is usually prepared in a kitchen using special cooking utensils. This methodology enables valid training data to be generated in a short period of time and used for supervised learning in frameworks such as Turi Create.

3.4 Classification

Turi Create is a framework for machine learning published by Apple in 2017. Turi Create supports the processing of texts, images, videos, audio content and sensor data and is intended for the creation of own machine learning models. (Turi, 2018)

The acquired sensor and camera data are merged using Turi Create and transferred to a model for later activity classification, which is able to recognize temporal characteristics in sensor data. This model can be exported to a suitable format for iOS and used in the context of real-time classification on an iPhone.

At first, primitive activities such as running, walking, etc. should be recognized. Based on this, patterns are to be recognized using clustering algorithms in order to be able to detect more complex activities in different contexts. This includes, for example, the means of transport used by a person or the type of event in which the person is currently attending. For this purpose, camera and audio data as well as sensor data are to be evaluated by the model. The aim is that the data fusion and the resulting model can be extended by any classes.

4. CONCLUSION

The developed system enables us to record sensor and video data and to store them on time as well as to combine them using time stamps and interpolate missing data. In current use, the new architecture proved to be extremely stable and high-performance.

At the moment, the first use of the newly developed technology is underway by students. Measured values of various activities both indoor and outdoor of different persons have already been recorded for several hours. The next step is end data collection, the annotation of the data, automatic extraction of additional meta information and finally the training of the classifier.

We plan to evaluate the system by means of scalability, usability, field and long-term tests. In the future, we will also use other test data sets for the created model in order to examine them in the context of accuracy. The resulting results are used to optimize the system. We will apply the evaluation procedures to both our own and freely available data sets in order to improve the methods for sensor acquisition and pre-processing, if necessary.

5. REFERENCES

Apple (2018) ARKit - Apple Developer. Retrieved May 03, 2018, from https://developer.apple.com/arkit/

Apple (2018) Machine Learning - Apple Developer. Retrieved May 01, 2018, from https://developer.apple.com/machine-learning/

Chen, Y., & Shen, C. (2017) Performance Analysis of Smartphone-Sensor Behavior for Human Activity Recognition. IEEE Access, 5, 3095–3110

Eum, S., Lee, H., Kwon, H., & Doermann, D. (2017) IOD-CNN: Integrating Object Detection Networks for Event Recognition. arXiv.org.

Henpraserttae, A., Thiemjarus, S., & Marukatat, S. (2011) Accurate Activity Recognition Using a Mobile Phone Regardless of Device Orientation and Location. Bsn, 41–46.

Hitachi (2017) DFKI and Hitachi jointly develop AI technology for human activity recognition of workers using wearable devices - News Releases. Retrieved June 4, 2018, from http://www.hitachi.com/New/cnews/month/2017/03/170308.html

Jalal, A., Kim, Y., Kim, Y.-J., Kamal, S., & Kim, D. (2017) Robust human activity recognition from depth video using spatiotemporal multi-fused features. Pattern Recognition, 61, 295–308

Núñez, J. C., Cabido, R., Pantrigo, J. J., Montemayor, A. S., & Vélez, J. F. (2018) Convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition. Pattern Recognition, 76, 80–94

Ordóñez, F., & Roggen, D. (2016) Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. Sensors, 16(1), 115

Turi (2018) turicreate - Turi Create simplifies the development of custom machine learning models. Retrieved May 06, 2018, from github.com/apple/turicreate