Master Thesis Proposal

Dumbbell movement recognition with a Deep Convolutional LSTM Neural Network and low frequency acceleration data

Stephan Lerner, Matr.-Nr. 6221045

November 17, 2016

1 Introduction

Human Activity Recognition (HAR) has emerged as a highly active field of research. Traditionally, computer vision was a dominant approach to monitor physical daily life activities of individuals. Efforts to unbound potential use cases from instrumented rooms shifted attention to wearable sensors, like accelerometers, gyroscopes and magnetometers [2]. These attempts were enabled by key technologies in three main building blocks of a wearable system:

- the sensing and data collection hardware to collect movement data
- the communication hardware and software to relay data to a remote center
- the data analysis techniques to extract relevant information

Related advances were the miniaturization of sensors and the tremendous progress in communication standards for low-power wireless communication. Another catalytic effect is based on the broad availability of smartphones as remote monitoring systems. Finally data analysis techniques have strongly developed in recent years [16].

The purpose of this research project is to evaluate the performance of deep neural networks within in a wearable system in recognizing different gestures with a dumbbell. Previous work of the author related to a wearable exercise tracking system are a survey with over 500 participants to investigate customer benefits and needs[1], a patent design to formulate potential claims[2] and a business concept to weigh influence factors on an economical success[3].

[1] see *Umfrageergebnisse.xlsx* in digital resources
[2] see *Patententwurf.pdf* in digital resources
[3] see *Seminararbeit.pdf* in digital resources

2 State of the Art

Recognition of sport-related activities as a subfield of HAR made extending progress in the last decade. Table 1 shows a collection of different activities to be recognized and associated methods. All shown application areas are related to sport, except for [15]. Gathered information regards recognition goals, experiment setup and Activity Recognition Chain (ARC) consisting of data acquisition, preprocessing, segmentation, feature extraction and classification [2]. The collection is merely intended as point of reference. A direct performance comparison is not possible for a variety of reasons. The classification problems reach from complex sporadic activities of daily living [15] to pure recognition of repetitions without typing an exercise [17]. The setup parameters include wide ranges regarding number of classes, size of dataset, sensor location, sensor characteristics, number of sensor channels and sample granularity. Even the evaluation metrics are different in respect of type and calculation basis[4].

Despite the variance, latest research approaches share common short falls, which could be clearly illustrated by an interesting fact: The majority of commercial products only report simple parameters, such as step counting, energy expenditure and sleep duration [22]. Even though the research results in recent years are impressive, provided solutions lack in terms of practicability and are not marketable. The often promised benefit for physical and psychological well-being (e.g. [3] or [20]) remain unfulfilled. Three subdimensions of practicability are defined: usability, efficiency and scalability. This division makes no claim to be exhaustive, but helps to describe researches' shortcomings.

One main potential of wearable systems in sport-related areas is usability improvement [9]. By automatically recognizing exercises unnecessary interaction with a device (e.g. a smartphone or notepad) can be reduced. For most of presented solutions in Table 1 this advantage is outweighed by sensors that users have to place on different body locations in addition to their smartphone. The systems provided in [12] and [17] are exceptions in this respect. In [12] acceleration data of a smartphone worn in an arm holster is used to recognize different types of exercises. Still training with dumbbells would require to place the arm holster alternately on both arms. [17] proposes to place the smartphone alternately on different body locations or the exercise machine to fully recognize exercises. The added value in terms of usability is therefore nullified.

[4] see *StateOfTheArt.xlsx* in digital resources

Table 1: Examples of fitness activity recognition in research (Sensor abbreviations: accelerometer, gyroscope, magnetometer: "aclm", "gyr", "mag"; Evaluation metrics abbreviations: F-measure, accuracy, precision, recall: "F", "acc", "prec", "rec"; Feature abbreviations: mean, standard deviation, variance, maximum, minimum, root mean square, duration, fluctuation, amplitude, kurtosis: "mn", "sd", "var", "max", "min", "rms", "dur", "fluc", "amp", "kur")

Reference	Activities	Classes [#]	Repetitions [#]	Sensor Placement	Sensor Types	Channels [#]	Sampling Rate [Hz]	Preprocessing	Segmentation	Feature Extraction	Classifier	Results
[3]	weight lifting exercises	9	4925	glove, waist	acc	6	80	Low-pass filter	sliding window	amp, cor, energy, velocity	Naive Bayes, Hidden Markov Model	85.0 acc
[12]	upper body weight lifting exercises	7	11000	upper arm	acc	3	100	Low-pass filter	sliding window	bandwith, cepstrum coefficients, fluc, frequency centroid, mn, sd, spectral fluc, spectral rolloff frequency	knn-Classifier, C4.5 decision tree, radial kernel Support Vector Machine	93.6 F
[20]	cardio and weight lifting exercises	16	-	glove, leg, chest	acc, heart rate sensor	10	100	-	sliding window	mn, var	Multivariate Gaussian Classifier	92.0 pre, 95.0 rec
[10]	weight lifting exercises	9	1610	glove	acc	3	100	Butterworth low-pass filter	peak analysis	-	Improved Dynamic Time Warping	98.4 acc
[13]	exercise at a leg press machine	3	-	leg press machine	Load cell, wirewound potentiometer	2	100	Low-pass filter, Butterworth low-pass filter	based on displacement values	acceleration, amp, decline, dur, fluct, incline, max, min, power, range, relations, velocity	Artificial Neural Network	-
[17]	weight lifting, resistance bands and body weight exercises	2	3598	ankle, wrist, stack of weights	acc	3	10	Linear Interpolation, Savitzky-Golay smoothing filter	peak analysis	Normalized Dynamic Time Warping distance, dur, max, min, mn, rms, sd	Logistic regression	99.3 F
[21]	unilateral dumbbell biceps curls	5	300	glove, armband, lumbar belt and dumbbell	acc, gyr, mag	36	45	-	sliding window	amp, kur, max, min, mn, sd, skewness, var	Ensemble of Random Forest classifiers	98.0 acc, 98.2 pre, 98.2 rec
[11]	weight lifting exercises	4	-	forearm	acc, gyr	6	50	Butterworth low-pass filter	sliding window	Auto-cor, integrated rms, interquartile range, kur, mn, power bands, rms, sd, variance	Support Vector Machine	99.4 F
[18]	upper body weight lifting exercises	6	2640	chest, left and right wrist, left and upper arm	acc	15	25	temporal alignment, uniform resampling	sliding window	amp, cor, min, max, mn, range, rms, sd	Radial kernel Support Vector Machine	86.1 acc
[15]	activities of daily living	17	1907	limbs, back, hip, feet	acc, gyr, mag, motion jacket	113	30	-	sliding window	-	Deep Convolutional and Long Short Term Memory Recurrent Neural Network	91.5 F

To avoid users bringing their own sensor devices or placing them alternately on different locations, sensors could be permanently placed on training devices (e.g. in a gym). However this requires an energetically efficient implementation, considering that limitations of currently available battery technology are still a challenge [16]. The high sample granularity and various sensor modalities of presented systems in Table 1 suggests that battery life time is on a scale of minutes and hours. One possible improvement is to store data temporarily on the sensor and send data with lowered frequency to the smartphone as shown in [20]. In this way the sensor could operate for 50 hours with a fully charged battery, what still could be considered as not sufficient for a typical gym[5] with around 100 training tools.

In a highly scalable system number of classes could be increased at runtime after the system is initially deployed. The system should also be incrementally expandable by environment-specific and user-dependent data. Table 1 shows a variety of engineered features for different approaches. Identifying relevant features is time consuming and makes a system hardly scalable. This dependency on "hand-crafted" features is dissolved in [15] by exploiting Convolutional Neural Networks for recognition of activities of daily living.

3 Motivation & Objectives

As mentioned initially, HAR with sensors was enabled by advances in related key technologies. This work is especially motivated by latest sensor releases based on Bluetooth Low Energy. Figure 1 shows an *Estimote Nearable* [4] compared to the size of a 50 cent coin. *Estimote* states that battery life time is up to one year. For maximum send frequency of 10 Hz battery life time should be rather in the order of months. For the proposed project only three-axis acceleration and identification data is needed. To address current researches' shortcomings send frequency should be artificially lowered after recording data with 10 Hz. In this way frequency impact on system performance will be investigated. The sensors will be placed on the dumbbell.

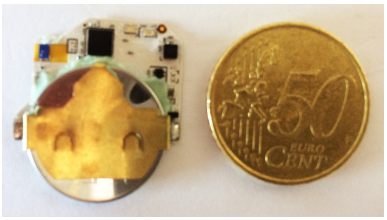


Figure 1: *Estimote Nearable* next to a 50 cent coin

This work is further encouraged by the novel use of Deep Convolutional and LSTM Recurrent Neural Network (DeepConvLSTM) in HAR [15]. The proposed framework outperformed all baselines in recognition of activities of daily living in the OPPORTUNITY challenge [14]. The authors attribute this to two main reasons. Convolutional layers are suitable to automatically

[5] see *DurchschnittlichesFitnessstudio.xlsx* in digital resources

extract features, that actual do relate - in contrast to statistical and frequency features - to semantically meaningful aspects of human motion. On the other hand the authors state that recurrent LSTM cells are fundamental to distinguish gestures of similar kind, which differ only by the ordering of sensor samples.

DeepConvLSTM were already successfully tested for activity recognition in video and achieved state of the art results on the Sports-1Mand UCF-101 dataset [24]. Other deep convolutional and recurrent approaches outperformed non-recurrent networks in video gesture recognition on Montalbano dataset [19], due to their capability to capture temporal information. These cases are closely related to the proposed project, since they are also addressing time series analysis.

Even if convolutional and recurrent approaches were already applied in video analysis and HAR, the introduced frameworks were not economical in terms of input data. This is proving not problematic for video analysis, since frames per second and resolution are usually sufficient for practical applications. On the contrary recording high-frequency data from several channels in HAR leads directly to impractical solutions as described in chapter 2. For this reason the proposed project mainly aims to develop a highly practicable wearable system in terms of usability, efficiency and scalability.

One common criticism of Neural Networks is their lack of transparency. There is still little insight how they achieve such good performance, what is deeply unsatisfactory from a scientific point of view. Visualizing and understanding the features of convolutional layers helped to improve performance on the ImageNet benchmark without being dependent on trial and error [25]. Also plotting activation values for the neurons in each convolutional layer in response to an image or video has proven to be informative and and lead to several surprising intuitions [23]. A similar approach might be applicable for the proposed project. Interpretability of the proposed system will also be enhanced by defining atomic gestures that cannot be further decomposed [1]. All gestures to be recognized are therefore atomic gestures, sequences or superimpositions of atomic gestures.

In specific the proposed work will deal with:

- The design of an infrastructure to easily collect labeled data
- The development of the Convolutional Long Short-term Memory architecture to recognize atomic gestures and thereof composed sequences or superimpositions
- The implementation and evaluation of experiments to measure performance, especially

for reduced sample frequency

- The visualization of results to derive insights and diagnose potential problems

4 Methodology

The proposed architecture is largely inspired by the framework presented in [15]. A sliding window as a *de facto* standard approach for realtime applications is used (see Table 1). Figure 2 outlines the proposed architecture for illustrative purposes. The actually used method may be different with respect to length and overlapping of sliding windows, length of kernels, convolutional layers, feature maps, recurrent layers, LSTM cells and output classes. Figure 2 shows nine channels and ten time steps within the sliding window. Two feature maps are extracted by temporal convolution of two different one-dimensional kernels (exemplarily shown in blue and green) with a length of four. The feature maps form the convolutional layer. The feature maps are then processed successively for each time step by an LSTM layer consisting of five LSTM cells, afterwards an output layer yields the classification outcome for three different classes (exemplarily shown in red). The network would be trained in a supervised manner by backpropagation.

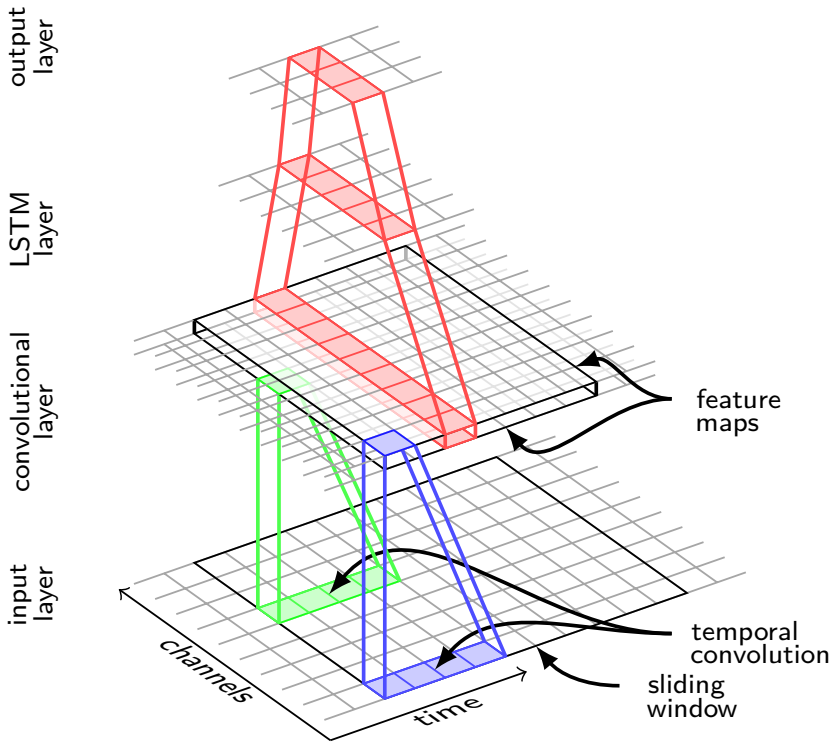


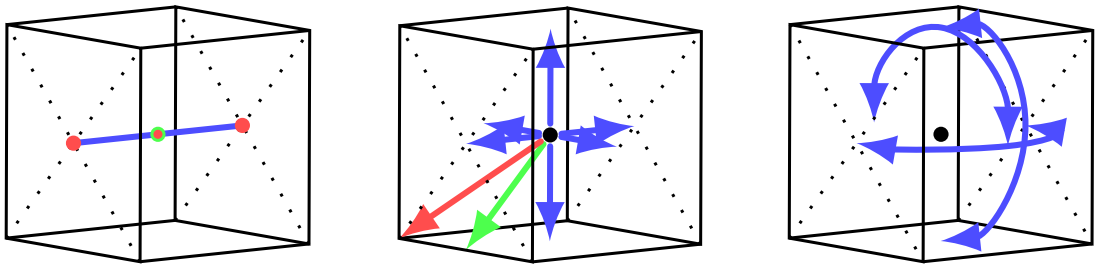
Figure 2: atomic gesture recognition architecture

A higher level of detail and parameterisation of the model is part of the proposed master thesis and should not be specified further at this point. Only several general remarks should be made:

- The number of channels describes the number of available one-dimensional raw acceleration data streams. No preprocessing will be applied.
- Since convolution is only computed, where the input and the kernel fully overlap, maximum possible number of convolutional layers increases with increasing length of sliding window and decreasing length of kernels. An extended sliding window on the other hand leads to delay in real time recognition, while a smaller kernel might not be suitable to capture temporal signal structure.
- Two-dimensional kernels were shown to be capable of capturing spatial dependencies of different sensors [5].
- A depth of at least two recurrent layers was shown to be beneficial [8].
- The number of output classes corresponds to the number of gestures to be recognized plus an extra NULL class that represents time spans without "interesting" activities
- The output layer provides a class probability distribution for every time step of a sequence within a sliding window. There needs to be further discussion, how the label of a sequence will be defined.

5 Experiments

The proposed system will be experimentally evaluated. Therefore a dataset will be collected. Figure 3a schematically shows a dumbbell (blue). Three *Estimote Nearables* (red) and one *Texas Instruments CC2650STK* [6] (green) are placed on it. The *Estimote Nearables* will send three-axis acceleration data with a frequency up to 10 Hz. These data channels will form the input for the proposed architecture. The *Texas Instruments CC2650STK* contains a 9-axis



(a) sensor placements (b) translational gestures (c) rotational gestures

Figure 3: experimental setup and atomic gestures with hereof composed superimpositions

inertial motion sensor and will send with a frequency of 10 Hz. The transmitted data will be used for analytical purposes, since the power consumption of the gyroscope and magnetometer [7] contradicts the idea of the proposed project.

Figure 3b and Figure 3c show atomic translational and rotational gestures which should be recognized (blue). Figure 3b also exemplarily shows superimpositions of two (green) respectively three (red) atomic translational gestures. For collecting labeled data an infrastructure will be designed, in which participants are visually and aurally guided to perform certain atomic gestures, sequences and superimpositions of atomic gestures with the prepared dumbbell. The gestures will be performed within typical time frames for dumbbell exercises. This process will be supervised to avoid incorrect data.

6 Novelty

As mentioned before DeepConvLSTM were already successfully implemented for recognition of activities of daily living with sensors. This project proposes a similar approach for movement recognition of a dumbbell. In contrast to [15] the proposed project aims for maximum practicability. To ensure usability sensors are placed on the dumbbell. To increase energy efficiency sensor modalities are limited to accelerometers, due to lower power consumption. Furthermore data is recorded and send to smartphone with a maximum frequency of 10 Hz. The evaluation will try to answer the question to what extent the frequency could be lowered, without restraining system functionality too much. To derive deeper insights into the trained network, convolutional layers should be visualized and analyzed.

7 Time Plan

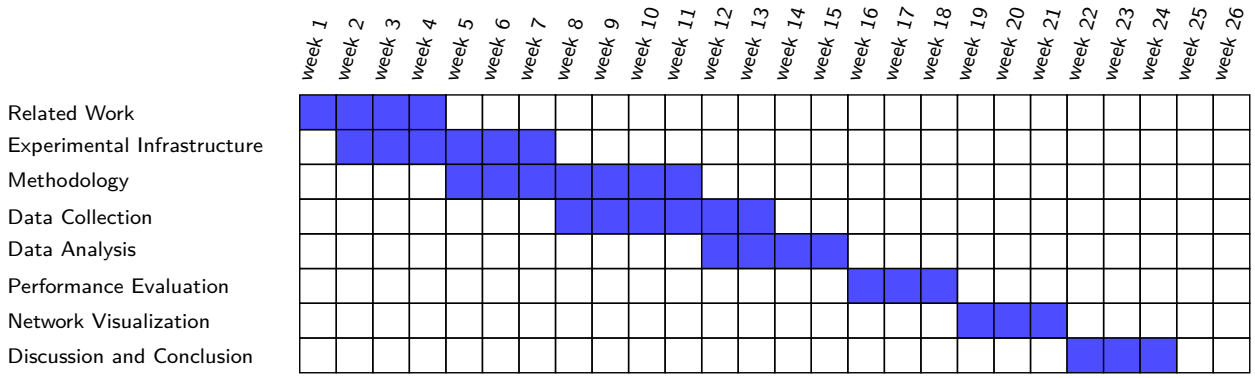


Figure 4: rough time schedule

8 References

[1] Benbasat, A. Y. and Paradiso, J. A. (2001). An inertial measurement framework for gesture recognition and applications. In *International Gesture Workshop*, pages 9–20. Springer.

[2] Bulling, A., Blanke, U., and Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput. Surv.*, 46(3):33:1–33:33.

[3] Chang, K.-h., Chen, M. Y., and Canny, J. (2007). *Tracking Free-Weight Exercises*, pages 19–37. Springer Berlin Heidelberg, Berlin, Heidelberg.

[4] Estimote (2014). *Nearable.* `http://developer.estimote.com/nearables/` *[Accessed: 11 November 2016].*

[5] Ha, S., Yun, J.-M., and Choi, S. (2015). Multi-modal convolutional neural networks for activity recognition. In *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, pages 3017–3022. IEEE.

[6] Instruments, T. (2016). *CC2650STK.* `http://www.ti.com/tool/CC2650STK?HQS=TI-null-null-mousermode-df-pf-null-wwe&DCM=yes` *[Accessed: 16 November 2016].*

[7] InvenSense (2016). *MPU-9250.* `https://www.invensense.com/wp-content/uploads/2015/02/PS-MPU-9250A-01-v1.1.pdf` *[Accessed: 16 November 2016].*

[8] Karpathy, A., Johnson, J., and Fei-Fei, L. (2015). Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

[9] Kranz, M., Möller, A., Hammerla, N., Diewald, S., Plötz, T., Olivier, P., and Roalter, L. (2013). The mobile fitness coach: Towards individualized skill assessment using personalized mobile devices. *Pervasive and Mobile Computing*, 9(2):203 – 215. Special Section: Mobile Interactions with the Real World.

[10] Li, C., Fei, M., Hu, H., and Qi, Z. (2012). Free weight exercises recognition based on dynamic time warping of acceleration data. In *International Conference on Intelligent Computing for Sustainable Energy and Environment*, pages 178–185. Springer.

[11] Morris, D., Saponas, T. S., Guillory, A., and Kelner, I. (2014). Recofit: using a wearable sensor to find, recognize, and count repetitive exercises. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 3225–3234. ACM.

[12] Muehlbauer, M., Bahle, G., and Lukowicz, P. (2011). What can an arm holster worn smart phone do for activity recognition? In *2011 15th Annual International Symposium on Wearable Computers*, pages 79–82. IEEE.

[13] Novatchkov, H. and Baca, A. (2013). Artificial intelligence in sports on the example of weight training. *Journal of sports science & medicine*, 12(1):27.

[14] Opportunity (2011). *Opportunity challenge.* ***http://www.opportunity-project.eu/challenge*** *[Accessed: 11 November 2016].*

[15] Ordóñez, F. J. and Roggen, D. (2016). Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115.

[16] Patel, S., Park, H., Bonato, P., Chan, L., and Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, 9(1):21.

[17] Pernek, I., Hummel, K. A., and Kokol, P. (2013). Exercise repetition detection for resistance training based on smartphones. *Personal and ubiquitous computing*, 17(4):771–782.

[18] Pernek, I., Kurillo, G., Stiglic, G., and Bajcsy, R. (2015). Recognizing the intensity of strength training exercises with wearable sensors. *Journal of biomedical informatics*, 58:145–155.

[19] Pigou, L., Oord, A. v. d., Dieleman, S., Van Herreweghe, M., and Dambre, J. (2015). Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *arXiv preprint arXiv:1506.01911*.

[20] Seeger, C., Buchmann, A., and Van Laerhoven, K. (2011). myhealthassistant: a phone-based body sensor network that captures the wearer's exercises throughout the day. In *Proceedings of the 6th International Conference on Body Area Networks*, pages 1–7. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[21] Velloso, E., Bulling, A., Gellersen, H., Ugulino, W., and Fuks, H. (2013). Qualitative activity recognition of weight lifting exercises. In *Proceedings of the 4th Augmented Human International Conference*, AH '13, pages 116–123, New York, NY, USA. ACM.

[22] Yang, C.-C. and Hsu, Y.-L. (2010). A review of accelerometry-based wearable motion detectors for physical activity monitoring. *Sensors*, 10(8):7772.

[23] Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

[24] Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4694–4702.

[25] Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer.