A Recurrent Neural Network for Hand Gesture Recognition based on Accelerometer Data

Philipp Koch, Mark Dreier, Marco Maass, Martina Böhme, Huy Phan, and Alfred Mertins

***Abstract*— For many applications, hand gesture recognition systems that rely on biosignal data exclusively are mandatory. Usually, theses systems have to be affordable, reliable as well as mobile. The hand is moved due to muscle contractions that cause motions of the forearm skin. Theses motions can be captured with cheap and reliable accelerometers placed around the forearm. Since accelerometers can also be integrated into mobile systems easily, the possibility of a robust hand gesture recognition based on accelerometer signals is evaluated in this work. For this, a neural network architecture consisting of two different kinds of recurrent neural network (RNN) cells is proposed. Experiments on three databases reveal that this relatively small network outperforms by far state-of-the-art hand gesture recognition approaches that rely on multi-modal data. The combination of accelerometer data and an RNN forms a robust hand gesture classification system, i.e., the performance of the network does not vary a lot between subjects and it is outstanding for amputees. Furthermore, the proposed network uses only $5\,\mathrm{ms}$ short windows to classify the hand gestures. Consequently, this approach allows for a quick, and potentially delay-free hand gesture detection.**

I. INTRODUCTION

Decoding hand gestures from biosignals is essential for a variety of different applications such as human machine interaction [1] and virtual reality [2]. Depending on their actual application, such hand gesture recognition systems have to meet various requirements. In general, cheap systems that detect hand movements with very low delay are desired. Often it is required that the hand movement detection systems are part of an embedded system or built as a mobile device.

Especially in the medical field, hand gesture detection systems find frequent application, e.g., in prosthesis control [3] or in the control of exoskeletons [4], [5]. These systems typically rely on surface electromyography (sEMG) signals exclusively. These sEMG signals are acquired by noninvasive electrodes and allow for the decoding of hand movements from electric fields that are caused by muscle contractions. The overall recognition pipeline is usually composed of a preprocessing step followed by a hand-crafted feature extraction and a conventional classifier such as a support vector machine or a random forest [6], [7], [8]. Recently, different kinds of deep learning methods including convolutional neural networks (CNNs) [9], [10] and recurrent neural networks (RNNs) [11], [12] have been employed. On the one hand, RNNs turn out to be a promising tool as even small networks with a very limited number of trainable parameters have shown outstanding classification performance. On the other hand, CNNs are particularly suitable for analyzing raw data because a feature extraction and a classifier are jointly learned in a single end-to-end trained network. For these reasons, these networks are particularly suitable for classifying hand gestures in sEMG data and have shown promising performance [9], [10]. However, in general, all of the mentioned approaches suffer from two drawbacks. First, to achieve satisfying classification results, long analysis windows are required resulting in long delays. Second, an expensive and complicated sEMG system is required for data acquisition.

An affordable and simple alternative to sEMG systems is the data acquisition with accelerometers. The sensors are placed around the forearm similar to the usually used sEMG electrodes. Each accelerometer can be used to measure the local skin motion caused by muscle contractions. Since the accelerometer data indirectly include information about the voluntary muscle contraction, they can be used to decode hand movements.

In this work, a stacked RNN-based architecture for hand movement recognition is proposed. This network combines the feature extraction abilities of CNNs and the sequential analysis capabilities of RNNs by using different kinds of RNN cells. The proposed network is capable of classifying over 40 different hand gestures given windows of length $5\,\mathrm{ms}$.

The suitability of the proposed approach for classifying hand movements is validated using three databases containing data recordings from able-bodied as well as amputated subjects. For all databases, the RNN-based system significantly outperforms state-of-the-art approaches even though it relies only on accelerometer data. The network attains satisfying classification results for all subjects. Furthermore, having in mind that usually window sizes around $200\,\mathrm{ms}$ are necessary to achieve satisfying results, the short window size is advantageous since the delay of the hand movement recognition system is minimal.

II. NETWORK ARCHITECTURE FOR ACCELEROMETER SIGNAL ANALYSIS

With standard approaches based on hand-crafted features and a conventional classifier such as random forest it is

Philipp Koch, Martina Böhme, and Alfred Mertins are with the Institute for Signal Processing, University of Lübeck, 23562 Lübeck, Germany {koch,boehme,mertins}@isip.uni-luebeck.de

Mark Dreier is with the University of Lübeck, 23562 Lübeck, Germany mark.dreier@student.uni-luebeck.de

Huy Phan is with the School of Computing, University of Kent, Canterbury, Kent, CT27NF, United Kingdom h.phan@kent.ac.uk

Marco Maass is with the Institute for Signal Processing and the Graduate School for Computing in Medicine and Life Sciences, University of Lübeck, 23562 Lübeck, Germany maass@isip.uni-luebeck.de

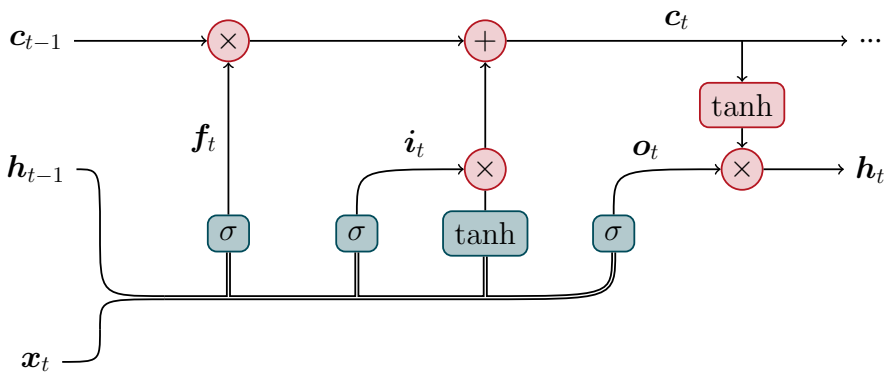

Fig. 1. Illustration of an LSTM cell.

not possible to reliably classify short windows. However, by exploiting the sequential nature of a signal, short windows are sufficient to recognize hand gestures correctly. Therefore, an RNN-based network is proposed. The input of the network is a sequence of three-dimensional matrices. The first dimension corresponds to the accelerometers (usually arranged in a circle around the forearm) and the second to the samples of the window. The third dimension represents the three axes of the accelerometer and is treated in the network like the channel dimension of a color image. The proposed architecture contains two different kinds of RNN cells, long-short term memory (LSTM) cells [13] and convolutional LSTM (ConvLSTM) cells [14]. These cells are stacked.

To extract features and to exploit the spatial information of the sensor position, a ConvLSTM cell is used. The corresponding hidden layer of the ConvLSTM cell can be described as $(\boldsymbol{H}_t, \boldsymbol{C}_t) = \mathcal{H}^{\text{conv}}(\boldsymbol{X}_t, \boldsymbol{H}_{t-1}, \boldsymbol{C}_{t-1})$, where the subscript t denotes the current time step, $\boldsymbol{H}$ the output vector, $\boldsymbol{X}$ the input vector, and C the cell state. The cell state is updated via

$$\begin{aligned}\boldsymbol{C}_t =& \boldsymbol{I}_t \odot \tanh\left(\boldsymbol{W}_{XC} * \boldsymbol{X}_t + \boldsymbol{W}_{HC} * \boldsymbol{H}_{t-1} + \boldsymbol{B}_C\right) \\ &+ \boldsymbol{F}_t \odot \boldsymbol{C}_{t-1},\end{aligned} \tag{1}$$

where $*$ denotes the convolution, $\odot$ the Hadamard product, and

$$\boldsymbol{I}_t = \sigma\left(\boldsymbol{W}_{XI} * \boldsymbol{X}_t + \boldsymbol{W}_{HI} * \boldsymbol{H}_{t-1} + \boldsymbol{B}_I\right) \tag{2}$$

and

$$\boldsymbol{F}_t = \sigma\left(\boldsymbol{W}_{XF} * \boldsymbol{X}_t + \boldsymbol{W}_{HF} * \boldsymbol{H}_{t-1} + \boldsymbol{B}_F\right). \tag{3}$$

All $\boldsymbol{W}$ matrices contain the trainable filter kernels and every $\boldsymbol{B}$ represents a trainable bias. The output of the cell is calculated by

$$\boldsymbol{H}_t = \boldsymbol{O}_t \odot \tanh\left(\boldsymbol{C}_t\right) \tag{4}$$

with

$$\boldsymbol{O}_t = \sigma\left(\boldsymbol{W}_{XO} * \boldsymbol{X}_t + \boldsymbol{W}_{HO} * \boldsymbol{h}_{t-1} + \boldsymbol{B}_O\right). \tag{5}$$

The second kind of RNN cell used in this work is the standard LSTM cell. An illustration of an LSTM cell is shown in Fig. 1. The mathematical formulation of an LSTM is very similar to that of a ConvLSTM and can be obtained by substituting the convolution by a matrix multiplication in the above equations and, consequently, replacing the bias matrices by vectors.

The proposed network is a stacked combination of above-mentioned RNN cells and is shown in Fig. 2. The first cell of the network is a ConvLSTM with 3×3 filter kernels. The number of filter kernels of a single convolution operation is 32. It is followed by a max-pooling with a kernel size of 4×4 and a stride of 2. After the pooling layer another ConvLSTM cell follows. This RNN cell has just 16 filter kernels per convolution layer but the same filter size as in the first ConvLSTM cell. The final RNN cell is a standard LSTM cell with a state size of 512. To obtain the actual classification the output of the last RNN cell is fed through a fully-connected layer followed by a softmax activation function.

III. TRAINING AND VALIDATION OF THE NETWORK

To generate enough training examples the training of the network is based on overlapping sequences of fixed length that are extracted from the training data. Each window of the sequence is classified by the network leading to a sequence of labels. However, only the classification of the final window is deployed to calculate the error of the network. Let $\hat{\boldsymbol{y}}_T$ denote the network's classification of the final window T. Then the loss function for optimizing the network is based on the cross-entropy given by

$$E\left(\Theta|\boldsymbol{X}, \boldsymbol{y}_T\right) = -\boldsymbol{y}_T \log\left(\hat{\boldsymbol{y}}_T\left(\Theta|\boldsymbol{X}\right)\right) \tag{6}$$

with $\boldsymbol{X}$ being the input sequence and $\boldsymbol{y}_T$ the ground truth of the final time step T represented as a one-hot encoded vector. The parameters of the network are denoted by Θ.

In contrast to the training procedure, the sequence length varies in the test case because all test examples are processed as one sequence. Unlike the training case, no sequences of fixed length are extracted. Consequently, the network estimates a class for each window of a test sequence. This test setting is close to the actual application of such classifiers because in a hand gesture system the classifier has to detect

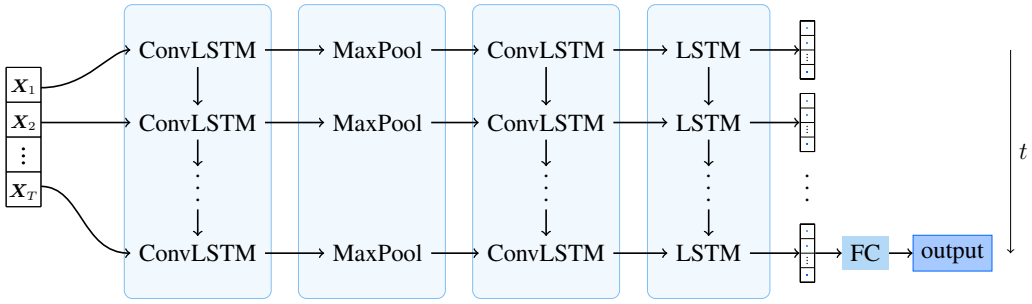

Fig. 2. Illustration of the proposed network architecture in training configuration. The network shown is enrolled over time. MaxPool denotes a two-dimensional max-pooling and FC represents a fully-connected layer that is used for classification.

hand gestures in each window coming to the system in realtime. To evaluate the performance of the network for each window the predicted class and the corresponding ground truth are compared and the accuracy is calculated.

IV. EXPERIMENTS

A. Databases

To validate the proposed network architecture, experiments on three different publicly available databases were conducted. The databases are DB2, DB3, and DB7 of the Ninapro project [6], [15]. The databases provide data to evaluate hand gesture recognition systems based on biosignals.

The experiments of all three databases followed mainly the same protocol. For the experiments, sensors containing at least an accelerometer and an sEMG electrode were used. If possible, 12 of these sensors were placed around the forearm of the subjects. The subjects performed different hand movements. Each movement was repeated 6 times.

The data of DB2 and DB3 were acquired with the Delsys® Trigno™Wireless sensors. Each of these sensors acquires sEMG signals and the data of a tri-axial accelerometer sampled at 148 Hz. In order to match the sampling frequency of the sEMG signals the accelerometer data were upsampled to 2 kHz by linear interpolation already within the Ninapro project. The number of performed hand gestures in both databases DB2 and DB3 is 50 (excluding rest). The hand gestures range from movements of single finger over wrist motions to complex motion sequences such as different grasps. DB2 contains experiments of 40 able-bodied persons. In contrast, DB3 includes the experiments of 11 amputees.

Database DB7 contains experiments of 20 able-bodied subject and 2 amputees. The subjects were asked to perform 40 (exclusive rest) different hand movements. The movements are a subset of the set of hand gestures performed in DB2 and DB3. For DB7, the Delsys® Trigno™IM Wireless System was used to acquire the data. The sensors of this system acquire sEMG data at 2 kHz as well and the signals of an inertial measurement unit (IMU) at 128 Hz. The IMU with 9-degree-of-freedom consists of tri-axial accelerometer, gyroscope, and magnetometer. In this work, only the tri-axial accelerometer of the IMU is used. Analogous to DB2 and DB3 the accelerometer signals were upsampled to 2 kHz.

TABLE I

ACCURACY COMPARISON OF ACCELEROMETER AND SEMG. THE REPORTED RESULTS FOR ACCELEROMETER WERE OBTAINED BY A NETWORK BASED ON A SINGLE LSTM CELL WITH A STATE OF SIZE 256. THE RESULTS OF [12] ARE THE BEST PUBLISHED RESULTS FOR DB2 AND DB3 AND WERE ACHIEVED ON 100 ms LONG WINDOWS THAT WERE REPRESENTED BY A FEATURE VECTOR AND CLASSIFIED BY AN RNN.

Database	Accelerometer	sEMG [12]
DB2	81.8 %	78.0 %
DB3	70.1 %	55.3 %

B. Preprocessing

The training and the test datasets were generated following the recommendations for the database. The signals had to be preprocessed before feeding them to the network. A normalization was performed for each of the 3 axes of the accelerometer individually by subtracting the mean and dividing by the standard deviation. All necessary statistics were calculated based on the training data exclusively. For the actual classification the signals were split into windows of 5 ms length.

C. Results

To allow comparisons with previous works, reported results for DB2 and DB3 are average accuracies that are calculated across all subjects of the corresponding database while for DB7 the median is reported.

The average classification accuracies obtained by the proposed network are shown in Table I. The results are compared with, to the best of your knowledge, best published results in both DB2 and DB3 [12]. These results were obtained based on sEMG data and an RNN. The sEMG signals were split into 100 ms long windows followed by a window-wise feature extraction to prepare the signals for the RNN classification. With the proposed approach the hand gestures of healthy subjects in DB2 could be classified with 3.8 % (absolute) higher accuracy than by the sEMG-based system. The difference is even more significant for amputees. The accelerometer-based RNN improves the average classification accuracy by almost 15 % absolute. The

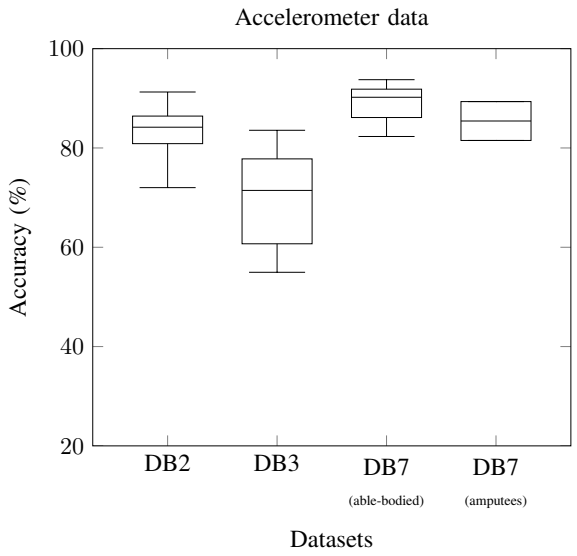

Fig. 3. Boxplots of the obtained results for all databases.

TABLE II

RESULTS OF THE PROPOSED NETWORK ON DB7. THE RESULTS OF THE PROPOSED NETWORK ARE COMPARED WITH STATE-OF-THE-ART RESULTS ACHIEVED WITH AN APPROACH THAT TAKES ALL IMU MODALITIES INTO ACCOUNT, NOT JUST THE ACCELEROMETER.

Data / Method	able-bodied	amputated
IMU / [15]	81.7 %	77.7 %
Accelerometer / proposed RNN	89.8 %	85.4 %

boxplots in Fig. 3 show that no severe outliers were obtained. Furthermore, for some amputees the robustness of the hand gestures recognition is at the same level as for abled-bodied subjects.

The results of the proposed network on DB7 are presented in Table II. The presented approach outperforms the state-of-the-art system [15] by about $8\,\%$ absolute for both abled-bodied subjects and amputees, even though the state-of-the-art system uses in addition to the accelerometer also the gyroscope and the magnetometer. Moreover, the state-of-the-art approach requires $256\,\mathrm{ms}$ long windows, whereas the presented network handles significantly shorter windows of length $5\,\mathrm{ms}$. The performance of the system in [15] varies significantly among the subjects. For several subjects the individual classification accuracy is roughly $20\,\%$ lower than the average accuracy calculated over all subjects. The boxplots in Fig. 3 reveal that the proposed network achieves similar but very good results for all subject, even for the amputees. Overall, the results indicate that the proposed network is more robust and more accurate than state-of-the-art approaches even though only accelerometer data are provided for classifying hand gestures.

V. CONCLUSIONS

In this work, an RNN hand gesture classification based on accelerometer data was proposed. A combination of ConvLSTMs and standard LSTM cells was used to exploit both temporal and spatial information in the accelerometer signals. This approach is preferable to state-of-the-art systems in terms of better classification accuracy and shorter window length requirements. The classification accuracy can be improved by nearly $15\,\%$ and the window length can be reduced to $5\,\mathrm{ms}$ compared to 100 / $200\,\mathrm{ms}$. Furthermore, the results of the experiments reveal that the proposed architecture works comparably well for all individual subjects. It seems to take the physiological characteristics of the individual subjects into account. Since the proposed approach showed promising results especially for amputees it would be interesting to evaluate such an accelerometer-based hand gesture classifiers in a real-world scenario.

REFERENCES

[1] J. Cheng, X. Chen, Z. Lu, K. Wang, and M. Shen, "Key-press gestures recognition and interaction based on sEMG signals," in *Proc. Int. Conf. Multimodal Interact. and Mach. Learn. Multimodal Interact.*, 2010.

[2] F. Muri, C. Carbajal, A. M. Echenique, H. Fernández, and N. M. López, "Virtual reality upper limb model controlled by EMG signals," *J. Phys. Conf. Ser.*, vol. 477, 2013.

[3] C. Cipriani, F. Zaccone, S. Micera, and M. C. Carrozza, "On the shared control of an EMG-controlled prosthetic hand: Analysis of user-prosthesis interaction," *IEEE Trans. Robot.*, vol. 24, no. 1, pp. 170–184, 2008.

[4] J. Rosen, M. Brand, M. B. Fuchs, and M. Arcan, "A myosignal-based powered exoskeleton system," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 31, no. 3, pp. 210–222, 2001.

[5] K. Kiguchi and Y. Hayashi, "An EMG-based control for an upper-limb power-assist exoskeleton robot," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 4, pp. 1064–1071, 2012.

[6] M. Atzori, A. Gijsberts, C. Castellini, B. Caputo, A.-G. Mittaz Hager, S. Elsig, G. Giatsidis, F. Bassetto, and H. Müller, "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Sci. Data*, vol. 1, no. 140053, 2014.

[7] X. Zhang, X. Chen, Y. Li, V. Lantz, K. Wang, and J. Yang, "A framework for hand gesture recognition based on accelerometer and EMG sensors," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 41, no. 6, pp. 1064–1076, 2011.

[8] K. Englehart and B. Hudgins, "A robust, real-time control scheme for multifunction myoelectric control," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 7, pp. 848–854, 2003.

[9] M. Atzori, M. Cognolato, and H. Müller, "Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands," *Front. Neurorobot.*, vol. 10, no. 9, 2016.

[10] W. Geng, Y. Du, W. Jin, W. Wei, Y. Hu, and J. Li, "Gesture recognition by instantaneous surface EMG images," *Sci. Rep.*, vol. 6, no. 36571, 2016.

[11] P. Koch, H. Phan, M. Maass, F. Katzberg, and A. Mertins, "Recurrent neural network based early prediction of future hand movements," in *Proc. IEEE Eng. Med. Biol. Soc. (EMBC)*, July 2018.

[12] P. Koch, H. Phan, M. Maass, F. Katzberg, R. Mazur, and A. Mertins, "Recurrent neural networks with weighting loss for early prediction of hand movements," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, September 2018.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, 2015, NIPS'15, pp. 802–810.

[15] A. Krasoulis, I. Kyranou, M. S. Erden, K. Nazarpour, and S. Vijayakumar, "Improved prosthetic hand control with concurrent use of myoelectric and inertial measurements," *J. Neuroeng. Rehabil.*, vol. 14, no. 71, 2017.