

WAVEGLOVE: TRANSFORMER-BASED HAND GESTURE RECOGNITION USING MULTIPLE INERTIAL SENSORS

Matej Králik, Marek Šuppa

Faculty of Mathematics, Physics and Informatics, Comenius University Bratislava, Slovakia

ABSTRACT

Hand Gesture Recognition (HGR) based on inertial data has grown considerably in recent years, with the state-of-the-art approaches utilizing a single handheld sensor and a vocabulary comprised of simple gestures. In this work we explore the benefits of using multiple inertial sensors. Using WAVEGLOVE, a custom hardware prototype in the form of a glove with five inertial sensors, we acquire two datasets consisting of over 11000 samples. To make them comparable with prior work, they are normalized along with 9 other publicly available datasets, and subsequently used to evaluate a range of Machine Learning approaches for gesture recognition, including a newly proposed Transformer-based architecture. Our results show that even complex gestures involving different fingers can be recognized with high accuracy. An ablation study performed on the acquired datasets demonstrates the importance of multiple sensors, with an increase in performance when using up to three sensors and no significant improvements beyond that.

Index Terms— hand gesture recognition, transformer, multi-sensor, inertial sensor

1. INTRODUCTION

Human-computer interaction is experiencing a significant increase in popularity due to the high availability and low cost of various sensors. Human Activity Recognition (HAR) is most often performed using vision-based sensors, which often need to be pre-installed in the environment, require favorable light conditions, specific position and rotation of the user and significant computing power. Inertial Measurement Units (IMUs) are a type of inertial sensor that provides a more pervasive experience. IMUs have lower power consumption, tend to be smaller and can work in a variety of environments.

Hand Gesture Recognition (HGR) – a sub-field of HAR – is of particular interest, because hand gestures are a natural method of expression for humans. Practical applications of HGR include device control [1], health monitoring, gaming, virtual and augmented reality and biometrics [2]. The state-of-the-art approaches utilize a single (often handheld) IMU sensor and a vocabulary of very simple gestures [3, 4, 5, 6]. The use of multiple sensors is largely unexplored, which can

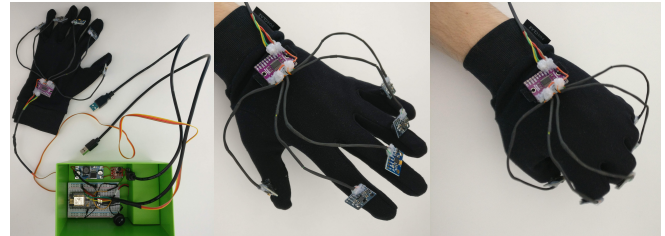


Fig. 1. The WAVEGLOVE prototype.

be attributed to their significant cost in the past. At the same time, various hand gestures used on a daily basis make distinctive use of each of the fingers, making it hard to distinguish these natural gestures using only one handheld sensor.

To explore the multi-sensor scenario we built a custom hardware prototype called WAVEGLOVE and designed two gesture vocabularies: WAVEGLOVE-single with 8 whole-hand movements and WAVEGLOVE-multi with 10 carefully designed gestures, which exhibit different movements for each of the fingers. Using the prototype we acquired 1000 and 10000 samples for datasets based on the WAVEGLOVE-single and WAVEGLOVE-multi vocabularies, respectively.

In an effort to put our work in context, we use a set of 11 datasets. It includes 2 datasets captured with WAVEGLOVE, 3 publicly available datasets we normalized as part of this work and 6 datasets normalized in [7]. These are used to evaluate a wide selection of Machine Learning approaches for hand gesture recognition. We reproduce multiple previously published methods and propose an improved approach based on the Transformer [8], which achieves state-of-the-art results.

To the best of our knowledge, our work is the first to specifically examine the impact of using multiple sensors for HGR. Through an ablation study we explore the accuracy implications of multiple sensors. We also compare the recognition performance of individual sensors and demonstrate a dependency between sensor position and gesture type.

In summary, we built a custom multi-sensor hardware prototype and used it to acquire datasets of over 11000 samples. We further present 11 datasets in a uniform, normalized format, use them to evaluate a number of recognition approaches. Finally, we propose a Transformer-based model and report state-of-the-art results in its evaluation on the pre-

sented datasets. Using this model we conduct an ablation study which shows that increasing the number of sensors leads to increased recognition accuracy, up to a point.

2. RELATED WORK

Several recent comprehensive surveys summarize research and define the state of the art in the field of HAR [9, 10]. The most commonly used recognition methods include Classic Machine Learning, although Deep Learning methods experience a significant rise in popularity lately [11].

Catal et al. [12] used classifier ensembles in combination with hand-crafted features. Convolutional Neural Networks were applied to the task of HAR in [13] and [14] demonstrated that 2D convolutions achieve better results than one dimensional convolutions. In [4], recurrent LSTM layers were used for online recognition and [15] has achieved state-of-the-art results using bidirectional LSTM and GRU layers. DeepConvLSTM [16] represents a combination of convolutional and recurrent layers. Inspired by the Transformer [8] (which established a new state-of-the-art for machine translation tasks), [17] introduces the self-attention mechanism in the field of HAR. Our work builds on these results and proposes an improved method based on the self-attention mechanism.

Hand Gesture Recognition was previously explored in [1, 4, 5, 6]. In [18] a similar glove prototype was built to recognize french sign language characters by estimating the rotation of each finger. A single-sensor glove prototype called GestGlove [19] was built to recognize a simple set of hand gestures allowing phone control. In [20], the concept of forked recurrent neural networks is applied to HGR.

One of the current issues in the field of HAR is the lack of a standardized benchmark. A base for a standardized benchmark with a predefined methodology and publicly released code has been laid in [7]. We build on top of this work and aim to extend the realm of reproducible HAR methods. From the dataset perspective, Opportunity [21], Skoda [22] and PAMAP2 [23] are three widely used HAR datasets, which use multiple sensors. In [7], 6 more HAR datasets are identified and normalized. The uWave [6] dataset presents one of the few available datasets which focus on hand gestures.

3. WAVEGLOVE

We built a custom hardware glove prototype called WAVEGLOVE (shown in Figure 1). It is based on a left-hand silken glove to which 5 IMUs are attached: one MPU6050 [24] module on each of the fingers. A multiplexer is present on the back of the palm, connected via a flexible and long enough cable to allow comfortable use.

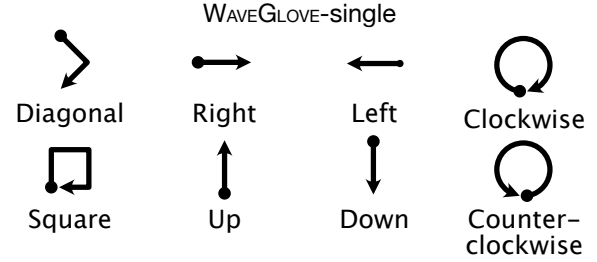


Fig. 2. Eight gestures from the WAVEGLOVE-single vocabulary.

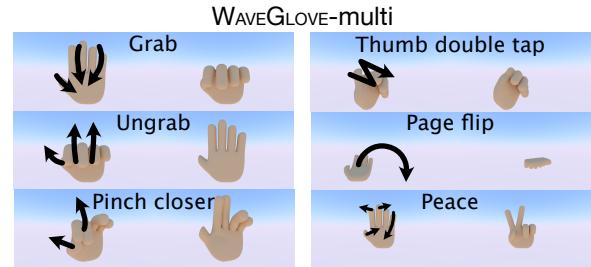


Fig. 3. A selection of the gestures from the WAVEGLOVE-multi vocabulary, which contains 10 gestures in total.

3.1. WAVEGLOVE datasets

Two gesture vocabularies were used to acquire two separate datasets.

The WAVEGLOVE-single vocabulary consists of 8 gestures and a null class – the state in which no gesture is performed. These gestures are simple movements, during which the whole hand performs a uniform motion (Figure 2). This set of gestures is motivated by [26], which identified it to be a popular and user-preferred set for device control. It is also one of the most often used gesture vocabularies [5, 6, 27]. WAVEGLOVE-single is primarily used as a reference point – it uses the same set of gestures as previous work, with the only difference being the use of five sensors instead of a single one.

On the other hand, WAVEGLOVE-multi was designed as a set of custom hand gestures which would showcase the potential of multiple sensors. These gestures consist of different movements for individual fingers and some gestures differ only in specific finger movements. This vocabulary is used to demonstrate that the use of multiple sensors located on the hand significantly expands the space of recognizable gestures and allows the hand gestures to become much more natural. It contains 10 gestures and a null class.

The key difference between the two gesture vocabularies lies in their level of complexity. While gestures from WAVEGLOVE-single were successfully classified using a single sensor in previous work, the gestures in WAVEGLOVE-multi are designed to fully utilize multiple sensors. We take a closer look at this difference in an ablation study in Section 6.

	WAVEGLOVE-single	WAVEGLOVE-multi	MHEALTH	USC-HAD	UTD-MHAD1	UTD-MHAD2	WHARF	WISDM	uWave	Average
Baseline Decision Tree	99.10	96.63	93.41	89.22	67.51	81.90	66.26	61.61	70.72	80.71
DeepConvLSTM (2016)[16]	98.05	99.30	81.01	83.97	67.29	86.50	67.98	91.23	98.10	85.94
DCNN Ensemble (2019)[25]	-	-	93.09	88.49	62.03	81.63	75.50	89.01	-	-
Transformer-based (proposed)	99.40	99.99	90.35	89.83	76.32	88.42	78.63	84.53	98.80	89.59

Table 1. Accuracy comparison of classification methods. The “-” symbol denotes that the given method was not applied on the dataset. Note that this table only showcases the most relevant methods and datasets. For the full set of results (over 12 methods and 11 datasets), exact model and hyperparameter description, as well as the code used to preprocess the input datasets and produce the results, please refer to <https://github.com/zajozor/waveglove>.

Several important factors affect the final quality of the dataset [28]. In order to make our samples more realistic, ensure the recordings stay independent and improve the reproducibility of our experiments, we took several measures. The gestures are recorded in short (approximately 2 minute) sessions. Each session consists of recording a few samples for each of the gestures in one of the datasets. In each session, the gestures are presented in a random order and the recording software requires a break (of a few seconds) between any two gestures. This improves gesture independence and allows the data captured during the breaks to be used as the null class.

Before recording, the test subjects are first presented with uniform instructions. The test subjects themselves mark the beginning and the end of a sample by pressing and releasing a button on the keyboard. In case the subject made a mistake in the process (eg. forgot to move the hand, or pressed the button too late), they are allowed to re-record a sample.

In total, the WAVEGLOVE-single dataset contains over 1000 samples and the WAVEGLOVE-multi dataset contains over 10000 samples. To the best of our knowledge WAVEGLOVE is the first device utilizing multiple inertial sensors used to record publicly available datasets of this scale.

4. DATASET NORMALIZATION

In order to ensure fair comparison of the presentend HAR and HGR models, we prepare a set of 11 datasets as a standardized benchmark.

We include the 6 datasets preprocessed in [7] using the Leave-One-Trial-Out (LOTO) folds, which ensure that all samples from a single trial are either in the training or the test set. The LOTO folds use semi-overlapping windows for sampling, while the rest of the datasets are sampled using non-overlapping windows. These 6 datasets include MHEALTH and USC-HAD – two commonly used HAR datasets. The uWave [6] dataset is a single-sensor HGR dataset with the same vocabulary as WAVEGLOVE-single. Finally, we include the Opportunity [21] and Skoda [22] datasets, along with the two WAVEGLOVE datasets, resulting in 11 datasets in total.

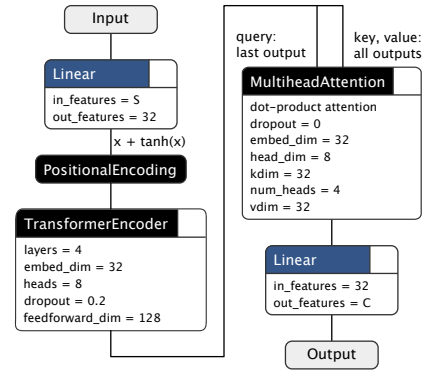


Fig. 4. Architecture of the proposed Transformed-based model.

5. METHODS

For classification we use methods ranging from Classical Machine Learning to approaches based on Deep Learning. To make our results comparable, we reproduce several previously published methods. We further propose a classification method which utilizes the self-attention mechanism and is based on the Transformer. A Transformer-based architecture was utilized for HAR in [17], upon which we further improve by simplification and by using a more commonly used dot-product attention.

Figure 4 shows the proposed architecture. The first linear layer transforms an input of size $B \times T \times S$ (B = batch size, T = temporal length, S = sensor channels) to $B \times T \times 32$, creating a constant-size sensor embedding. Next, positional encoding[8] and 4 Transformer encoder layers are used to provide an input ($T \times B \times 32$) for a dot-product attention layer (the final encoder output in the temporal dimension is used as the query). The final linear layer yields a one-hot encoding for target classes ($B \times 32 \rightarrow B \times C$, C = classes).

A hyperparameter sweep has been performed yielding the final, best-performing parameters. Final model accuracy comparison can be found in Table 1.

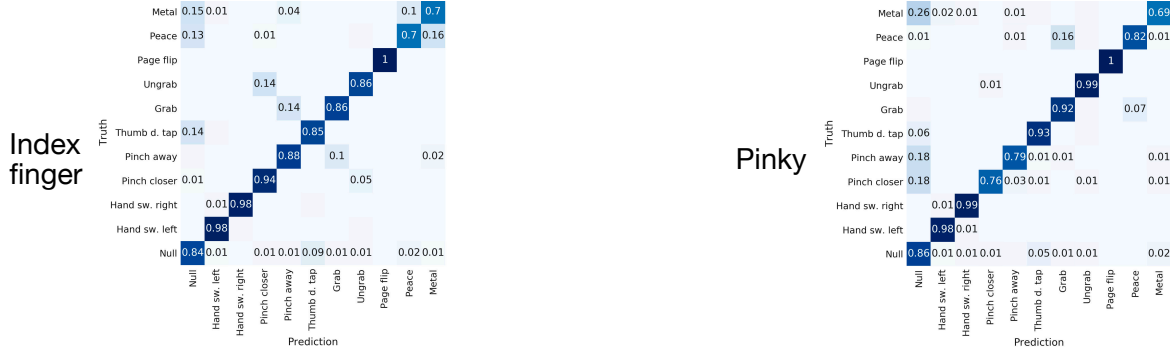


Fig. 5. Confusion matrices of a baseline model trained solely on **index finger** (left) and **pinky** (right) sensor data

6. ABLATION STUDY

Most of the evaluated methods achieved near-perfect accuracy on WAVEGLOVE datasets. Given the multi-sensor setup, this opens up an opportunity to conduct an ablation study to examine the effect of various factors on the final performance.

6.1. Finger attribution

In Figure 5 we compare the confusion matrices obtained by training a model on data from only a single sensor. It allows us to observe a few distinguishable patterns.

Using the index finger sensor, the *Metal* and *Peace* gestures are classified as the *Null* class since they do not include index finger movement. The gesture pairs *Ungrab*/*Pinch closer* and *Grab*/*Pinch away* include the same index finger movement, resulting in higher inter-class confusion. When using the pinky sensor, the *Peace* and *Grab* gestures are often confused. Several gestures are confused with the *Null* class, because they do not involve significant pinky movement.

This experiment helps us demonstrate that sensor placement on different fingers can have significant effect on classification results, particularly when various fingers are featured prominently in the gestures.

6.2. Impact of multiple sensors

Figure 6 depicts how the number of used sensors and the training set size affect the final accuracy¹. With WAVEGLOVE-single, the accuracy does not increase with the addition of multiple sensors. This is what we would expect, as the gestures in this vocabulary do not make use of multiple sensors.

Conversely, for the WAVEGLOVE-multi dataset we observe a clear positive impact of using more sensors on the accuracy. The improvements can be considered significant when using up to three sensors. Beyond that the increase in performance is marginal. This positive impact is even more significant when the training set size is limited.

¹When there are multiple combinations for a certain amount of sensors, we plot the mean accuracy and the sleeves show the range.

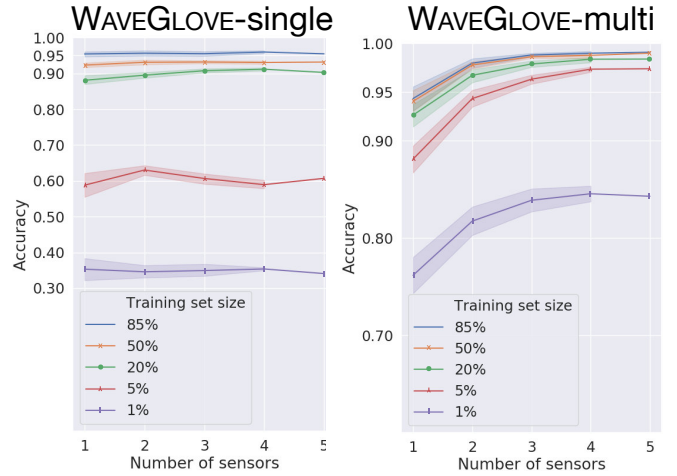


Fig. 6. Average accuracy based on the amount of sensors used and the training set size as measured on the two WAVEGLOVE datasets.

7. CONCLUSIONS

In this work we present a custom multi-sensor hardware prototype, which we used to acquire a dataset of over 11000 gesture instances. To the best of our knowledge this is the largest publicly available dataset of multi-sensor hand gestures. For a fair comparison with prior work we use 11 HAR datasets, implement several previously published methods and compare their performance. We further propose a Transformer-based network architecture, which shows promising classification results across multiple datasets.

In an accompanying ablation study, we identify that (only) relevantly designed gestures benefit from the use of multiple sensors. By comparing classification performance for two different sensor locations we demonstrate that the recognition accuracy has strong dependence on the location of the sensor and the type of classified gestures.

8. REFERENCES

- [1] Zhenyu He et al., “Gesture recognition based on 3d accelerometer for cell phones interaction,” in *APCCAS 2008*, 2008, pp. 217–220.
- [2] Qin Zou et al., “Deep learning-based gait recognition using smartphones in the wild,” *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 3197–3212, 2020.
- [3] Vasileios Sideridis et al., “Gesturekeeper: Gesture recognition for controlling devices in iot environments,” *CoRR*, vol. abs/1903.06643, 2019.
- [4] Alessandro Carfi et al., “Online human gesture recognition using recurrent neural networks and wearable sensors,” in *RO-MAN 2018*. 2018, pp. 188–195, IEEE.
- [5] Gabriele Costante et al., “Personalizing a smartwatch-based gesture interface with transfer learning,” in *EU-SIPCO, 2014*, 2014, pp. 2530–2534.
- [6] Jiayang Liu et al., “uwave: Accelerometer-based personalized gesture recognition and its applications,” *Pervasive and Mobile Computing*, vol. 5, no. 6, pp. 657–675, 2009.
- [7] Artur Jordao et al., “Human activity recognition based on wearable sensor data: A standardization of the state-of-the-art,” *CoRR*, vol. abs/1806.05226, 2018.
- [8] Ashish Vaswani et al., “Attention is all you need,” *CoRR*, vol. abs/1706.03762, 2017.
- [9] Henry Friday Nweke et al., “Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges,” *Expert Syst. Appl.*, vol. 105, pp. 233–261, 2018.
- [10] Kaixuan Chen et al., “Deep learning for sensor-based human activity recognition: Overview, challenges and opportunities,” *CoRR*, vol. abs/2001.07416, 2020.
- [11] Florenc Demrozi et al., “Human activity recognition using inertial, physiological and environmental sensors: a comprehensive survey,” *CoRR*, vol. abs/2004.08821, 2020.
- [12] Cagatay Catal et al., “On the use of ensemble of classifiers for accelerometer-based activity recognition,” *Applied Soft Computing*, vol. 46, 01 2015.
- [13] Yuqing Chen others, “A deep learning approach to human activity recognition based on single accelerometer,” in *SMC 2015*, 2015, pp. 1488–1492.
- [14] Sojeong Ha et al., “Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors,” in *IJCNN 2016*, 2016, pp. 381–388.
- [15] Ce Li et al., “Deep fisher discriminant learning for mobile hand gesture recognition,” *Pattern Recognition*, vol. 77, pp. 276–288, 2018.
- [16] Francisco Javier Ordóñez et al., “Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition,” *Sensors*, vol. 16, no. 1, pp. 115, 2016.
- [17] Saif Mahmud et al., “Human activity recognition from wearable sensor data using self-attention,” *CoRR*, vol. abs/2003.09018, 2020.
- [18] Chaithanya Kumar Mummadi et al., “Real-time embedded recognition of sign language alphabet fingerspelling in an imu-based glove,” in *iWOAR 2017*. 2017, pp. 11:1–11:6, ACM.
- [19] Sonu Agarwal et al., “Gestglove: a wearable device with gesture based touchless interaction,” in *AH 2017, Mountain View, CA, USA, March 16-18, 2017*. 2017, p. 3, ACM.
- [20] Philipp Koch et al., “Forked recurrent neural network for hand gesture classification using inertial measurement data,” in *ICASSP 2019*. 2019, pp. 2877–2881, IEEE.
- [21] Paul Lukowicz et al., “Recording a complex, multi modal activity data set for context recognition,” in *ARCS ’10*. 2010, pp. 161–166, VDE Verlag.
- [22] Piero Zappi et al., “Activity recognition from on-body sensors: Accuracy-power trade-off by dynamic sensor selection,” in *EWSN 2008*. 2008, vol. 4913 of *Lecture Notes in Computer Science*, pp. 17–33, Springer.
- [23] Attila Reiss et al., “Introducing a new benchmarked dataset for activity monitoring,” in *ISWC 2012*. 2012, pp. 108–109, IEEE Computer Society.
- [24] *MPU-6000 and MPU-6050 Product Specification*, 2013, Revision 3.4.
- [25] Jessica Sena et al., “Human activity recognition based on wearable sensors using multiscale dcnn ensemble,” in *Anais Estendidos da XXXII Conference on Graphics, Patterns and Images*. 2019, p. 112–118, SBC.
- [26] Juha Kela et al., “Accelerometer-based gesture control for a design environment,” *Pers. Ubiquitous Comput.*, vol. 10, no. 5, pp. 285–299, 2006.
- [27] Matthias Rehm et al., “Wave like an egyptian: accelerometer based gesture recognition for culture specific interactions,” in *BCS-HCI 2008*, 2008, pp. 13–22.
- [28] Emiro De la Hoz et al., “Sensor-based datasets for human activity recognition – a systematic review of literature,” *IEEE Access*, vol. PP, pp. 1–1, 10 2018.