




Article

Performing Realistic Workout Activity Recognition on Consumer Smartphones

Biying Fu ^{1,*}, Florian Kirchbuchner ¹ and Arjan Kuijper ^{1,2}

¹ Fraunhofer Institute for Computer Graphics Research IGD, 64283 Darmstadt, Germany; florian.kirchbuchner@igd.fraunhofer.de (F.K.); arjan.kuijper@igd.fraunhofer.de (A.K.)

² Mathematical and Applied Visual Computing, TU Darmstadt, 64283 Darmstadt, Germany

* Correspondence: biying.fu@igd.fraunhofer.de

Received: 16 July 2020; Accepted: 2 November 2020; Published: 6 November 2020



Abstract: Smartphones have become an essential part of our lives. Especially its computing power and its current specifications make a modern smartphone a powerful device for human activity recognition tasks. Equipped with various integrated sensors, a modern smartphone can be leveraged for lots of smart applications. We already investigated the possibility of using an unmodified commercial smartphone to recognize eight strength-based exercises. App-based workouts have become popular in the last few years. The advantage of using a mobile device is that you can practice anywhere at anytime. In our previous work, we proved the possibility of turning a commercial smartphone into an active sonar device to leverage the echo reflected from exercising movement close to the device. By conducting a test study with 14 participants, we showed the first results for cross person evaluation and the generalization ability of our inference models on disjoint participants. In this work, we extended another model to further improve the model generalizability and provided a thorough comparison of our proposed system to other existing state-of-the-art approaches. Finally, a concept of counting the repetitions is also provided in this study as a parallel task to classification.

Keywords: ubiquitous sensing; ultrasonic sensing; mobile sensing; human activity recognition; proximity sensing; exercise recognition

1. Introduction

Quantified-self describes individuals committed to self-tracking of physical or behavioral information [1], like for example step counts per day, sleep rhythms or statistics of performed sportive activities. The most popular gadgets to perform this kind of activity collection are wearable devices, such as smartwatches or smartphones. However, in order to perform more precisely and accurately, applications are relying on acceleration data, which then requires the user to wear it directly on the body. Thus, wearable devices provoke the constraints of body-worn sensors.

Quantified-self is more than only tracking simple daily activities such as step counts or time duration of doing outdoor activities such as running, bicycling or walking. It also includes physical activities, such as performing exercises. Physical exercise can help people to maintain physical fitness and overall health. Exercise is a subset of physical activity that is planned, structured and repetitive.

The following work is an extended version of our earlier contribution: Unconstrained Workout Activity Recognition on Unmodified Commercial off-the-shelf Smartphones in the Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments, © ACM, 2020. <http://dx.doi.org/10.1145/3389189.3389195> [2], where we focused on using a commercial smartphone device to recognize eight realistic workout exercises.

A modern smartphone with multiple integrated sensors is well suited for the task of human activity recognition. By leveraging the integrated smartphone loudspeaker to emit a continuous signal

of 20 kHz, we actively turn the device into an active sonar sensor. The internal microphone is used to receive the echo signal modulated by the body movement close to the sensing device. By analyzing the echo signals and extracting features from the transformed frequency time spectrum, we are able to train several carefully designed end-to-end learning classifiers. The sequence model as well as the finetune model both show superior results on this kind of activity data. Finally, we examined the few-shot method to further increase the generalization ability of the performing model. We further develop a way to count the repetition in addition to the classification task.

The eight different workout activities, such as *push-up*, *sit-up*, *squat*, *segmental rotation*, *trunk rotation*, *swim*, *bridge*, and *quadruped*, are illustrated in Figure 1. The contributions of this work extended to our previous work [2] are concluded in the following aspects:

1. Improved model generalization to reduce the challenge of user diversity by applying few-shot classification learning;
2. Comparison of our proposed model to other existing state-of-the-art solutions and state the advantage and disadvantage of this application;
3. Count the exercise repetition with peak detection algorithm on pre-processed Doppler spectrum to build useful user exercise profiles.

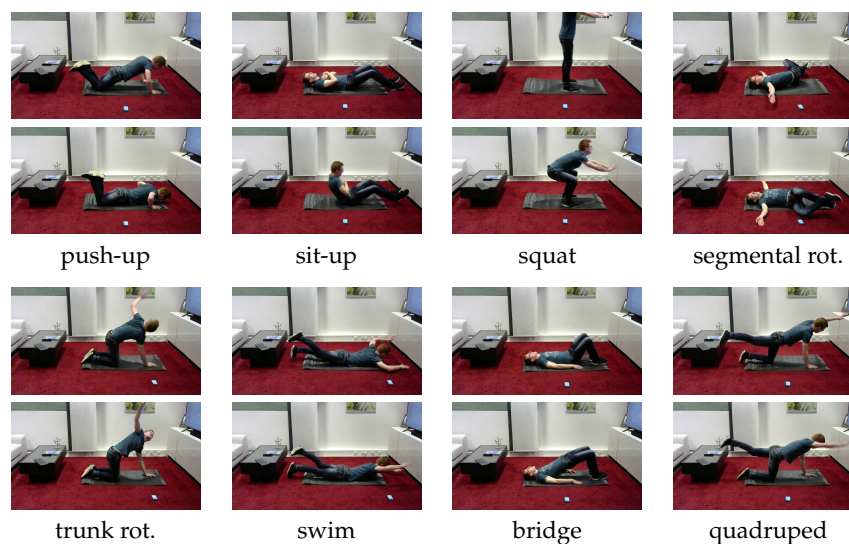


Figure 1. The eight workout exercises we collected in our living laboratory are visualized here. The figure also illustrates the position of our sensing mobile device with respect to the performing user body.

The overall structure is grouped in the following sections. We first introduce some related works focusing on strength-based exercise recognition in Section 2 and applications using mobile ultrasound to measure human activities. In Section 3 we shortly explain the physical sensing principle and provided the detailed processing pipeline in Section 4. In Section 5 we propose two different end-to-end learning architectures and justify our design choices. To further improve the generalization ability on the small data amount, we examine one approach from the few-shot classification learning. The setup for the final evaluation is discussed in detail in Section 6. In Section 7, a conceptual method of the repetition counting is proposed. It is then followed by a thorough comparison with other existing works in Section 8. Based on the design and the conducted tests, we discuss the challenges and viable solutions in Section 9 and finally conclude our work in Section 10.

2. Related Work

In this section, we first review various ways of performing physical exercise recognition commonly used in the human activity recognition (HAR) field. By identifying the disadvantages of such systems,

we provide a solution to overcome these limitations by using an unmodified commercial smartphone. We then introduce viable applications based on the same sensing principle of using a smartphone as an active sonar device in the context of HAR.

2.1. Applications for Physical Exercise Classification

GymCam proposed by Khurana [3] is a fixed installation of a camera system to leverage computer vision techniques on the task of sport exercise recognition in a public gym with multiple users. The system is able to unobtrusively and simultaneously recognize, track and count fitness exercises performed by multiple persons. The processing pipeline includes, image segmentation, exercise recognition and user tracking based on raw RGB images. By leveraging the motion information extracted from the optical flow method, they were able to achieve a segmentation accuracy of 84.6% and around 93.6% accuracy in recognizing the type of exercises. Despite its high accuracy, the fixed installation and the non-portability limit its field of application and makes it not perfectly suitable for quantified-self applications. Furthermore, even in public domains, using a camera-based system may often raise privacy concerns due to the visual inputs.

Fu [4] proposed a yoga mat embedded with eight capacitive proximity sensors to recognize eight workout exercises partially similar to that proposed by Sundholme. The difference opposed to a pressure-based sensing technique is that proximity sensors allow measurement up to 15 cm in the vicinity rather than applying direct force to it. By leveraging end-to-end training with convolutional neural networks, they achieved a user independent recognition accuracy of 93.5% and a user dependent recognition accuracy of 95.1% based on a test study with nine participants each performing two full sessions. A correlation-based matching method is used to count repetitions resulting in a user independent counting accuracy of 93.6%.

Most of the popular applications for quantified-self task are either wearable devices, which requires the user to wear it on the body, or needs external hardware setups, such as camera installations or smart textiles. Fixed camera installations in a public gym may further induce privacy concerns. Flexible textiles are indeed mobile, but adopted the drawback of easy deformation. This often leads to drops in the recognition accuracy. These challenges are constraining the concept of performing exercises anywhere at any time. One solution is to use a commercial smartphone in a stationary setup, as people are already carrying a smartphone everywhere. Use-cases with a modern smartphone is diverse.

2.2. Applications with Commercial Smartphones in the Context of Human Activity Recognition

Using smartphones to recognize human activities is nothing new, in this Subsection, the most prominent research works using a smartphone for HAR is introduced. Nandakumar [5] showed that it is possible to use the smartphone to measure the respiration rate, based on the chest movement. By measuring the distance profile of this periodical movement from the chest, they were able to detect the breathing cycle and a medical issue caused by irregular breathing cycle, called sleep apnea.

A smartphone emitting an active sonar signal can be further leveraged to detect mid-air hand gestures. *Dolphin* [6] was a project to detect fine-grained hand gestures close to the smartphone. No additional hardware is required besides the integrated microphone and loudspeaker. A continuous sound signal of 21 kHz is emitted by the integrated loudspeaker of the smartphone. Echo signal reflected from the motion executed in the vicinity of the device is used to extract Doppler motion information. They were able to recognize a set of predefined hand gestures in various environments despite different surrounding noise.

Though there exists related works on acoustic sensing based gesture recognition and physiological signal measuring, to the best of our knowledge, there exists no application leveraging this sensing method for these targeted sets of whole-body activity recognition. In our previous work [7], we have shown the possibility of using the smartphone to classify three vastly different exercises. In this paper, we extended it to a set of more complex activities, such as proposed by Sundholm [8],

but based on a fully different sensing principle by only leveraging the integrated sensors from a commercial smartphone.

3. Sensing Theory

The sensing method is based on Doppler sensing. We used the smartphone, (Samsung Galaxy A6 2018) and turned the device into an active sonar system by emitting a continuous sound wave with a carrier frequency of 20 kHz. The integrated loudspeaker is used to emit the signal, while the integrated microphone is used to receive the echo signals. This operating frequency is chosen above the audible hearing and is according to the physical definition of the lower frequency bound of the ultrasonic sound wave. With an audio sampling frequency of 44.1 kHz, we are able to reconstruct echo signals with an upper limit of 22.05 kHz. This makes a Doppler frequency range of up to 2.05 kHz possible.

A discretely sampled input wave file received by the device internal microphone represents the time series encoded with the repetitive motion patterns from the workout exercises performed in the vicinity of the sensing device. For each time series, a frequency time spectrum is calculated to reveal the Doppler profile over time. The resolution in time and frequency can help the classifier to better model the data. Since the motion speed for workout exercises are fairly small compared to hand gestures, the corresponding Doppler shift in frequency is thus minor. In order to have a high frequency resolution, a large observation time window is therefore required. This leads to a coarse time resolution and a large response time of our application, which makes it difficult to build a system with nearly real-time feedback.

We used the zero padding approach to resolve this issue. We increase the frequency resolution while keeping the time resolution as dense as before. This trick enables us to have both a fine-grained frequency and time resolution. This smooth and fine resolution in the frequency domain allows us to better detect the Doppler shift caused by relatively slow body motions.

An overview of some technical details is provided in Table 1. The audio sample frequency of the smartphone is 44.1 kHz. For each 4096 time samples, a fast Fourier transformation is calculated. With an overlap of 50%, we achieve a time resolution of 46.5 ms. We use the zero padding for the entire time window to have 12,288 values, which corresponds to a frequency resolution of 3.6 Hz of each frequency bin. This results in a relative speed resolution of 3 cm/s, thus enables us to have a fine resolution even for a slow motion speed.

Table 1. The table shows the hardware and software parameters of our proposed system.

Term	Meaning	Values
f_s	sample frequency	44.1 kHz
Δt	time resolution	46.5 ms
f_0	carrier frequency	20 kHz
v_0	speed of sound wave	340 m/s
N_{FFT}	number of FFT points	12,288
Δf	frequency resolution	3.6 Hz
Δv	speed resolution	3 cm/s

4. Data Processing

In this section, we will introduce a series of processing steps to prepare our data for the classification networks. For the data acquisition task, we developed an android application to get the exercise data and its corresponding labels. The processing pipeline is illustrated in Figure 2.

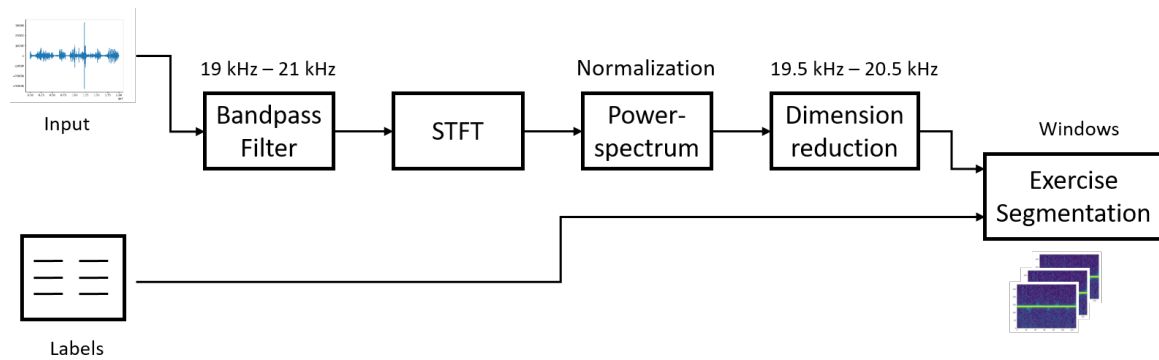


Figure 2. The processing pipeline is depicted here starting from the raw audio input to the segmentation step.

A Butterworth bandpass filter with the sixth order is applied on the raw input signal to filter out natural speech and only focusing on the frequency ranges close to the center frequency. Then the short time Fourier transformation (STFT) is applied on the filtered time signal to convert the 1D time series to 2D frequency over the time signal. The Hanning window of $n = 4096$ samples and zero padding are applied to the segmented time windows to reduce the spectra leakage of the fast Fourier transformation. The output of the STFT contains the magnitude and phase information. Here, we only use the magnitude information to construct our Doppler profile.

In each exercise, there are 10 repetitions included, while the swim class contains data of 25–30 s duration each. Dimension reduction is to limit the frequency bins from 19.5 to 20.5 kHz. In order to reduce the computation cost, we focus on the reduced spectrum bands containing the region of interest. Here the maximum Doppler of 500 Hz corresponds to a maximum speed of ± 4.25 m/s. The power spectrum is normalized to the median power by applying Equation (1).

$$S_{STFT} = 10 \cdot \log_{10}(|X_{STFT}|^2) - 10 \cdot \log_{10}(\text{median}(|X_{STFT}|^2)) \quad (1)$$

The segmentation part is the central part of the entire pipeline. The time window is set to 6 s and with an overlap of 50% for the sliding window approach. This parameter is set according to the offline processing with respect to system performance. To reduce the computation cost, only segment containing activity is used in the training process.

5. Classification Methods

The input training samples are the segmented spectrograms with a dimension of 279×129 , where 279 samples correspond to the frequency bins (from 19.5 to 20.5 kHz) and 129 samples represent the 6 s time steps. In Figure 3 a sample spectrogram of each workout exercise is depicted. These 2D spectra construct the base signal to the classifier models.

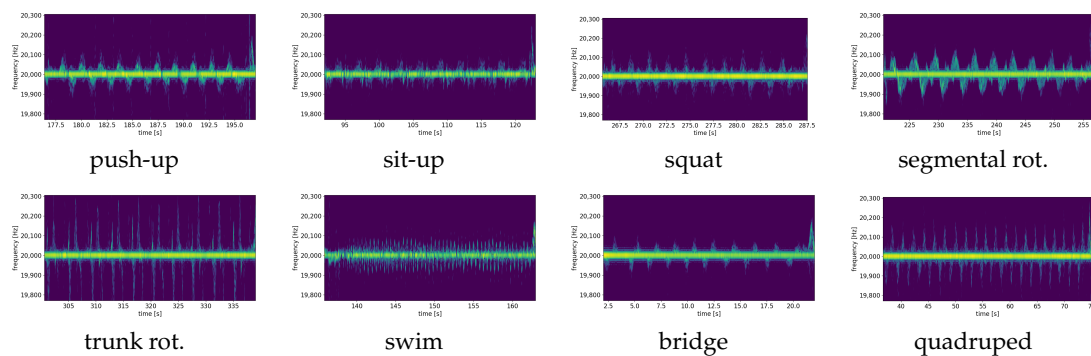


Figure 3. Figure depicts the spectrogram of the sample activities. Each exercise with 10 repetitions can be easily observed. The Doppler profile is distinctive around the center frequency.

Without including much of domain knowledge a priori, we evaluated our data on two end-to-end neural network architectures. The first architecture is the finetune model with VGG16 [9] weights for the base feature extraction layer. The second inference model is a sequence model called bidirectional LSTM. In the following section, we will introduce the model architectures and the hyperparameters used in the individual architectures. The hyperparameters were finetuned using 5-fold cross-validation. The models were built using the Pytorch [10] framework and trained on a GeForce RTX 2080 module. The weights of the finetune model VGG16 were directly downloaded from the Pytorch model zoo. We further improved the generalization ability of the inference model by using a few-shot classification method on unseen test data.

5.1. M1: VGG16 Plus Global Average Pooling Layer

We aim to improve the recognition accuracy by applying the finetune model of the VGG16 network. The finetuning allows us to exploit the base knowledge extracted from the ImageNet [11] task. The finetuning allows us to use knowledge extracted from a large supervised pre-training task as a backbone to our specific task, especially for the automatic low-level feature generation task. The lower convolution layers are intended to automatically extract useful features for the task of object recognition in two dimensional images. We fixed the weights in the pre-trained lower feature extraction layers. The decision layer is replaced by a global average pooling (GAP) layer combined with a softmax layer to output the class probability of each exercise. The GAP layer was used to reduce the over-fitting problem, due to our limited amount of input training data. The hyper-parameters of a GAP layer is much smaller compared to the fully connected layer. Instead of using the $7 \times 7 \times 512$ features to the fully connected layer, we reduced the output to $1 \times 1 \times 256$ features, which then fully connected to the class outputs with a softmax layer. The network architecture is displayed in Figure 4 (M1).

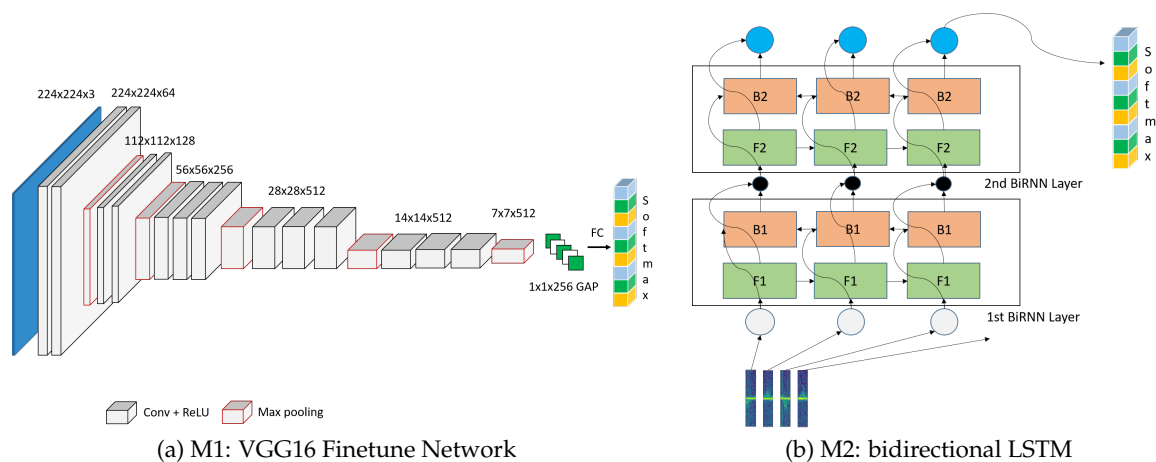


Figure 4. (a) It depicts the model architecture of the VGG16 finetune network with an additional global average pooling layer to reduce the model complexity and perform the classification. (b) Each long short term memory (LSTM) cell (B_i, F_i) contains 128 hidden nodes and two stacked layers are used to build the biLSTM network. For each input node, a slice of the frequency bands (ranging from 19.5 to 20.5 kHz) from a time step resolution (46.5 ms) is provided to the network.

An Adam optimizer with a learning rate of 0.003 was used to minimize the cost function. The objective was to minimize the weighted cross entropy loss and the l2-regularization on the model's parameters with a weight factor of 0.015. The weight parameters for the classes are based on the class distribution as well as the data sampler. Each sample has its own draw probability according to the class it belongs to. A batch size of 100 was chosen and we trained 100 epochs for the model to converge. We applied an instance normalization to the input layer to restrict the input samples to the same input range. This step can be considered as another regularization step to prevent the model

from over-fitting. The normalization of the input data helps the model to converge faster, in which case the input range is restricted between 0 and 1.

5.2. M2: Bidirectional LSTM Architecture

The long short term memory (LSTM) model is mostly used for sequence modeling or sequence tagging [12], such as natural language modeling. Recently it has been adapted to work for image classification tasks as well. The network architecture of our proposed model is depicted in Figure 4 (M2). The architecture of the bidirectional structure is rendering the network the ability to look into the future and past in order to better understand the whole context. This architecture should be able to cope better with the problem of inter-class similarity. The windowed sample spectrum was sliced to feed into the biLSTM network.

The input was instance normalized to convert the input range between 0 and 1. One important step to reduce over-fitting for the LSTM network is the dropout layer applied to the input before feeding to the LSTM layer. The ratio was set to 0.2 to avoid losing too much input information. This step prevents the LSTM network from simply memorizing our input data. A batch size of 100 is chosen to be trained for 100 epochs. An Adam optimizer with a learning rate of 0.003 was selected to minimize the cost function. The network consists of two LSTM layers each with 128 hidden nodes. The output of the bidirectional LSTM was directly fed to a fully connected layer with the class probability as output. Gradient clipping is also applied to reduce the inherent problem of exploding or vanishing gradient from LSTM networks. The objective was to minimize the weighted cross entropy loss based on the underlying class distribution and the weighted l2 regularization for the model parameters. A data sampler was used to draw the batches and also corresponds to the underlying sample distribution for each class. The cross entropy loss was weighted according to the class distribution.

5.3. M3: Siamese Few-Shot Learning

As stated before, the human activity data from the sensory output are difficult to acquire in comparison to vision-based data. To overcome the problem of the small data amount and to increase the model generalization ability, few-shot learning classification is leveraged. Based on knowledge extracted on a few samples named as support samples, the network is able to generalize on similar unseen samples without retraining the inference model. This is possible under the assumption that similar samples have similar embeddings located closer together. Here, we propose a modified Siamese network architecture to perform this multiclass classification task.

Commonly, the Siamese network is used for comparing the similarity between two sample inputs. The objective is to close up the distance of similar object pairs and enlarge the distance of dissimilar object pairs. Here, we modified the infrastructure of the network to simultaneously working on all pairs of the query sample to all different multiclass samples at once. The training objective of the modified Siamese network is thus to close up the distance of the query sample towards the correct support sample class.

The Siamese network consists of two identical feature extraction base networks with shared weight parameters. We learn in general the distance between the query input against all other support samples from different classes. The class category with the closest mean distance metric towards the unknown sample is selected to be the correct class. The network architecture is illustrated in Figure 5 M3. The designed structure aims at learning the optimum separation between all multiple classes at once.

The internal structure of the ConvNet is constructed of three stacked convolutional layers with pooling layers to reduce the input dimension. The main task of this ConvNet is to construct feature embeddings from input images. A similarity metric using euclidean distance is used to classify the unknown target sample to the known support samples of different classes. The optimization is based on minimizing the cross entropy loss of the classification task.

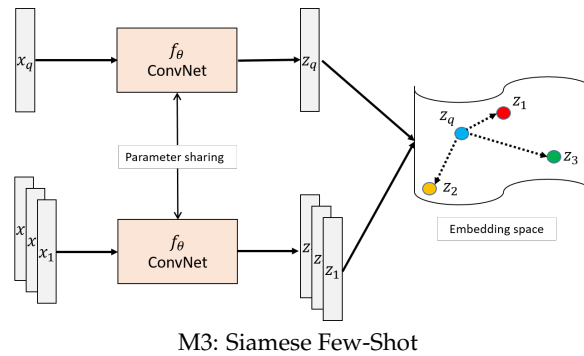


Figure 5. The Siamese network is modified to adapt to the few-shot classification task. The structure of shared parameters for the ConvNet embeddings aims at learning the optimum separability for the multiclass classification problem by using supports. The output label corresponds to the highest similarity score towards the correct class category. Here a learned distance metric is applied to determine the similarity score.

During the training phase, each batch consists of 9 classes and 15 query samples in each class. That makes in total 135 samples per batch training. The support set consists of 5 support samples each class and makes in total 45 support samples. An Adam optimizer is used to train the Siamese network parameter with a learning rate of 0.0005 with 100 epochs.

6. Evaluation

To evaluate our proposed system, we conducted a test study collecting exercise data from 14 individuals in our living laboratory. The group consists of 4 females and 10 males with an average height of 165.5 cm for the female group and 182.3 cm for the male group. Some general statistics about the population distribution are provided in Table 2.

Table 2. The statistics of the test population are provided in the table.

Males	Age	Height	Females	Age	Height
Number	10	-	Number	4	-
Min	21	172	Min	21	157
Max	33	193	Max	32	172
Average	25.4	182.3	Average	24.75	165.5

To acquire the user data, we placed a yoga mat on the carpet under a constrained environment and placed our sensing device 50 cm away, aligned to the exercising body part. The microphone of the sensing device was directly facing the exercising participant. The smartphone was aligned with the hip, except for the *swim* and *trunk rotation*, where we aligned the device with the shoulder to better catch the micro Doppler motion from the waving arm movement. For each individual, we collected two full sessions in two successive recording sessions. Each exercise was performed ten times each in every session, except the class *swim*, which was collected for around 25 to 30 s, in order to acquire enough data samples comparable to other exercise types. The participants were asked to label their data by using our recording app on the mobile device, as a way to pose less intervention on the natural action.

In the segmentation stage, we used the user-defined labels to cut periods of exercises. We further discarded the first and last 2 s of each exercise at the beginning and end to remove the handling of the labeling process. A sliding window of 6 s is applied to cut the spectrum for each exercise class with an overlap of 50% for the data augmentation purpose. In addition, we carefully designed network regularization schemes to avoid over-fitting. For each sample time window, we determined the upper and lower Doppler broadening profiles and kept only the windows with large variance in the Doppler envelope indicating the presence of true activities.

We conducted two sets of evaluation to investigate the robustness and the generalization ability of our proposed application design. Our first evaluation was conducted on the cross-subject performance. Thereby, we split the entire dataset into 70% training and 30% test by using a stratified splitting mechanism to maintain the same distribution of the underlying class in both splits. For each training set, a 5-fold cross-validation approach was used to finetune the classification models. In the second evaluation phase, we intended to measure the generalization ability of our classification models on disjoint participants. For this purpose, we keep out all sessions of four randomly selected individuals as the holdout set to be used in the test split, while the remaining 10 individuals were used as training data. Again, 5-fold cross-validation was applied to finetune the inference models.

The weighted F1-score was used as the evaluation metric. It is a better measure balancing the precision or recall scores, especially in the face of unbalanced class distributions. This measure provides a harmonic mean of precision and recall, compensating for the precision favors the majority and recall favors the minority class.

6.1. Cross Individual Classification

As described in the previous section, a stratified 70%:30% split is applied on the entire dataset. The same split for the training and test dataset, as well as the 5-fold cross-validation was used on the VGG16 Finetune and biLSTM models to maintain comparability across different inference models.

The weighted F1 score for the 5-fold cross-validation is provided in Table 3. The F1 score is a balanced score between precision and recall and indicates the performance of the inference model. The variance across the 5-fold cross-validation indicates the stability of the inference model against noise in the data distribution.

Table 3. For the cross subjects case, it depicts the F1 score for the 5-fold cross-validation results on the two inference models. Bold marks the fold with the best F1 score.

Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	σ^2
M1: VGG16	81.62%	88.70 %	85.16%	83.70%	82.21%	2.53
M2: biLSTM	86.05%	86.27%	81.83%	80.88%	88.86 %	2.98

Sequence modeling (M2:biLSTM) performs even better than VGG16 finetune for our given task. The confusion matrix with the highest F1 score is shown for biLSTM in Table 4 and the one for the VGG16 finetuned model in Table 5. Derived from the results of the confusion matrix, we see the challenging classes, which the individual model has the most difficulty distinguishing.

In the case of the biLSTM, the variance across the eight workout exercises is slightly smaller compared to the VGG16 model, which has minor problems for interpreting similar classes. In case of the class *sit-up* in the VGG16 model, a strong misclassification tends towards the class *push-up*, and *bridge*. Both classes have a false positive rate of 10%. This is explainable due to the similar upper body movement. Those exercises are ground-bounded while the user is lying on the ground and the sensing device is placed on the same position. Thus, the main reflections in the signal are from the same upper body part. The pooling layer from the convolutional network architecture disregards the sequence information in favor of a larger field of view and thus makes these similar exercises hard to distinguish.

The sequence model biLSTM has more problems for the class *swim* and *squat*. The class *swim* tends to be confused with the class *sit-up*. The class *swim* includes very small and faster arm movements, which leads to smeared spectral patterns causing misclassification. VGG16 Finetune outperforms the biLSTM model in this case by about 18 percentage points on accuracy, due to its ability to observe the local and the global connected features at once. The class *squat* performs comparably worse in both models, as the distance of the performing body part is quite distant from the sensing device.

Table 4. Confusion matrix for biLSTM in the case of cross subjects training is depicted here. Number corresponds to 1: *bridge*, 2: *idle*, 3: *push-up*, 4: *quadruped*, 5: *segment rotation*, 6: *sit-up*, 7: *squats*, 8: *swim*, and 9: *trunk rotation*.

		Predicted								
True	1	2	3	4	5	6	7	8	9	
1	0.96	0.04	0	0	0	0	0	0	0	
2	0	0.92	0	0.02	0	0	0	0.04	0.02	
3	0.09	0	0.83	0	0	0.04	0	0.04	0	
4	0	0.01	0	0.92	0	0	0.01	0	0.06	
5	0	0.02	0.01	0	0.91	0	0.06	0	0	
6	0	0.04	0.09	0	0.02	0.76	0.04	0.04	0	
7	0	0.06	0	0	0	0.22	0.72	0	0	
8	0	0.03	0.06	0	0	0.16	0	0.74	0	
9	0	0	0	0	0	0	0	0	1	

Table 5. Confusion matrix for VGG16 in the case of cross subjects training is shown here.

		Predicted								
True	1	2	3	4	5	6	7	8	9	
1	0.83	0.02	0	0	0	0.13	0	0.02	0	
2	0	0.90	0	0.07	0	0	0	0	0.02	
3	0.04	0	0.82	0	0	0.08	0	0.06	0	
4	0	0.06	0	0.94	0	0	0	0	0	
5	0	0.04	0	0	0.96	0	0	0	0	
6	0.1	0.12	0.1	0.02	0	0.60	0	0.07	0	
7	0	0.03	0	0.1	0.03	0.1	0.73	0	0	
8	0.08	0	0	0	0	0	0	0.92	0	
9	0	0	0	0	0	0	0	0	1	

6.2. Generalization Ability on Holdout Individuals

In this experiment, we intended to study the generalization ability of the trained inference models. For this purpose, we selected four individuals and used their entire sessions for the test dataset. The remaining disjoint 10 individuals were considered to build the training dataset. The same 5-fold cross-validation split was applied across all inference models to maintain comparability of the model performance. The class distribution in the training and test dataset are closely equal.

The evaluation result is provided in Table 6. The expected performance drop is observed in this specific setup. This performance drop is explainable by the diversity of the collected data. Since human activities, especially the targeted exercise classes, are highly complex and diverse, we need lots of diverse data to train a model that can cope with all possible situations. This is hardly possible.

Table 6. For the generalization test with holdout test participants, it depicts the F1 score for the 5-fold cross-validation results on the two inference models. Bold marks the fold with the best F1 score.

Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	σ^2
M1: VGG16	77.96%	75.98%	74.69%	71.31%	74.52%	2.17
M2: biLSTM	75.23%	76.96%	81.37%	76.91%	79.52%	2.17

The corresponding confusion matrices for the best performing VGG16 and biLSTM network are depicted in Tables 7 and 8. The biLSTM model caught the structural sequence information better than the VGG16 finetuned model, since the off-diagonal elements are slightly smaller compared to the VGG16 finetune model with the exception of certain classes, such as *sit-up* and *swim*. Several exercise classes are quite similar, such as *sit-up* and *bridge*, as can be seen in Figure 3. The VGG16 model with the pooling layer thus makes it harder for the network to distinguish both classes, while the sequence

model without modifying the sequence information with pooling layers is still able to distinguish both exercises. However, the overall performance for both models is quite similar compared to each other. The trend for the performance of each class reflects the results from the cross-subject method discussed in the previous subsection.

Table 7. Confusion matrix for VGG16 in the case of holdout subjects training is depicted here. Number corresponds to 1: bridge, 2: idle, 3: push-up, 4: quadruped, 5: segment rotation, 6: sit-up, 7: squats, 8: swim, and 9: trunk rotation.

True	Predicted								
	1	2	3	4	5	6	7	8	9
1	0.62	0	0.09	0	0	0.19	0	0.1	0
2	0.02	0.67	0	0.04	0.01	0.15	0.02	0.01	0.08
3	0.09	0	0.54	0.07	0	0.25	0	0.05	0
4	0	0.04	0	0.68	0	0.09	0.15	0	0.04
5	0.03	0.01	0.01	0	0.93	0	0.01	0	0
6	0.02	0	0.12	0.02	0	0.76	0	0.07	0
7	0	0	0	0	0	0.22	0.67	0.11	0
8	0	0.03	0	0	0	0.1	0	0.87	0
9	0	0.02	0	0	0	0	0	0	0.98

Table 8. Confusion matrix for biLSTM in the case of holdout subjects training is shown here.

True	Predicted								
	1	2	3	4	5	6	7	8	9
1	0.79	0.01	0.13	0	0	0	0	0.07	0
2	0	0.81	0	0.12	0	0.03	0	0.04	0
3	0.09	0	0.70	0	0.04	0	0	0.17	0
4	0	0.13	0	0.82	0	0	0.04	0.01	0
5	0.04	0	0.01	0.03	0.91	0.01	0	0	0
6	0	0.23	0.07	0	0	0.46	0.05	0.19	0
7	0	0.05	0	0	0.09	0.14	0.73	0	0
8	0	0.03	0	0	0.03	0.21	0	0.73	0
9	0	0	0	0.01	0	0	0	0	0.99

6.3. Results on Few-Shot Learning

We noticed that we can further increase the classification performance by leveraging few-shot classification learning method. We observed the performance drop in the case of individual holdout training is mainly due to the user diversity problem. The inference model trained with limited variations of user data does not extrapolate well on unseen test data. The problem leads back to high variance in human activities. As for such a high dimensional problem, the data we collected to train the data-driven model is thus quite limited. This leads conventional end-to-end model to have low bias, but high variance results. In order to resolve this issue, we examined the approach of using a few-shot classification method with a modified Siamese network.

In this work, we examined the model generalization ability by including a few unseen samples. With only five support samples from each class during classification, we can improve the final performance on unseen test data. In both individual experimental setups (cross-subject case and with holdout users), we observe an increase of more than 7–10 percentage points on average for both best working models proposed in this work (VGG16 and biLSTM). The results are listed in Table 9.

To conclude, the Siamese network for few-shot classification is well suited for the problem of improved inference on new test data based on knowledge extracted from a few support samples. This approach allows us to increase the generalization ability of the network without the need to

retrain the inference network. Taking benefit from a few support samples to perform the classification task further reduced the need of training with large amount of training samples.

Table 9. Classification accuracy for both setups are listed here. For the evaluation, 5 support samples from each class are used. Compared to other proposed classifiers, we observed an increase of at least 7–10 percentage points. The reason is because, by including knowledge from a few known samples, the few-shot classification task is especially suited for learning with limited data amount.

Method	Setting	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	σ^2
M3: Siamese	cross subjects	93.50%	94.11%	93.77%	93.62%	93.87%	0.21
M3: Siamese	holdout	85.70%	86.04%	85.41%	85.50%	85.16%	0.30

7. Repetition Counting

Exercise counting can be viewed as a parallel task besides classification. In order to build a useful user exercise profile, the counts of each performed exercises is considered to be a useful statistical value. Here a short conceptual view of the counting method is proposed. For counting the exercises, we extract envelopes from both the positive and negative Doppler profiles from the pre-processed spectrogram. The envelope stretches from the middle frequency component and broadens to both directions as the amplitude falls below a minimum threshold. The envelope signal is further smoothed with a Gaussian kernel of size three to suppress noisy signals. A peak detection algorithm is applied on the Doppler envelop with finetuned minimum peak distance for suppressing multiple successive peak detection and to ensure clear separable peaks from the exercises.

For exercises with left and right variations, such as *quadruped* and *trunk rotation*, the negative envelope of the Doppler profile performs better compared to positive envelope. Due to micro-Doppler motion from the arms and legs, multiple peaks are detected for one repetition. As for *quadruped*, *segment rotation* and *trunk rotation*, a high peak followed by a lower peak is resolved as one repetition, since the high peak indicates the main reflection, while the lower peak represents the remote Doppler movement from the opposite body part. As for clear defined repetitions such as for *bridge*, *push-up*, *sit-up* and *squat*, both the positive and negative envelopes can be used to count repetitions. In Figure 6, a conceptual view of the exercise counting can be viewed with the asterisk indicating the detected peaks of each repetition.

In Table 10, the mean error count and the standard deviation compared to the reported ground truth count for the given 14 test participants are provided for each exercise. In Figure 7, the mean counting errors in relation to the reported true count are depicted for each test participant individually (marked with a cross) and the mean error across all test participants is marked with a diamond symbol.

Table 10. Mean error count compared to the reported ground truth count is given for the 14 test participants.

Error Measure	Bridge	Push-up	Quadruped	Seg Rot	Sit-up	Squat	Trunk Rot
mean overall	1.04	1.29	1.18	2.64	2.86	1.79	1.45
standard deviation	1.15	1.41	0.89	1.63	1.41	2.01	1.18

To summarize the performance of counting, floor-based exercises without left and right variations perform the best, due to the main Doppler reflection from the upper body part. This can be observed for the exercise classes *Bridge*, *Push-ups* and *Squats*. Exercises including left and right variations, such as *Quadruped* and *Segment Rotation*, sometimes have issues with the micro-Doppler motion causing multiple false peaks and should be further improved to increase the overall performance of counting.

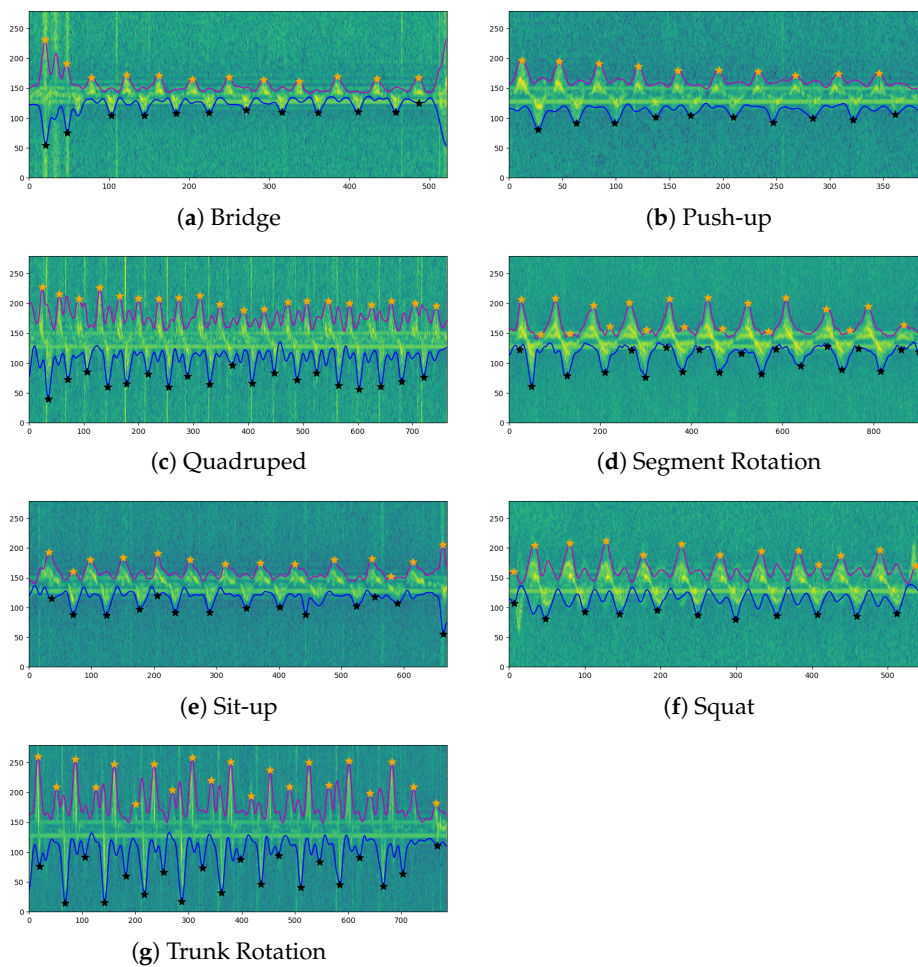


Figure 6. For each exercise, the positive and negative Doppler profiles are depicted. A peak detection algorithm is applied to detect the number of repetitions. There are 10 reported repetitions in each exercise. For exercises with left and right variation such as *quadruped*, *segment rotation* and *trunk rotation*, the negative Doppler profile depicts a more repetitive pattern with higher maxima followed by lower maxima caused by micro-movement from the limbs and arms.

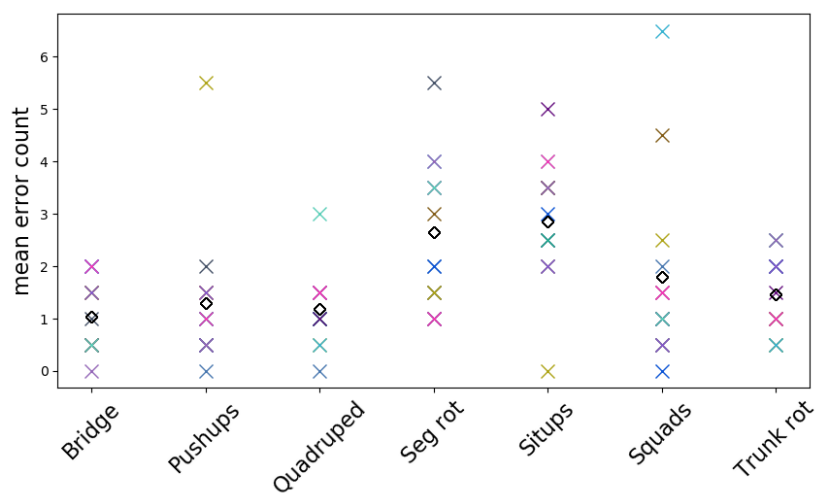


Figure 7. For each exercise, the mean counting error for each test participant is depicted with a cross, where each color represents the ID of the individual test participant and the mean error count over all test participants is depicted with a diamond symbol.

8. Comparison with Other Works

In this section, we compare our developed application to other existing applications targeting strength-based exercise recognition. We distinguish between customized design and modifying available infrastructure as proposed in this work. Also, we identify the merit of using multiple sensor fusion in comparison to using only one single sensor to accomplish this recognition task.

8.1. Comparing to Single Sensor Application

Using a single sensor component, such as using an acceleration sensor in wearable devices, is most common for activity monitoring. Tracking outdoor exercises like running, walking or cycling in combination with GPS location information are widespread in commercial fitness trackers. High precision and accurate position tracking can help athletes to create detailed user profiles.

On the opposite, stationary and strength-based exercises place strong restrictions on the sensor placement in order to achieve high performance. Exercises with stable wrist movement, such as *bridge* and *segment rotation* make it hard to extract distinctive features from acceleration data placed on the wrist. This fact poses quite a challenge to the current smartwatch solution to track strength-based exercises.

A camera-based solution provides a remote approach to target this issue. Using remote camera installation, the user is not restricted while performing exercises. However, it raises the issue of privacy. In the public or private domain, visual input with video streams could reveal much more information than the target exercises. Furthermore, the computation with video streams requires more processing effort compared to time series. Therefore, other remote solutions based on time series data are more desirable.

8.2. Comparing to Multiple Sensor Fusion

To overcome the above-mentioned restrictions of dormant wrist movement of certain exercises, a remote stationary sensing method is more preferable to enable a direct interaction between user and the sensing system. Several applications are addressing this issue by using multiple sensor fusions. Sundholm et al. [8] used piezoelectric elastic fiber sheets of the dimension 80×80 cm to measure forces applied from body movement. Instead of acceleration data, they applied a mesh of stripes with conductive foils to measure the force distribution from multiple pressure resistive sensors. However, they had similar issues for the class *squat* as in our proposed application due to the same reason of increased sensing distance. The confusion matrix for the person dependent use-case can be seen in Table 11 cited from [8]. One major drawback of their approach is the deformation of the flexible sensing surface after stretching and folding, which leads to a strong performance drop and thus the degradation of usability.

Other methods using capacitive proximity sensing improved the performance of certain exercises by further embedding the ability of proximity sensing instead of enforcing direct pressure. Fu et al. [4] equipped a commercial fitness mat with eight capacitive proximity sensors to recognize the same set of exercises as in our proposed use-case. Such an application includes sensor contexts from various locations across the sensing mat and has therefore a larger coverage. This leads to a more superior detection performance as compared with only using one single sensor. The accuracy for the person independent use-case can be seen in Table 11.

However, those applications require the design of external hardware and prototype. The advantage of using our proposed method is thus to avoid designing and setup external hardware. With the price of reduced accuracy, we benefit from a broad application area and ease of use, such as no need to carry additional prototypes. Using finetuning and model adaptation techniques, we can further increase the performance of our proposed application as shown for the few-shot classification case. This work thus intends to provide another view of approaching the problem of strength-based activity recognition.

Table 11. Accuracy from person independent evaluation is depicted here for three different applications. Value range lies between 0 and 1. Our proposed method performs well for exercises with left and right variation, such as *Seg Rot* and *Trunk Rot*. However, it shows reduced accuracy for exercises having similar appearance, such as *Bridge*, *Push-up* and *Sit-up*. Similar worse performance can be seen on the exercise *Squat* due to large distance and low contact area.

Other Applications	Bridge	Push-up	Quadruped	Seg Rot	Sit-up	Squat	Swim	Trunk Rot
Exertrack [4]	0.98	0.92	0.99	0.99	0.94	0.9	0.93	0.97
Sundholm [8]	0.76	0.80	0.92	0.77	-	0.63	-	-
Our method	0.79	0.70	0.82	0.91	0.46	0.73	0.73	0.99

9. Discussion

Since quantified-self can lead to a healthy life style, we propose a novel application of using a commercial smartphone to recognize eight whole-body workout exercises. Our application aims at mobile and remote sensing to enable practice everywhere at any time. By using the integrated hardware of the smartphone, we turn it into an active sonar device to measure the Doppler profile caused by moving body in the vicinity of the sensing device. Carefully designed processing and segmentation steps help us to work with even weak reflections. However, we identified several challenges during our evaluation phase, which can be improved in future applications.

Limited to the signal strength of ultrasound measurement with a commercial smartphone, the sensing distance is restricted below 50 cm. For a larger distance, the signal power is a trade-off to the power efficiency. In addition, the Doppler profiles for several classes are very similar. This is the problem of inter-class similarity. To cope with this problem, we identify the stacked bidirectional LSTM model to be more appropriate. The shape and rotation invariance introduced by the pooling layer of the convolutional model makes it sometimes even more difficult to distinguish between these classes, as can be seen in the off-diagonal elements from the confusion matrix in the evaluation Section 6.

Finally, the inter-person variability caused a relatively strong performance drop as observed in the holdout subjects for the testing case. This problem is inherent to the complex nature of human activities. Different people show different affinity towards physical exercises. People regularly perform workouts intuitively have a different signal shape than those who do not participate in sport on a regular basis. Careful design should be applied to resolve this challenging issue. Ensemble models can be used for different similar groups of users. We could first cluster different users into similar groups with its individual classification model. Then based on the ensemble learning, the final decision can be fused from the ensemble outputs.

Another way to address the problem of inter-person variability or complexity in data is to use few-shot classification learning tasks. By leveraging the modified Siamese few-shot classification, we improved the overall performance in both experimental setups by at least 8–9 percentage points on average for both best working models. Especially in the holdout experiment, by only including a small portion of the unknown samples, we achieved a large increase by 7–10 percentage points in the classification performance. Usually, we can not train conventional deep learning methods by applying such a small amount of samples. In this case, we benefit from the objective of the few-shot classification by increasing the generalization ability through knowledge extracted from support samples from different categories. By mapping objects from the same category closer together in the embedding space and measuring a metric distance, unknown objects from the same category can be easily determined in comparison to other categories.

10. Conclusions

In this paper, we showed the first results of using a commercial smartphone (in this paper, we used Samsung Galaxy A6 2018) to remotely detect eight more realistic and complex whole-body exercises. The integrated hardware is adequate to turn the most current commercial smartphone into an active

sonar device to measure remote body motion. We leverage the Doppler motion profiles, caused by human motion and especially the micro Doppler motions caused by the limb movement to catch the delicate features across the eight exercises. The aim is to build a mobile application allowing the user to practice anywhere at any time without the need to carry any extra hardware setups or wearables.

In our previous work [2], we presented the evaluation results on various end-to-end classification methods targeting the recognition of strength-based exercises. We showed that sequence model, such as the bidirectional LSTM network is more suitable for this kind of problem. Convolutional neural networks with pooling layers increase the global ambiguity of similar exercises. In this extended work, we try to improve the inference on disjoint user data by applying a modified Siamese network for few-shot classification.

A few-shot classifier further improves the performance by more than 7–10 percentage points on average for both best working models, in both experimental setups by leveraging the metric information in the feature embedding space. Only with the knowledge extracted from a few support samples, the generalization on samples from similar classes is possible to achieve. We further proposed a way to count the repetition of the exercises in addition to the classification task. Combining both classification and counting, the final application can be deployed on a user's smartphone for strength-based exercise recognition tasks.

We further added a more thorough investigation on the presented application in comparison to other existing state-of-the-art applications. Based on this comparison, we showed the advantages and disadvantages of our proposed work compared to other solutions requiring additional hardware prototypes. Accepting a reduction in performance by applying this single sensor solution, we gain the ease of use by benefiting from the existing infrastructure.

Finally, a basic concept of repetition counting is proposed using a peak detection algorithm. The task counting is useful, in addition to the classification task, to help the user to build a useful exercise profile. The first evaluation results showed the usefulness of the proposed approach, but there is room for improvement, especially for exercises with left and right variations to reduced false peak detection caused by micro-Doppler.

Author Contributions: Conceptualization and methodology, software, hardware, evaluation and data acquisition, writing—original draft preparation and final version, B.F., writing—review and editing, F.K., and A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN convolutional neural network
LSTM long short term memory

References

1. Swan, M. The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data* **2013**, *1*, 85–99. [[CrossRef](#)]
2. Fu, B.; Kirchbuchner, F.; Kuijper, A. Unconstrained workout activity recognition on unmodified commercial off-the-shelf smartphones. In Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments, Corfu, Greece, 30 June–2 July 2020; pp. 1–10.
3. Khurana, R.; Ahuja, K.; Yu, Z.; Mankoff, J.; Harrison, C.; Goel, M. GymCam: Detecting, Recognizing and Tracking Simultaneous Exercises in Unconstrained Scenes. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *2*, 1–17. [[CrossRef](#)]
4. Fu, B.; Jarms, L.; Kirchbuchner, F.; Kuijper, A. ExerTrack—Towards Smart Surfaces to Track Exercises. *Technologies* **2020**, *8*, 17. [[CrossRef](#)]

5. Nandakumar, R.; Gollakota, S.; Watson, N. Contactless Sleep Apnea Detection on Smartphones. In Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services, Florence, Italy, 18–22 May 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 45–57. [[CrossRef](#)]
6. Yang, Q.; Tang, H.; Zhao, X.; Li, Y.; Zhang, S. Dolphin: Ultrasonic-Based Gesture Recognition on Smartphone Platform. In Proceedings of the 2014 IEEE 17th International Conference on Computational Science and Engineering (CSE), Chengdu, China, 19–21 December 2014; pp. 1461–1468. [[CrossRef](#)]
7. Fu, B.; Gangatharan, D.V.; Kuijper, A.; Kirchbuchner, F.; Braun, A. Exercise monitoring on consumer smart phones using ultrasonic sensing. In Proceedings of the 4th International Workshop on Sensor-Based Activity Recognition and Interaction, Rostock, Germany, 21–22 September 2017; pp. 1–6.
8. Sundholm, M.; Cheng, J.; Zhou, B.; Sethi, A.; Lukowicz, P. Smart-Mat: Recognizing and Counting Gym Exercises with Low-cost Resistive Pressure Sensing Matrix. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing—UbiComp’14 Adjunct, Seattle, WA, USA, 13–17 September 2014; Brush, A.J., Friday, A., Kientz, J., Scott, J., Song, J., Eds.; ACM Press: New York, NY, USA, 2014; pp. 373–382. [[CrossRef](#)]
9. Qassim, H.; Verma, A.; Feinzimer, D. Compressed residual-VGG16 CNN model for big data places image recognition. In Proceedings of the 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 8–10 January 2018; pp. 169–175.
10. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Vancouver, BC, Canada, 2019; pp. 8024–8035.
11. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–26 June 2009; pp. 248–255.
12. Graves, A.; Jaitly, N.; Mohamed, A.R. Hybrid speech recognition with deep bidirectional LSTM. In Proceedings of the 2013 IEEE workshop on automatic speech recognition and understanding, Olomouc, Czech Republic, 8–12 December 2013; pp. 273–278.

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).