Performance assessment and feedback of fitness exercises using smartphone sensors

Nino van Hooff
1685848
Juli 2013

Master Thesis
Human-Machine Communication
University of Groningen, The Netherlands

Primary supervisor:
Dr. Marco Wiering (Artificial Intelligence, University of Groningen)

Secondary supervisor:
Dr. Fokie Cnossen (Artificial Intelligence, University of Groningen)

university of groningen / faculty of mathematics and natural sciences / peperzaken

ABSTRACT

Where GPS-based apps are popular for tracking outdoor fitness activities, no automated solutions exist for strength training. We use an off-the-shelf Android smartphone to provide users with feedback on tempo, movement range, and number of repetitions. Accelerometer signals are averaged into an exercise profile during calibration, after which new data can be compared with the created profile. Because exercise profiles are created by the user, our solution is suitable for many free-weight exercises. We use dynamically set thresholds to recognize repetitions. This approach is computationally efficient, and information on tempo and movement extent is retained. Feedback is given through auditory, visual and haptic modalities. Results indicate that repetition counting performance is on-par with earlier research, where performance on exercises with a rotational movement (98% correct) is higher than on exercises with a linear movement (91% correct). Trainers graded participants who received feedback significantly higher than those who did not. When directly measuring tempo and movement extent, however, the effect of the given advice on participant performance was not significant. We conclude that our app may help people perform their exercises better and more safely, but that tempo and movement range are insufficient predictors for a correctly performed exercise.

ACKNOWLEDGMENTS

Quite a number of people helped me to make this project a success.

Firstly, I would like to thank Marco Wiering and Fokie Cnossen for their guidance. When using your own idea as a basis for a Master's project, many things can go wrong. I thank Marco for his insights on machine learning and the many spelling and grammar corrections. Fokie mainly supervised the interface part, and offered some helpful advice on the structure of this thesis. I thank them both for guarding the scientific relevance of this project.

I had a great few weeks at Peperzaken. Although I was able to use the wisdom of almost every employee at Peperzaken, I am especially indebted to Eltjo, Mark and Lennart. Without them, the interface of SenseFit would not have been the same. I really liked the informal atmosphere at the office and the trip to the CeBit expo.

The SenseFit app was tested in a real-world environment. Fit 4 Free Groningen generously allowed me to use their gym for my experiment. I would like to thank all the trainers for instructing participants and grading the participant's performance.

Because studying at home can get boring and distracting at times, I would like to thank my fellow graduate students for their support and lunch-time chats. Good luck with finishing your studies!

Finally, I would like to thank André Miede for the Classicthesis template I used. I must admit I have sinned and used bold face in a few places, but the result looks like it is ready to be printed. I will make sure you receive a postcard shortly ;-)

Thanks again to all of you!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

ODP	Overhead Dumbbell Press	6
BOR	Bent-over Row	32
pp	percentage point	48
PC	Peirce's Criterion	28
HMM	Hidden Markov Model	3
NBC	Naïve Bayes Classifier	6
DTW	Dynamic Time Warping	3
PCBA	Principal Components Breakdown Analysis	8
WIMP	Windows, Icons, Mouse, and Pointer	58
BMD	Bone Mineral Density	4

CONCEPTS

movement
The atom of an exercise. An exercise repetition is usually comprised of two movements: the forward and backward movement.

movement features
The three properties we store in an exercise profile for each movement: duration, start amplitude, and end amplitude.

rep(etition)
One completion of an exercise.

correlation score
A measure for how well a movement is performed. The correlation between one movement and its corresponding movement in an exercise profile.

set, series
A sequence of repetitions before a period of rest.

major axis
Position and movement sensors usually measure in three axes. The major axis is the axis which shows the most significant fluctuation of values caused by the exercise.

gravity effect
The influence of gravity on recorded accelerometer signals. When a device rotates with respect to the ground, gravity shifts from one axis to another. This effect can be exploited to detect movement boundaries.

sensor coordinate system
The frame of reference for most Android sensor values. The frame of reference is relative to the device, and rotates along with the device.

dynamic rule-based decision
The algorithm used for our prototype. New movements are compared to a prototype using thresholds which are derived from that same prototype.

performance score
A measure for repetition counting performance. The percentage of correctly counted repetitions.

Part I

INTRODUCTION

1 INTRODUCTION

Resistance training [1], when appropriately prescribed and supervised, has favorable effects on muscular strength, endurance, cardiovascular function, metabolism, and psychosocial well-being [45]. Due to the myriad benefits, major health organizations recommend resistance training regimens, in addition to endurance training, at a frequency of at least twice per week [28, 18]. Although resistance training is advised to healthy populations, particular high-risk groups can especially benefit from regular exercise. Risk factors such as coronary heart disease, diabetes, hypertension, and lower-back pain are positively affected by resistance training [62].

Proper exercise execution is important to achieve desired results and to avoid injury, however. Especially over-extension and improper loads are dangerous. Over-extension may cause strains and sprains [48]. It is also important to gradually increase the strain imposed on muscles, as over-exertion is another cause of injury [60, 34]. Long repetition durations are effective for building strength, while short repetition durations are more effective at training muscles for explosive action. It is thus important to perform exercises at a specific tempo, depending on the training goal [22].

GPS-enabled smartphone apps such as RunKeeper [51] can assist in tracking endurance training progress and performance. For fitness machines, some commercial solutions exist, such as Fitlinxx [21]. For free-weight exercises, apps exist to help with keeping record of performance gains and periodization, but they require manual data entry. We found no popular automated alternatives.

The current objective is to create an app which records the user's movements while performing free-weight exercises and gives qualitative feedback on performance in terms of tempo, movement range, and repetition count. Smartphones contain a multitude of movement and position sensors, of which the accelerometer is most popular in related work. It was also shown to be of sufficient accuracy and resolution for tracking sports activities [55]. We will use this sensor for our application, although others, such as the gyroscope, are also considered.

Since we use a smartphone without additional sensors, we have only one measuring point. We chose to attach our smartphone to the forearm, which is a stable basis with large movement range in most free-weight exercises. We use individual exercise profiles be-

[1] A form of physical exercise in which weights or body mass is used to structurally overload muscles or muscle groups. Resistance training is more commonly called weight training or strength training.

cause proper execution is dependent on the training goal of the user, and movement range may be limited by injury or disability. It should also provide users with the flexibility to record custom exercises.

Because we need to retain qualitative information such as tempo and movement range, we cannot use classification methods such as Hidden Markov Models (HMMs) [47]. Dynamic Time Warping (DTW) allows tagging of parts of the signal such as the start and end of a repetition [4]. However, DTW is resource intensive, which makes it unsuitable for our cause. The solution we chose uses thresholds for amplitude and duration, which are based on characteristics of the exercise profile.

Feedback is given by comparing the user's current movements against those stored in the active exercise profile. We do not limit ourselves to the touch screen for output. A ubiquitous computing solution such as the current one calls for the use of other feedback modalities. Haptic, auditory, and visual feedback modalities are compared in a user preference study. The optimal visual design is iteratively determined.

This thesis is structured in several parts. In the remainder of this part, we first elaborate on the benefits of physical fitness training. Then, earlier work is presented, as well as commercially available fitness solutions. The research questions are stated next, along with the way they will be tested.

The next part is about the machine learning involved in this Master's project. Chapter 2 describes the algorithms used to process raw signal data into exercise profiles and compare those profiles to incoming data. The performance of these algorithms is presented and discussed in Chapter 3. Part iii describes the process of developing user feedback for our system. A display study which compares different representations of tempo and movement extent is presented in Chapter 4. In Chapter 5, we finalize the display design and design non-visual feedback. In Part iv, we combine the conclusions from both the machine learning and interface parts to answer the research questions.

1.1 BENEFITS OF PHYSICAL FITNESS TRAINING

Free-weight training is a type of resistance training, which means that weights are used to counter the work performed by the muscles. In resistance training, muscles generate energy primarily by an anaerobic process called glycolysis. Anaerobic rather than aerobic – which uses oxygen as the main fuel – exercises are intense, short-burst movements. When the goal is to increase strength, resistance training is performed at one's maximum capacity. Muscle fibers are traumatized, to which the body reacts by increasing the amount and size of contractile proteins. Because the muscle capacity increases as a result of training, periodization is important. Periodization involves increasing the weights to keep overloading the muscle as it grows stronger and allowing for enough rest to recover from the trauma caused by exercise.

There is an abundant amount of scientific studies investigating the relation between physical fitness and health. Some compared muscle strength between sedentary and active populations, while other focused on treatment and prevention of various diseases or studied the effects on mental health.

BONE MINERAL DENSITY is the percentage of bone minerals present on a x-ray scan (grams/cm^2). Low Bone Mineral Density (BMD) is an indication of brittle bones, which increases the chance of fractures. BMD is maintained by exerting force upon the bone. In an otherwise sedentary lifestyle, exercise is required to provide the required pressure. The brief, high intensity pressure associated with resistance training appears to be more effective towards this end than lower intensity activities. Nelson et al. [38] studied 39 women aged 50-70 during a 1-year high intensity strength training programme. Their BMD increased by 1% for the femoral neck bone and 10% for the lumbar spine, while BMD decreased by −2.5% for femoral neck bone and −1.8% for lumbar spine in the controls.

BLOOD PRESSURE Earlier research [31, 26] shows a small but significant decrease of about 3 mm Hg in both systolic and diastolic blood pressure for people with slightly elevated blood pressure at the start of a resistance training experiment. Control groups did not show a significant effect. When hypertensive populations are used, results are mixed [27].

BODY FAT Although aerobic training is usually prescribed for the reduction of body fat, there are clear indications that anaerobic training provides additional benefits that will help patients to maintain a lower body fat percentage. Since aerobic exercise for the purpose of losing weight is accompanied by a decrease in caloric intake, metabolic rate decreases which makes it difficult to lose more weight and which increases the chance of regaining weight when one stops dieting. Anaerobic training in combination with no or slightly decreased caloric intake promotes metabolic rate and muscle growth, and a more promising long-term effect can be attained [62, 45]. Physical activity in general is associated with better control of body weight and fat loss. For those who engage in physical activity, body fat is more favorably distributed [46].

FUNCTIONAL ABILITY Resistance training has myriad and spectacular benefits for elderly people. Apart from a higher BMD, muscle strength increases of over 100% are possible with a 10 week exercise programme which results in practical improvements of functional ability such as a 12% increase in walking speed and 28% increase in stair climbing power [19].

TREATMENT AND PREVENTION Back pain is one of the most prevalent causes for health care claims in the US. Strengthening the lower-back muscles significantly reduces the chance of complications [62]. Mooney et al. [35] asked a population of miners to use a lower-back training machine. With just one set performed once a week, strength increased between 54 and 104% and health claim costs decreased from $14,430 to $380 per person per month. Physical inactivity has been widely associated with coronary heart disease [46]. For those who smoke or are hypertensic, physical activity is an effective treatment [41]. Resistance training improves mechanisms in glucose metabolism, which makes it a suitable treatment and prevention therapy for diabetes and heart disease [62].

FREQUENCY To maintain bone structure the American College of Sports Medicine recommends resistance training at a frequency of 2 or 3 times a week, especially in older adults [28]. Feigenbaum and Pollock [18] state that single set exercises provide much of the health benefits gained from multiple set schedules. They advice to train all major muscles twice a week with a single set of up to 15 repetitions. This results in 15-20 minute sessions. Intensity (the used weight) is most important for developing muscle strength while the total training volume (intensity × sets × repetitions) is most important for developing muscle mass and endurance [18].

Although we focus on strength training in this thesis, it must be noted that endurance training is also recommended because any kind of exercise contributes to maintenance of body weight and overall fitness [46].

1.2 PREVIOUS WORK

Because smartphones with motion sensors and enough computing power for real-time data processing have only been around for a couple of years, earlier research in this field is relatively scarce. The papers we found generally focus on signal processing and machine learning, while the user interface is not discussed. Consumer products which aim to aid users with their workout start to emerge, however, and are supported by big brands such as Nike and Apple. In many cases, interesting and original choices have been made for the user interface. Apart from the academic work, we will discuss a selection of the available products to show the diversity in possible interfaces for ubiquitous sporting devices.

1.2.1 *Academic research*

Chang et al. [8] conducted an exercise recognition and repetition counting experiment in a gym setting. They compared the performance of HMMs with a Naïve Bayes Classifier (NBC). NBCs predict the class of an item by combining the prior probability of a class, the probability of finding the item's feature values, and the probability of those values given the class in question by using Bayes' theorem [49]. The classes could be, for example, a well-performed exercise, a badly-performed exercise, or noise. Chang et al. provided a taxonomy of free-weight exercises which cover a full-body workout. We also use these exercises for our pilot, which should allow for easy comparison of results. To capture movement data, they used two measurement points. One accelerometer was attached to a workout glove, and another to a belt clip. The latter was used to detect posture. According to Chang et al., this was necessary to discern between Overhead Dumbbell Press (ODP) and bench press, although the accelerometer traces they provide show largely differing movement ranges. This makes us skeptical as to whether the belt clip is a necessary addition.

Chang et al. note that the majority of the energy in free-weight exercises can be found in 1 of the 3 axes, which they call the *major axis*. *major axis* When considerable energy is found in two axes, one of them is redundant. We found this to be true for our data as well, and will use this concept in the remainder of this thesis. Another signal characteristic they found in their data pertains to the difference between exercises with a rotational movement, such as the biceps exercise, and those with a linear movement, such as the bench press. For rotational exercises, gravity is a large component of the signal which shifts between axes during the exercise. This *gravity effect* will be further explained on page 20.

Results from their study indicate that HMMs and NBCs perform equally well, but that HMMs need a lot more training examples. Particularly, when training and testing on data acquired from the same user, the use of HMMs produced unacceptable results while NBCs performed better on a user-specific (95% recognition accuracy) than on a leave-one-out (85% recognition accuracy) protocol. Since we expect to use personal profiles, it might thus be unwise to use HMMs for our project.

Pernek et al. [44] use Dynamic Time Warping (DTW) to find occurrences of a pre-defined repetition in a continuous data stream of smartphone accelerometer data. DTW can be used to compare time series which are not temporally aligned to produce a mapping between

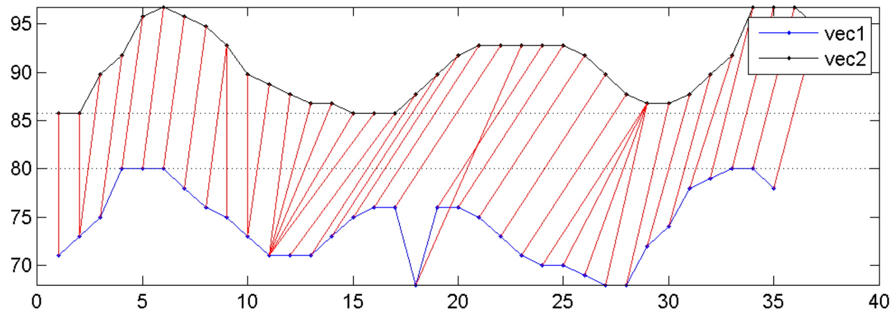


Figure 1.1: Dynamic Time Warping. Values in vec1 are matched to the most similar point in vec2 [2].

them in such a way that the distance between them is minimized [4]. Figure 1.1 illustrates the concept. The duration of a repetition can be determined by annotating start- and endpoints in the reference pattern. By comparing the points in a candidate pattern which are mapped to these start- and endpoints, the duration can be determined.

Unfortunately, performing DTW on a continuous data stream is resource intensive. To resolve this issue, likely candidates of exercise repetitions are selected first. Pernek et al. use a derivative-based peak detection algorithm to find the peaks that are within 1/3 of the reference's magnitude. Once a candidate is found, a part of the data stream is selected which is twice the length of the reference, and centered around the peak. DTW is then performed to extract features such as duration and normalized DTW distance. Repetition candidates are finally classified by a logistic regression model. Logistic regression models construct a linear formula of features with corresponding coefficients. This formula can be used to linearly separate classes. The classes in this case are 'repetitions' and 'noise'. To reduce computational demands further, only the major axis is considered in these calculations.

The algorithm was not only tested in a gym environment, but also outdoors. In both environments, repetition counting results were very promising, with a 1% miscount rate. The overall median error on duration estimation was 11%. On average, this error was lower for the unconstrained outdoor environment. The authors suggest that this may be due to the acceleration patterns being of higher intensity for the unconstrained environment.

Kranz et al. [29] propose a smartphone-based solution for assessing performance of balance board exercises called 'VMI Fit'. The used balance boards can tilt in one direction, which simplifies the tracking problem considerably. As data source, both the accelerometer and magnetometer (tri-axial compass) were used. Two approaches

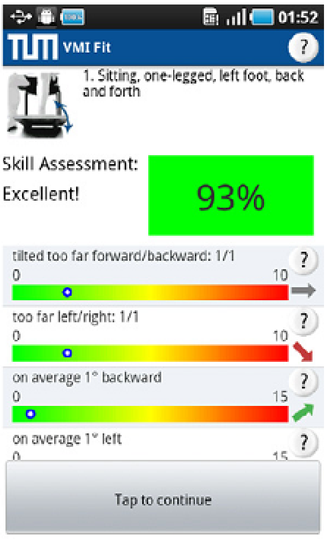
[2] http://mirlab.org/jang/books/dcpr/example/output/dtwBridgePlot02.png

for data processing were considered. The first uses Principal Components Breakdown Analysis (PCBA). PCBA is a technique for compressing a set of feature values in a smaller set of feature values, in such a way that the smaller set's expressiveness, or amount of explained variance, is maximized. Kranz et al. use a fixed target dimensionality. They assume that an exercise repetition with noise or deviations from a golden standard can be less accurately captured by this feature set than an exemplary repetition. As a result, the amount of explained variance by the reduced feature set is a measure of performance.

How well the algorithm performs is not reported, but they do note that an approach with more fine-grained assessment than just an overall similarity score is required.

Their second algorithm segments the data stream into movements on zero-crossings. From these movements, several features are extracted which are defined by the experimenters. These features are domain specific, such as whether the board touched the ground, but more general features such as pace and amount of repetitions are used as well. The performance scores given by VMI Fit are compared against expert assessments. When using magnetometer data, the assessment error was $< 20\%$ in 94% of the cases.

Using accelerometer data yielded an assessment error $< 20\%$ in 90% of cases. These figures pertain to exercises where participants were tasked to rock back and forth on the board. When the task was to keep balance, accelerometer and magnetometer data performed equally well. Assessment error was $< 15\%$ in about 98% of cases.



1. Sitting, one-legged, left foot, back and forth

Skill Assessment:
Excellent!
93%

tilted too far forward/backward: 1/1
0 10

too far left/right: 1/1
0 10

on average 1° backward
0 15

on average 1° left
0 15

Figure 1.2: VMI Fit interface

The feedback display is shown in Figure 1.2. An overall score is given in the form of a percentage, and feedback on individual aspects is given by placing a marker on a green-yellow-red gradient. This indicator is accompanied by textual information and an arrow indicating the direction of improvement. Participants (n = 6) reported individual exercise feedback as very important (5.0 on a 5-point Likert scale). Concerning usability, Kranz et al. further suggest to minimize interaction with the device, for example by recognizing which exercise the user is performing. This way, the user does not have to manually select it.

1.2.2 *Consumer products*

Kranz et al. [29] conducted a small comparative study of Android apps available at the time (2011). They found 3 categories of apps in their results: GPS trackers, workout planners and exercise books.

GPS-trackers annotate outdoor activity with quantitative information such as the route taken, distance traveled, time taken, average pace, and an estimation of burnt calories. Popular examples are Endomondo [16], Runtastic [53] and RunKeeper [51]. All three apps have similar features. The route can be visualized on a map, a history of earlier activities is kept, and it is possible to play music via earbuds. A motivational aspect is added by sharing workout summaries via social media and setting goals for the user. Runtastic allows reading heart rate information from a Bluetooth-connected device. Interfaces are mostly text based, showing various metrics. Some apps provide navigation, to allow one to follow a predefined route. Some of these apps allow voice feedback, which is mixed with background music.

Where the GPS trackers use the GPS information to keep track of progress, the workout planners and exercise book apps require manual data entry. These apps mainly facilitate periodization and recording progress. After a set is completed, users have to manually indicate this by pressing a button. These apps are used for strongly goal-directed workouts such as body building.

Finally, exercise books are references for beginning sporters, with instructional videos and tips on proper exercise execution and injury prevention.

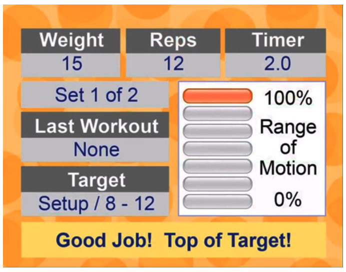
For sit ups and free-weight training, Kranz et al. did not find apps which actively monitor the user's progress.

Currently (2013), the offering of Android apps is largely unchanged. A remarkable exception is the 'pro' series of workout apps by Northpark [40], and the Runtastic workout apps [52]. In the squats pro app, users hold their phones with their arms extended in front of their chest while performing squats. The acceleration sensor is used to translate the up- and downward motion into a repetition count, which allows the user to concentrate on the exercise. Judging from user reviews on Google Play and personal experience this works very well, although even the slightest up- and downward motion is counted as a squat. Note that squats can only be performed without weights, since the user holds his phone in his hands. The approach taken by Runtastic is largely similar. Still, only repetitions are counted. Other exercise properties such as tempo and how 'deep' one squats are not measured.

The company Six To Start [57] uses an entirely different way to motivate users to work out. Their app immerses the user into a story of a zombie invasion. Whenever the user's tempo needs to increase according to the underlying interval training schedule, music is interrupted and they are told by a voice actor to 'run for their lives'. The app received a lot of praise for making running fun and providing an engaging experience.

Stand-alone devices for fitness tracking are gaining popularity. Nike+ Fuelband is a bracelet which is equipped with an accelerometer [39]. It displays a proprietary measure called Nike Fuel, which is to be a general measure for daily activity. The only user interface it has is a LED display which shows the number of Fuel points scored that day and progress towards the daily goal. The accompanying smartphone app does not provide additional info and is only used for setting the goal and interact with social media. The acceleration patterns are matched against a database, to guess the activity which would have caused the pattern. This, in turn, is translated into oxygen expenditure, which Nike considers a measure for exertion. By translating any physical activity into a number, goals can be set and friends can be challenged [17]. This seems a good example of a ubiquitous computing device. It does not require any interaction and does not distract users from other tasks. The performance of activity recognition is under debate, however. To some users it seems that points are awarded arbitrarily and one user even noted that eating a slice of pizza generated more points than climbing a flight of stairs [63].

Fitlinxx [21] is a system targeted at gyms. It is used to track tempo, range, and repetitions, as well as programme adherence. Each workout machine has a touch screen display mounted on it and sensors attached to the weight stack. Since the weights are restricted to travel in one direction, tracking is easy. The display is shown in Figure 1.3. When working out, the weight stack indicator moves from the top of the range scale to the bottom and back. When the user over-extends, a text warning is displayed at the bottom of the screen.


Weight 15
Reps 12
Timer 2.0
Set 1 of 2
Last Workout None
Target Setup / 8 - 12

Good Job! Top of Target!

Figure 1.3: Fitlinxx interface

The Pebble smartwatch is an interesting new product for ubiquitous interface designers [42]. It can connect to smartphones to show navigation information, incoming calls, texts, and other information provided by apps installed on the smartphone. Because it is equipped with an accelerometer, it is not necessary to interact with the smartphone directly during a workout to receive feedback. Via the small e-ink display, it could also give feedback on reps and sets performed.

1.3 CURRENT STUDY

Current products most closely related to the one we aim to create are the series of fitness apps available from NortphPark and Runtastic [40, 52]. The disadvantage of their approach is that, for each exercise, a separate app needs to be downloaded. This is a nuisance to users [40, 52], and limits the usefulness of these apps unnecessarily. With the exception of the push-up apps, NorthPark and Runtastic use the accelerometer to count repetitions for all exercises. Technically, the difference between these apps is a setting which determines which accelerometer axis to use for peak counting, or perhaps the signal thresholds. Instead of downloading a new app for each exercise, it should thus be sufficient to download an exercise profile. We feel that it is even possible to learn such a profile from individual user data.

In this study, we will create an app which can learn a new exercise from a calibration session. Characteristics like duration and movement range are stored in a profile. We regard movements as atoms. By segmenting a data stream into movements rather than repetitions, we can give specific feedback like 'perform the upward movement a little bit slower'. Feedback will be provided by on-line segmentation of a data stream into movements and comparing those movements against the movements stored in the exercise profile.

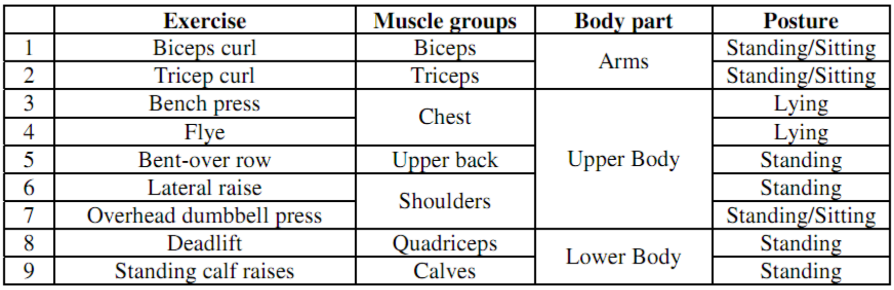
We will not only design the signal processing algorithm, but also the feedback users receive. The final design will be based on user preference studies, and will not be limited to visual feedback. Tactile and auditory modalities will be considered as well.

1.3.1 *Exercises*

To determine whether our app performs well in a realistic setting, a representative set of exercises is required. This set should not only cover the range of exercises which are commonly performed. It is also important to cover the diversity in signal characteristics that can be expected from an actual training session. Looking at the diversity of exercises performed in the gym, both exercises with a rotational movement and exercises with a linear movement must be included. Ideally, we also want to select a set of exercises which is used in previous research, so that the performance of our app can be related to the state of the art.

Chang et al. [8] constructed a 'Taxonomy of free-weight exercises', which is reproduced in Table 1.1 . They divided the body into the muscle groups Arms, Upper body and Lower body, identified the muscles which are most commonly trained during free-weight exercise, and selected one or two exercises which are used to strengthen those muscles. This appealed to us, since this approach ensures that a

Table 1.1: The taxonomy of free-weight exercises. These are the exercises used in the pilot phase. This taxonomy was created by Chang et al. [8] as a representative set of exercises that as a whole provide a full body workout.



	Exercise	Muscle groups	Body part	Posture
1	Biceps curl	Biceps	Arms	Standing/Sitting
2	Tricep curl	Triceps		Standing/Sitting
3	Bench press	Chest	Upper Body	Lying
4	Flye			Lying
5	Bent-over row	Upper back		Standing
6	Lateral raise	Shoulders		Standing
7	Overhead dumbbell press			Standing/Sitting
8	Deadlift	Quadriceps	Lower Body	Standing
9	Standing calf raises	Calves		Standing

full body workout is covered. Below, instructions on how to perform the exercises responsibly is given. We refer to Figure A.1 for the starting and ending positions of each exercise. The following terms are used to indicate the rotation of the arm.

- Neutral: No arm rotation. When held alongside the body, palms face inward toward the legs.
- Supination: When held alongside the body, palms face outward.
- Pronation: When held alongside the body, palms face forward.

biceps curl This exercise is performed while seated on a workout bench. Arms are stretched so that the hands point down towards the ground with the thumbs facing outward (supination pose). The upward movement is performed by lifting the weight while the elbow is fixated in the flank. When the arm points slightly upward, one slowly lowers the weight back to the starting position. The arm should not be completely extended, there should be a little tension on the muscle at all times. Apart from the biceps, the upper arm muscle is trained as well.

triceps hammer curl For our experiment, this exercise was performed while seated using only 1 arm at a time. The forearm is held horizontally behind the head in the starting position (thumbs toward the ground). Then, the arm is stretched so that the hand points upwards, after which the weight is brought back to the starting position.

bench press This exercise is performed while lying down on a bench. While this exercise is usually performed with a bar bell, we use free weights. We ask participants to use both hands, as training one hand at a time would make this a strenuous balancing exercise. The arms are held vertically above the chest, with the thumbs facing each other (pronation pose). Then, the forearm keeps pointing upward while the upper arm is rotated so that it is facing outwards horizontally from the shoulder, with the elbow in a 90° angle. This exercise trains the chest, delta, and triceps muscles.

flye This exercise is performed while lying on a bench. The arms are held upright over the chest, weights parallel with the torso (arms in neutral position). In contrast to the bench press, the elbow is bent only slightly while lowering the weights, and the whole arm ends up facing outward from the shoulder. The flye trains the chest muscles.

bent-over row This exercise is usually described as 'sawing a log of wood'. One leg and one hand are placed on the edge of a bench. The other leg rests on the ground beside the bench and the other hand holds the weight straight down with the arm in neutral position. The exercise is executed by moving the forearm up and down. The elbow should bend in this process, since this exercise targets the back muscles, not the shoulder. Care should be taken to maintain a straight or slightly hollow back throughout the exercise.

lateral raise Standing upright with the arms in neutral position alongside the body, a weight is lifted by rotating the upper arm so that the the arm points slightly upward and perpendicular to the direction one faces. This exercise primarily trains the middle deltoid.

overhead dumbbell press This exercise is performed while standing. The weights are raised over the shoulders so that the arms point straight up with the palms facing forward. The elbows are brought to shoulder level while keeping the forearms pointing straight up. Since keeping balance is involved, a large range of muscles is activated. The main muscles being the pectorals, deltoids and triceps.

deadlift This exercise is performed while standing upright. The starting position is the same as used for the lateral raise. The arm muscles are not used, however. The exercise is performed by successively squatting and standing up. One should bend over slightly and keep a straight back when squatting. The main muscles targeted are the quadriceps and hamstrings.

calf raise This exercise is performed while standing on a step with only the toes. The heel is lowered below the level of the toes and then raised up to maximal height. The muscles involved are located in the lower leg: gastronemius, tibialis posterior and soleus.

As instructed by many fitness coaches, all exercises are executed at a 1-2 pace. An exercise has a positive and a negative movement. The positive movement is the one wherein the weight is pushed or rotated away from the ground. The negative movement is the movement wherein the weight is lowered or rotated back to the ground in a controlled fashion. Surprisingly to most, the negative movement actually is most effective at training the muscle, which is why it is performed slower (lasting 2 seconds) than the positive movement (which lasts 1 second) [11] [3] .

1.3.2 *Research questions*

This project encompasses both machine learning and interface design. Therefore the main research question is rather broad:

MAIN RESEARCH QUESTION: How can sensor-equipped handheld devices facilitate correct execution of fitness exercises?

For both machine learning and interface design, we pose a sub-question. Each sub-question is treated in a separate part of this thesis.

SUB-QUESTION 1: How to use handheld device sensors to assess fitness exercise performance?

Topics addressed to answer this question:

1. Which sensors are most suitable?
2. Is user-specific calibration required to reliably assess performance?
3. How to use machine learning with only positive examples?

We hypothesize that:

1. The accelerometer is most suitable for free-weight and resistance training.
2. User specific calibration *is not* required for reliable exercise recognition using our algorithm.
   User specific calibration *is* required for reliable repetition counting and feedback on the user's performance when using our algorithm.
3. Our dynamic thresholding algorithm allows for reliable exercise recognition and repetition counting without the need for large amounts of training data and negative examples required by algorithms such as HMMs.

[3] The reason why the negative – or excentric – movement is more effective, has to do with the way muscles are strengthened. During the negative movement, muscles are stretched which is traumatizing for the fibers. The body reacts to this trauma by increasing muscle mass and strength.

The first topic is treated in Section 2.1.4, in which we compare different Android sensors based on accuracy, reliability, and power consumption. The second topic is separately considered for exercise recognition and repetition counting. In Section 2.7, we will look at how well our algorithm can select the currently performed exercise from a set of 8 free-weight exercises. In Section 3.2, repetition counting is discussed. In this section, performance is also compared to research which employed negative examples.

SUB-QUESTION 2: How should feedback about fitness exercise performance be designed?

Topics addressed to answer this question:

1. Which feedback modalities are available to an Android smartphone?
2. Which modalities are suitable for fitness environments?
3. Which modality do users prefer for the different pieces of feedback we want to provide?
4. How should the feedback be designed?
5. Can the advice given by the device effectuate a better exercise execution?

Because this sub-question is of an exploratory nature, we do not state hypotheses for topics 1-4. We hypothesize that users who receive feedback from our app will be able to perform their exercises in a way that is more consistent to a recorded profile than when they do not receive feedback. We also think that users who receive feedback get higher grades from both fitness professionals and our app than those who do not receive feedback.

The feedback modalities available to the device we use as our prototype are discussed in Section 2.1. To assess which of these modalities are preferred by users and are suitable for use in a fitness environment, results of the main usability study are discussed in Chapter 5. As a basis for the display design, the results from the display design study in Chapter 4 are used. The auditory and haptic feedback design will be described in section Section 5.3 and Section 5.4, respectively. In Section 5.5, we will see whether the advice given by our app improved exercise performance and in Section 5.6 we will see whether this advice agrees with the advice given by fitness professionals.

Part II

MACHINE LEARNING

In which we answer the research question

"How to use handheld device sensors to assess fitness exercise performance?"

2 SYSTEM & METHODS

The prototype of our mobile fitness coach was built using a smartphone as a starting point. It is attached to the forearm with a wristband to construct a quite complete testing platform. It has capabilities for data capture, user input, user output and data processing. This section describes considerations for the hardware and algorithms used to assess the user's exercise performance. The user interface will be discussed in a subsequent chapter.

We will first describe the hardware platform we chose for the sensor and data processing. Since it is rather novel that we use only one measurement point for tracking movement, we will give a short motivation for this choice. This chapter contains many figures that show time plots of sensor data. To interpret them, it is important to understand the coordinate system used by our app, which is explained next. Subsequently, we answer the research question of which sensors would be most suitable for tracking fitness exercise movements. Section 2.2 describes the processing pipeline from raw data to repetition count and performance scoring. Next, the elaborate experimental setup of the main experiment is described.

To determine the optimal parameters for our algorithm, a small pilot was conducted which focused on data processing. We asked volunteers to perform two consecutive sets of 10 repetitions for a single exercise which they had not done that day. Standing calf raise was left out of this study because of discouraging earlier results. The data recorded during this pilot will be used to describe the characteristic patterns for each exercise. Next, we will pick one data file and guide the reader through the processing stages. Finally, the best parameter results are discussed.

2.1 APPARATUS

Figure 2.1: The Samsung Galaxy S II smartphone used for this project.

The smartphone used is shown in Figure 2.1. It is a Galaxy S II manufactured by Samsung Electronics [54]. It is a highly successful model targeted at a wide audience. As such, it is a model that is representative of a device that might already be in the possession of our target audience.

The operating system is Android. The programming language used to write Android apps is essentially Java, with a superset which provides interfaces for the smartphone-specific hardware such as touch screen, sensors and camera. The advantage of Java for us is that the parts of the implementation written in pure Java can be executed on a desktop machine if necessary. We gratefully used this opportunity for determining the most optimal parameter values for our algorithms. The device weighs 116 grams and measures 125,3 x 66,1 x 8,5 mm. The low weight in combination with a width smaller than most people's forearm allows it to be worn comfortably around one's wrist.

2.1.1 *Sensor placement*

Where most other projects use at least 2 measurement points on the body [8, 2], our project uses only one. Care should be taken to determine what the most effective spot would be to attach the sensor. Since it would be inconvenient for the user to re-position the sensor for every exercise, we aim to pick one measurement location for the complete workout session. For all free-weight exercises under consideration, the forearm moves at least to some degree. Although the ankle might provide a more stable measurement point for the calf raise exercise, it was found that even at this point, the measured acceleration values were too small for reliable use. Since sensor placement on the back of the hand would allow us to record data more precisely (including wrist-rotation), we considered a glove, as used by Chang et al. [8]. Wrist rotation does not play a significant role in any of the exercises, however. As such we consider it a liability of rotation noise. Apart from the reasons mentioned above, the forearm was chosen as the location for sensor placement, because it povides a stable base for the rather long smartphone.

SENSOR	UNIT	POWER DRAW (mAh)
Accelerometer	m/s^2	0.23
Gyroscope	ω/s	6.10
Magnetic field	μT	6.80
Orientation	°	13.13

Table 2.1: Android sensor overview. The Orientation sensor is a 'virtual sensor', which incorporates data from the 3 other sensors to calculate orientation relative to magnetic North [23].

2.1.2 *Wristband*

At the moment, no wristbands are commercially available for smartphones. A solution was found by using an armband. This armband was originally intended for wearing a smartphone on the upper arm while jogging. By shortening the strap, the neoprene armband can be firmly fixed around the forearm. The elasticity of the neoprene ensures a tight fit while the wristband is still easy to put on.

2.1.3 *Sensor coordinate system*

Sensor coordinate system

The coordinate system used for most Android sensors is the Sensor coordinate system. It is defined relative to the device's frame of reference rather than to the world's frame of reference, see Figure 2.2. This means the signal is not influenced by the wind direction the user is facing. When worn as shown in Figure 2.2, the x axis points to the right, the y axis points to the hand and the z axis points toward the sky [23].

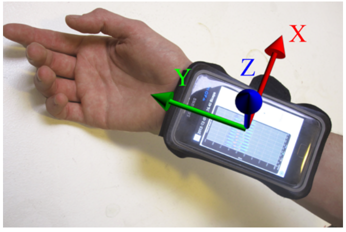


Figure 2.2: The device's coordinate system. The axes have a fixed orientation relative to the device.

2.1.4 *Sensors*

Table 2.1 shows an overview of the orientation and movement sensors available to the Android Framework. For our application, we could use any or multiple of these sensors as a data source. We will now discuss our considerations for using each of these sensors.

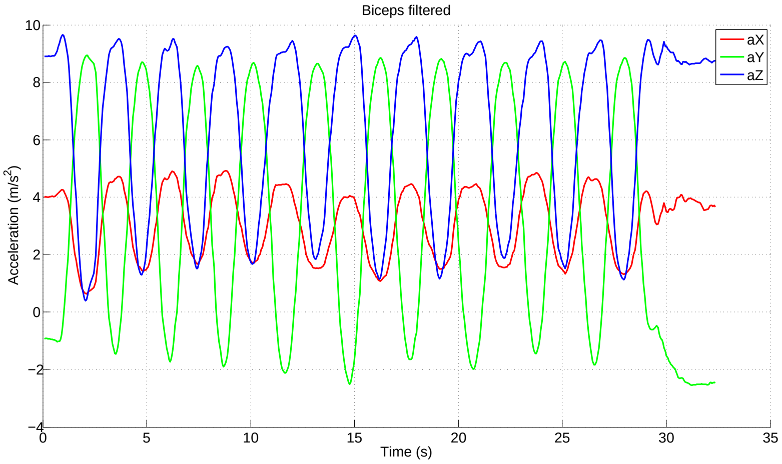


Figure 2.3: Filtered biceps data. 10 repetitions are shown. In the start position of this exercise, the y-axis (green) is parallel to the ground and reads approximately 0 m/s^2. In the end position, the y-axis is perpendicular to the ground and reads approximately 9.8 m/s^2. The inverse is true for the z-axis (blue). The x-axis (red), shows the left-right movement of the forearm and should be kept as steady as possible.

2.1.4.1 *Accelerometer*

The accelerometer measures the acceleration applied to the device, including the force of gravity in three directions. Therefore, the sum of all three acceleration signals will be 9.81 m/s^2 when the device is at rest. The difference of the actual sum from the gravitational constant is the linear acceleration. For our application, this would be the acceleration of the forearm, zero based at rest (being stationary). When using this sensor, one's first intuition might be to remove the gravity component. A zero-based signal would certainly be easier to interpret. This could be done by applying a high-pass filter [23]. A disadvantage would obviously be the inherent delay, but there is a more important disadvantage. The shift of gravity magnitude from one axis to another is actually a very reliable signal which can be used to monitor, for example, a biceps movement. At rest, when the forearm is held horizontally, the gravity is applied to the z axis. When the user completes the forward movement, the forearm is held vertically and the gravity now applies to the y axis, see Figure 2.3 and Figure 2.2. The transformation between the two positions results in a signal as shown in Figure 2.3. The signal is smooth and has a very large signal to noise ratio. Using this principle, elbow rotation can be measured. Different poses make different axes vertical to the ground. Chang et al. [8] call this the *gravity effect*, a term we will continue to use in the rest of this thesis.

Gravity effect

Although removing the gravity effect is unwise, we initially did perform a calibration step at the start of each exercise. During this calibration, in which the user is asked to keep the device steady at the exercise's resting position, a vector of $[x, y, z]$ resting values was averaged over a 1 second interval. This calibration vector was subsequently subtracted from all following samples to provide a signal which is zero at the resting position while retaining the gravity component. It was now easy to tell whether the user started or ended an exercise by checking whether the signal moved away from this baseline or towards it. This approach has two major drawbacks.

Firstly, the calibration step could be annoying for the user. Calibration data is recorded in a 1 second interval. The message asking to hold the device steady is displayed for a longer period of time (8 seconds), however, to give the user enough time to read the message and assume a starting position. In practice this is even more troublesome, because holding heavy weights stationary for such a long time can be exhausting. Secondly, it caused data from different training sets to be hard to compare. Consider two training sessions. Let us say the raw peak y value for both is $7.5\,m/s^2$. For the first training session, the user kept his forearm perfectly horizontal, resulting in a baseline y value of $0\,m/s^2$. For the second training session, the user's forearm pointed slightly downward, causing a baseline value of $-1\,m/s^2$. While the raw values are equal, the second session's corrected peak value now is 8.5, while the first session's value is $7.5\,m/s^2$. Because the knowledge of which values are indicative of a user at rest is very valuable information, the following solution was found.

Since the zero-based data itself is only useful for more legible data when plotted, the baseline was no longer subtracted from the raw data. Because we also wanted the baseline from different training sessions to be comparable we simply decided to record a baseline once for each exercise and use it as a standard for all other sessions. This also remedies the first drawback of annoyance at having to run calibration for each exercise at each session.

The accelerometer has several advantages. It is able to measure linear acceleration as well as the gravity effect. Its power consumption is exceptionally low. This is important for a mobile device. In practice, the system as a whole loses about 20% battery charge in an hour. And, it is available in almost all Android devices released since the platform went to market.

2.1.4.2 *Gyroscope*

The gyroscope measures rotation velocity in radians per second. It is most useful for measuring precise rotations such as hand gestures. As explained in Section 2.1.1, we are not interested in this kind of movement. Chang et al. [8] follow the same reasoning and noted that Minnen et al. [33] achieved results using gyroscopes + accelerometers that are comparable to those of Chang et al., who use only accelerometers. An additional drawback is the high energy consumption.

2.1.4.3 *Magnetic Field*

The geomagnetic field sensor measures the strength of the magnetic field around the device in three axes. It is used to compute a rotation vector, which in turn is used to determine bearing for navigational purposes. For our application this sensor might be relevant to make the app orientation independent. This way, exercises could be recognized even when the device is used upside-down while the training data was recorded with the right side up. This could work by multiplying movement data by a rotation vector. The resulting data would be aligned to magnetic North, regardless whether the device was used upside down or otherwise rotated.

Despite the name however, the sensor also picks up perturbations of the earth's magnetic field caused by electronics or heavy metal objects. A field test showed that the magnetic influence of metal objects in a living room was so strong that the signal was completely unreliable. A needle compass showed the same results, excluding the possibility of a device fault. Since a gym is full of heavy metal objects, we decided to disregard this sensor. In practice, we do not expect the device to be used in different orientations, because the wristband limits the number of possible ways the device can be fixed to the forearm. For all the ways the device can be attached to the forearm, the screen is only usable when the device is worn in the way as intended (Figure 2.2).

2.1.4.4 *Orientation sensor*

The orientation sensor is a virtual sensor. Its values are calculated by combining data from all of the above sensors, and the power consumption is the sum of these sensors. It has been notoriously unreliable, and it is marked as deprecated in the Android documentation [23]. Although the accuracy should have been improved by incorporating the gyroscope since Android version 4.0, the deprecation status still applies. This, together with the large power draw, has made us decide not to use this sensor.

SUMMARIZING: We have adopted an approach using only accelerometer data. This sensor has very low energy demands, can measure both linear acceleration and the *gravity effect*, and is available in all iPhones and virtually all Android devices.

2.2 ALGORITHMS

For exercise recognition and repetition counting, a few proven methods are available. Most earlier work considers at least the use of Hidden Markov Models (HMMs) [2, 15, 33]. Another preferred method is the Naïve Bayes Classifier (NBC) [8, 2]. Because we want to be able

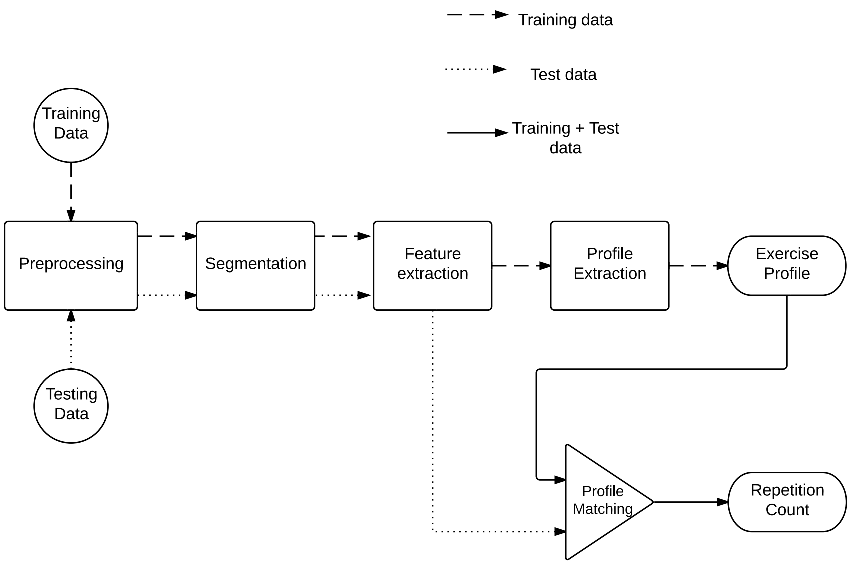


Figure 2.4: Data processing pipeline.

to provide the user with rich feedback on his/her performance, algorithms with hidden states or processes seemed less suitable. Examples of these algorithms are HMMs and Neural Networks. Our approach is best described as 'dynamic rule-based decision'. It is described in Section 2.2.4.

The data processing pipeline is shown in Figure 2.4. It follows the general data processing pipeline as commonly used in pattern recognition [14]. In the preprocessing stage, raw data is first smoothed to eliminate unwanted noise. The filtered data is then separated into *Movements* in the feature extraction stage. These are the atoms that will be used in the remainder of the process. From the sequence of movements, a subset is selected which appear to be most prototypical for the exercise (profile extraction). The resulting profile can be matched against new data for exercise recognition and repetition counting. These stages are described in more detail below.

2.2.1 *Preprocessing*

The purpose of preprocessing is to transform the data in such a way that it is most convenient to extract the components that are of interest to the application. We are interested in free-weight exercise related movements. Ideally, the preprocessed signal would be zero when the user is not exercising, and would contain only acceleration data that is directly caused by the movement of the muscles which are involved in the exercise the user is performing.

98% of the energy in walking at a regular pace is in the 0–10 Hz spectrum [1]. For the 8 exercises in our pilot, we do not expect to encounter higher target frequencies. For many free-weight exercises, people are instructed to perform them at a 1–2 pace, meaning that the *forward movement* takes 1 second, and the *backward movement* takes twice as long (see page 14). For our purpose, a low-pass filter seems most suitable. Wang et al. [61] compared multiple filtering algorithms for the purpose of movement analysis. Butterworth [7] seems unsuitable because of large delays. Since we build an on-line application, delays are to be kept to a minimum. Median filters have a delay of half the window length. Kalman filters are at an advantage because they are dependent only on the previous sample [61].

> "The KF [Kalman Filter] is a state estimator that works on a prediction-correction basis. This means that it computes a belief in a certain state estimate by first making a prediction based on the dynamics of the system and later correcting this prediction using measurements of the system." [37]

Rather than providing an estimator that transforms the data directly without any knowledge of the data's origin, the KF uses two models that allow for a more accurate estimation of whether the data it receives is reliable. The *prediction* step uses the following models to predict the current state.

THE PROCESS MODEL describes the source of the data. The simplest model assumes an unchanging signal:

$$x_k = x_{k-1} \tag{2.1}$$

where x_k is the current signal value and x_{k-1} is the previous value. In theory, a noise component is usually added, but since the value of the noise is unknown and assumed to be zero-mean, this term can be omitted. We will later account for its variance. Since the value of the current state only depends on the previous state, this model satisfies the *Markov assumption*. This model seems too simple for our process; the movement of the human body. Because of the spring-like properties of muscles, human movements can be nicely modeled with a sine wave. Fitting incoming data to a sine function is impractical however, since we do not know a-priori what its phase and amplitude could be. So instead we simulate the alternating dampening and increasing speed of the signal by modeling a parabolic process. Parabolas have a linear derivative and a constant second derivative. Since we do not have future values at our disposal, we cannot apply central differences. When we speak of the derivative in point x'_{k-1}, we mean de backward difference $x_{k-1} - x_{k-2}$. Since the second derivative is constant, it does not have an index and is expressed as x''.

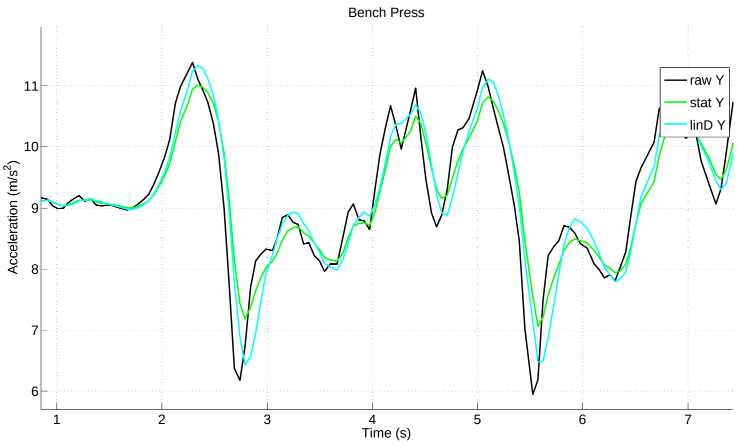


Figure 2.5: Comparison of process models. The green stat Y line uses the process model in Equation 2.1. The turquoise linD line uses the process model in Equation 2.3.

$$\begin{aligned} x_k &= x_{k-1} + x'_k \\ &= x_{k-1} + x'_{k-1} + x'' \\ &= x_{k-1} + x'_{k-1} + x'_{k-1} - x'_{k-2} \\ &= x_{k-1} + 2(x_{k-1} - x_{k-2}) - (x_{k-2} - x_{k-3}) \\ &= 3x_{k-1} - 3x_{k-2} + x_{k-3} \end{aligned} \tag{2.2}$$

This model did not perform better than the model for an unchanging signal. Instead, noise was extrapolated, which produced a high-frequency signal with a very large amplitude. By changing the coefficients in the last line of Equation 2.2 we arived at a model that predicts a reversal of the current trend in the signal. It thus has a dampening effect:

$$x_k = x_{k-1} - x_{k-2} + x_{k-3} \tag{2.3}$$

Using this model, the loss in peak amplitude is minimized compared to the model in Equation 2.1. There is no overshoot, because the current trend is predicted to reverse, see Figure 2.5. Note that because we include more past time steps than the immediately preceding time step, the Markov assumption does not hold for this model. The advantage could be that if it will be necessary in future work to integrate the signal, a more accurate result is available.

The uncertainty of our prediction is increased with a constant value at each time step. σ_ω^2 indicates the process noise variance. The update function is simply:

$$\sigma_k^2 = \sigma_{k-1}^2 + \sigma_\omega^2 \tag{2.4}$$

THE SENSOR MODEL describes the dynamics of the sensor. If there is some systematic flaw in the sensor or its calibration, this model can account for it. The sensor we use is a tri-axial acceleration sensor. When the device is placed on a table, The x and y values are approximately zero and the z value is close to the gravitational constant. We have no equipment to evaluate the precision while in movement; we assume the sensor value to be equal to the actual acceleration, plus some noise. The magnitude of the noise is assumed constant, since each measurement is independent of any previous measurement. It should be noted that the hardware we use is popular among consumers, but low budget smartphones might not perform as well.

CORRECTION. Using the process- and sensor model, we can make a prediction. The predictions provided by both models are combined into updated predictions for the current value and its uncertainty as a weighted average.

$$\begin{aligned} x_k^+ &= \frac{\sigma_v^2}{\sigma_k^2 + \sigma_v^2} x_k + \frac{\sigma_k^2}{\sigma_k^2 + \sigma_v^2} z_k \\ &= \quad x_k + \frac{\sigma_k^2}{\sigma_k^2 + \sigma_v^2} (z_k - x_k) \end{aligned} \tag{2.5}$$

z_k indicates the current sensor value and σ_v^2 denotes the sensor noise. We can also update the uncertainty with the new sensor data:

$$\frac{1}{\sigma_k^{2+}} = \frac{1}{\sigma_k^2} + \frac{1}{\sigma_v^2} \tag{2.6}$$

which can be rewritten as

$$\sigma_k^{2+} = \sigma_k^2 - \frac{\sigma_k^2}{\sigma_k^2 + \sigma_v^2} \tag{2.7}$$

The weighting factor which appears both in equation (2.5) and (2.7) is called the *Kalman gain*

Kalman gain

$$K = \frac{\sigma_k^2}{\sigma_k^2 + \sigma_v^2} \tag{2.8}$$

It is a measure for how much certainty we have in the new measurement relative to our most recent value estimation. When the uncertainty in the new measurement is large, the denominator in Equation 2.8 becomes larger and the less of z_k in Equation 2.5 gets included in the new value estimation. For a more elaborate introduction to Kalman filters, we refer to Negenborn [37], from which most of the equations in this section were adapted.

2.2.2 *Feature extraction*

After the preprocessing stage we have got a signal which has been smoothed to attenuate most noise. But the signal still needs to be segmented in meaningful chunks, which can be compared to each other. We call these chunks *movements*. A movement corresponds to the extension or contraction of a muscle in the target exercise. Exercises usually consist of a forward movement, which corresponds to contraction of the muscle and a backward movement, which corresponds to extension of the muscle. Since all data processing is done on a per-axis basis, one movement by the user can produce a movement in multiple axes. This is the case when the device is rotated during the exercise, causing a gravity effect. The forward movement in the biceps exercise is visible in both the y and the z axes, for example. The forward movement in Overhead Dumbbell Press (ODP) does not involve rotation, and is only visible in the y signal.

Movement

The signal is segmented based on the magnitude of the signal's derivative. We essentially perform peak detection by watching for sign changes in the signal. This was described by [8] as the most effective method for their repetition counting goal. To eliminate the detection of movements in noise and small or blunt peaks, two conditions must be met. To eliminate peaks which are very short in duration, there can only be 1 movement detected every $minSignSpacing$ milliseconds. When the user holds his arm steady, the signal never has a derivative which is exactly equal to 0. Therefore, we define a $zeroDerivativeThreshold$. The derivative must be larger, in the absolute sense, than this value to be recognized as part of a movement. If it is not, the derivative is clamped to 0. Thus, we consider 3 distinct states for the signal's derivative: downward, steady and, upward. When a change of state is detected, a new movement is defined by the following movement features:

Movement features

1. startAmplitude: The signal's value at the start of the movement
2. endAmplitude: The signal's value at the end of the movement
3. duration: the amount of time in ms between the start and end of the movement.

$minSignSpacing$ and $zeroDerivativeThreshold$ are parameters. The optimal values differ per exercise, we discuss the optimal values in a later section. At the end of a data recording session, the movements are stored in chronological order in the JSON data format [25]. We highly recommend this format for its versatility and wide support of platforms.

2.2.3 *Profile extraction*

The result of the previous step is a set of quantitative data in the form of movements which can be compared to new data. To make this process easier and less resource intensive, we would like to create a prototype of the exercise, which is described by the data. The prototype should be an average of the extracted movements. When comparing this prototype to other data we may receive in the future, we will never get an exact match. That is why we also need to model how much deviation from the prototype is allowable to still be classified as an instance of the exercise. Two reasons why we would *not* want to classify new data as an instance of the target exercise is because the data is generated by noise, or an exercise other than the target exercise.

Since we want a prototype for both the forward and backward movement of the exercise, we first split the movements into two lists, one for upward and one for downward movements. For both lists, we remove outliers using Peirce's Criterion (PC) for outlier detection [43].

Peirce's Criterion (PC)

Ross [50] gives an insightful and practical manual for applying the criterion. He also points out that it is more rigorous than the much more popular criterion by Chauvenet [9]. Chauvenet's method assumes one outlier in the entire data set, while PC can accommodate for multiple outliers and multiple observed quantities. PC is derived from probability theory. Observations should be rejected when the standard deviation obtained by retaining them is less than that of the standard deviation obtained by their rejection multiplied by the probability of having that particular number of outliers.

The rejection criterion for PC is

$$|x_i - x_m| > R * \sigma \tag{2.9}$$

Where x_i is the data value. Since we use PC to reject data for all of our three movement features separately , x_i represents either a single value of start amplitude, peak amplitude or duration. x_m is the mean of the data set, and σ is the standard deviation of the data set. R is the maximum allowable ratio of sample deviation from the data set's standard deviation. It depends on the size of the data set and the amount of assumed outliers. The calculation of R is quite complicated. A table which lists values for data sets of size 3 through 60 and 1 through 9 doubtful observations is listed in [50]. One starts out by assuming 1 outlier. When 1 or more observations are rejected by the criterion in Equation 2.9, the amount of assumed outliers is incremented by 1. The original data set's standard deviation and size are retained, but the value of R is updated. This process is iterated until no more data points are eliminated.

We eliminate a movement when at least one of its three feature values described on page 27 is marked as an outlier. This successfully eliminates preparatory movements, usually having abnormal amplitude, and random jerks that usually have either abnormal duration or amplitude.

Next, we assume that the 10 movements with the largest amplitude correspond to the 10 repetitions of the target exercise we asked the user to perform. Because it is possible that one or more of the repetitions themselves were performed incorrectly, PC is applied to this subset again. The resulting set is stored in an exercise profile as a set of 6 normal distributions [1].

We have thus achieved our goal of creating a model of the target exercise which also models the amount of variation that can be expected.

2.2.4 *Profile matching*

Let's refer back to the processing pipeline in Figure 2.4. By passing our raw data through the preprocessing, segmentation, feature extraction and profile extraction stages, we have a prototype of the movements which describe the target exercise. To transform new data in a format which can be used to compare to the prototype, the new data is passed through the first three stages. Our initial strategy was to check how likely it is that all three feature values in the new data were drawn from the probability distribution in the prototype. We accept the new movements when this likelihood is larger than 95% for all three movement features.

$$|x_i - x_\mu| < R * \sigma \tag{2.10}$$

Where x_i again is a movement feature value, x_μ is the mean of distribution in the prototype, and σ is the standard deviation of the distribution in the prototype. R is the ratio of acceptable deviation from the mean. $R = 2$ corresponds to a 95% confidence interval. Whether this interval is suitable can be debated. Firstly, since there are 3 movement features, the likelihood that a movement will be falsely rejected is $1 - (0.95)^3 = 14.3\%$ in the worst case. That is, assuming the 3 feature values are statistically independent, which they most probably are not [2]. Secondly, we found that there is a large variability in how consistently users perform an exercise. The result was a large variation in the prototype's standard deviations. A fixed value of $2\,\sigma$ of acceptable margin was sufficient for some, while $10\,\sigma$ was required for others. A fixed ratio thus seems unsuitable. One idea would be to use R as a user-defined difficulty setting. The lower the value, the stricter the algorithm would be.

[1] One for each of the three movement features multiplied by two (forward and backward) movements.

[2] If the measured amplitude is larger than the amplitude in the prototype, the duration will likely be abnormal too.

Because we prefer an application which requires as little user input as possible, we propose an alternative criterion for accepting a movement, which does not consider the variance at all:

$$|x_i - x_\mu| < R * x_{range} \tag{2.11}$$

Where x_{range} is the range of all the movements in a prototype, i.e. the largest absolute difference between the start and end amplitude of a movement. R is a predefined constant between 0 and 1. Initial tests pointed towards a significantly improved result when compared to the inclusion criterion with standard deviation (Equation 2.10). If desirable, R can still be used as a difficulty setting. As the 'default' setting for R, we considered values of 0.3, 0.4, and 0.5. We call this *dynamic rule-based decision* , since a movement is accepted if the criterion holds, and the criterion itself is dynamic because x_{range} is calculated during profile extraction.

dynamic rule-based decision

If the movement is accepted by the above criterion, it is stored in a list for later scoring. One expects to add an instance of the forward and then backward movement alternatively. It is possible however, that either movement was not recognized or not accepted by the above criterion. The very first movement has a high likelihood of not being recognized because it is preceded by preparatory movements. Likewise, the last movement has a high likelihood of not being recognized because it is mixed with a movement associated with unstrapping the device, for example. To update the count of executed repetitions, we have to reconstruct the original movement sequence from a corrupted sequence. We do this by increasing the counter whenever we encounter the forward movement or when we receive a movement which was unexpected given the previous detected movement. Given the sequence

1:Backward, 2:Forward, 3:Backward, 4:Backward, 5:Forward

We would increase the counter at position

- 1, because we would expect a Forward movement first
- 2, because it is a Forward movement
- 4, because we would expect a forward movement after 3
- 5, because it is a Forward movement

We would *not* increase the counter at position 3, since the backward movement is expected to follow the forward movement at position 2, and is thus not an indication of a new repetition. Our final count becomes 4 repetitions. Another reason for keeping a list of accepted movements is that we will be able to analyze and score the user's performance on a series of repetitions. Since feature values are retained, users can receive specific feedback on tempo, start position and end position, for both forward and backward movements.

2.2.5 *Scoring*

After the matching process, we have a list of the target movements the user performed. We want to give the user an overall score that indicates how well he/she is doing. This score is based on a correlation score between the feature values of the actual movements and the ideal values in the model. The exact transformation from correlation ratio to performance scoring has to be fine-tuned in collaboration with fitness instructors, as there is currently no way to tell whether a correlation score of, say, $+0.8$ is acceptable or not. In the remainder of this thesis, we speak of a deviation score since the ideal score where movement and model correspond completely, is 0.0. Our initial implementation is as follows.

deviation score

$$\begin{aligned} D_i &= (\mathrm{Actual}_i - \mathrm{Model}_i)/(\mathrm{Model}_i * R) \\ S_i &= 1 - |D_i| \end{aligned} \tag{2.12}$$

Where D_i is the deviation score for feature i and S_i is the *correlation score* for feature i. Actual_i is the feature value in the data while Model_i is the ideal value. R is the same margin ratio as used in Equation 2.11. The value of D_i is capped to [-1,1] so that S_i can never be below zero.

correlation score

The total score S could be defined as the mean of all S_i. From post-experiment interviews, we may conclude that duration is more important than amplitude, for example. This could be represented by a weighting factor between the individual feature scores and the total correlation score.

Note that we assume that an exercise consists of 2 movements, which are executed alternatively. This means that in practice, the end amplitude is almost equal to the start amplitude of the second movement. This is reflected in the design of the interface, which shows only one (the peak) amplitude indication for each movement.

2.3 EXPERIMENTAL SETUP

In our main experiment, we will investigate the performance of our repetition counting algorithm. Additionally, we would like to know whether the use of our app enables users to better perform their exercises. For this purpose, we apply a repeated measures design. During the control condition, the user performs exercises unsupervised, while during the experimental condition, real time and post-task summary data is available. We are interested to know whether this feedback allows the user to perform the exercise in a better way. Due to time constraints, we had to select 2 exercises for inclusion in the main experiment. We reasoned that we should have one exercise with a rotational movement, and one without a rotational movement. We chose biceps because it allows users to view the screen while practicing, and the Bent-over Row (BOR) because it is an exercise that is easily explained to novices.

2.3.1 *Participants*

Our application aims to aid anyone who would like to do resistance training. Earlier research indicated that people with beginning to intermediate experience might benefit most from the type of system we built [44]. Because we are interested in how experts perform in comparison to beginners, and perhaps use them as a benchmark, we aim to recruit a representative sample of gym attendees. Trainers and people who have no experience with resistance training will also be included.

2.3.2 *Procedure*

We employ a repeated measures design with four blocks. The first block is meant to record a profile, which will be used as a representation of a 'perfect' example. The second block is used to collect a baseline performance without feedback from our app and in the final two blocks data is collected with feedback enabled. Because we want to prevent a ceiling effect where the participant remembers the exercise perfectly during block 2 through 4, and to simulate a more realistic scenario, we impose the requirement that there should be at least 1 day between the administration of block 1 and 2.

BLOCK 1: After a short introduction of the experiment we explain that it is necessary to record data at two different time periods and an appointment is made. Depending on the time available, the participant may perform one or both exercises. The exercise is explained by a professional (trainer or therapist) and the participant is asked to perform 3 repetitions to see if he/she can perform the exercise in a reasonable manner. We select a weight which the participant is easily able to lift with a pause between each set of 10 repetitions. Fatigue can distort the outcome of the result in the sense that participants will perform better on earlier blocks. The order of the baseline and feedback block can not be balanced because the benefit of the feedback given in the feedback block may carry over to the baseline block.

After the participant has learned to perform the exercise, he/she is asked to perform the exercise 10 more times in order to record a 'golden standard' profile. We record the following variables for each participant: gender, age and experience with free weights in 3 categories (none to sporadic, free-weight exercise as part of a balanced workout, and free weight exercise as the main or only part of workout).

After we remind the participant of the appointment for the second measuring block, we say goodbye. The expert gives a baseline score for the golden standard. To make processing easier, the expert is asked to evaluate the exercise in the same way as our app does. He/she records an overall score on a scale of 1 to 10 (10 being perfect), and a rating on a 7-point Likert scale for tempo of the upward and downward movement (too slow – too fast). The extent to which a participant flexes and extends is likewise recorded (too short – too far). Scores are given for the upward and downward movement separately. Finally, the amount of executed repetitions is noted, in case the subject miscounts. This makes for 6 measures recorded per block, both by the expert and our app.

BLOCK 2: The participant is asked to perform one set of 10 repetitions of the same exercises(s) as he/she did during block 1. Although the participant is wearing the apparatus and movement data is recorded, no feedback is given during the exercise. The expert scores the performance in private so as to not influence the participant. The app's assessment is also recorded.

Up until this point, the participant has not been given any feedback by our app. Between blocks 2 and 3 we take some time to explain the various components and modalities of the feedback our app will provide. We let the user get used to the feedback by encouraging them to perform the exercise incorrectly. We refer to Part iii to see in what way feedback is provided.

BLOCK 3: At this point, feedback is enabled. After every repetition, the participant gets an auditory cue when he/she *should* have ended the repetition. When the cue is heard before the actual repetition has been completed, the repetition was performed too slow. When the cue is heard when the next repetition has been started, the repetition was performed too fast, and the pace of the next repetition should be adapted accordingly.

BLOCK 4: Summary data for the performance during block 3 is shown on screen. The values used for this screen represent the average over all 10 repetitions of block 3. See Figure 5.4 for an example. With live feedback enabled, another 10 repetitions are performed. Again the expert is asked to score the participant's performance and the app's feedback is recorded. Since the experiment is now at an end, the participant may view the expert's score.

CONTROL GROUP: When people were assigned to the control group, they would go through blocks 1 to 4 while wearing the device but without receiving feedback from the device. This means that they were only given instructions on how to perform the exercise correctly before the start of block 1.

2.4 EXAMPLE DATA

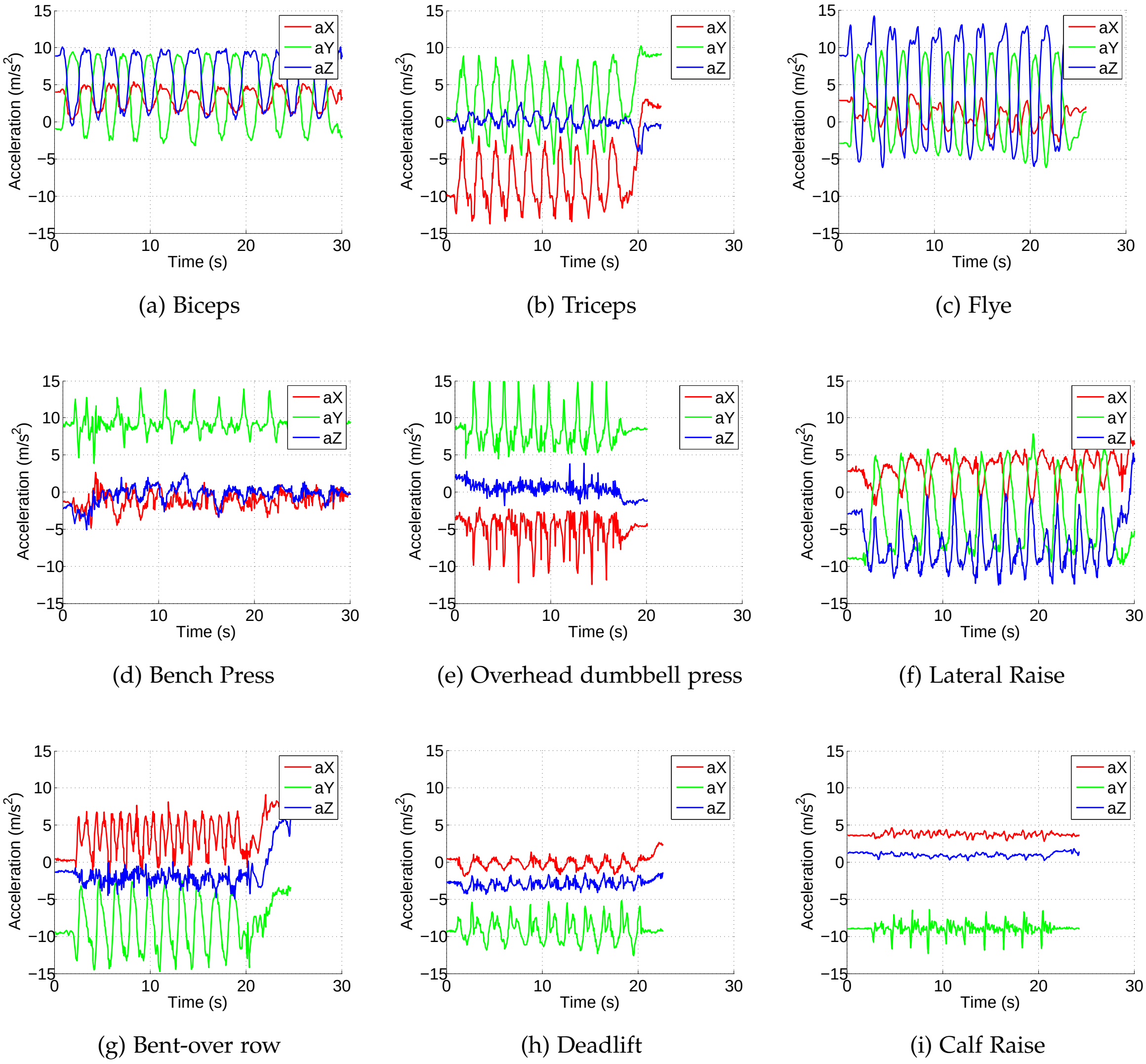
Figure 2.6 shows a plot of the data as it is presented to the processing algorithm described in Section 2.2. It is captured at 100 Hz and then downsampled to 20 Hz by a 5 point averager. Working at a lower sample rate is less resource intensive and the downsampling might average out very high frequency noise. As we saw earlier, we do not expect to encounter higher target frequencies than 10 Hz [1]. Thus, the recording frequency of 20 Hz. should be sufficient.

THE BICEPS exercise data is characterized by a sinusoidal signal of equal amplitude in both the y and the z axis. The difference in phase is exactly half a period. This is because of a very strong *gravity effect*. Refer to Figure 2.2 and Figure A.1a: in the starting position, gravity pulls in the z direction. As the arm is rotated, orientation of the sensor shifts so that gravity pulls in the y direction. That is why the y and z signal are in complete anti-phase. The weaker sinusoidal movement in the x direction is presumably caused by the wrist being slightly rotated inwards, which causes a right to left movement when rotating the elbow.

THE TRICEPS signal can be explained in much the same way as the biceps signal. Where gravity shifts from the z tot the y axis in the biceps exercise, it shifts from the x to the y axis in the triceps exercise. Another difference is that for the triceps exercise, both signals have the same phase. This is because the x signal is negative in triceps' resting position while the z signal is positive in biceps' resting position [3].

THE FLYE is another exercise with a gravity effect. y and z are in anti-phase, just like they are in the biceps exercise. Our algorithm should still be able to discern between these two exercises, because the range in the z axis is much larger. We saw a surprising variety in the actual data we collected during the pilot. For one participant, it was the x axis which was in anti-phase with the z axis. We can only explain this when assuming this particular participant wore the wristband on the side of the forearm rather than on the inside, or performed the exercise with rotated wrists.

[3] When the sensor would be attached to the left hand, the triceps signal is actually positive in the resting position, and x and y *will* be in anti-phase



(a) Biceps

(b) Triceps

(c) Flye

(d) Bench Press

(e) Overhead dumbbell press

(f) Lateral Raise

(g) Bent-over row

(h) Deadlift

(i) Calf Raise

Figure 2.6: Raw acceleration data from each of the 9 exercises considered in this study. Data was captured at 100 Hz and downsampled to 20 Hz. The figures show the typical patterns for each exercise.

THE BENCH PRESS shows a different kind of signal. It is an exercise with a linear movement. Because there is no rotation involved, no gravity effect is observed. As can be seen in figure Figure 2.6d , there is one axis (y) which has the strongest fluctuation in the signal. The other two signals are mostly noise. The peaks are much sharper than the sinusoidal signals in the biceps data.

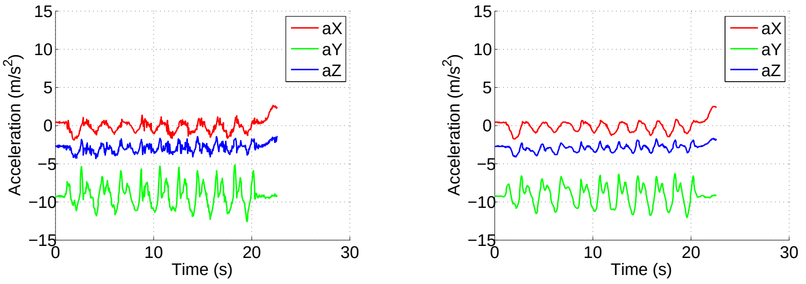
THE OVERHEAD DUMBBELL PRESS is another example of an exercise with a linear movement. There is, however, a lot of energy in a secondary axis too. The x axis shows peaks because when the user stretches his arm, the forearm moves sideways to the user's chest, which is the x direction.

THE LATERAL RAISE is an exercise with a gravity effect. The y and z axes are in phase. In the z axis an additional, smaller peak is visible, presumably because participants stop to balance the weights between each repetition.

THE BENT-OVER ROW exercise is sometimes described as 'sawing a log of wood'. It is a vertical or slightly diagonal movement from a spectator's point of view, although it is purely performed in the sensor's y direction. The signal value goes up when the user pulls the weight up and goes down when the weight is lowered to the ground. Gravity constantly pulls in the x direction, which has to be countered by the user. This causes a significant fluctuation in the x axis.

THE DEADLIFT signal is a very noisy one because the whole body is in motion. We see a periodic movement in the z direction, which corresponds to the arm moving slightly to and from the body. The big dip in the y signal comprises the two target movements, the signal moving down when the user squats. An additional fluctuation between repetitions is visible at around $-7\ m/s^2$. This is caused by the user finding his/her balance when completing a repetition and preparing for the next one.

As said earlier, the standing calf raise will not be considered in the remainder of the study because of the poor signal. Although the peaks in the y signal are pronounced, the signal-to-noise ratio is low and the peaks vary greatly in amplitude. The data shown in Figure 2.6i was recorded from a fitness instructor. Signals recorded from regular users varied much more. Even when the sensor was attached to the ankle rather than the wrist, this problem was not alleviated.



(a) Raw

(b) filtered

Figure 2.7: Raw and filtered deadlift data. The used filter parameter values are the ones which perform best on all exercises.

2.5 PROCESSING EXAMPLE

In this section, we focus on a single data file and walk the reader through the processing pipeline. We have chosen the deadlift example for our exercise, as this is an exercise with a linear movement. Linear movement data is more noisy and less periodic in general and thus poses more of a challenge to our algorithm. If we would have taken biceps for example, the raw data would not need any smoothing for proper results.

PREPROCESSING: The filter values we used were determined by maximizing the overall repetition counting score for all exercises. These values were also used for the main experiment. The values were 0.02 for the process noise and 0.8 for the sensor noise. A comparison of the raw and preprocessed data is shown in Figure 2.7. Local minima are eliminated or attenuated to an extent that they will be ignored by the feature extraction stage. The Kalman filter is especially effective in the non-major axes x and z, although this is not really useful for our project. The delay introduced by this parameter combination is $50\text{-}100\,\mathrm{ms}$.

FEATURE EXTRACTION In Figure 2.8, the black vertical lines show the movement boundaries for our example. At first glance, there are a lot of false positives, for example at $t = 12$ for the y axis. These will be removed in the profile extraction stage. What is important however, is that the target movements are captured correctly. Target movements should not be broken up into multiple smaller movements and the markers should be placed at or near the peaks of the movements. Although the markers are placed one sample too late for the higher peaks at around $-6\,\mathrm{m/s}$, the delay is acceptable and the movements are not interrupted.

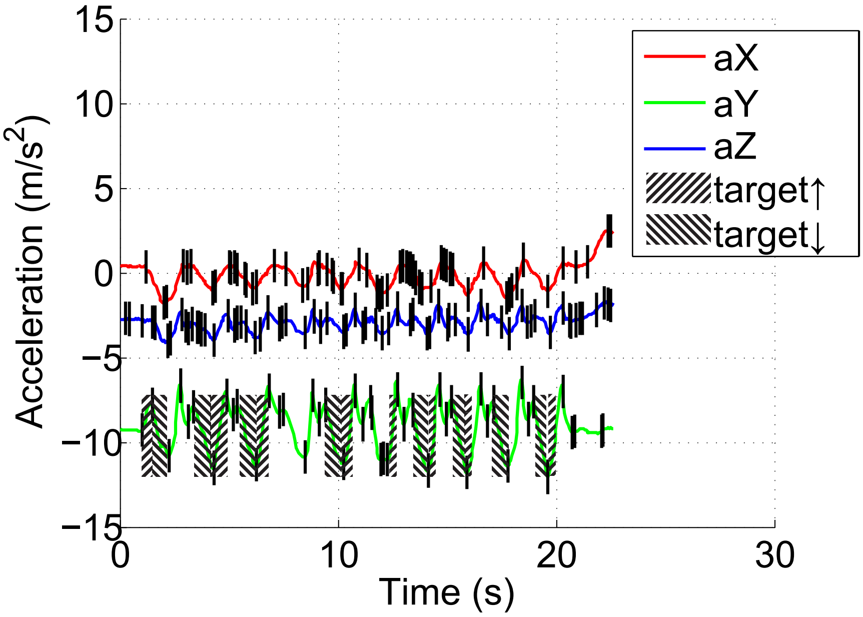


Figure 2.8: Filtered deadlift data with automatically labeled target movements. The areas marked with a bottom-left to top-right diagonal are parts of the signal which were identified during the profile matching phase as an upward movement. The top-left to bottom-right diagonals are identified as downward target movements. The black vertical lines show movement boundaries, including non-target movements.

PROFILE EXTRACTION During the profile extraction stage, we aim to construct a prototype of the forward- and backward movements. In order to do this, the non-target movements are discarded. We expect to find 10 repetitions in the signal [4]. We assume that the target movements are the ones which have the largest amplitude. Therefore, we sort all movements based on amplitude and select the 10 movements with the largest positive amplitude and the 10 movements with largest negative amplitude. For each set of ten movements, outliers are removed using Peirce's criterion and the remaining movements are averaged to create the prototype. The part of the profile corresponding to the y axis is shown in Listing 1. The prototype is an array of 2 movements. The $startA$ and $peakA$ correspond to start and end amplitudes of the target movements. Note that the start amplitude of the first movement is close to the end amplitude of the second movement. The $rangeA$ is the range of the signal in that axis, calculated as $startA + peakA + startSD + peakSD$. dT is the average duration in milliseconds. The movements that contribute to the prototype for our deadlift example are marked with diagonal lines in Figure 2.8. The

[4] When data is recorded for a profile, the user is asked to perform 10 repetitions of the exercise. When the user miscounts, the created profile is somewhat deteriorated, but still acceptable.

Listing 1: Deadlift profile, extracted from data shown in Figure 2.8. Only the y axis is shown. See main text for explanation of fields.

```json
{
  "axes": {
    "aY": {
      "baseline": -9.275943613052368,
      "pattern": [
        {
          "timestamp": [
            2280,
            12290,
            10269,
            4306,
            6264,
            14143,
            15917,
            19647
          ],
          "startA": -11.429187453730062,
          "peakA": -6.935757361572085,
          "dT": 562.25
        },
        {
          "timestamp": [
            18996,
            15215,
            17126,
            3382,
            13442,
            5411,
            9418,
            1478
          ],
          "startA": -7.793103842298285,
          "peakA": -11.471142929042756,
          "dT": 767.125
        }
      ],
      "type": "movementPattern",
      "rangeA": 6.0274379024559535
    }
  },
  "version": 3
}
```

$\mathtt{timestamp}$ values in Listing 1 correspond to the onset times of the hatched areas.

Two target movements were removed for both the upward and downward movements. The upward movement at $t = 8.5$ was removed because its duration of 358 ms was outside the allowable range (mean = 562, allowable deviation = 144) [5]. The upward movement at $t = 17.8$ was removed because its peak amplitude of -6.48 was outside the range (mean = -6.92, allowable deviation = 0.40). For the downward movements, both removed movements were outside the duration range (mean = 756, allowable deviation = 306). The values for the movements at $t = 7.5$ and $t = 11.6$ were 1019 and 407 ms, respectively.

We feel that the maximum allowable deviation of 0.40 m/s^2 for the upward movement's peakA is very strict. The reason for this strict value is that the standard deviation of the 10 largest upward movements is very small: only 0.21. Thus, it does not take much to consider any movement an outlier. The allowable deviations for duration are reasonable however. We are tempted to relax the allowable deviation for peakA, but for most other examples we have seen, this is not necessary. Since only the mean is used in the rest of the process, and not the standard deviation, we feel this is insufficient reason to add another parameter to the process.

PROFILE MATCHING To count repetitions, the created profile is matched against new data. Figure 2.9 shows data which was captured from the same participant. For the major axis, all three features (duration, peak amplitude and range) must be within a certain ratio (0.5 for our example) of the profile mean. For non-major axes, the signal only needs to be in the expected range. When these conditions are met, this repetition is accepted. As usual, the first and last movement are not recognized because they are mixed with preparatory/ending movements. Those first and last repetitions do contribute to the repetition count however, because the other movement in that repetition *is* recognized. Two additional movements are rejected. At $t = 4.8\,\mathrm{s}$ we see that the dip in the signal is not as deep as the other movements. From $t = 12\,\mathrm{s}$ onwards, we see a decline in amplitude in the x signal. At $t = 15.8\,\mathrm{s}$, the x signal is out of range, invalidating that movement. Since two exercise repetitions are rejected with a total of 8 performed, we would expect a repetition count of 8. The actual repetition count is 9, however.

The root cause for this higher than expected repetition count lies in the profile extraction. The movement whose peak is farthest from the baseline is considered the first movement, and the movement with

[5] We report the maximum allowable deviation from the mean for 1 doubtful observation according to Peirce's criterion. When discarding additional observations, the allowable deviation decreases.

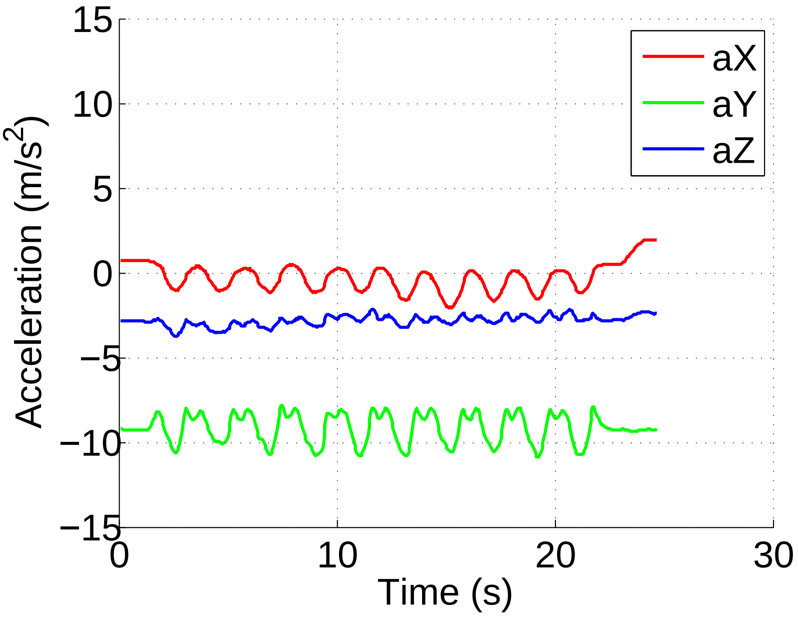


Figure 2.9: Filtered deadlift data used to match against the profile data in Figure 2.7b. The filter parameter values used are the ones which perform best on all exercises.

the peak closest to the baseline is considered the second movement. For exercises involving linear movement, the baseline of the major axis lies in the middle of the two extremes of the range. In the case of the profile extracted from the data in Figure 2.7b, the baseline is at $-9.3\,\mathrm{m/s^2}$. The peak of the downward movement lies at $-11.4\,\mathrm{m/s^2}$ and the peak of the upward movement lies at $-6.9\,\mathrm{m/s^2}$. Since the upward movement's peak is farthest from the baseline, this is incorrectly classified as the first movement.

Now, the first movement recognized in the signal from Figure 2.9 is the downward movement. Since this is the second movement according to the profile, we assume we failed to recognize the first movement and the repetition counter is set to 1. Since the next movement is the first movement *according to the profile*, the repetition counter is set to 2 when in fact the first repetition has just been completed. Adding the 7 remaining recognized repetitions, the total comes to 9.

Note that in the final implementation used during the main experiment, the definition of which movement was the first in an exercise was annotated in exercise-specific meta-data.

Table 2.2: Repetition counting scores using parameters which maximize the average repetition counting score over all exercise files while minimizing its standard deviation. **N**: the amount of data files for this exercise. All data files contain recordings of a series of 10 repetitions. **N10**: the amount of data files for which exactly 10 repetitions were counted. **repCount**: the average repetition count. **repSd**: The standard deviation of the repetition count. **score**: The average repetition counting score (see main text). **scoreSd**: The standard deviation for the score measure.

Exercise	N	N10	repCount	repSd	score	scoreSd
biceps	10	10	10	0	100%	0%
flye	8	7	10.13	0.35	98.8%	3.5%
BOR	12	10	10.17	0.39	98.3%	3.9%
lateralRaise	14	11	10.21	0.43	97.9%	4.3%
triceps	10	8	9.9	0.74	97.0%	6.7%
deadlift	8	5	9.88	0.64	96.3%	5.2%
ODP	10	5	10	0.94	9.4	7.0%
benchPress	10	4	9.5	1.51	89.0%	11.0%
All exercises	82	60	9.99	0.75	96.5%	6.5%

2.6 parameters & repetition counting

Our algorithm uses a number of parameters for which we would like to find optimal values. One could try to find optimal parameters for each exercise, the overall best over all data, or make a distinction between exercises with linear movement versus exercises with a rotational movement in the major axis. We are interested in finding a default setting for the main experiment. Therefore, we will look for the overall best result. An additional constraint we impose is that for all data, the signal-to-noise ratio in the filtered data should be high enough to extract a profile. Using brute force search we tested all combinations of the following parameters: The Kalman process- and sensor noise, zero-derivative threshold, minimum sign spacing, and margin ratio. For a detailed discussion of these parameters, we refer back to Section 2.2. For each parameter we used 5 to 8 levels in a range which showed promising results. Since all participants performed an exercise two times, we train a model on one file and test on the other, and vice versa. A total of 82 data files (820 repetitions) were used.

For 32 out of $10,103$ parameter combinations, a profile could be extracted from all data files. From these 32 combinations, we selected the parameter combinations with the highest repetition counting score. The ideal repetition count is 10. We give equal penalty to false positives and false negatives, so that both 8 and 12 counted repetitions

Table 2.3: Confusion matrix for exercise classification using the repetition count measure. The numbers on the diagonal indicate correctly classified exercises. Off-diagonal values indicate exercises which were incorrectly classified.

		Given Classification								
		biceps	flye	BOR	lateral raise	triceps	deadlift	ODP	bench press	**Precision %**
Correct Classification	biceps	8			2					80
	flye	1	7							87,5
	BOR	1	1	9	1					75
	lateral raise	5			7			2		50
	triceps	1	1			8				80
	deadlift			2	1		5			62,5
	ODP					2		2	6	20
	bench press							3	7	70
	Recall %	50	77,8	81,8	63,6	80	100	28,6	53,8	

result in a score of 8. This is expressed as a percentage of the total amount of repetitions. The parameter combination with the highest score is: process noise = $0.02\,\mathrm{m/s^2}$, sensor noise = $0.8\,\mathrm{m/s^2}$, minimum sign spacing = $50\,\mathrm{ms}$ and zero derivative threshold = $0.04\,\mathrm{m/s^2}$. The repetition counting performance using this parameter combination is summarized in Table 2.2. For 60 out of 82 data files, the perfect repetition count of 10 repetitions was obtained. The mean repetition count was 9.99 with an sd of 0.75. This indicates that there was a balanced amount of false positives and false negatives. The mean score was 9.65 with an sd of 0.65. A mean of 9.65 indicates that one out of every $100\%/3.5\% = 28.6$ repetitions will be either missed or double counted. An sd of 0.65 indicates that for 95% of all 10-repetition series, at least $9.65 - 2 * 0.65 = 8.35$ repetitions will be correctly counted.

2.7 Exercise recognition

So far, we have matched data against a single exercise model to see how many repetitions of the modeled exercise can be counted in that data. We could, however, apply the same data to all exercise models at once. Movement files were matched against all the models which were *not* from the same user. By looking at which model matches the data best, we could try to recognize which exercise was performed. To do this, we need a measure to express how well the data matches the model. We considered two options. Firstly, we calculated the correlation score as explained in Section 2.2.5. This gives us a correlation score based on how similar the duration, range and peak amplitude of the data's movements are to that of the model. Secondly, we considered the amount of counted repetitions when applying the model to the data. We assume that a model trained on exercise X would count more repetitions in a data file produced from that same exercise than it would in a data file produced from another exercise. Since it is possible that two models would result in the same repetition count, we used the matching score as a fallback in case of a tie between models. Both approaches performed equally well. With the counting measure, 53 exercise out of 82 were classified correctly, versus 50 for the score measure. See Table 2.3 for a confusion matrix of the recognition count results. $53/82 = 64.6\%$ correct is significantly above chance; random allocation would result in a performance of $100\%/8$ categories $= 12.5\%$ correct. Performance is way too poor for practical use, however. As we expected, misclassified exercises with a gravity effect are usually confused with other exercises with a gravity effect (eg. biceps and flye). Misclassified exercises without a gravity effect are usually confused with other exercises without a gravity effect (eg. bench press and ODP). Chang et al. [8] postulated that an additional sensor would be required (a belt clip in their case) to discern between bench press and ODP. These results support that postulation.

When matching movements against models from the same user, a recognition score of 100% is obtained. This is not surprising since only a few participants completed more than 1 exercise. The only choice for classification is the correct one, in that case. There is one participant, the experimenter, who performed all the exercises, however.

2.8 SUMMARY

We have discussed the various motion sensors available to the Android smartphone. We have chosen to use only data from the accelerometer. The smartphone is attached to the inside of the forearm by means of a custom made wrist strap, allowing the user to hold weights while using the device. We have explained the way in which we employ Kalman filters for preprocessing and have explained other components of the data processing pipeline. The way we match newly encountered data to exercise models can be described as dynamic rule-based decision. We have introduced the term *gravity effect*, which will be used throughout the remainder of this thesis. The main experiment will employ a repeated measures design with an interval of 1 week between blocks. This allows us to measure how much an exercise repetition varies over time without renewed instruction. It also allows us to see if feedback given by our app can incite participants to bring their performance level back to the standard that was set in the first week.

We conducted a small-scale experiment to get a rough estimate of the algorithm's performance. Repetition count results were promising, with a mean score of 96.5% over all exercises. Exercises with a gravity effect are easier to count, with the biceps exercise repetitions being counted perfectly and 1 repetition out of 80 being missed for flye. Performance on exercises with linear movements such as Overhead Dumbbell Press (ODP) and bench press was worse with a score of 94% and 89%, respectively.

Exercise recognition was sub-par when compared to contemporary research [30, 58, 36]. 64.4% of the performed series were correctly recognized, which is insufficient for practical use. Exercise recognition will not be further investigated in the main experiment.

The following chapter illustrates the results from the main experiment, in which the winning parameter combination from this pilot will be used.

3 RESULTS & DISCUSSION

Our algorithm for isolating and counting target movements from a stream of acceleration data was tested on a representative sample of gym attendees. We look at the performance of repetition counting and what factors might influence this performance, such as the type of exercise, interval between training and testing, and parameter values. We will also analyze the data to see how much the recorded profiles differ between participants and if there is a difference in the recorded profiles between participants who received high grades from trainers and those who received lower grades. After discussing the results we will suggest several ways in which our algorithm can be improved.

3.1 PARTICIPANTS

71 participants agreed to perform at least 1 free-weight exercise while wearing our SenseFit prototype. 16 of the participants in the experimental (device feedback) group and 25 participants in the control group (no feedback) returned a week later to complete the experiment. 63% of the participants in the experimental group and 60% of the participants in the control group was male. This difference was not significant. Age distribution did not differ significantly between groups either. The mean age in the experimental group was 27.6 years, and in the control group 30.2 years. Sd for both groups was 10.3 years.

The average experience is much lower for the control group (mean $= 1.8, sd = 0.90$) than for the experimental group (mean $= 2.38, sd = 0.81$). This is a significant difference ($t_{28} = -7.5, p < 0.001$). We do not expect this to be problematic however, because those with more experience did not perform in a more consistent way than users with less experience.

We might want to know whether both groups were instructed to do the exercises in the same manner. To test this, we compared the block 1 files, which were recorded after the exercises were explained, between groups. We compared the durations, start and end amplitudes and the range of the movements. All differences between groups were highly insignificant for BOR. For biceps the difference in mean range and duration of the backward movement approached statistical significance ($t_{29} = -1.05, p < 0.07$ for range and $t_{22} = 195, p < 0.09$ for the backward movement duration). Since no means were significantly different, we can say that both groups performed comparably during block 1.

Table 3.1: Repetition counting results. Results are shown separately for each exercise or both and the interval between profile recording and data matching. **Ntotal**: the amount of files tested. **Nmatched**: the amount of files from which a profile could be extracted. **Nperfect**: the amount of files for which the repetition count was exact. **score**: see main text. **sd**: the score's standard deviation

	Ntotal	Nmatched	Nperfect	score	sd
Short interval					
Biceps	82	79	65	98.2%	3.8%
BOR	76	68	41	91.3%	15.8%
Both	158	150	104	95.3%	10.7%
Long interval					
Biceps	82	76	52	95.3%	11.7%
BOR	76	71	29	85.2%	19.8%
Both	158	149	62	89.1%	16.5%

3.2 REPETITION COUNTING

In this section we discuss the algorithm's ability to count the amount of repetitions performed in a set. The performance could vary depending on different factors. We would like to see whether the amount of time between recording the profile and matching it against new data is of influence, whether the movement in the exercise in question is linear or rotational, and to what extent the chosen parameters influence the outcome.

Table 3.1 shows the repetition counting performance. The listed numbers are averages of participants in both the experimental and control groups. The reported data is grouped by interval. For the long interval, we used the data files from block 1 and block 4, which were recorded approximately 1 week apart. For the short interval, we used data from block 2 and 4, which were recorded withing a few minutes from each other. We used all pairs of data files twice: one as the model and the other as data file to match against, and vice versa. By comparing the $\mathtt{Ntotal}$ and $\mathtt{Nmatched}$ columns, one can see from how many of the data files a profile could not be extracted. A profile cannot be extracted when the signal-to-noise ratio is too low or when repetitions were paused halfway through a movement. The score measure is the same as used in the pilot (see page 42). It is the absolute difference between actual repetition count and the count produced by the algorithm, expressed as a percentage of the total amount of repetitions. For each row in the table, a distinct set of parameters was used. Parameters were optimized to the exercise shown in the first column. The chosen parameters maximized the score and minimized its sd. Note that the figures reported for 'Both' are not the averages over the biceps and BOR rows, but the results obtained with a single parameter set which performs best when applied to both biceps and BOR.

The overall score is 95.3% ($sd = 11.7\%$). Performance on BOR is lower than performance on Biceps, both in terms of mean and standard deviation. The difference is 6.9 percentage point (pp) for mean and 12 pp for sd.

INTERVAL Our algorithm performs better on files which were recorded shortly after another than on files which were recorded one week apart. The difference for biceps is 2.9 pp ($sd = 7.9$). This difference is significant ($t_{90} = 2.1, p < 0.04$). The differences for BOR are greater.

EXERCISE Repetitions are counted more reliably for biceps than for BOR. When using the overall best parameters, the score difference in pp (mean = 5.5, sd = 1.8) is significant ($t_{82} = 3.1, p < 0.004$) when model and data file are recorded shortly after another.

COMPARED TO PILOT To guard for an overfit to the data, we have used files from block 1 and block 4 both as model and test files. This setup is less powerful than k-folding cross-validation, for example. It might be informative to compare how the algorithm performs when using parameters from the pilot to process the main experiment data. Comparing the optimal parameters should also give an indication of robustness.

Although no demographics were recorded during the pilot, we can safely assume the populations are different. The pilot was conducted at a university sports center, so the participants were all of student age. Also, most of the pilot participants had moderate to extensive experience with weight training, where this was more evenly balanced in the main experiment. Also, the majority of participants was male. Since data and model files were acquired on the same day during the pilot, we should compare pilot data (Table 2.2) to the 'short interval' performance in the main experiment.

Where the biceps repetitions were all counted correctly during the pilot, the score is 97.5% ($sd = 0.4\%$) in the main experiment when using the pilot's parameters. For BOR, the performance on pilot data ($score = 98.3\%, sd = 3.9\%$) was better than on the main experiment ($score = 85.6\%, sd = 21.3\%$) as well. It seems that performance on exercises which are harder to count in one environment (BOR) shows higher degradation when moving to a more difficult environment than performance on easier exercises (biceps) does. When exercise-specific parameters are used, degradation is much smaller. For BOR, the performance is 99.0% on the pilot and 91.3% on the main experiment. This is a performance degradation of 7.7 pp for the exercise-specific parameters versus 12.7 pp for the overall best parameters.

COMPARED TO EARLIER RESEARCH Chang et al. [8] compared the perfomance of Naïve Bayes Classifiers (NBCs) with Hidden Markov Models (HMMs) on tracking the same free-weight exercises we used. They found that HMMs could not be used on single-user data, so we will compare our results with the results they achieved using NBCs. The reported scores range from 83.6% to 99.5%. These results were attained with parameters fitted separately for each exercise. When using data from all users, HMMs appeared to score comparable to NBCs.

For the fairest comparison, we should compare these results to our pilot data, which included the same exercises, and use parameters optimized for each exercise. Analysis shows that our scores range from 96.0% to 100%. The exercises on which the worst scores were achieved matched between our algorithm and the NBCs. Bench Press achieved the lowest scores and biceps the highest. Apart from biceps, our algorithm scored 100% on flye and deadlift. Our approach thus seems superior to NBCs or HMMs.

Pernek et al. [44] use Dynamic Time Warping (DTW) in combination with thresholds for pre-selecting repetition candidates. The results they report cannot be directly compared to our own because they use precision, recall, and F-score, neither of which we calculated. The reported F-score is 99%. Since no exercise specific parameters are used, this figure should be compared to our overall score of 95, 3%. It seems that DTW performs better than our dynamic rule-based decision. We cannot be sure however, as the F-score produces a more favourable figure than our performance score measure.

PARAMETERS Table 3.2 shows the parameter values used in both pilot and main experiment. The first thing that stands out is that during the pilot, the biceps repetitions were very robustly counted. Only the amplitude margin ratio affects the results. Not visible from the table is that parameters did not influence biceps results much in the main experiment either. The second observation we make is that the 'spacing', the minimum amount of milliseconds for which the direction of the signal must be different from the current trend before a peak is detected [1] , is quite high. This can be a problem in real-life situations because a spacing value of 300 ms means that feedback given to the user is delayed by an additional 300 ms. On top of the delay of about 50-100 ms introduced by the Kalman filter, that is. Given

[1] Let us say the signal changes direction from upward to downward. When minimumSignSpacing| is 150 milliseconds, the signal must continue to move downward for 150 ms before we assume that this is caused by a movement being ended. This is a way to eliminate local minima, or noise.

Table 3.2: Optimal parameters per exercise for the pilot and main experiment. Main experiment results are shown separately for training and test data recorded on the same day (short) and one week apart (long). **process** and **sensor** are noise parameters for the Kalman filter. **spacing**: Minimum amount of time between peaks (ms). **zeroDthr**: absolute threshold for the instant derivative below which a signal is considered steady (m/s^2). **ampMargin**: the ratio between model and candidate movement within which the candidate is accepted. * indicates that any value produces equally good results, + indicates that the preceding value and any higher value produces identical results.

	process	sensor	spacing	zeroDthr	ampMargin
Pilot					
Biceps	*	*	*	*	0.4+
BOR	0.01	6.4	150	0.05	0.6
Overall	0.02	0.8	50	0.04	0.5
Main Short					
Biceps	0.01	0.4	100	0.05	0.5
BOR	0.02	0.8	250	0.05	0.5
Both	0.01	0.2	200	0.03	0.5
Main Long					
Biceps	0.01	0.05	300	0.05	0.6
BOR	0.01	0.4	300	0.04	0.6
Both	0.01	0.4	300	0.04	0.7

a movement duration of 1–1.5 seconds, this would be unacceptable. Note that in practice this has not been a problem since a value of 50 has been used for both pilot and main experiment. Why the spacing value of the optimal parameters from the pilot is 50 ms is not immediately clear. Possibly because a high spacing value has the potential to partially discard target movements which contain noise around their peaks. But, more importantly, a low spacing value has the potential to break up a target movement in two separate movements when a little noise is introduced. The pilot, unlike the main experiment, included 8 different exercises. Many of the linear movement exercises produce quite a noisy raw signal. We would thus expect a large spacing value in the pilot's overall best parameter combination. When looking at exercises such as deadlift and Overhead Dumbbell Press (ODP), optimal values are 200 and 250 respectively. Exercises with a rotational movement are quite robust and do not suffer from high spacing values. Of course, the effect of one parameter on performance is not independent of other parameters.

The sensor noise and process noise values determine the way in which the raw signal is filtered. An interesting observation is that there is a very strong linear relationship between these two parameters for equal scores. That is, when the process noise is set to 0.01 and sensor noise is set to 0.02, the resulting signal is the same as when we set process noise to 0.1 and sensor noise to 0.2 (sensor $= 2 \times$ process). To test this interaction effect, we calculated 'multiplier' as sensor noise divided by process noise. We fixed all other parameters, to eliminate confounding factors. A one-way ANOVA showed a strong effect of multiplier on score: $F(9, 21) = 2147$, $p < 0.001$, $MSerror < 0.01$.

If we used a static model (Equation 2.1) for both the process and sensor noise, there would be absolutely no difference in the filtered signal when the ratio of both parameters is fixed. This is because Kalman Gain is constant in such a case (Equation 2.8). This implies that sensor and process noise are redundant parameters when a static model is chosen for both. Given the very small variance in score for equal ratios in our data, the same can be said for our scenario, in which a linear derivative model is used for the process noise and a static model for noise.

$ampMargin$ was originally used as a threshold ratio for the movement amplitude. In contrast to the other parameters discussed, this parameter is not used for profile extraction, but for matching a profile against incoming data. When the amplitude of a movement is within $ampMargin \times range$, the movement would be further evaluated, and discarded otherwise. $range$ being the absolute difference between the start and end amplitude in a model. For the duration threshold, a fixed value of 0.33 was used. Early tests showed that the algorithm performed better when using $ampMargin$ for all thresholds. A value of 0.5 means that the movement's duration must be within a ratio of 50% of the mean duration in the model. When comparing values for the short versus long interval, we see that higher values tend to perform better on long interval matching than on short interval matching. This is to be expected. Two data files which were recorded shortly after each other tend to contain movement data which are more similar than data files which were recorded a week apart.

3.3 PROFILE VARIANCE

In this section we answer the research question:

> Is user-specific calibration required to reliably assess performance?

When the exercise profiles show little variance between subjects, we can say that these profiles can be universally applied. When they do significantly differ from each other, there are two possibilities. Either some of the profiles do not actually represent an ideally executed exercise, or the ideal execution significantly varies per person. To exclude the former, we will only use data which trainers scored with an 8 or higher on a 10-point scale to build our average profile. We used the files from block 1 as models and the files from block 4 as test files and the corresponding optimal parameters.

Let's first look at the variation in profile properties. We consider the amplitude of the signal when an exercise repetition is started ($startA$), the amplitude when the forward movement is completed ($peakA$), the difference between $startA$ and $peakA$ ($range$) and the durations of both movements. Are there differences between properties for participants who scored below 8 out of 10 in a professional's judgment versus people who scored at least an 8?

Surprisingly, none of these properties were significantly different between high and low scoring participants for the biceps exercise. $startA$ differed the most for biceps. The high scoring group produced a higher $startA$ ($mean = -6.3\,m/s^2, sd = 1.8$) than the lower scoring participants ($mean = -7.2\,m/s^2, sd = 1.6$). This corresponds to an elbow rotation of only 8° [2]. This difference is not significant ($t_{37} = 1.6, p < 0.12$). For BOR, we found one significant difference: the range. The high scoring group produced a range of smaller magnitude ($mean = -3.3\,m/s^2, sd = 1.2$) than the lower scoring participants ($mean = -4.3\,m/s^2, sd = 1.2$). This difference is significant ($t_{33} = 2.15, p < 0.02$). A range with lower magnitude indicates a more controlled movement. Levene's test for equality of variances indicated no significantly different variances for any of the properties.

Since properties are not significantly different between high scoring and low scoring groups, it seems that trainers base their judgment on various additional factors, which we do not extract from the signal. Some cannot be extracted, such as whether one performs the BOR with a straight back. One of the trainers noted that this is very important to prevent injury. Whether users perform the BOR with a hollow or straight back does not influence the movement of the arm, however, and we cannot detect it.

The difference in profiles is small between high and low scoring groups. What about individual differences? The mean of the $peakA$ is $5.2\,m/s^2$ ($sd = 1.4\,m/s^2$). The sd seems problematic because this means that to capture 95% of variation you have to recognize $peakA$'s between 2.4 and $8.0\,m/s^2$, a range corresponding to a 51° angle. But

[2] For an exercise with a gravity effect, gravitational acceleration is the major component in the recorded signal. When the axis is perpendicular to the ground, the reading is approximately $9.8\,m/s^2$. When the axis is parallel to the ground, this is $0\,m/s^2$. Since the relation between angle and acceleration is linear, a difference of $0.9\,m/s^2$ corresponds to $0.9/9.8 * 90° = 8.2°$

this is in fact no problem when considering our matching criterion. Our algorithm only matches a movement when:

$$
\begin{gathered}
|MpeakA - PpeakA| < |R \times Prange| \\
\& \\
|Mrange - Prange| < |R \times Prange| \\
\& \\
|Mduration - Pduration| < R \times Pduration
\end{gathered}
$$

Here, variables preceded by P denote the values in the profile and M denote variables in the movement which we are trying to match. The margin ratio R is set at 0.7 since we use the optimal parameters for long intervals between model and test file recordings. $Prange = PpeakA - PstartA = 5.2 - (-6.9) = 12.1$. So the threshold for the difference between the movement's $peakA$ and the profile's $peakA$ is a huge $0.7 \times 12.1 = 8.5\,m/s^2$, or 6 standard deviations. For biceps, the criterion for the duration of the backward movement is the most strict. The mean is $1453\,ms$ ($sd = 311ms$). Expressed in standard deviations the allowed deviation from the profile is $(0.7 \times 1453)/311 = 3.3\,sd$. The criterion is quite broad, which gives us a likely explanation of why the exercise *recognition* did not perform that well, since a lot of false positives may occur when testing a profile for exercise A on data from exercise B.

When the profiles are not significantly different between individuals or high and low scoring groups, one would expect that an average profile should fare quite well. We constructed biceps and BOR profiles which are averages of those which were scored an 8 or higher, so that we know these profiles are representative of properly executed exercises. We compared the performance of these average profiles against block-4 data for each participant. The personal models yielded an average score of 87.2% ($sd = 20.0\%$) [3].

Our algorithm makes a distinction between the primary axis and non-primary axes. Signals in the primary axis are subject to the above matching criterion while non-primary axis values only have to lie within 2 sd's from the mean of the whole recording session. At first the average models yielded a score of 76%. Inspection of files on which the algorithm counted hardly any repetitions showed that the distribution of these non-primary axis varied greatly. By relaxing the threshold for these axes, the average profile yielded an average score of 82.1% ($sd = 22.1\%$). This is not significantly worse than the results for personal models ($t_{74} = 1.8, p < 0.08$).

[3] The alert reader might notice that this percentage differs from the 89.1% listed in Table 3.1. This is because for those results, files from block 4 were also used as model files.

Since performance using an average profile was not significantly worse than when using personal profiles, we conclude that it is not necessary to use personal profiles for repetition counting. The only thing we did to adapt the algorithm for average profiles was relaxing the thresholds for non-primary axes, which yielded an improvement in score from 76 to 82 percent. We feel that we can narrow the gap with the personal profiles's score another percent or two by further analysis.

We do advise to maintain the ability to record a personal profile however, since an average profile is probably only suitable for physically fit users. People hindered by physical disabilities may need to record a personal profile under supervision of a physiotherapist, to avoid injury and because target movements would likely not match an average profile. We also stress that average profiles can be safely used for repetition *counting*. That the counting performance is not significantly different between average and personal profiles does not mean that the profiles themselves are comparable too. Remember that the spread of starting amplitudes within 95% of the startA distribution represented a 51° angle. When using profiles to give advice about how far one should extend the elbow, the advice may not be accurate for that particular person.

3.4 FURTHER RESEARCH

For our experiment, we asked participants to choose a weight that they could comfortably lift multiple series of ten repetitions. It would be interesting to see in what way signal characteristics will change when a profile is recorded using a certain weight, and is tested on a file recorded with higher or lower weight. Over longer periods of time it would be interesting to see how the signal changes when the weight is kept constant while the user's strength increases. Possibly, an indicator of difficulty can be based on these characteristics, and advice could be given about adjusting the exercise program by adding weight or increasing the amount of repetitions. Another way in which robustness to everyday variations could be tested is by training a model where users lift with both hands and test on a case where the user is alternating between left and right. Also, our prototype was always attached to the same side of the body for the duration of the experiment. It should be interesting to see how the algorithm could adapt to usage on both arms while recording only one profile, and how the used arm can be detected. For many exercises, we expect signals to be similar between left-handed and right-handed use, such as the biceps and BOR. For triceps however, the x direction will be mirrored when changing hands (see Figure A.1 and Figure 2.2).

Pernek et al. [44] tested their smartphone-based prototype in a variety of scenarios. They also tried placing the smartphone on the weight stack of a workout machine and achieved good results. This is a controlled scenario since the direction of travel is strictly vertical. A disadvantage could be that the acceleration values can be quite low when the distance traveled is small due to mechanical transmissions. We would like to know how our algorithm fares in this scenario.

Parameters have been pre-set to the values which are optimal for all exercises under consideration. Performance would have been better if we used parameter values which are optimal for a specific exercise and probably even better when parameter values are tailored to individual sessions. From the pilot data we know that the more noisy signals benefit from stronger smoothing. Perhaps Fast Fourier Transform could be applied to increase smoothing strength when high frequencies are detected, or when the spectral distribution is wide, indicating a high signal to noise ratio. This was previously suggested by Chang et al. [8].

While we were adapting our algorithm to work with an average profile per exercise instead of a single profile per participant, we achieved good results by relaxing the matching criterion for non-primary axes. It may be that there was not only a large variance in these axes between subjects but also within subjects, for example when comparing files from block 1 versus block 4. In that case, performance on personal profiles may also be improved (albeit to a lesser extent) when applying this looser matching criterion.

3.5 SUMMARY

In this chapter we have discussed the results from our main experiment. Our algorithm did fairly well on repetition counting. There was a significant difference of interval on the repetition counting score. When an exercise profile was tested on unseen data from the same participant on the same day, performance scores between 91 and 98% were achieved. Judging by the results reported by [8], this indicates that approach performs better than NBCs and HMMs. When compared to DTW, our algorithm seems to perform worse, although our results cannot be directly compared to the results in [44].

When the training and test moments were a week apart, performance was significantly lower at 89 to 95%. The BOR repetitions were harder to count than those of the biceps exercise. When taking pilot results into account, we conclude that exercises with linear movement are harder to count than exercises with a rotational movement. We also tested how robust the performance is to changing environments while keeping the same parameter set. When using the parameters of our pilot study on the main experiment data, degradation of performance seemed dependent on the exercise. Exercises which are harder to count in one environment are harder to count in the other, and performance degrades faster.

When calculating the average of all profiles recorded for each exercise and using that to count repetitions, performance is not significantly lower for average profiles than for personal profiles. We are reluctant to say that average profiles are sufficient for other purposes such as giving advice on the extent to which one should extend or flex. The spread in start and end angles for the biceps is quite high, for example.

As directions for further research we have suggested several everyday variations on the controlled way in which our prototype was tested, such as varying weights and right- versus left-handed usage. We would also like to see if it is possible to determine processing parameters based on the signal characteristics of a particular session, rather than have pre-set parameters. The decision for which parameters to use might be based on the FFT of the incoming signal.

In the next part of this thesis, we will look at how the matched movement data can be used to provide people with feedback on their performance during free-weight exercise. Feedback will be given not only on the repetition count, but also on tempo and the range of the movements.

Part III

INTERFACE

In which we answer the research question

“How should feedback about fitness exercise performance be designed?”

4 DISPLAY PILOT

In the previous part, we explained the algorithm used to process captured acceleration signals into a prototype of a fitness exercise. We also explained another algorithm to isolate repetitions of that exercise by comparing incoming data to this prototype. Characteristics such as duration and the extent to which a person flexes and extends are recorded. When the recorded prototype is considered a golden standard, this information can be used to provide the user with corrective feedback. For example, when the first movement in a prototype has a duration of 1200 ms and a corresponding movement is detected with a duration of 900 ms in a live signal feed, we would like to advise the user to perform this movement more slowly.

The subject of this part of the thesis is our second research question:

> How should feedback about fitness exercise performance be designed?

As explained in the introduction to this thesis, it was difficult to find earlier research that provided clear guidelines or paradigms to base our interface on. This is because interface design still takes place within the Windows, Icons, Mouse, and Pointer (WIMP) paradigm, or uses a system's main modality of in- or output. Smartphone interfaces rely heavily on the touch screen for input, while speech is also an option. The touch screen is the main modality for output too, while sound, speech and vibration are available as well. Most research on interface design assumes that the user is interacting with the system using his or her full attention. The smartphone's screen is not visible to the user while most exercises are performed. Live visual feedback is unfeasible for all exercises except biceps, for which the screen is visible at the end of the forward movement. Considering less often-used feedback modalities is thus a necessity.

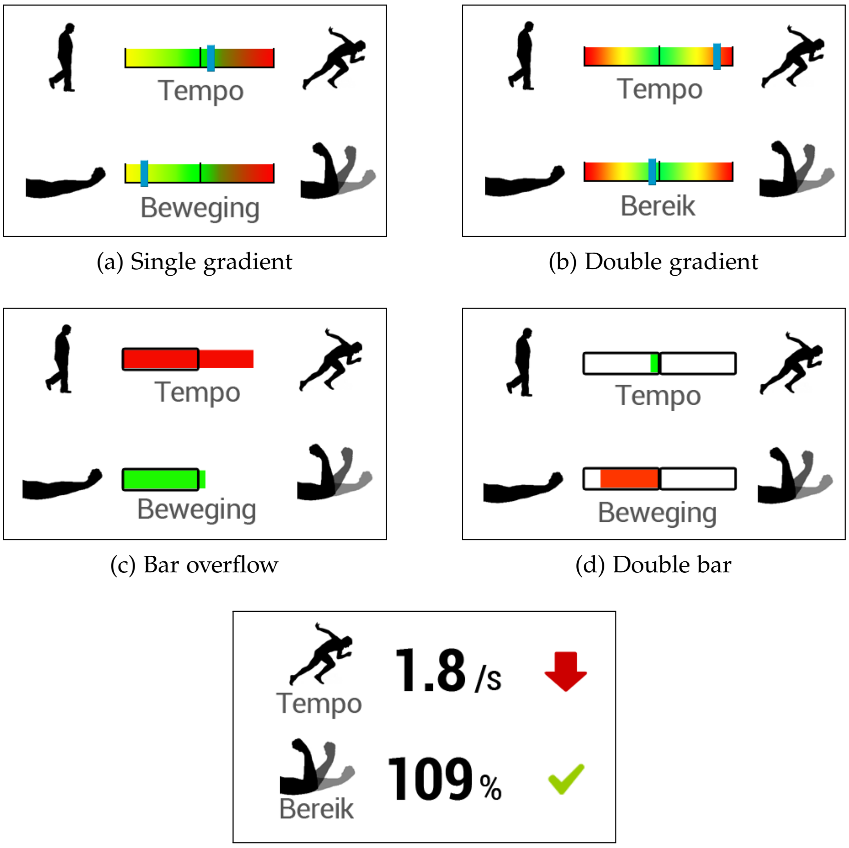
Despite this, we will still use visual feedback to present a performance summary after each set of 10 exercise repetitions. To determine which visual metaphor is best to represent the user's performance on the exercise properties tempo and range, we conducted a pilot. The main questionnaire, described in the next chapter, is a follow-up on this visual feedback pilot which also covers preferences for other feedback modalities and the form factor of the device.

4.1 methods

Because the display of our prototype will not be visible during exercise, we focused on a display design that informs users about their performance each time they complete a set. While resting, people can review their performance on the completed set and reflect on how to adjust their movements for the next set. We chose not to use the approach taken by Pernek et al. [44], which listed each repetition number along with a cross or check mark and a short comment on the speed of the exercise. We expected that it is more informative to report averages over series rather than per repetition because people may not remember at what speed they did each individual repetition. Since the screen is visible when performing the biceps exercise, we added a simple repetition counter to the top of our display. It represents a division of current repetition and target repetition count. The text 6/10 indicates that the user is working on the sixth repetition out of a total of ten. The current repetition is printed in a larger font size to attract attention and increase legibility.

From a short interview with our fitness instructors we learned that people often perform the forward and backward movements at the same speed, while in fact the backward movement should be performed at a slower pace. We decided to split the display in two parts, one for each movement, so that feedback could be more specific. To support recognition, we used an underlined section header as prescribed in the official Android design guidelines [24]. For each movement we have an indicator for tempo and the extent of the movement. The latter correlates with the peak amplitude of the movement signal. For biceps, this is linearly related to the rotation angle of the elbow. For tempo we used a silhouette of a running man as a metaphor and for movement extent an arm with slightly bulging biceps. Because bulging muscles might be associated with power rather than motion, we accentuated the idea of motion by adding two forearms at intermediate angles with lower opacity. To indicate how the user scores on these measures we made a few variations which were compared in an A/B usability study.

Our first alternative uses a gradient from yellow through green to red, see Figure 4.1a. On both sides of the gradient bar, a depiction of the two extremes of the property in question is placed. A blue rectangle indicates the user's performance on each property. The example in Figure 4.1a shows a state where the extent of the movement is much too short. Depending on which movement is depicted, the user should either flex or stretch further. Figure 4.1b is similar but uses a green-yellow-red gradient on both sides. We thought that red might be a stronger signal color than yellow, which might lead users to believe that stretching too little is less harmful than stretching too



(a) Single gradient

(b) Double gradient

(c) Bar overflow

(d) Double bar

(e) Text with arrows

Figure 4.1: Five different display designs used in the display usability study. a) uses a yellow-green-red gradient to indicate a scale from too low to too high, where green is centered in the display and indicates the optimum. A blue rectangle indicates the user's performance. b) uses a double gradient to indicate that both extremes are undesired. c) uses the overflow metaphor, where a black outline indicates the optimum level and a bar indicating the user's performance can over- or underflow this bar. d) uses two bars of which one fills from the center outwards. e) shows the values corresponding to the properties. Values larger than 100% are too high and under 100% too low. The word 'bereik' and 'beweging' are Dutch for range and movement, respectively.

much [1]. Figure 4.1c is inspired by the bar overflow metaphor used by Michels et al. [32]. An outlined bar serves as a reference for the ideal value. A colored bar either under- or overflows this reference horizontally. In Figure 4.1c, the exercise is performed at a pace which is too high. The next mockup shows a double bar design in which either bar fills up from the center outwards. In case of an ideal state, the bars are empty. When one does not flex enough, the left bar fills up, as shown in Figure 4.1d.

Inspired by the Concept 2 rowing machine (Figure 4.2), we also included a display that shows the measured values directly (Figure 4.1e). We do not expect this approach to be very popular with our respondents since the numbers in themselves are not intuitive. An additional difference with the other designs is that we included a status icon which could either be a check mark for acceptable state or a downward or upward facing red arrow for too high or too low values, respectively.

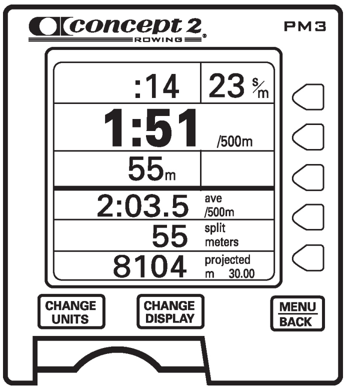


Figure 4.2: The concept2 textual display

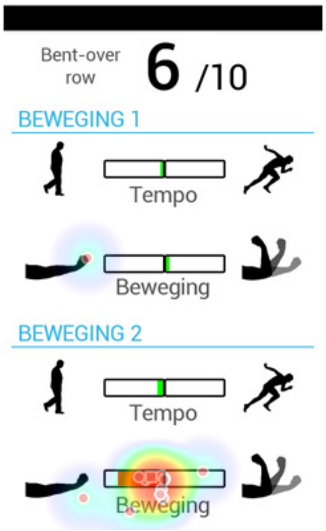


Figure 4.3: Usabilla sample heat map.

We use the Usabilla usability testing tool [59] to determine which of these mockups will be developed and included in our main experiment. Usabilla allows researchers to upload images and give respondents a task that can be performed with the mouse. The results can be downloaded in the form of heat maps. One of the tasks we included was 'click on the spot where you see a deviating value'. A sample result is shown in Figure 4.3. We pitted these designs against each other by asking the participants to click on the design they thought was most clear. We also pitted the single gradient versus the double gradient designs and the bar overflow versus the double bar designs. Respondents were recruited by sharing a link via Twitter and Facebook on both the business account of a mobile software developer and personal accounts of its employees.

[1] This might be a correct assumption, but we want to motivate the user to perform the exercise perfectly, and want to discourage under- and overstretching in equal measure. This should also help with processing the data from our main experiment.

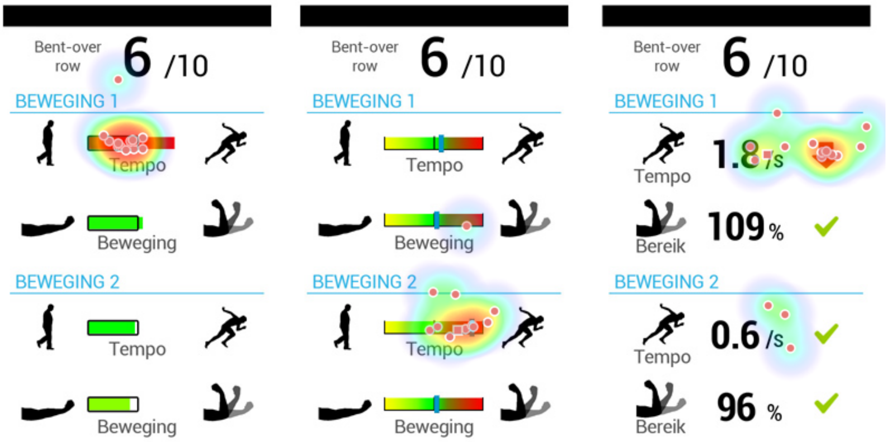


Figure 4.4: Result for the task 'Click on the spot where you see a deviating variable.'

4.2 RESULTS & DISCUSSION

The number of respondents was limited to 20 by the pricing plan of our testing tool. After discarding one dummy respondent used for testing, 19 actual respondents remained. These respondents were recruited via Twitter and Facebook and were all Dutch, but more specific demographics were not recorded. The ratio of men and women is about equal for these social media in the Netherlands, with Facebook having slightly more female members where Twitter has slightly more male than female members. 76% of Twitter users is younger than 30, while about 72% of Facebookers are below 45 [3, 56].
The first task we gave the participants was to indicate the area on the display where a deviating value could be seen. A simple task, included to get participants accustomed with the unorthodox procedure. All participants completed this task successfully.

One interesting result is that for the textual display, only 3 participants clicked on the measurement value while the remainder clicked on the red arrow alongside it, see the right-most display in Figure 4.4.

We should note that the concept of the bar overflow design might not have been clear to the participants, since the deviating value in the presented display underflows, see Figure 4.5. The concept might have been more clear when we had used the example in Figure 4.1c, which overflows

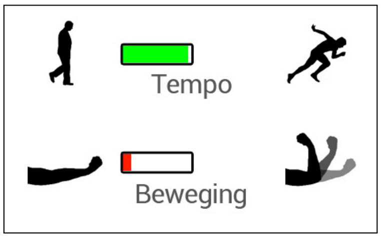


Figure 4.5: The bar overflow display used in the display pilot.

[2]. Another reason for the bar overflow design not receiving a favorable amount of clicks was that the display is not symmetrical in the ideal state, which might make it less clear or pleasant to look at.

When combining these findings, we can say that a single gradient design is preferred over a double gradient design. Overall the numerical display was preferred, but this seems due to the status indicator at the right side of the display, which is very glanceable and provides enough information in itself to correct one's movements.

Because it is still unclear whether the double bar is preferred over the single gradient or vice versa, these two designs will be pitted against each other in the main usability study, described in the next chapter. They will be combined with the status icons that proved popular in this study, after which we can determine the final design, which we will use in the main experiment.

[2] Figure 4.1c *was* actually used in this study, in one of the 'click the deviating value' questions, but we can assume that the participants were less engaged with the metaphors while performing that task and used a simple strategy of finding the red spot in the display.

5 MAIN USABILITY STUDY

We have three properties of an exercise about which we want to inform the user: repetition count, duration (tempo), and the range of the movement. For each of these properties, we wanted to know which of the available feedback modalities users prefer. To answer this question, we conducted an online questionnaire. Because the results of the display design pilot indicated that either a single gradient or double bar metaphor was preferred for performance feedback, these two concepts were further investigated. Because we want to explore the possibilities for turning our prototype into a marketable product, we also inquired about preferences for the type of device in which our technology was to be implemented. This led us to subdivide our questionnaire into six parts: general, counting, execution (form), tempo, device, and display. The 'general' part of the questionnaire contained questions about the demographics gender, age, and fitness experience.

The procedure for the main experiment has been described in Section 2.3.2. At the end of the experiment, users in the experimental (those who receive feedback) group are given a post-task questionnaire in which they are asked how useful each of the feedback types was to them. We are interested to see whether the answers to these questions are comparable to the results from the main questionnaire.

Since feedback for improving exercise performance was given after block 3 and during block 4 of the experiment, we naturally want to know whether this feedback was picked up by participants and resulted in improved performance. We will both compare duration and acceleration data per series (block 3 versus block 4) and per repetition.

5.1 QUESTIONNAIRE METHODS

TARGET AUDIENCE Our application could be useful to novices and could be used outside the gym. Despite this, we decided to select only people who have at least some experience with free weights for our study. They should have a better feeling of whether any sound produced by the device is annoying to their environment, or whether a device attached to the arm would hinder execution of the exercises. We did not use any additional criteria for selecting participants. For the Dutch formulations of the questions, see Appendix B.

MODALITY PREFERENCE The following questions were included for each of the counting, execution and tempo aspects, where care was taken to maintain the same style and tone.

We do not immediately ask respondents what modality they prefer, to prevent negative responses from people who are not interested in receiving feedback from our app. First, we pose the question of whether one succeeds in completing the exercise properly with regard to the aspect in question. We posed this question in a positive manner, for example: 'Do you succeed in maintaining the proper tempo throughout the exercise?'. We expect that people will be more comfortably with admitting that they occasionally make mistakes than when we would ask them if they are having difficulty completing an exercise at the proper tempo. Those who indicate to at least occasionally having difficulty with an aspect were given a simple follow-up question inquiring what the reason for this trouble was.

Next, we asked whether they would appreciate it when a device would help them with any of these aspects. Only if the answer was 'yes' the final question of which feedback modality was most suitable was presented.

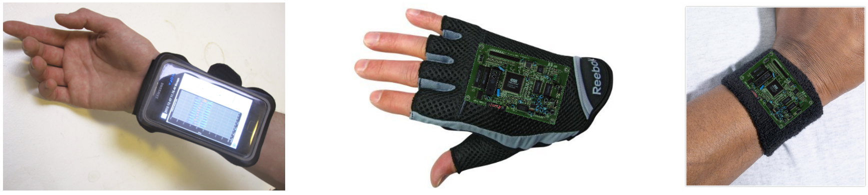
The main question of which feedback modality was most useful for a particular feedback aspect was asked in the form of a set of Likert questions. We used a 7-point scale. On the left side the label 'not useful' was placed and on the right side the label 'useful' was printed. Note that the use of an uneven number of bullet points allows for a neutral response. The five modalities were:

- A short beep
- A symbol or number on the screen
- Voiced advice
- Vibration similar to that produced by a cellphone
- An on-screen summary after completing the exercise

The wording of these modalities were adapted to the feedback aspect. For execution, the wording of the beep modality was 'A short beep, every time you do the exercise. The pitch of the beep tells you whether you performed the exercise correctly or incorrectly.'. For counting the phrase was 'A short beep, every time you do the exercise.'.

DEVICE As explained earlier, our prototype consists of a smartphone strapped to the arm with a screen on the inside of the forearm. A disadvantage is that the screen is mostly invisible during exercise. One idea is to separate the sensor from the rest of the device so that the screen can be fixed in an always-visible position [13].

We included a question concerning which of two possible alternatives to our prototype would be preferred, if any. The first alternative

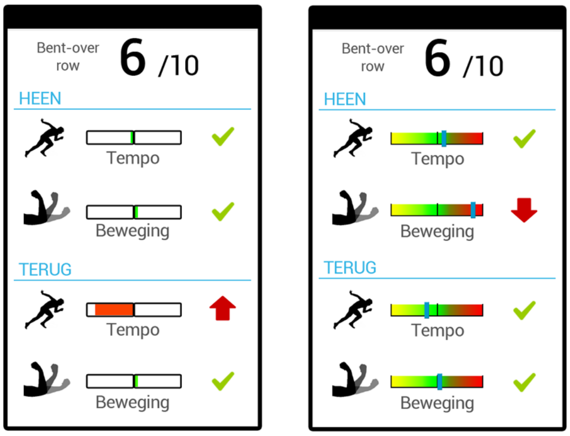


(a) Wrist band with device

(b) Workout glove with sensor

(c) Sweat band with sensor

Figure 5.1: The three device configurations included in the main questionnaire.



(a) double bar

(b) gradient

Figure 5.2: Questionnaire A/B study. The double bar (a) was compared to the single gradient(b). Both displays have status indicators.

is a workout glove with a sensor attached to it (Figure 5.1b). This is similar to the prototype used by Chang et al. [8]. A practical disadvantage to using this device configuration would be that the wrist has many more degrees of freedom than the forearm, resulting in less-predictable signals. Our second alternative (Figure 5.1c) does not suffer from this disadvantage. We asked respondents to indicate their preferred device configuration, with an option to suggest an idea of their own.

To estimate how intrusive the production of sound would be, we asked respondents whether they thought any sound produced would bother others. We also investigated whether wearing headphones or earbuds would be a problem.

DISPLAY From the display design pilot described in Chapter 4 we know that ticks and arrows are preferred over raw values as indicator of state. A question which remained unanswered is whether a double-bar design is more suitable than a gradient bar as a more precise indicator of how one should correct the exercise to perform it perfectly. We included these designs side by side. Figure 5.2a shows the double bar design. The ideal state is indicated by two empty bars with a black outline. Abnormal values are indicated by a filled bar, where the left bar indicates 'too low' and the right bar indicates 'too high'. Bar color is a redundant indicator of the severity of the deviation from the ideal state. Figure 5.2b shows the gradient bar. A continuous gradient scale is used to indicate values which are too low (yellow), normal (green) or too high (red). A blue rectangle superimposed over this scale indicates the current state. A black vertical line in the center is added as a reference point for the ideal state.

Firstly, people were asked what the red arrow means in this display. Because we expected the represented property (tempo) in display Figure 5.2a to be unambiguous, three options were given: 'I am too slow, I need to speed up', 'I am working at a proper pace' and 'I am too quick, I need to slow down'. In Figure 5.2b , however, the deviating property is the extent of the forward movement. Because we were unsure that the accompanying icon and label successfully conveyed the meaning of movement extent, we decided to not provide any options and include an open-ended question.

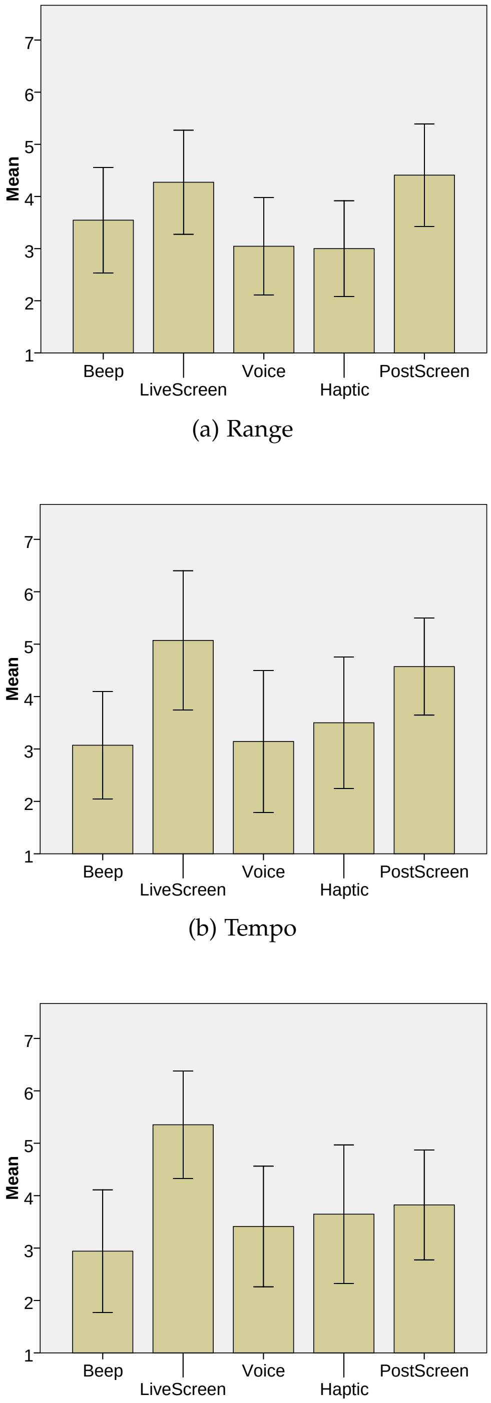
For each of the two displays, two Likert-scale questions were included. One about whether the screen is easy to understand and the other about whether it is possible to quickly determine how the performance of the exercise can be improved (whether it is 'glanceable'). Finally, as an overall indicator of preference, respondents are asked to indicate which of both screens they would rather use.

5.2 QUESTIONNAIRE RESULTS

PARTICIPANTS Out of 41 respondents, 31 were male. Participants are quite young (mean = 25.0, sd = 3.2 years). 29 had experience with free-weight exercise, 24 had experience with cardio training and 1 person was unsure of what kind of training he has experience with.

MODALITY PREFERENCE Modality preference results are shown in Figure 5.3. Because of an unfortunate error in the questionnaire text, the question of which feedback modality is preferred for range pertains to tempo. Therefore, the results for range preference have to be discarded. In the remainder of this section we will discuss the feedback preferences for tempo and repetition counting.

The overall impression is that there is a preference for visual feedback over the other types of feedback. Two types of visual feedback were included. 'LiveScreen' refers to a scenario where indicators such as those in Figure 4.5 represent the user's performance on the previous repetition, and is thus updated after every movement. 'PostScreen' refers to a scenario where the same indicators are presented at the end of a set and represent the user's average performance over the set.



(a) Range

(b) Tempo

(c) Repetition count

Figure 5.3: Feedback modality preferences. Anwers were given on a 7-point Likert scale. Because of an error in the questionnaire text, in the ‘range’ section of the questionnaire the question pertains to tempo. As a result, none of the modality preferences significantly differ between a and b. c) Feedback preference for repetition count.

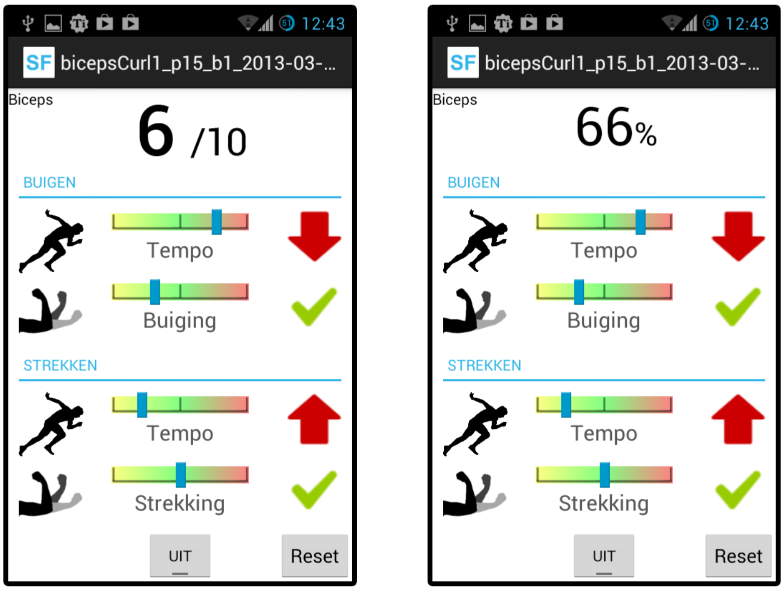
We averaged the Likert-scores of 'LiveScreen' and 'PostScreen' into a 'Screen' variable and the other three modalities into a 'NoScreen' variable. A paired-samples t-test showed a significant preference for 'Screen' over 'NoScreen' for feedback on tempo ($t_{13} = 3.2, p < 0.01$), but not for feedback on repetition counting.

Looking more specifically at individual modalities, feedback for repetition count is deemed more useful during the exercise rather than after it ($t_{16} = 2.7, p < 0.02$), which we feel is only logical. Contrary to our expectations, voiced feedback does not receive significantly lower appreciation than haptic feedback does. We expected haptic feedback to be less intrusive and therefore preferable. Also, out of 34 people who answered questions regarding sound annoyance, 10 indicated that sound produced by a device would certainly annoy others, and another 22 indicated that it might annoy others when sound volume is higher than ambient sound (such as the gym's background music). Furthermore, 7 out of 34 would not want to wear headphones or earbuds during their exercise.

DISPLAY PREFERENCE From the display pilot we know that a single green-yellow-red gradient is preferred over a double gradient as a performance scale. The 'double bar' is preferred over the 'bar overflow' concept (Figure 4.1). We also found a strong preference for a three-state performance indicator indicating a value which is either too low, normal, or too high. To determine our final design, the single gradient and double bar were both combined with a three-state indicator, so that the only difference in the design is the scale indicator (Figure 5.2). The three-state indicator may be ambiguous, since a red arrow which points downward could either mean that the corresponding value is a) too low, one should try to bring it up, or b) too high, and one should bring it down. Another thing that could be ambiguous is the exercise property expressed by the icon on the left of the screen in combination with the text underneath the scale. Because the label 'tempo' (which has the same meaning in Dutch as it has in English) in combination with the silhouette of a running man is a very strong description of the (inverse of) movement duration, we used a multiple choice question about the meaning of the red arrow in Figure 5.2a. The arrow proved to be very clear. 30 people thought that the tempo was too low, 1 thought it was all right, and 2 thought it was too high. The first interpretation is the one we chose for our final design. We were unsure whether the metaphor for movement extent was just as clear. Therefore the open question 'What do you think the red arrow in [Figure 5.2] means ?' was added.

The metaphor for movement range was indeed ambiguous. 20 out of 33 respondents gave an answer which was related to tempo. 9 gave an answer which was related to flexing or extending muscles. 4 gave a different answer, mostly related to load or weight.

Clearly the representation of movement performance had to be changed. We decided to specify separate labels for biceps and Bent-over Row (BOR). The label for the first movement of the biceps exercise was 'bending' while it was 'upwards' for BOR. Also, where questionnaire respondents received scarcely any explanation of the interface ('this display shows fitness exercise performance'), the interface was explained to participants of the main experiment before they could use it, and we made sure the concepts were clear. The final display is shown in Figure 5.4.



(a) Live feedback

(b) Summary screen

Figure 5.4: The final interface used during the main experiment. a) The feedback screen during exercising. Values represent performance on the previous repetition, and a repetition counter is shown at the top. b) Summary display representing the average performance over one set. The percentage shown at the top of the screen is a correlation score between golden standard and current values.

FORM FACTOR The prototype we used uses the smartphone's display. The advantage of this approach is that it is a self-contained device. No connections with sensors or eternal displays have to be made. The disadvantage, however, is that the display is not constantly visible. For the BOR exercise in particular, the display is invisible during performance of the exercise. We were interested to know whether there was a preference for such a self-contained device, or that people would rather use a system for which the sensor, processing unit and display are separated. Examples of accessories in which an acceleration sensor could be embedded are watches, sweat bands and workout gloves. The majority of respondents (26 out of 33) preferred an approach with separate sensors.

5.3 AUDITORY FEEDBACK

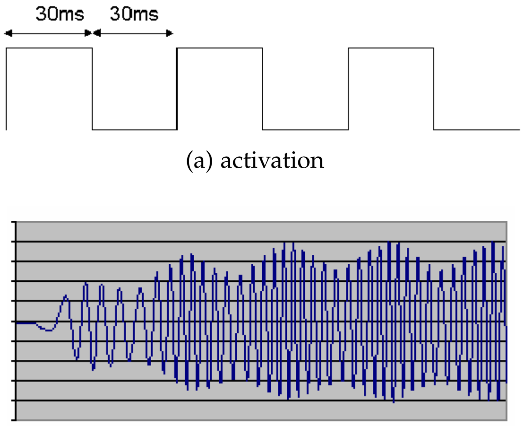
Auditory feedback can be provided through the headphones which are attached to the smartphone during the experiment. Initially we considered variations in the interval between two tones, variations in pitch, or playing a reference tone followed by another tone which is altered in either pitch or volume, or the interval between the reference tone and the cue tone is varied. Many of these concepts are used by Crease and Brewster [10] in their auditory progress bar, where a reference note of C_2 (65 Hz.) is played followed by a note which starts at C_4 (261 Hz.) and moves towards C_4 (130 Hz.) as the task progresses. For our application, the main problem would be that there would be insufficient time to present these tone combinations. Let's say we present a tone combination where the difference in pitch indicates the extent of the movement performed. We can only start playing this auditory feedback while the user is performing the next movement, which could be confusing. Furthermore, these 'earcons' can be hard to learn [12], which could cause an interference between the task of performing the exercise and interpreting the feedback. We decided to play a note at the moment when users *should* have finished their repetition. This way, the tone is the reference against which the user's movement is compared. By using only one tone, distraction from the exercise is minimized. To indicate a completed set, we played two trumpet chords after the tenth repetition has been completed [1].

In our first implementation, we used a chromatic scale to indicate progress, where for every repetition the next note in the scale would be played. Both tempo and repetition count were represented with a single note this way. After some preliminary tests we were afraid that the increasing pitch would entice users to perform the exercise with an ever increasing tempo until the trumpets were finally played. To prevent this, a single note of the same pitch was played every repetition in the final implementation.

5.4 HAPTIC FEEDBACK

The capabilities for haptic feedback are rather limited when using only the smartphone's built-in vibration motor. The only way we can control it is to send a series of on-off intervals to its controller. There is no way controlling the vibration frequency intensity, which appeared to vary greatly between devices. The variation in temporal

[1] We used the default completion sound from Windows 3.11 called 'tadaa.wav'. Despite the archaic nature of this sound, many participants recognized it and it put a smile on their face.



(a) activation

(b) waveform

Figure 5.5: The startup time of a simple vibration motor can be exploited to create several levels of frequency and perceived 'roughness'. a) The activation sequence of the device. The resulting vibration pattern measured by a laser vibrometer. The shown pattern is perceived as 'very rough'. Source: Brown and Kaaresoja [6].

accuracy varied too. While our device could make a 50 ms vibration perceivable, the vibration motor of a cheaper device had a startup time of 100 ms. By using special on/off patterns, several levels of perceived intensity, or 'roughness', can be achieved. Brown and Kaaresoja [6] show that constant vibration causes 'smooth' sensation. The latency of vibration motors can be exploited to achieve amplitude modulation, however, which causes a 'rough' sensation. The principle is shown in Figure 5.5. Although we could simulate roughness, we were afraid that it would be difficult to convey messages with so little control over the hardware. Concerns for auditory feedback also apply to haptic feedback and we were also afraid that it would be difficult to perceive the vibrations through the neoprene wrist band which holds the device in place. We ended up using a stimulus of 100 ms which was applied after each repetition should have ended. The actual onset was fine-tuned so that the perceived maximum intensity of the signal was reached at the target time. To achieve this, we had to take factors into account such as the signal processing delay and the startup time of the vibration motor. It provides information which is fully redundant with the auditory feedback. This way, we could directly sample the users' preferences for either modality.

5.5 EFFICACY OF TRAINING ADVICE

In this section we address a particular topic of our second research question:

> Can the advice given by the device effectuate a better exercise execution?

To determine whether our app effectuates an improvement in exercise performance, we compare user performance on block 3 (before the advice was given) with performance on block 4 (after advice was given). We compare performances per repetition and per series of 10 repetitions. Advice is given after each completed movement. In the case of the per-repetition performance comparison, participants have to rely on this 'live' feedback. In the case of the per-series comparison, a summary overview of performance on the earlier set is also available.

Let us first look at the series data. We split the data by movement, so that all forward movements are considered separately from backward movements. The reason for this is that the backward movements are usually performed worse than the forward ones, and patterns may be visible more clearly this way. Of course, we also look at both biceps and BOR separately. To compare the experimental group with the control group in the per-series case, we first compute the change in duration and amplitude as the difference of these properties between block 3 and 4.

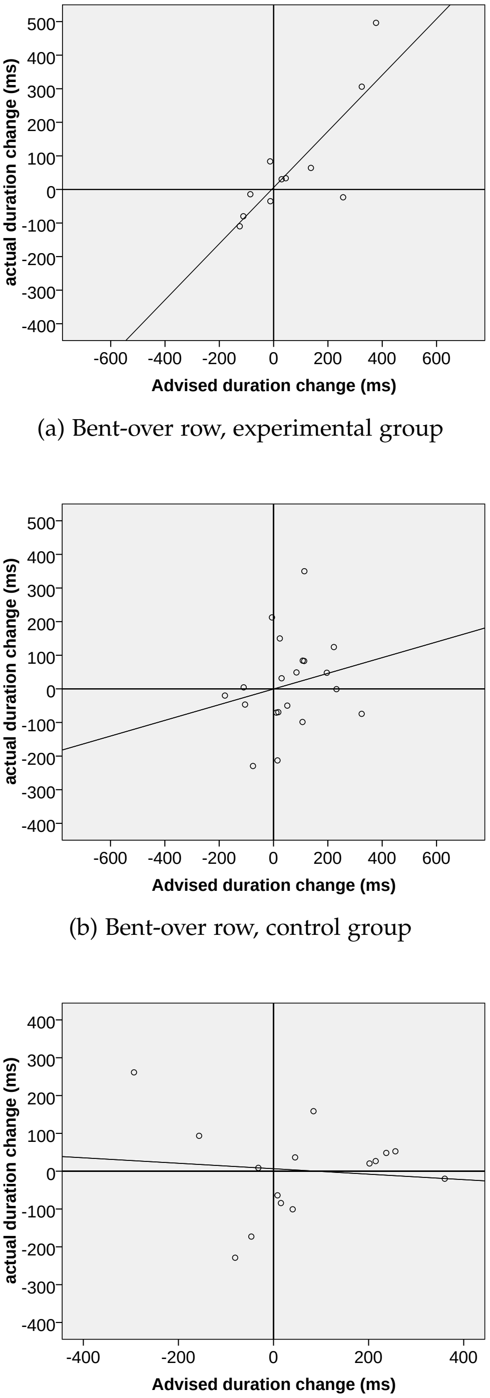
For example:

1. The target duration is 1200 ms (extracted from block 1).
2. The duration recorded from block 3 is 1450 ms.
3. The duration **advice** is $1200 - 1450 = -250$ ms.
4. The duration from block 4 is 1300 ms.
5. The duration **change** is $1300 - 1450 = -150$ ms.

We now compute correlations between the given advice and the changes. Finally, when these correlation scores are positive and significantly larger for the experimental than for the control group, we can assume that an improvement in exercise performance is caused by our app. We test for significantly different correlations using a two-tailed Fisher's r-to-z transform [20].

Overall, The correlations between advice and actual change are low, both for duration and amplitude (movement extent). Correlations between advice and change do not significantly differ between experimental and control groups. Only the correlation between duration advice and change of the backward movement for the BOR exercise is significantly better for the experimental group ($r(9) = 0.84$) than for the control group ($r(17) = 0.22$). $z = 2.3, p < 0.03$. See figure Figure 5.6a and 5.6b for a comparison.

It is interesting to see that the duration correlation for the biceps' backward movement gives a different impression. The correlation itself is even marginally negative ($r(12) = -0.1$). When looking at figure Figure 5.6c, however, it seems that there are two sub-populations in the data, of which one shows a strongly positive correlation. We can not be sure of what the difference between these populations is, but it might be that those who did not improve their performance



(a) Bent-over row, experimental group

(b) Bent-over row, control group

(c) Biceps, experimental group

Figure 5.6: On the horizontal axis are the differences between participants's movement durations and target durations. Data is shown for the second (backward) movement only. The effectuated duration differences after a) feedback and b) no feedback are on the vertical axis. c) shows the correlation between advice and effect for the biceps exercise's second movement. Only the correlation in a) is significant.

needed more time to get accustomed to the interface, or were not motivated to use it.

We expected our app to be of more benefit to the execution of the backward movement than to the forward movement, since the backward movement is usually performed worse than the forward movement according to the trainers. When more correction is needed, the feedback is stronger and we expect participants to be more enticed to adjust their movements.

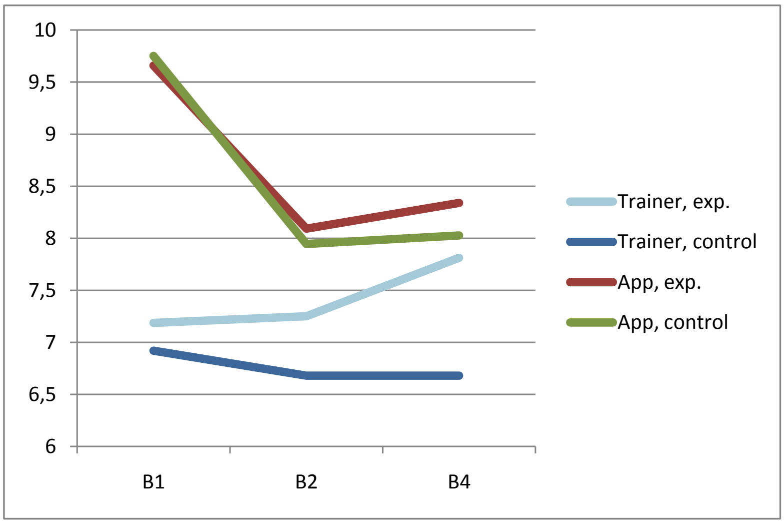
One last observation we make is that the correlation between movement extent advice and actual change is virtually 0 for BOR ($r(10) = 0.02$). This is to be expected since it was not possible during this prototype stadium to provide participants with this kind of feedback for the BOR exercise. For comparison, the same correlation for biceps was larger ($r(12) = 0.31$). Both are not statistically significant, however.

The per-repetition data shows weaker correlations, none of which are significantly better for the experimental condition than for the control condition. This may be because a change in duration does not consistently follow an advice in the previous movement. Users will likely need some time to interpret the feedback before they act on it, thus they will only react to feedback when they see a trend for repetitions $1-3$, and only act upon the given advice while performing the fourth repetition, for example. The match between advice and change for repetition 4 would boost the correlation, but it is moderated by the absent change for repetitions $2-3$. Also note that the display is not in the participant's line of sight while performing the BOR, but summary information between series can be comfortably viewed. This could be another explanation for why per-series data show higher correlations than per-repetition data.

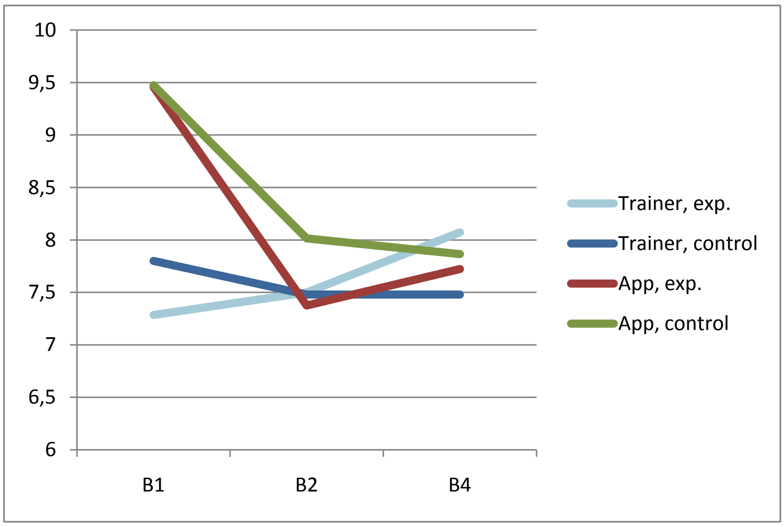
5.6 INTER-RATER AGREEANCE

Independently from our app, fitness instructors have given their assessment of the participants' performance. For a virtual fitness coach to be effective and safe to use, our app's assessment should agree with the professional assessments.

Figure 5.7 shows the valuations of exercise performance given by both our app and the trainers. Scores are given on a scale between 1 (lowest) and 10 (highest). When possible, subjects were scored by the same trainer for all blocks. After block 1, a rating was given for the quality of the exercise model. Block 2 was administered one week after block 1, so we assess the subjects again without reminding them of the proper execution of the exercises. This way, we could get an insight into the extent to which participants remember the instructions for proper exercise execution. It also provides us with a fresh baseline to compare the block 4 data against. Between block 3 and



(a) Biceps



(b) Bent-over row

Figure 5.7: Inter-rater agreeance. After block 1, a rating is given for the quality of the exercise model. Block 2 is administered one week after block 1, so we assess the subjects again without reminding them of the proper execution of the excercises. Between block 3 and 4, subjects either received feedback (exp.) or not (control). The App's assessment are a correlation score between each block and block 1, which is the reason why the app gives such high scores for block 1. Scores are given on a scale between 1 (lowest) and 10 (highest).

4, subjects either received feedback (exp.) or not (control). The App's assessments are a correlation score between each block and block 1. Block 1 is compared against block 1, which is the reason why the app gives such high scores for that block (auto-correlation).

For those who received feedback, a significant improvement in trainer scores is found between block 2 (mean = 7.3, sd = 1.0) and block 4 (mean = 7.9, sd = 0.9). $t_{28} = -2.7, p < 0.02$. For the control condition, performance on block 2 was not significantly different from performance on block 4 ($t_{47} = -0.6, p = 0.54$).

The correlation scores our app produced did not show a significant difference between block 2 and 4 when a two-tailed paired-samples t-test is used. One could argue that a one-tailed test is appropriate, since we do not expect our app to have a negative influence on exercise performance. A one-tailed test would indicate a significant difference ($t_{28} = -1.97, p < 0.03$).

For participants in the control condition, the trainer scores deteriorated significantly between block 1 and 2 ($t_{47} = -2.5, p < 0.02$), while the score for participants in the experimental condition did not ($t_{28} = -0.54, p = 0.60$). The app scores do not show this trend.

This could mean several things. One possibility is that the trainers were not independent and that the group to which a participant was assigned influenced their judgment. The experimental setup was not double blind. Both the trainer and experimenter were informed about the group each participant was assigned to. Organizing a double blind experiment would introduce some additional difficulty. The display must not be visible to the trainer, since it either shows feedback or not, from which one can derive to which group the participant was assigned. He/she would thus have to be standing some distance from the participant, which would complicate the grading of exercise performance.

Another possibility is that, while the app strictly compares performance against block 1, trainers may use their knowledge of sports physiology to grade exercises, which means that the same exercise, executed in two different ways, could be awarded the same grade. Or that both groups performed equally well on the measures tempo and movement range, but the experimental group performed better on additional factors only the trainers took into account. Still, this does not explain why group should be a factor. The procedure is exactly the same for both groups until block 3, after all. We thus recommend to organize a double blind experiment to re-evaluate these results.

Earlier, we saw that app and trainer score distributions are significantly better for block 4 than for block 2 on average. But is there also a correlation between individual differences? Figure 5.8 shows correlations for score differences between block 4 and block 2. The correlation between trainer and app differences is positive, but not

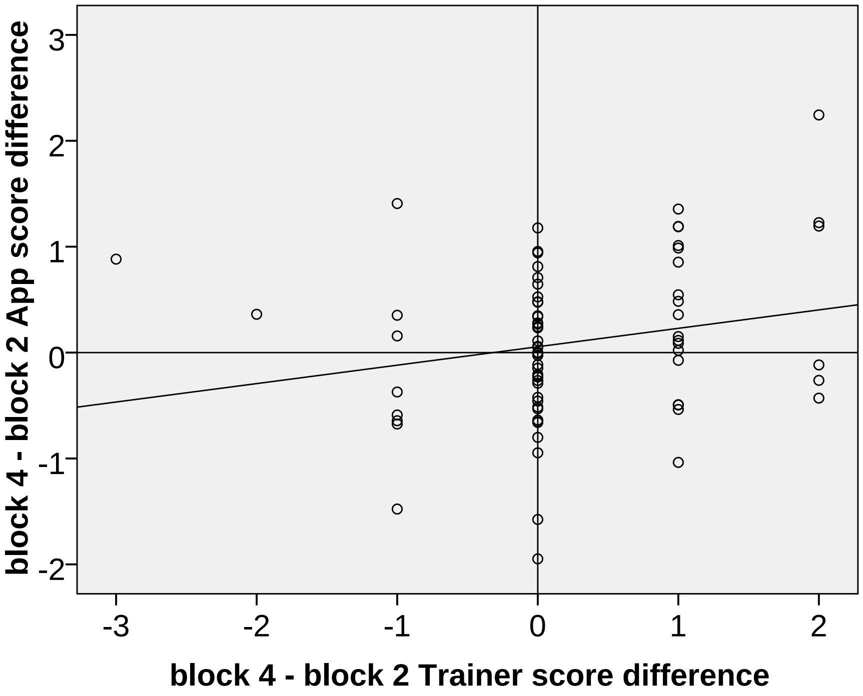


Figure 5.8: Inter-rater correlation between app scores and trainer scores. On the horizontal axis, the difference in trainer grade between block 4 and block 2 is shown. Positive numbers indicate higher grades on block 4. App scores are on the vertical axis. The correlation is not significant.

significant ($r(73) = 0.21, p = 0.07$). In Chapter 3, we saw that the values of duration and amplitude do not significantly differ between high and low scoring groups. We suggested that this could mean that trainers base their judgment on additional features of exercise performance, and that the recorded variables we use are insufficient to make a proper performance assessment. This result supports that hypothesis. Another suggestion for improving the experimental setup, is to make sure that participants are able to perform the exercises very well before administering block 1. The data from this block is used as a golden standard, after all. In our experiment, the exercise was demonstrated once by the instructor. The participant was asked to demonstrate a few repetitions, after which the instructor would give additional tips, if neccesary. The profile was recorded immediately thereafter. It might have been better to ask instructors to keep working with the participant until he/she was confident that the performance of the participant was worth an eight out of ten. Some participants did not have the flexibility to perform an exercise perfectly, which also contributed to the low mean trainer score on block 1.

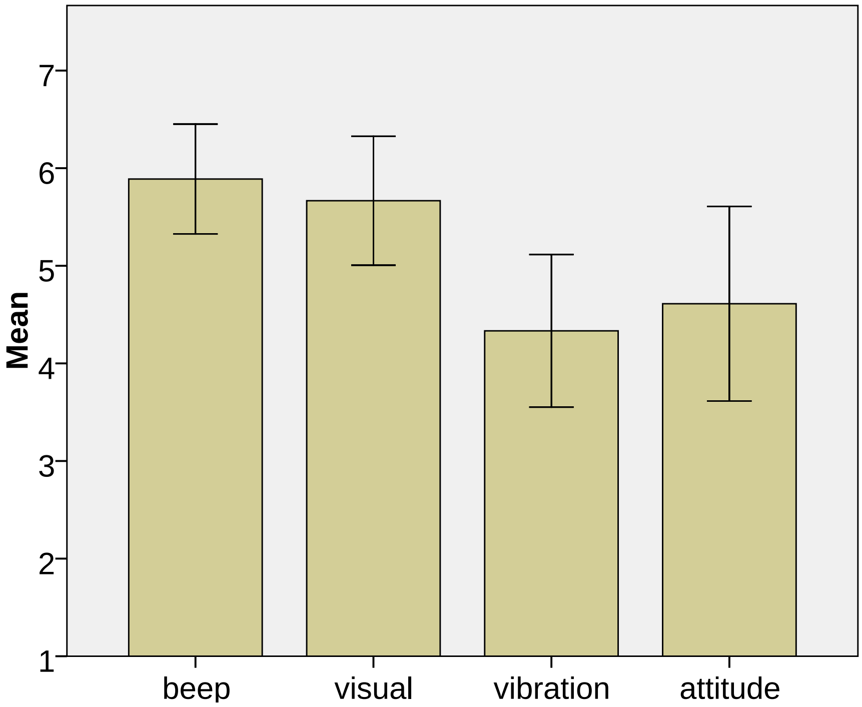


Figure 5.9: Valuations of the multiple feedback types received by participants in the experimental group. Answers were given on a 7-point Likert scale. Attitude is the overall willingness to use the app in its current form during exercise.

5.7 post-task questionnaire

During our main experiment, several feedback modalities were used. At the end of the experiment, subjects in the experimental group were asked how useful they thought these types of feedback are. Results are shown in Figure 5.9. When comparing these results against those collected in the main questionnaire (Figure 5.3), we see that especially the beep feedback was valuated higher during our experiment than we expected from the main questionnaire results. Participants of our experiment have a much more concrete - 'hands-on' - conception of what this feedback sounds like and what benefits it has for their exercising. Maybe this is because the main questionnaire was administered in a relatively early stage of development. The final implementation differs from the way this modality was explained in the questionnaire. We described beep feedback as 'Sound a beep, every time you complete a movement. The pitch tells you whether you did the exercise right or wrong'. In practice, pitch was always the same and the moment of beep onset carried information on tempo performance. It seems that this design decision paid off.

We expected visual feedback to receive high valuations, which proved to be the case. We cannot compare the ratings for haptic feedback since its intensity was too low to be felt by most participants. Some of the participants were asked to be especially observant of the vibration feedback, after which they reported that they could indeed feel it. We would be interested to know what the reason for this phenomenon is. Auditory and haptic feedback were redundant in the sense that they were given at the same time and coded for the same information. Is this a case of 'multi-modal masking'?

What also became apparent from informal interviews, is that the performance of the movement detection algorithm should be very high. When one movement is missed, the auditory feedback is not given. Since the auditory feedback is used as a cue for tempo, some of the participants paused and waited for a beep that would not be presented. Understandably, this caused annoyance among participants.

The last bar in Figure 5.9 shows the respondent's willingness to use our system in its current form. Many of the experienced participants (including trainers) commented that they thought this app could be useful for novices but that they themselves would not use it. When comparing those with extensive experience in free-weight training with those who are either novices, we see that the less experienced seem more positive towards using our system (mean = 5.25, sd = 1.9, n = 8) than experienced participants (mean = 4.10, sd = 2.0, n = 10). The difference is not significant, but we feel that this, as well as the large sd, is due to the small sample size.

5.8 further research

Since live visual feedback was the most preferred feedback method, it deserves some further investigation. To achieve visual feedback which is visible at all times during and after exercise, the display should be separated from the rest of the system. Since most free-weight exercises are performed standing or sitting in front of a mirror, we propose to construct a smartphone sleeve which holds a thin magnet in place along the backside of the device. By magnetizing the mirror, the smartphone's screen can be placed comfortably at eye level, much like magnets can be attached to a refrigerator door. To still be able to collect acceleration data, a wrist band or glove with integrated sensor could communicate over a low power communication protocol such as Bluetooth [5].

We presented a beep at the moment on which a repetition should be completed according to the profile. The delay for this beep was calculated when the forward movement was completed. Thus, when the forward movement was missed, the beep would remain absent, which is frustrating to users. To remedy this issue, it might be better to present feedback at fixed intervals, like a metronome. This also allows for feedback per movement instead of per repetition. One difficulty which would have to be dealt with however, is synchronization.

A best-of-both solution could be to use a metronome which is started when the first movement is detected, synchronized when any consecutive movement is detected, and stopped when the target amount of repetitions has been reached.

What we have called 'live feedback' thus far is actually feedback which is updated with every detected movement. It is technologically possible however, to make feedback truly live. Fitlinxx [21] is an example of such a system for weight stack machines, which displays the current position of the weight over a scale of the target movement range (see Figure 1.3). For rotational exercises, such as biceps, this same metaphor could be used. An animation of a flexing and extending biceps might even be more appropriate. For exercises with linear movement however, collecting data on movement range would be a considerable challenge.

It seems that the data we have collected during our experiment, or at least the features we have extracted, are insufficient to calculate an exercise grade which correlates with that of professionals. Using expert interviews and physiological literature, it might be informative to see if we can determine what the factors are on which trainers base their judgment and whether we can monitor these factors with additional sensors.

5.9 SUMMARY

Our display design study indicated that an arrow pointing up or down is the best way to signal a deviating variable. For a more precise indication of how much a variable deviates, we suggested several scales. A yellow-green-red gradient, in which yellow indicates a value which is too low and red indicates a value which is too high, was preferred by most. The final display design combined these elements.

When comparing auditory, haptic and visual feedback modalities, visual feedback was most popular. Live visual feedback during training was preferred over summary information between each set of 10 repetitions. When using our prototype however, the screen is not always visible during exercise, which is problematic since users have a preference for visual feedback presented during training. We suggest to use a system wherein the sensor is a separate component. The screen could be attached to the wall so that it is visible at all times. To visualize range of movement, we suggest to take inspiration from the existing Fitlinxx system for weight stack machines, which displays the current position of the weight over a scale of the target movement range.

Whether our app was able to effectuate an improved exercise performance is still open for debate. Participants who used our app receive significantly higher grades from trainers than those in the control group. Grades given by our app are only significant at the $\alpha = 0.05$ level when a one-tailed test is used, however. When looking at the individual exercise properties tempo and range, only the tempo of the backward movement for the bent-over row exercise is significantly improved by our app. Since the correlation between trainer and app grades was not significant, we suggest to perform a double blind experiment to determine whether the higher grades received by participants in the experimental group are actually caused by the advice our app gives.

We conducted two questionnaires regarding feedback modalities in this study. The main questionnaire was conducted during an early stage of development, the post-task questionnaire at the end of the experiment. The preference for live visual feedback was correctly predicted by the main questionnaire. Because auditory feedback using beeps of different pitch received low ratings in the main questionnaire, we decided to use beeps of constant pitch with different intervals instead. This type of feedback received favorable ratings in the post-task questionnaire.

Part IV

CONCLUSION

6 CONCLUSION

In this thesis, we have described the process of creating a working prototype for the purpose of tracking and evaluating free-weight exercises. Where earlier work mainly focused on signal processing and collecting quantitative information such as repetition counts, our aim was to provide users with qualitative feedback on their performance. By using this feedback, users would be able to exercise more effectively and responsibly than when they would train unaided. Commercial applications such as those by NorthPark [40] are separately written for each exercise. Because our app learns exercise profiles from a calibration session, our solution is much more flexible. The smartphone app we created was tested in a gym environment. Because our application is an example of ubiquitous computing, the requirements for the user interface are different from those of a desktop application. Mouse and keyboard are not available for input, nor can we expect to have the user's full attention at all times. For this reason, interaction design was also part of this study. The advice our app provided was compared with the advice given by professional fitness instructors.

6.1 APPROACH

As a platform for our project, we used an off-the-shelf smartphone. The accelerometer was used to capture movement data because it does not only measure linear acceleration, but also the acceleration imposed by the force of gravity. When an exercise includes a rotational movement, this *gravity effect* shifts from one axis to another, which can be reliably exploited to detect exercise repetition boundaries. To smooth raw accelerometer data, we used a Kalman filter. Our process model expects data of a parabolic nature, which proved suitable for the spring-like properties of muscles. When compared to the default, static, process model, much more of the amplitude in the raw signal was retained, without introducing additional delay.

We employed derivative based peak-detection to segment a data stream into movements. By considering movements as the atoms for further processing, we are able to provide feedback specifically for each movement that makes up an exercise. We could say, for example, that the tempo of the forward movement was correct, while the backward movement was performed too slowly.

To classify movements as being part of an exercise we used *dynamic rule-based decision*. The rules are based on prototypes defined in an exercise profile. A movement will be classified as belonging to an exercise when tempo, peak amplitude, and range are within a ratio of 50% of the values in the profile.

The visual display design was determined iteratively. Five different display designs were compared and the elements of the 2 most popular designs were combined. In another usability study, we investigated what feedback modalities were preferred, and whether the feedback type (repetitions, tempo, or range) was a factor which influenced these preferences. The final prototype made use of on-screen information and redundant auditory haptic cues.

6.2 FINDINGS

When new movement data was recorded shortly after the exercise profile was recorded, 95.3% of the repetitions were correctly counted. We found that the algorithm performed better on exercises with a gravity effect (98%) than on exercises with a linear movement (91%). These results are better than those achieved using peak counting in combination with either Naïve Bayes Classifiers or Hidden Markov Models [8]. The performance of Dynamic Time Warping seems superior to ours, although performance measures used in [44] are not directly comparable to our own.

When the interval between profile and test data recordings was about 1 week, counting performance was 89.1% on average. When using a single profile for each exercise on all data, performance was not significantly lower (82.1%). We do advice against using average profiles however, because the ideal tempo and movement range is dependent on an individual's training goals and capabilities.

The final display design is shown in Figure 5.4. The most popular display element was found to be a three-state indicator, which represents a value which is too low, correct, or too high, respectively. The summary screen which shows the user's average performance over a set of multiple repetitions was rated 5.7 on a 7-point Likert scale. When comparing feedback modalities, we found that the visual modality was preferred over the auditory and haptic modality. We found no effect of feedback type. An auditory cue was played at the moment on which the user should have completed a repetition, and is thus a reference for tempo, as well as an indicator of a successfully performed repetition. This type of feedback was also well-received, with a 5.9 on a 7-point Likert scale. The vibration feedback could not be evaluated because most subjects did not perceive it.

Participants who received feedback from our app received significantly higher grades for their exercise execution than participants in the control group. This is true for both the grades given by fitness instructors as for the grades given by our app. The inter-rater correlation between app and fitness instructors was not statistically significant, however.

6.3 FURTHER RESEARCH

It was difficult to detect the first and last movement in a series of repetitions, because these movements were mixed with other movements. We would be interested to see how our algorithm holds up when users alternate repetitions between the right and left arm because every repetition would be mixed with preparatory movements.

We would also like to see how signal characteristics change when different weights are used, and whether we can exploit these to determine how strenuous the exercise is for the user. This would allow us to adjust the training regimen accordingly.

For exercises with a linear movement, we could not provide participants with movement range feedback, because start and end positions can not be determined from acceleration signals. Maybe the distance traveled can be determined by integrating the signal, but this is not a complete solution.

We used movements as atoms for our performance assessments. We found a strong preference for live visual feedback however, and we think that a movement range indicator such as employed in the Fitlinxx system (see Figure 1.3) would be most intuitive. For rotational exercises, this kind of feedback could be easily implemented into our system.

We would like to know why we did not find a correlation between grades given by instructors and those given by our app. Through interviews and literature study, we may be able to learn what factors instructors use when evaluating exercises. If we could measure these factors, the automated grading of exercises might be significantly improved.

6.4 FINAL WORDS

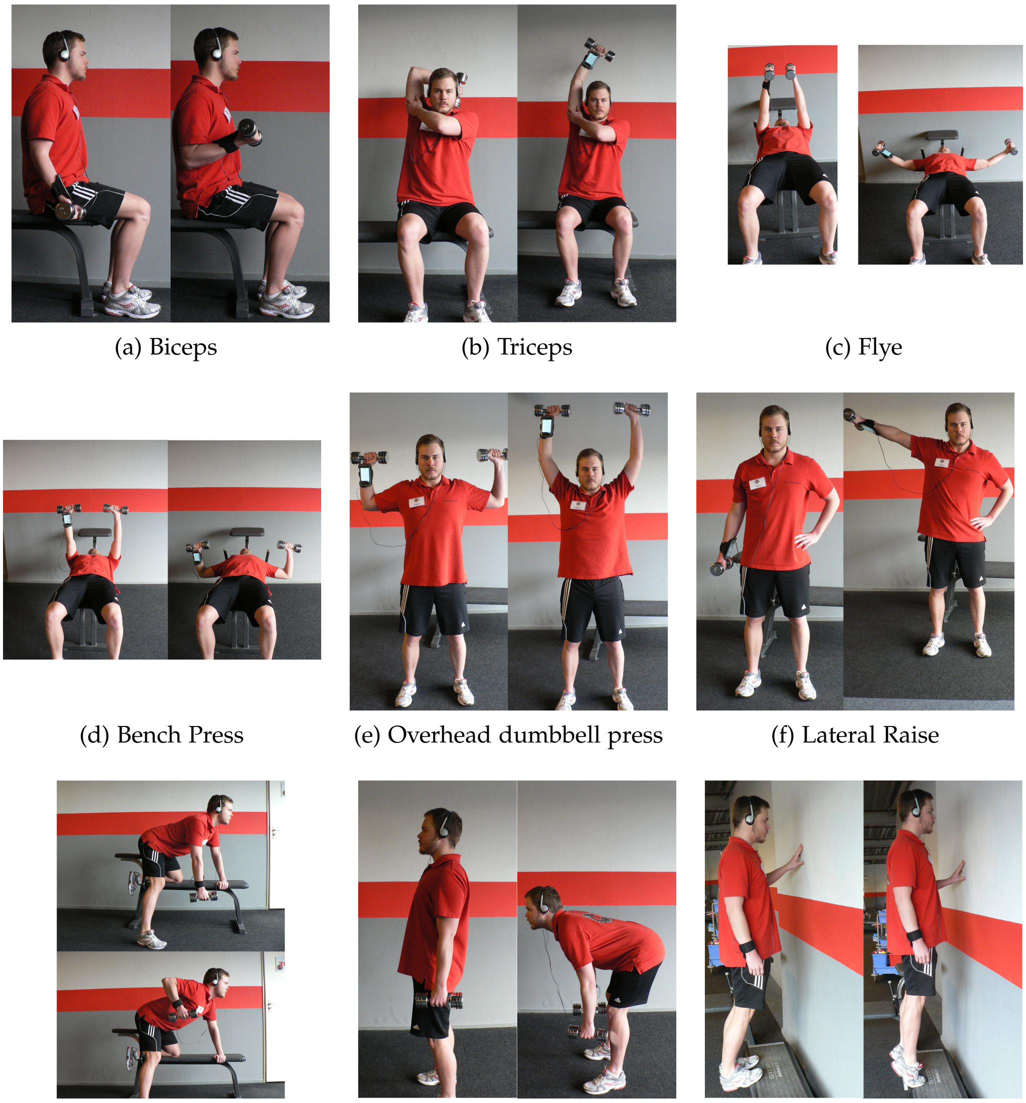
We were able to achieve comparable performance to contemporary research when counting repetitions, while using much simpler algorithms. Our approach is computationally efficient, and information on tempo and movement extent is retained. We did not only track fitness exercises but also provided feedback on how to improve exercise performance. This feedback was highly valued by users. Both according to fitness instructors and measures recorded by our app, gym attendees who use our app perform significantly better on free-weight exercises than those who do not. We think we have provided a promising starting point for automated qualitative feedback of fitness exercises.

Part V

APPENDIX

A

EXERCISES



(a) Biceps

(b) Triceps

(c) Flye

(d) Bench Press

(e) Overhead dumbbell press

(f) Lateral Raise

(g) Bent-over row

(h) Deadlift

(i) Calf Raise

Figure A.1: Start and end poses for each exercise considered in this study. Note that only biceps and bent-over row were included in the main experiment. Calf raise was also excluded from the pilot, because of a poor signal-to-noise ratio.

B QUESTIONNAIRE (IN DUTCH)

The next pages show the questions of the main questionnaire described in Chapter 5. Questions about device price are not discussed because of a missing option in question 18.

Enquête Fitness App

Let op: Deze enquête is bedoeld voor mensen die aan fitness doen in een sportschool of thuis trainen met gewichten. Als u alleen aan fietsen of hardlopen doet kunt u deze enquête beter niet invullen.

Deze enquête is een onderdeel van mijn Master scriptie. Het doel van de scriptie is het ontwikkelen van een hulpmiddel dat gebruikt kan worden bij het fitnessen. De enquête bestaat uit 25 vragen. Het invullen van de enquête duurt ongeveer 15 minuten.

Bij de meeste vragen kunt u een keuze maken uit meerdere antwoorden. Dit doet u door het hokje (□) aan te kruisen bij het antwoord dat het beste bij u past. Vul altijd minstens één antwoord in. Bij sommige vragen is het mogelijk meerdere hokjes aan te kruisen, dit wordt dan bij de vraag vermeld. Op een stippellijn (..........) kunt u zelf een antwoord invullen.
In sommige gevallen mag u een aantal vragen overslaan. Soms wordt u gevraagd om door te gaan met een volgend onderdeel. Onderdelen kunt u herkennen aan de **vet gedrukte kopjes** bovenaan de pagina.

Deze enquête is volledig anoniem, u hoeft uw naam niet in te vullen.

Bedankt voor uw medewerking!

Nino van Hooff

(Contact data removed)

De vragen beginnen op de volgende bladzijde →

Onderdeel: Algemeen

1. Wat is uw geslacht?

   □ Man

   □ Vrouw

2. Wat is uw leeftijd?

   …………. Jaar

3. Wat voor soort training doet u? U kunt meerdere antwoorden aankruisen.

   □ Krachttraining met apparaten

   □ Krachttraining met losse gewichten

   □ Cardio training (roeien, lopen, fietsen).

   □ Weet niet.

Onderdeel: Tellen

Bij veel oefeningen is het nodig om deze een vast aantal keer uit te voeren. Bijvoorbeeld 3 series van 10 keer. U doet de oefening dan in totaal 30 keer, met een pauze na elke 10^{e} keer.

4. Raakt u weleens de tel kwijt?

□ Ja, bij meer dan de helft van de series die ik doe.
□ Ja, bij ongeveer de helft van de series.
□ Ja, een enkele keer
□ Nee, ik weet altijd precies hoe vaak ik de oefening gedaan heb. → Ga door naar vraag 6.
□ Nee, ik doe geen oefeningen waarbij je moet tellen. → Ga door naar het onderdeel "Uitvoering".

5. Waarom raakt u de tel kwijt?
U kunt meerdere antwoorden aankruisen.

□ Mijn gedachten dwalen af, of ik word afgeleid door mijn omgeving (radio,tv, andere sporters etc.).
□ Ik raak verveeld.
□ Anders, namelijk …………………………………………………..

6. Stel dat er een hulpmiddel zou zijn dat u zou helpen bij het tellen van het aantal oefeningen dat u gedaan heeft. Zou u dit willen gebruiken?

□ Ja
□ Nee → Ga door naar het onderdeel "Uitvoering".

7. Op welke manier zou dit apparaat het best aan kunnen geven hoe vaak u een oefening gedaan heeft? U hoeft hier geen rekening te houden met andere sporters, die misschien last zouden kunnen hebben van het geluid dat het apparaat maakt. U kunt eventueel ook een eigen idee aandragen.

Een piepje laten horen, elke keer dat u de oefening doet.	onhandig	○ ○ ○ ○ ○ ○ ○	handig
Een getal op het beeldscherm dat het aantal keren toont.	onhandig	○ ○ ○ ○ ○ ○ ○	handig
Een stem die het aantal keren telt.	onhandig	○ ○ ○ ○ ○ ○ ○	handig
Een trilling die u voelt, zoals het trilsignaal van een mobiele telefoon.	onhandig	○ ○ ○ ○ ○ ○ ○	handig
Een getal op het scherm ná het uitvoeren van de oefening.	onhandig	○ ○ ○ ○ ○ ○ ○	handig

Eigen idee, namelijk ………………………………………………………………………………………….

Onderdeel: Uitvoering

Voor een effectieve training is het van belang dat u een oefening op de juiste manier uitvoert.

8. Lukt het u de oefening op de juiste manier uit te voeren?

□ Ja, dat gaat mij altijd goed af. → Ga door naar vraag 10.
□ Bij een enkele oefening is dat lastig.
□ Bij meerdere oefeningen is dat lastig.
□ Weet ik niet → Ga door naar vraag 10.

9. Waarom gaan één of meerdere oefeningen lastig?
U kunt meerdere antwoorden aankruisen.

□ Het is lastig om te onthouden hoe de oefening moet.
□ Het is lastig om te onthouden op welke stand een apparaat ingesteld moet worden.
□ Het instellen van het apparaat zelf is lastig of zwaar.
□ Mijn gedachten dwalen af, of ik word afgeleid door mijn omgeving (radio,tv, andere sporters etc.).
□ Als ik vermoeid raak ga ik de oefening slordiger uitvoeren.
□ Anders, namelijk …………………………………………………..

10. Stel dat er een hulpmiddel zou zijn dat u zou helpen om bij te houden of u een oefening goed uitgevoerd heeft? Zou u dit willen gebruiken?

□ Ja
□ Nee → Ga door naar het onderdeel "Tempo".

11. Op welke manier zou dit apparaat het best aan kunnen geven of u een oefening goed gedaan heeft? U hoeft hier geen rekening te houden met andere sporters, die misschien last zouden kunnen hebben van het geluid dat het apparaat maakt. U kunt eventueel ook een eigen idee aandragen.

Een piepje laten horen, elke keer dat u de oefening doet. Aan de toon kunt u horen of u de oefening goed of fout doet.	onhandig	○ ○ ○ ○ ○ ○ ○	handig
Een symbool op het scherm, dat bijvoorbeeld aangeeft dat u de oefening rustiger moet doen.	onhandig	○ ○ ○ ○ ○ ○ ○	handig
Een stem die advies geeft.	onhandig	○ ○ ○ ○ ○ ○ ○	handig
Een trilling die u voelt, zoals het trilsignaal van een mobiele telefoon.	onhandig	○ ○ ○ ○ ○ ○ ○	handig
Een overzicht op het scherm ná het uitvoeren van de oefening.	onhandig	○ ○ ○ ○ ○ ○ ○	handig

Eigen idee, namelijk ………………………………………………………………………………………….

Onderdeel: Tempo

Voor een effectieve training is het van belang dat u een oefening op het juiste tempo uitvoert.

12. Lukt het u om de oefening op het juiste tempo uit te voeren?

□ Ja, dat gaat mij altijd goed af. → Ga door naar vraag 14.
□ Bij een enkele oefening is dat lastig.
□ Bij meerdere oefeningen is dat lastig.
□ Weet ik niet. → Ga door naar vraag 14.

13. Waarom hebt u moeite met het tempo?
U kunt meerdere antwoorden aankruisen.

□ Mijn gedachten dwalen af, of ik word afgeleid door mijn omgeving (radio,tv, andere sporters etc.).
□ Ik raak verveeld.
□ Als ik vermoeid raak gaat het lastiger, vooral de laatste paar keer is lastig.

□ Anders, namelijk …………………………………………………..

14. Stel dat er een hulpmiddel zou zijn dat u zou helpen om bij the houden of u de oefening te snel of te langzaam uitgevoerd heeft Zou u dit willen gebruiken?

□ Ja
□ Nee → Ga door naar het onderdeel "Apparaat".

15. Op welke manier zou dit apparaat het best aan kunnen geven of u een oefening op het juiste tempo gedaan heeft? U hoeft hier geen rekening te houden met andere sporters, die misschien last zouden kunnen hebben van het geluid dat het apparaat maakt. U kunt eventueel ook een eigen idee aandragen.

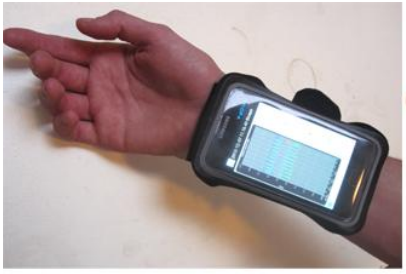
Een piepje laten horen, elke keer dat u de oefening doet. Aan de toon kunt u horen of u de oefening goed of fout doet.	onhandig	○ ○ ○ ○ ○ ○ ○	handig
Een symbool op het scherm, dat bijvoorbeeld aangeeft dat u de oefening rustiger moet doen.	onhandig	○ ○ ○ ○ ○ ○ ○	handig
Een stem die advies geeft.	onhandig	○ ○ ○ ○ ○ ○ ○	handig
Een trilling die u voelt, zoals het trilsignaal van een mobiele telefoon.	onhandig	○ ○ ○ ○ ○ ○ ○	handig
Een overzicht op het scherm ná het uitvoeren van de oefening.	onhandig	○ ○ ○ ○ ○ ○ ○	handig

Eigen idee, namelijk ……………………………………………………………………………………….

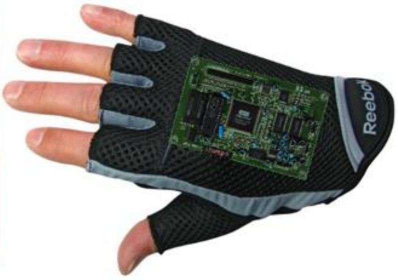
Onderdeel: Apparaat

Om te beoordelen of u een oefening goed uitvoert, moet het apparaat de bewegingen van uw onderarm kunnen meten.

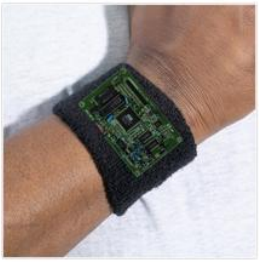
16. In wat voor soort accessoire zou u dit apparaat het liefst zien? Ga er vanuit dat alle onderstaande apparaten evenveel wegen. LET OP: Heeft u een eigen idee, kruis dan ook het hokje aan van één van de drie gegeven ideeën die u het meeste aanspreekt.



A



B



C

□ A: Een polsband waar u het apparaat inschuift (met beeldscherm)
□ B: Een vingerloze sporthandschoen, waarbij de apparatuur op de rug van uw hand zit (zonder beeldscherm). De bediening gaat via een apart apparaat.
□ C: Een zweetbandje waar de apparatuur in verwerkt zit (zonder beeldscherm). De bediening gaat via een apart apparaat.

□ Ik heb een ander idee, namelijk…………………………………………………………………………

17. Stel dat er een app was die u kan helpen met het tellen, de uitvoering, en het tempo van uw oefeningen. U gebruikt dit dan met uw smartphone en een polsband zoals in afbeelding A hierboven. Hoeveel geld zou u voor de app over hebben?

□ Ik heb geen smartphone / geen behoefte aan.
□ € 0 (gratis)
□ € 0,10 – 1,00
□ € 1,10 – 3,00
□ Meer dan € 3,00.

18. Stel dat er een accessoire zoals in afbeelding B en C nodig is. Hoeveel geld zou u hier voor over hebben?

□ Geen behoefte aan.
□ € 0 (gratis)
□ € 0 – 10
□ € 21 – 30
□ € 31 - 50
□ Meer dan € 50.

19. Als dit apparaat een geluid zou maken, denkt u dat dat storend zou zijn voor andere sporters in uw omgeving?

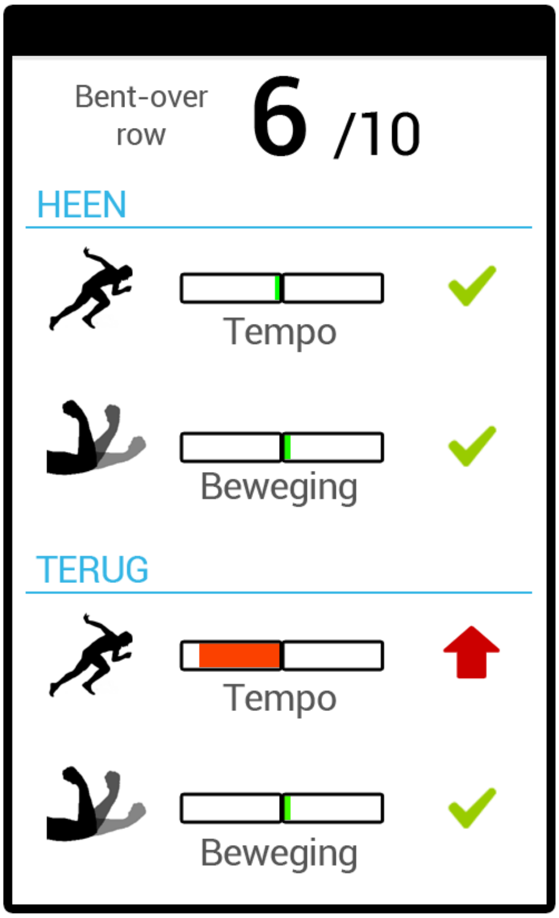
□ Ja, het is namelijk vrij stil waar ik sport.
□ Ja, wel als het geluid harder is dan de muziek of ander geluid dat te horen is waar ik sport.
□ Nee, er zijn geen mensen in de buurt die er last van kunnen hebben.

20. Zou u het vervelend vinden om een koptelefoon of oordopjes te dragen tijdens het sporten?

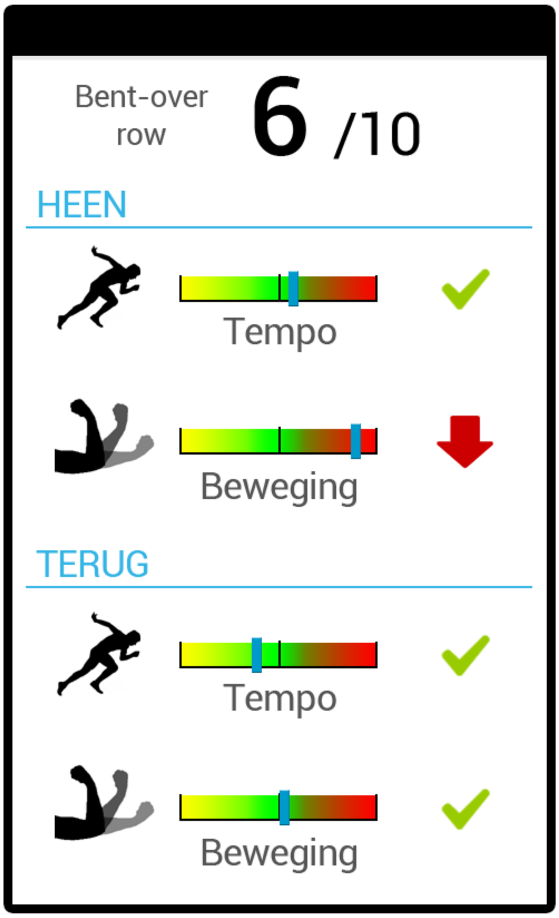
□ Ja
□ Nee, maar op dit moment draag ik geen koptelefoon of oordopjes
□ Nee, ik draag nu al een koptelefoon of oordopjes tijdens het sporten

Onderdeel: scherm

Wanneer u de oefening uitvoert worden uw bewegingen gevolgd. Dit onderdeel gaat over de manier waarop het apparaat u feedback geeft.



A



B

Het apparaat is voorzien van een beeldscherm dat u inzicht geeft in hoe goed u de oefening uitvoert. Hiervoor hebben wij twee ontwerpen gemaakt die hierboven met 'A' en 'B' aangegeven zijn. Beide schermen geven een andere situatie weer.

21. Stel u ziet **scherm A.** Wat denkt u dat de **rode pijl** betekent?

□ Ik ga te traag, ik moet het sneller doen
□ Ik doe dit op het goede tempo.
□ Ik ga te snel, ik moet het trager doen.

22. Stel u ziet **scherm B**. Wat denkt u dat **de rode pijl** betekent?

....

23. Kruis per regel één cirkeltje aan. Als u neutraal staat tegenover de stelling kunt u het middelste cirkeltje aankruisen.

Ik vind **Scherm A** makkelijk te begrijpen.	oneens	○ ○ ○ ○ ○ ○ ○	eens
Met scherm **Scherm A** kan ik snel zien hoe ik mijn training kan verbeteren.	oneens	○ ○ ○ ○ ○ ○ ○	eens
Ik vind **Scherm B** makkelijk te begrijpen.	oneens	○ ○ ○ ○ ○ ○ ○	eens
Met scherm **Scherm B** kan ik snel zien hoe ik mijn training kan verbeteren.	oneens	○ ○ ○ ○ ○ ○ ○	eens

24. Als u zou moeten kiezen tussen **Scherm A of Scherm B**, welke zou u dan het liefste gebruiken?

25. Zijn er nog functies die u graag terug zou willen zien in een apparaat voor bij het fitnessen die nog niet eerder genoemd zijn in deze enquête?

……

……

……

……

……

……

Dit is het einde van de enquête. Bedankt!

BIBLIOGRAPHY

[1] E. K. Antonsson and R. W. Mann. The frequency content of gait. *Journal of Biomechanics*, 18(1):39–47, 1985. ISSN 0021-9290. doi: 10.1016/0021-9290(85)90043-0. URL http://www.sciencedirect.com/science/article/pii/0021929085900430. (Cited on pages 24 and 34.)

[2] L. Bao and S. Intille. Activity recognition from user-annotated acceleration data. *Pervasive Computing*, page 1–17, 2004. URL http://www.springerlink.com/index/9AQFLYK4F47KHYJD.pdf. (Cited on pages 18 and 22.)

[3] M. Baten. Cijfers en statistieken van Facebook in Nederland [Figures and statistics on Facebook in the Netherlands], 2012. URL http://www.socialmediaacademie.nl/cijfers-en-statistieken-van-facebook-in-nederland/, accessed 2013-06-06. (Cited on page 62.)

[4] D. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, page 359–370, 1994. URL http://www.aaai.org/Library/Workshops/1994/ws94-03-031.php. (Cited on pages 3 and 7.)

[5] Bluetooth SIG. Bluetooth Technology Website, 2013. URL http://www.bluetooth.com/Pages/Bluetooth-Home.aspx, accessed 2013-06-12. (Cited on page 80.)

[6] L. M. Brown and T. Kaaresoja. Feel who's talking: using tactons for mobile phone alerts. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, page 604–609, New York, NY, USA, 2006. ACM. ISBN 1-59593-298-4. doi: 10.1145/1125451.1125577. URL http://doi.acm.org/10.1145/1125451.1125577. (Cited on page 72.)

[7] S. Butterworth. On the theory of filter amplifiers. *Wireless Engineer*, 7:536–541, 1930. (Cited on page 24.)

[8] K.-H. Chang, M. Y. Chen, and J. Canny. Tracking free-weight exercises. In *Proceedings of the 9th international conference on Ubiquitous computing*, UbiComp '07, page 19–37, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-74852-6. URL http://dl.acm.org/citation.cfm?id=1771592.1771594. (Cited on pages 6, 11, 12, 18, 20, 21, 22, 27, 44, 49, 55, 66, and 85.)

[9] W. Chauvenet. *A manual of spherical and practical astronomy: Embracing the general problems of spherical astronomy, the special applications to nautical astronomy, and the theory and use of fixed and*

portable astronomical instruments, with an appendix on the method of least squares, volume 2. JB Lippincott, 1908. (Cited on page 28.)

[10] M. Crease and S. A. Brewster. Making progress with sounds-The design and evaluation of an audio progress bar. In *Proceedings of ICAD*, volume 98, 1998. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.125.9589&rep=rep1&type=pdf. (Cited on page 71.)

[11] F. Delavier. *Strength training anatomy*. Human Kinetics, Champaign, IL, 3rd ed edition, 2010. ISBN 9780736092265. (Cited on page 14.)

[12] T. Dingler, J. Lindsay, and B. N. Walker. Learnability of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech. In *Proceedings of the 14th International Conference on Auditory Display, Paris, France*, 2008. URL http://www.icad.org/Proceedings/2008/DinglerLindsay2008.pdf. (Cited on page 71.)

[13] S. Drenthen. Provide reliable feedback during fitness derived from online activity matching. In *Proceedings of the 15th Twente student conference on IT*, 15, pages 49–54, Enschede, 2011. University of Twente. (Cited on page 65.)

[14] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley, New York, 2001. ISBN 0471056693 9780471056690 0471429775 9780471429777. (Cited on page 23.)

[15] S. Eickeler, A. Kosmala, and G. Rigoll. Hidden Markov model based continuous online gesture recognition. In *Fourteenth International Conference on Pattern Recognition, 1998. Proceedings*, volume 2, pages 1206 –1208 vol.2, August 1998. doi: 10.1109/ICPR.1998.711914. (Cited on page 22.)

[16] Endomondo. Endomondo Sports Tracker PRO, 2013. URL https://play.google.com/store/apps/details?id=com.endomondo.android.pro, accessed 2013-06-19. (Cited on page 9.)

[17] D. Fankhauser. The Tiny, Powerful Brain Inside Nike's FuelBand, 2013. URL http://mashable.com/2013/01/31/nike-fuelband/, accessed 2013-06-19. (Cited on page 10.)

[18] M. S. Feigenbaum and M. L. Pollock. Prescription of resistance training for health and disease. *Medicine and Science in Sports and Exercise*, 31:38–45, 1999. URL http://www.ais.up.ac.za/med/sportcert/prescription1a.pdf. (Cited on pages 2 and 5.)

[19] M. A. Fiatarone, E. C. Marks, N. D. Ryan, C. N. Meredith, L. A. Lipsitz, and W. J. Evans. High-intensity strength

training in nonagenarians. *Jama*, 263(22):3029–3034, 1990. URL http://faculty.fullerton.edu/leebrown/PDFFiles/Academic/Fiatarone-strengthtrainingoldpeople.pdf. (Cited on page 5.)

[20] R. A. Fisher. *Statistical methods for research workers*. Oliver and Boyd, Edinburgh, 14th ed., revised and enlarged edition, 1970. ISBN 0050021702. (Cited on page 73.)

[21] Fitlinxx. FitLinxx Professional, Fitness Tracking Website, Kiosk, 2013. URL http://www.fitlinxx.net/fitlinxx-professional-overview.htm, accessed 2013-06-12. (Cited on pages 2, 10, and 81.)

[22] S. J. Fleck and W. J. Kraemer. *Designing Resistance Training Programs-3rd*. Human Kinetics 1, 2004. URL http://books.google.nl/books?hl=nl&lr=&id=ylsfDoufD_4C&oi=fnd&pg=PR9&ots=f58QJKueCN&sig=_GY_dz6yj0H6bRioltUYPsMqLAk. (Cited on page 2.)

[23] Google. Sensor | Android Developers, 2012. URL http://developer.android.com/reference/android/hardware/Sensor.html, accessed 2012-07-24. (Cited on pages 19, 20, and 22.)

[24] Google. Design | Android Developers, 2013. URL http://developer.android.com/design/index.html, accessed 2013-06-03. (Cited on page 59.)

[25] JSON. JSON, 2012. URL http://json.org/, accessed 2012-12-23. (Cited on page 27.)

[26] G. A. Kelley and K. S. Kelley. Progressive resistance exercise and resting blood pressure a meta-analysis of randomized controlled trials. *Hypertension*, 35(3):838–843, 2000. URL http://hyper.ahajournals.org.proxy-ub.rug.nl/content/35/3/838.short. (Cited on page 4.)

[27] W. L. Kenney and E. J. Zambraski. Physical Activity in Human Hypertension. *Sports Medicine*, 1(6):459–473, 1984. URL http://link.springer.com.proxy-ub.rug.nl/article/10.2165/00007256-198401060-00005. (Cited on page 4.)

[28] W. J. Kraemer, K. Adams, E. Cafarelli, G. A. Dudley, C. Dooly, M. S. Feigenbaum, S. J. Fleck, B. Franklin, A. C. Fry, J. R. Hoffman, R. U. Newton, J. Potteiger, M. H. Stone, N. A. Ratamess, T. Triplett-McBride, and American College of Sports Medicine. American College of Sports Medicine position stand. Progression models in resistance training for healthy adults. *Medicine*

and science in sports and exercise, 34(2):364–380, February 2002. ISSN 0195-9131. PMID: 11828249. (Cited on pages 2 and 5.)

[29] M. Kranz, A. Möller, N. Hammerla, S. Diewald, T. Plötz, P. Olivier, and L. Roalter. The mobile fitness coach: Towards individualized skill assessment using personalized mobile devices. *Pervasive and Mobile Computing*, 2012. ISSN 1574-1192. doi: 10.1016/j.pmcj.2012.06.002. URL http://www.sciencedirect.com/science/article/pii/S1574119212000673. (Cited on pages 7, 8, and 9.)

[30] J. R. Kwapisz, G. M. Weiss, and S. A. Moore. Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter*, 12(2):74–82, 2011. URL http://dl.acm.org.proxy-ub.rug.nl/citation.cfm?id=1964918. (Cited on page 45.)

[31] G. F. Martel, D. E. Hurlbut, M. E. Lott, J. T. Lemmer, F. M. Ivey, S. M. Roth, M. A. Rogers, J. L. Fleg, and B. F. Hurley. Strength training normalizes resting blood pressure in 65-to 73-year-old men and women with high normal blood pressure. *Journal of the American Geriatrics Society*, 47(10):1215, 1999. URL http://www.ncbi.nlm.nih.gov.proxy-ub.rug.nl/pubmed/10522955. (Cited on page 4.)

[32] P. Michels, D. Gravenstein, and D. R. Westenskow. An integrated graphic data display improves detection and identification of critical events during anesthesia. *Journal of Clinical Monitoring*, 13(4):249–259, July 1997. ISSN 0748-1977. URL http://www.ncbi.nlm.nih.gov/pubmed/9269619. PMID: 9269619. (Cited on page 61.)

[33] D. Minnen, T. Starner, M. Essa, and C. Isbell. Discovering characteristic actions from on-body sensor data. In *Wearable Computers, 2006 10th IEEE International Symposium on*, page 11–18, 2006. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4067720. (Cited on pages 21 and 22.)

[34] M. Moffat, S. Vickery, and American Physical Therapy Association (1921-). *Book of body maintenance and repair*. Henry Holt, New York, 1999. ISBN 0805055711 9780805055719. (Cited on page 2.)

[35] V. Mooney, M. Kron, P. Rummerfield, and B. Holmes. The effect of workplace based strengthening on low back injury rates: a case study in the strip mining industry. *Journal of Occupational Rehabilitation*, 5(3):157–167, 1995. URL http://link.springer.com.proxy-ub.rug.nl/article/10.1007/BF02109956. (Cited on page 5.)

[36] M. Muehlbauer, G. Bahle, and P. Lukowicz. What can an arm holster worn smartphone do for activity recognition? In *2011 15th Annual International Symposium on Wearable Computers (ISWC)*, pages 79–82, 2011. doi: 10.1109/ISWC.2011.23. (Cited on page 45.)

[37] R. Negenborn. Robot localization and Kalman filters. Master's thesis, Utrecht University, 2003. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.9672&rep=rep1&type=pdf. (Cited on pages 24 and 26.)

[38] M. E. Nelson, M. A. Fiatarone, C. M. Morganti, I. Trice, R. A. Greenberg, and W. J. Evans. Effects of high-intensity strength training on multiple risk factors for osteoporotic fractures. *JAMA: the journal of the American Medical Association*, 272 (24):1909–1914, 1994. URL http://jama.ama-assn.org/content/272/24/1909.short. (Cited on page 4.)

[39] Nike. Nike+ FuelBand, 2013. URL http://www.nike.com/us/en_us/lp/nikeplus-fuelband, accessed 2013-01-19. (Cited on page 10.)

[40] NorthPark. Apps by NorthPark, 2012. URL https://play.google.com/store/apps/developer?id=NorthPark&hl=en, accessed 2012-07-24. (Cited on pages 9, 11, and 84.)

[41] R. S. Paffenbarger, R. T. Hyde, A. L. Wing, C. Bouchard, R. J. Shephard, T. Stephens, J. R. Sutton, and B. D. McPherson. Physical activity and physical fitness as determinants of health and longevity. In *Exercise, fitness, and health: a consensus of current knowledge: proceedings of the International Conference on Exercise, fitness and health, May 29-June 3, 1988, Toronto, Canada*, page 33–48, 1990. URL http://www.cabdirect.org/abstracts/19921892114.html. (Cited on page 5.)

[42] Pebble. Pebble, 2013. URL http://getpebble.com, accessed 2013-06-19. (Cited on page 10.)

[43] B. Peirce. Criterion for the rejection of doubtful observations. *The Astronomical Journal*, 2:161–163, 1852. URL http://adsabs.harvard.edu.proxy-ub.rug.nl/full/1852AJ......2..161P. (Cited on page 28.)

[44] I. Pernek, K. A. Hummel, and P. Kokol. Exercise repetition detection for resistance training based on smartphones. *Personal and Ubiquitous Computing*, pages 1–12, 2012. ISSN 1617-4909, 1617-4917. doi: 10.1007/s00779-012-0626-y. URL http://link.springer.com.proxy-ub.rug.nl/article/10.1007/s00779-012-0626-y. (Cited on pages 6, 7, 32, 49, 55, 59, and 85.)

[45] M. L. Pollock, B. A. Franklin, G. J. Balady, B. L. Chaitman, J. L. Fleg, B. Fletcher, M. Limacher, I. L. Piña, R. A. Stein, M. Williams, and T. Bazzarre. Resistance exercise in individuals with and without cardiovascular disease benefits, rationale, safety, and prescription. An advisory from the Committee on Exercise, Rehabilitation, and Prevention, Council on Clinical Cardiology, American Heart Association. *Circulation*, 101(7):828–833, February 2000. ISSN 0009-7322, 1524-4539. doi: 10.1161/01.CIR.101.7.828. URL http://circ.ahajournals.org/content/101/7/828. PMID: 10683360. (Cited on pages 2 and 4.)

[46] S. K. Powers. *Exercise physiology: theory and application to fitness and performance*. McGraw-Hill Higher Education, New York, NY, 7th ed edition, 2009. ISBN 9780073376479. (Cited on pages 4 and 5.)

[47] L. Rabiner and B. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4 –16, January 1986. ISSN 0740-7467. doi: 10.1109/MASSP.1986.1165342. (Cited on page 3.)

[48] R. K. Reeves, E. R. Laskowski, and J. Smith. Weight training injuries. Part I. *Phys Sportsmed*, 26:67–83, 1998. URL http://www.worldclassbodybuilding.com/forums/f484/weight-training-injuries-part-1-a-1531/. (Cited on page 2.)

[49] I. Rish. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, page 41–46, 2001. URL http://www.cc.gatech.edu/home/isbell/classes/reading/papers/Rish.pdf. (Cited on page 6.)

[50] S. M. Ross. Peirce's criterion for the elimination of suspect experimental data. *Journal of Engineering Technology*, 20 (2):38–41, 2003. URL http://classes.engineering.wustl.edu/2009/fall/che473/handouts/OutlierRejection.pdf. (Cited on page 28.)

[51] RunKeeper. RunKeeper - GPS Track Run Walk - Android-apps op Google Play, 2013. URL https://play.google.com/store/apps/details?id=com.fitnesskeeper.runkeeper.pro, accessed 2013-06-18. (Cited on pages 2 and 9.)

[52] Runtastic. Apps by Runtastic, 2013. URL http://play.google.com/store/search?q=pub:runtastic, accessed 2013-06-19. (Cited on pages 9 and 11.)

[53] Runtastic. Runtastic PRO - Android Apps on Google Play, 2013. URL https://play.google.com/store/apps/details?id=com.runtastic.android.pro, accessed 2013-06-19. (Cited on page 9.)

[54] Samsung. Samsung Galaxy S2 (S II) - SAMSUNG UK - OVERVIEW, 2012. URL http://www.samsung.com/uk/consumer/mobile-devices/smartphones/android/GT-I9100LKAXEU, accessed 2013-01-03. (Cited on page 18.)

[55] K. Sato, S. L. Smith, and W. A. Sands. Validation of an accelerometer for measuring sport performance. *Journal of Strength and Conditioning Research*, 23(1):341–347, January 2009. ISSN 1064-8011. doi: 10.1519/JSC.0b013e3181876a01. URL http://journals.lww.com.proxy-ub.rug.nl/nsca-jscr/Abstract/2009/01000/Validation_of_an_Accelerometer_for_Measuring_Sport.50.aspx. (Cited on page 2.)

[56] N. Schoonderwoerd. Wie zijn die 313.852 Nederlandse twitteraars? [Who are those 313.852 Twitter users?], 2010. URL http://nl.twirus.com/details/blog/713/Wie-zijn-die-313.852-Nederlandse-twitteraars?, accessed 2013-06-06. (Cited on page 62.)

[57] Six To Start. Apps by Six to Start, 2013. URL https://play.google.com/store/apps/developer?id=Six+to+Start, accessed 2013-06-19. (Cited on page 9.)

[58] S. Thiemjarus. A device-orientation independent method for activity recognition. In *2010 International Conference on Body Sensor Networks (BSN)*, pages 19 –23, June 2010. doi: 10.1109/BSN.2010.55. (Cited on page 45.)

[59] Usabilla. Usabilla - A new standard in website feedback, 2013. URL https://usabilla.com/index2, accessed 2013-06-03. (Cited on page 61.)

[60] S. Vickery and M. Moffat. *The American Physical Therapy Association Book of Body Maintenance and Repair*. Holt Paperbacks, 1 edition, April 1999. ISBN 0805055711. (Cited on page 2.)

[61] W. Wang, Y. Guo, B. Huang, G. Zhao, B. Liu, and L. Wang. Analysis of filtering methods for 3D acceleration signals in body sensor network. In *Bioelectronics and Bioinformatics (ISBB), 2011 International Symposium on*, page 263–266, 2011. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6107697. (Cited on page 24.)

[62] R. A. Winett and R. N. Carpinelli. Potential health-related benefits of resistance training. *Preventive Medicine*, 33(5):503–513, November 2001. ISSN 0091-7435. doi: 10.1006/pmed.2001.0909. URL http://www.sciencedirect.com/science/article/pii/S0091743501909090. (Cited on pages 2, 4, and 5.)

[63] J. Wortham. Nike Fuelband tracks physical activity inconsistently. *The New York Times*, July 2012. ISSN 0362-4331. URL http://www.nytimes.com/2012/07/29/technology/nike-fuelband-tracks-physical-activity-inconsistently.html. (Cited on page 10.)