Recognizing Gym Exercises Using Acceleration Data from Wearable Sensors

Heli Koskimäki and Pekka Siirtola
Computer Science and Engineering Department
P.O. BOX 4500, FI-90014, University of Oulu, Oulu, Finland
Email: heli.koskimaki@ee.oulu.fi, pekka.siirtola@ee.oulu.fi

***Abstract*—The activity recognition approaches can be used for entertainment, to give people information about their own behavior, and to monitor and supervise people through their actions. Thus, it is a natural consequence of that fact that the amount of wearable sensors based studies has increased as well, and new applications of activity recognition are being invented in the process. In this study, gym data, including 36 different exercise classes, is used aiming in the future to create automatic activity diaries showing reliably to end users how many sets of given exercise have been performed. The actual recognition is divided into two different steps. In the first step, activity recognition of certain time intervals is performed and in the second step the state-machine approach is used to decide when actual events (sets in gym data) were performed. The results showed that when recognizing different exercise sets from the same occasion (sequential exercise sets), on average, over 96 percent window-wise true positive rate can be achieved, and moreover, all the exercise events can be discovered using the state-machine approach. When using a separate validation test set, the accuracies decreased significantly for some classes, but even in this case, all the different sets were discovered for 26 different classes.**

***Keywords*—Activity recognition, event recognition, gym data, wearable sensors, accelerometer**

I. Introduction

During the last decade, many practical constraints related to carry-on sensors like accelerometers, magnetometers and GPS-receivers have been solved. This has enabled monitoring and classification of human activity based on information from wearable sensors to become a growing research area of pattern recognition and machine learning. The grounds for that lie in high commercial potential of context-aware applications and user interfaces. Moreover, activity recognition can be utilized to attack some of the serious societal challenges like rehabilitation, sustainability, elderly care and health, but also in industry, for worker monitoring and proactive instruction systems. The sensors are also included in todays smart phones constituting an extensive group of possible leisure users.

The first activity recognition study based on wearable sensors was published in 2000 by Van Laerhoven [1]. In the study, the daily activities were recognized was based on two acceleration sensors attached to pants. The daily activity recognition is still one of the most studied problems in the area (e.g. [2], [3], [4], [5]) but on the other hand, the promising results have been deployed in implementing commercial products like Polar Loop [6], Nike Fuel [7] and Fitbit [8]. For example, in Polar Loop, five different activity zones are recognized (resting, sitting, low, medium and high). Moreover, also the mobile phone sensors based studies have increased. For example, in [9] a mobile phone was used to recognize walking, running, cycling, riding a car and idling, and also Google [10] and Apple [11] have released their own versions of mobile phone-based activity recognition which can be implemented to any mobile application.

Another interesting trend in activity recognition is the utilization of the developed methods into sport activity recognition in general level [12] or in sport specific areas like in swimming [13]. A very interesting application area in sports is related to gym exercises. The gym approach provides multiform data but also enables to test the developed activity recognition methods and further apply them into other activity recognition problems. On the other hand, there are high commercial potential if reliable enough approaches are discovered.

There are quite a few studies concentrating on wearable sensors based gym exercise recognition like [14], [15] and [16]. The results achieved in those are very promising but the data sets are quite sparse compared to high variety of possible gym exercises. In those studies, the person independent approach has been the main research interest on the behalf of exercise variability. From the exercise point of view, the recognition is based only on nine or ten exercises.

In this study, a person specific recognition approach using gym data, including continuously measured acceleration information from 30 different upper-body exercises, is introduced. Thus, the data set also includes the so-called NULL-data consisting of data between the actual exercises which is not needed to be recognized. The actual recognition was divided into two different steps enabling the system also to work online. In the first step, activity recognition of certain time intervals is performed and in the second step the state-machine approach is used to decide when actual events (=sets in gym data) are performed. The actual recognition rates are presented as window-wise accuracies to three different classifiers and to different sensor combinations using metrics introduced in [17]. Also event-based accuracies are modified based on the article.

II. Data collection

The data were collected using a GeneActiv 3D accelerometer [19] at a frequency of 100 Hz. The accelerometer can be used as a wrist-worn or it can be attached with straps to other body locations. In this study, two accelerometers were used, one in the left wrist and another in torso (using a chest-strap like as in heart-rate monitors). There are studies where the optimum amount of sensors and their locations are investigated

978-1-4799-4518-4/14/$31.00 ©2014 IEEE

TABLE I. EXERCISES, THE NAMES CONSISTENT WITH A WEBPAGE [18].

Class	Name	Muscle group	Posture	One-arm, both or alternate	Equipment
1	Close-Grip Barbell Bench Press	Triceps	On back	Both	Barbell
2	Bar Skullcrusher	Triceps	On back	Both	Barbell
3	Triceps Pushdown	Triceps	Standing	Both	Cable rope
4	Bench Dip / Dip	Triceps	Weight on hands	Both	Own weight
5	Overhead Triceps Extension	Triceps	Standing	Both	Barbell Plate
6	Tricep Dumbbell Kickback	Triceps	Bent Over	One-arm	Dumbbell
7	Spider Curl	Biceps	Sitting	Both	E-Z Curl Bar
8	Dumbbell Alternate Bicep Curl	Biceps	Standing	Alternate	Dumbbell
9	Incline Hammer Curl	Biceps	Seated inclined	Both	Dumbbell
10	Concentration Curl	Biceps	Seated	One-arm	Dumbbell
11	Cable Curl	Biceps	Standing	Both	Cable Bar
12	Hammer Curl	Biceps	Standing	Alternate	Dumbbell
13	Upright Barbell Row	Shoulders	Standing	Both	Barbell
14	Side Lateral Raise	Shoulders	Standing	Both	Dumbbell
15	Front Dumbbell Raise	Shoulders	Standing	Alternate	Dumbbell
16	Seated Dumbbell Shoulder Press	Shoulders	Seated	Both	Dumbbell
17	Car Drivers	Shoulders	Standing	Both	Barbell Plate
18	Lying Rear Delt Raise	Shoulders	On stomach	Both	Dumbbell
19	Bench Press	Chest	On back	Both	Barbell
20	Incline Dumbbell Flyes	Chest	Seated inclined	Both	Dumbbell
21	Incline Dumbbell Press	Chest	Seated inclined	Both	Dumbbell
22	Dumbbell Flyes	Chest	On back	Both	Dumbbell
23	Pushups	Chest	On hands & knees	Both	Own weight
24	Leverage Chest Press	Chest	Seated	Both	Machine
25	Seated Cable Rows	Middle Back	Seated	Both	Cable
26	One-Arm Dumbbell Row	Middle Back	Bent Over	One-arm	Dumbbell
27	Wide-Grip Pulldown Behind The Neck	Lats	Sitting	Both	Cable
28	Bent Over Barbell Row	Middle Back	Bent Over	Both	Barbell
29	Reverse Grip Bent-Over Row	Middle Back	Bent Over	Both	Barbell
30	Wide-Grip Front Pulldown	Lats	Sitting	Both	Cable
6b	Tricep Dumbbell Kickback	Triceps	Bent Over	Counter arm	Dumbbell
10b	Concentration Curl	Biceps	Seated	Counter arm	Dumbbell
26b	One-Arm Dumbbell Row	Middle Back	Bent Over	Counter arm	Dumbbell
96	Cross-trainer	Warm-up	Standing	Alternate	Cross-trainer
97	Rowing seated	Warm-up	Seated	Both	Row Machine
98	Walking	Warm-up	Standing	Alternate	Treadmill
99	NULL				

[20], but for this study, two was considered as a maximum amount of sensors that a person would be comfortable with when exercising. Also the location of sensors was decided based on the end-user view. The mobile phone usage was ruled out, while a sensible and comfortable position, not e effecting to the exercise, was not found out in a quick survey.

The data were collected from a single person from 30 different exercises, each of them consisting of three sets of ten repetitions. The exercises were mostly done using free weights, and for every upper body muscle group, data from six different exercises were collected. In addition, data from walking, cross- trainer and rowing for warming up and cooling down were collected. Similar data sets were collected twice; the first set for training and the second for independent testing. Altogether, there were 36 activities to be recognized, when considering the data from the counter arm resting as a class of its own in some exercises.

The actual exercises are introduced in detail in Table I. The numbering is formed so that the basic exercises are numbered from 1 to 30 and the counter arm as b-version (rows 31 - 33). The warming up and cooling down sequences are numbered from 96 to 98 to separate them from the basic exercises.

While the data set was gathered as a continuous signal, the data set constituted also data between every exercise set in which the subject moved around at the gym, changed weight, stretched or just stayed still. In this study, this data is called NULL-data but also the term other activities could be used. In Table I, this is labeled with 99. Altogether, there were six hours of data (three for training and three for independent testing), where approximately 65 percent was considered as NULL-data. This data is also available for further studies from [21].

III. METHODS

A. Window-wise recognition

The recognition was approached as an online system, and thus, a sliding window method was used to divide the time series signals into smaller sequences. Results were obtained by the window length of two seconds with a slide of 0.5 seconds between two sequential windows. Using the sliding window method, altogether 23,000 windows were obtained for training and 21,000 for testing. For feature extraction, widely used features were selected as in [22]. For every window, 42 different features were calculated consisting of simple statistical values calculated for every channel of acceleration separately, frequency domain features, as well as correlation features between different channels.

For window-wise classification, three basic classifiers were utilized: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and the k nearest neighbors classifier (KNN). The LDA and QDA are parametric classifying methods relying on estimation on multivariate normal densities [23]. The main difference between the methods is that LDA tries to find linear class boundaries, while class boundaries in QDA

are nonlinear. The KNN classifier is a non-parametric method which relies heavily on the training data requiring more memory and calculation capacity in the actual classification phase compared to the parametric methods. Nevertheless, it is widely used for classification due its capability to create multiform class boundaries. The basic idea of KNN classification is quite simple: a data point is classified into the class where most of its k nearest neighbors belong [24]. The nearest neighbors are defined using, for example, the Euclidean distance measure [25]. In cases where there is a tie with the majority classes the class with a smaller average distance is selected.

The window-wise classification rates are shown in this article using TPR, FPR and PR metrics proposed in [17]. The actual equations used were: *True Positive Rate*, (TPR)= True Positives / (True Positives + False Negatives), *False Positive Rate*, (FPR)= False Positives / (True Negatives + False Positives), and *Precision* (PR)= True Positives / (True Positives + False Positive).

B. Event-wise recognition

For the event-wise classification, state-machine approach presented in [26] was used to decide when actual events (=sets in gym data) were performed. The state-machine estimates were based on the classifications achieved in the previous step.

In a nutshell, an event is decided as started when some predefined amount of windows having the same activity estimate inside a certain time frame is found and the event is marked as ended when a certain amount of windows having a different activity estimate from than the latest event are achieved. With this kind of approach, the overall recognition system is allowed to work with a lag of few seconds.

For events recognition rates a slightly modified version of event summary from [17] was used. The eight different types of event error scores are presented together with correct events (C) as event analysis diagram (EAD) in Figure 1. Four of the error scores are applied to ground truth events: deletions (D), fragmented (F), fragmented and merged (FM) and merged (M). The remaining four are applicable to returned events: merging (M), fragmenting and merging (FM), fragmenting (F) and insertions (I). The modification of this diagram in this study arises from the fact that all the exercises and sets of gym data are considered in a single diagram instead of using dedicated own diagrams for different exercises.

Algorithm 1 Transitions of a state machine

E = event estimate
W_s, W_e = amount of windows needed that event can be reliable recognized as started or ended
p = minimum percentage of windows of same class in a sequence
a_i = window i
$c(a_i)$ = class of a_i
$U(a_i)$ = union of windows having the same class $c(a_i) \neq null$ and being close enough with each other

for for each window a_i **do**
  if $\exists\, U(a_j) \neq \emptyset$ so, that $c(a_i) = c(a_j) \forall\, a_j \in U(a_j) \wedge \exists\, a_j \in U(a_j)$ so, that $d(a_i, a_j)) < W_e$ **then**
    $U(a_i) = U(a_j) \leftarrow a_i$
  else
    if $c(a_i) \neq null$ **then**
      create union $U(a_i) \leftarrow a_i$
    end if
  end if
  if $\#U(a_i) \geq W_s \wedge \frac{max(d(a_i,a_j))}{\#U(a_i)} \geq \frac{p}{100}, a_j, a_i \in U(a_i) \wedge E \neq c(a_i)$ **then**
    $E = c(a_i)$
    Event start $\leftarrow \min_i U(a_i)$
  end if
  if $E \neq null \ \wedge (i - \max_j U(a_j)) \geq W_e$ **then**
    $E = null$
    Event end $\leftarrow \max_j U(a_j)$
  end if
end for

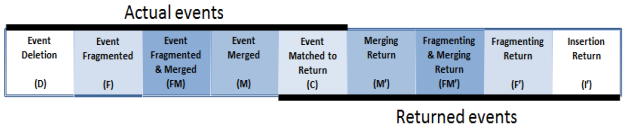

Fig. 1. Event analysis diagram (EAD)

IV. Classifier Training Phase

A. Window-wise recognition

In the first step of the study, the different classifiers, LDA, QDA and KNN, as well as the optimal value for k were compared. The data set in this phase consisted of altogether 23,000 windows obtained from the first data collection. The data set from the second data collection was left out as a separate validation set. Three-fold cross-validation was used to divide the first data set into training and testing data sets; every set of ten repetitions was used as testing data separately. A random division of data sets was considered but due the overlap in the sliding window method, it was rejected to avoid approximately the same instances from belonging to both sets.

1) Optimal value of k: The optimal value for k was studied at first. Values from 1 to 11 were tested and the value of 9 was selected because it corresponded to the highest true positive rate with the lowest standard deviation within the three-fold cross-validation.

2) Comparison of different classifiers: The achieved classification results for linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and the k nearest neighbors classifier (KNN) are shown in Table II. As it can be seen, the average of true positive rates is clearly highest for the KNN classifier. The false positive rate is naturally highest in the NULL-data class due to the proportion and versatility of NULL-data in the data set. Thus, the precision rates are more informative when comparing actual exercise classes. In the precision rate, the QDA classifier clearly outperforms the KNN classifier. In a short, KNN finds better than QDA the performed exercises, but also makes more miss-classifications into those classes. The overall results of LDA are significantly inferior and thus it is ruled out from further discussions.

When studying the results from Table II in more detail, several interesting phenomena can be found. When using the KNN classifier, the overall classification accuracies are quite good. Within 22 classes, the average true positive rate has been over 90 percent and only within seven classes the average TPR has been below 80 percent (Dumbbell Alternate Bicep Curl (8), Seated Dumbbell Shoulder Press (16), Bench Press (19), Incline Dumbbell Press (21), Wide-Grip Front Pulldown (30), Counter Arm Concentration Curl (10b) and Cross-trainer (96)).

TABLE II. TRUE POSITIVE, FALSE POSITIVE AND PRECISION RATES FOR ALL THE ACTIVITIES USING DIFFERENT CLASSIFIERS

Class	KNN			LDA			QDA		
	TPR	**FPR**	**PR**	**TPR**	**FPR**	**PR**	**TPR**	**FPR**	**PR**
1	86.2 ± 10.4	0.5	63.2±11.1	92.5 ± 6.0	1.1	42.2 ± 2.9	86.8 ± 12.1	0.1	89.5 ± 3.8
2	98.0 ± 1.4	0.1	93.4±6.5	86.7 ± 1.2	0.2	85.5 ± 10.8	91.3 ± 12.1	0	100 ± 0
3	93.1± 2.6	0.4	70.3± 21.0	91.1 ± 5.7	1.6	32.2 ± 6.0	89.2± 9.2	0.0	97.0 ± 2.1
4	98.1± 1.8	0.1	87.0± 11.9	98.1 ± 1.8	1.1	36.5 ± 5.5	80.7± 13.8	0.0	95.3 ± 8.1
5	98.0± 2.3	0.1	91.6± 2.4	93.2 ± 5.2	0.2	76.4 ± 6.8	95.3± 2.6	0	100 ± 0
6	96.6± 4.2	0.0	98.6± 1.2	98.6 ± 1.2	0.3	70.9 ± 1.0	84.3± 3.0	0	100 ± 0
7	98.5± 2.7	0.1	92.4± 5.6	72.9 ± 2.7	0.5	64.6 ± 7.8	92.6± 4.6	0.0	97.5 ± 2.9
8	79.8± 6.1	0.2	82.8± 5.9	76.3 ± 15.7	0.8	46.0 ± 3.5	88.7± 4.2	0.3	73.4 ± 12.0
9	91.6± 6.9	0.3	79.2± 0.9	66.5 ± 6.3	0.5	60.2 ± 12.2	94.4± 5.6	0.1	88.9 ± 3.9
10	93.0± 12.2	0.1	89.2± 13.3	75.4 ± 18.1	0.4	62.2 ± 6.9	95.5± 5.2	0.0	98.1 ± 6.2
11	100± 0	0.2	84.9± 12.7	77.0 ± 5.8	0.4	67.6 ± 2.7	98.7 ± 2.3	0.0	98.8 ± 1.3
12	87.0± 14.2	0.1	83.2± 5.3	80.4 ± 15.8	0.9	44.4 ± 17.6	91.1 ± 7.0	0.0	97.8 ± 1.9
13	97.0± 5.2	0.1	90.8± 3.1	93.3 ± 8.6	1.1	38.5 ± 9.6	89.5± 5.9	0.0	99.0 ± 1.2
14	98.8± 2.1	0.1	93.0± 4.3	81.1 ± 7.8	0.9	39.0 ± 1.6	81.9± 14.9	0.0	98.7 ± 1.6
15	93.1± 3.0	0.3	81.0± 10.6	75.3 ± 5.6	2.2	32.6 ± 10.3	97.5± 1.8	1.3	47.0 ± 22.9
16	66.8± 3.8	0.4	67.9± 6.3	78.2 ± 9.5	0.9	50.1 ± 5.7	75.7± 4.3	0.1	85.6 ± 2.8
17	86.4± 1.2	0.0	93.5± 9.4	86.6 ± 7.9	0.4	56.9 ± 11.0	69.1± 18.3	0.0	97.9 ± 3.6
18	89.2± 11.0	0.1	86.8± 8.2	95.4 ± 0.1	0.4	57.6 ± 9.0	45.4 ± 12.8	0	100 ± 0
19	60.7± 25.9	0.3	66.4± 4.5	9.1 ± 6.0	0.3	19.0 ± 8.1	61.8 ± 51.6	0.1	85.3 ± 29.2
20	93.6± 4.0	0.4	77.9± 24.2	88.7 ± 4.3	0.6	62.6 ± 8.0	92.9 ± 6.1	0.0	94.6 ± 4.7
21	74.8± 7.5	0.4	67.5± 6.2	73.8 ± 5.7	0.6	60.1 ± 10.3	90.8± 1.5	0.2	84.2 ± 0.9
22	89.7± 2.5	0.2	89.8± 10.7	45.6 ± 10.3	0.5	55.8 ± 16.5	85.1± 4.9	0.2	86.5 ± 9.4
23	87.6± 2.9	0.1	90.3± 5.0	91.3 ± 4.2	0.7	50.7 ± 13.8	22.9± 36.3	0	NaN (100)
24	97.0± 0.4	0.6	68.5 ± 24.3	98.9 ± 1.1	0.2	85.2 ± 12.4	88.5± 7.3	0.0	96.7 ± 5.7
25	95.6± 1.9	0.3	78.1± 4.5	98.5 ± 1.7	10.1	10.7 ± 2.8	95.1± 6.5	0.1	88.4 ± 5.6
26	93.3± 7.0	0.2	78.1± 20.8	97.1 ± 2.5	2.8	22.5 ± 9.4	68.2± 33.3	0.0	97.6 ± 3.6
27	83.4± 5.9	0.2	58.7± 13.5	78.2 ± 8.8	1.0	49.5 ± 11.6	86.6± 12.2	0.4	79.4 ± 22.2
28	97.3± 3.0	0.1	92.1± 13.7	83.5 ± 0.9	0.2	71.8 ± 6.8	72.9± 7.9	0	100 ± 0
29	95.4± 2.9	0.2	83.2± 18.1	90.7 ± 4.7	0.6	51.6 ± 12.3	81.3± 3.6	0.0	98.8 ± 1.4
30	68.2± 44.0	0.3	68.0± 13.1	72.3 ± 27.7	1.3	38.9 ± 3.1	58.8± 51.0	0.2	82.9 ± 50.1
6b	90.7± 6.3	0.2	72.3± 3.0	92.1 ± 5.8	2.1	23.4 ± 7.4	39.1± 44.7	0	NaN (100)
10b	75.7± 17.9	0.2	76.4± 16.7	93.3 ± 9.0	5.0	17.8 ± 7.3	85.0± 10.0	0.5	71.9 ± 26.5
26b	86.1± 10.5	0.2	75.6± 25.9	96.2 ± 2.2	2.9	21.0 ± 8.5	58.5± 40.6	0.0	98.3 ± 3.2
96	75.8± 33.9	0.2	87.3± 18.2	54.6 ± 36.7	0.6	61.3 ± 28.6	52.8± 40.5	0.0	96.8 ± 12.1
97	84.1± 27.1	0.2	88.5± 8.2	84.5 ± 26.4	0.7	71.1 ± 4.4	78.7± 34.2	0	100 ± 0
98	97.4± 2.2	0.0	95.9± 2.5	97.4 ± 2.2	0.8	56.9 ± 11.4	93.1± 6.3	0	100 ± 0
99	91.4± 1.9	5.8	96.7± 0.6	41.9 ± 2.2	0.9	98.9 ± 0.3	96.1± 1.2	15.9	90.2 ± 1.6
mean	**88.9 ± 1.4**	**0.4**	82.2 ± 2.4	81.3 ± 1.0	0.7	51.1 ± 1.4	79.9 ± 2.1	0.5	**90.8 ± 2.8***

For three of these (8, 16 and 21), the poor recognition accuracy has been consisted with all the three testing sets while the standard deviation has been low. This could mean that the used features and classifier are not optimal for these activities. As a confirmation to this assumption, the QDA classifier accuracies can be seen to be remarkably higher.

On the other hand, for the other four poorly classified classes (19, 30, 10b and 96), the classification accuracies have varied remarkably between different test sets. The accuracies have changed within three testing sets to those classes between [31.0, 78.9], [17.6, 96.7], [56.7, 92.2] and [37.0, 99.4], respectively. For the cross-trainer, the change is due to the change in grip from top of the handle to the middle of the handle, but for the upper body exercises, the different rates correlate with weight change, change in posture or in exercise order (muscles getting tired) when doing the exercise. Similar variation can also be found from six other classes. Nevertheless, this variation should not affect the final classification while the variation between sets is normal during gym exercises.

If considering the QDA classifier results, similar findings can be noticed. Approximately in half of the instances, the standard deviation of TPR between different testing sets have been more than ten (in eight, even more than 30) reflecting to the high variance between different exercise sets. When considering the true positive rates and precision rates between the KNN and QDA, the KNN classifier considered to outperform the QDA. Thus, it was selected to be used in the further steps.

3) Comparison of sensors separately and jointly: As a second comparison of the study, the data from different sensors separately and jointly was classified using KNN. The results for these are introduced in Table III. As assumed, the average activity recognition rates by using merely chest sensor are significantly lower with every metric compared to the results from the wrist sensor. Moreover, by utilizing the data from both sensors, a very high overall recognition rate is achieved.

In a closer study, it can be seen that using the data from both sensors, the class-wise accuracy has been improved in almost every class. An exception to this is the exercise class 29 (Reverse Grip Bent-Over Row) and NULL-data, but the decreases are quite small. Moreover, only with three classes, the true positive rate is below 90 percent (Dumbbell Alternate Bicep Curl (8), Bench Press (19) and Cross-trainer (96)).

Also an interesting aspect is the fact that the true positive rates using both sensors can improve significantly compared to the separate results. For example, inside classes Seated Dumbbell Shoulder Press (16) and Wide-Grip Front Pulldown (30), the classification rates from wrist data are 66.8 % and 68.2 % and from chest data 56.6 % and 35.3 %, respectively, but the combined accuracy is over 90 percent.

Thus with this data set, the window-wise recognition rates of gym data are very promising using two sensors.

TABLE III. TRUE POSITIVE, FALSE POSITIVE AND PRECISION RATES FOR ALL THE ACTIVITIES USING THE SENSORS SEPARATELY AND JOINTLY. THE RESULTS FOR WRIST SENSOR ARE THE SAME PRESENTED ALSO IN TABLE II.

Class	wrist			chest			both sensors		
	TPR	**FPR**	**PR**	**TPR**	**FPR**	**PR**	**TPR**	**FPR**	**PR**
1	86.2 ± 10.4	0.5	63.2±11.1	77.1± 19.8	0.2	84.1 ± 3.1	91.2 ± 8.5	0.2	80.5 ± 13.7
2	98.0 ± 1.4	0.1	93.4±6.5	97.7 ± 1.2	0.5	76.6 ± 27.0	98.8 ± 2.1	0.1	91.6 ± 13.4
3	93.1± 2.6	0.4	70.3± 21.0	45.0 ± 17.9	0.2	66.2 ± 17.5	95.2± 1.7	0.2	82.8 ± 8.2
4	98.1± 1.8	0.1	87.0± 11.9	91.7 ± 5.5	0.4	73.6 ± 26.3	98.1± 1.8	0.1	87.6 ± 12.9
5	98.0± 2.3	0.1	91.6± 2.4	70.6 ± 5.5	0.2	77.9 ± 4.2	99.0± 0.9	0.1	91.3 ± 6.4
6	96.6± 4.2	0.0	98.6± 1.2	99.7 ± 0.6	0.5	67.6 ± 30.1	98.6± 1.2	0.1	91.9 ± 10.8
7	98.5± 2.7	0.1	92.4± 5.6	85.1 ± 8.8	0.4	67.3 ± 3.8	99.6± 0.7	0.5	75.1 ± 24.0
8	79.8± 6.1	0.2	82.8± 5.9	56.3 ± 14.6	0.8	45.1 ± 12.1	86.1± 5.7	0.3	72.3 ± 8.7
9	91.6± 6.9	0.3	79.2± 0.9	98.4 ± 0.5	0.6	77.8 ± 32.2	99.2± 0.7	0.5	81.0 ± 28.7
10	93.0± 12.2	0.1	89.2± 13.3	96.7 ± 4.1	0.1	91.5 ± 5.0	98.7± 2.3	0.1	92.1 ± 13.6
11	100± 0	0.2	84.9± 12.7	76.2 ± 11.8	0.5	63.3 ± 15.4	100 ± 0	0.2	97.6 ± 19.1
12	87.0± 14.2	0.1	83.2± 5.3	48.7 ± 13.8	0.7	38.4 ± 13.0	99.4 ± 1.0	0.2	81.3 ± 13.0
13	97.0± 5.2	0.1	90.8± 3.1	57.3 ± 13.1	0.2	71.0 ± 13.2	100± 0	0.1	85.4 ± 12.7
14	98.8± 2.1	0.1	93.0± 4.3	51.4 ± 19.9	0.4	49.9 ± 16.3	99.4± 1.1	0.1	86.3 ± 3.3
15	93.1± 3.0	0.3	81.0± 10.6	79.4 ± 14.1	0.7	67.8 ± 12.9	98.6± 1.6	0.3	82.9 ± 10.7
16	66.8± 3.8	0.4	67.9± 6.3	56.6 ± 11.4	0.6	56.4 ± 16.9	97.5± 2.5	0.1	91.8 ± 9.3
17	86.4± 1.2	0.0	93.5± 9.4	55.8 ± 14.4	0.2	69.5 ± 14.9	92.9± 5.0	0.1	91.8 ± 8.7
18	89.2± 11.0	0.1	86.8± 8.2	96.6 ± 3.0	0.2	80.8 ± 17.7	97.7 ± 2.3	0.1	86.2 ± 19.9
19	60.7± 25.9	0.3	66.4± 4.5	62.1 ± 8.6	0.7	52.7 ± 11.6	78.9 ± 17.2	0.3	74.6 ± 19.1
20	93.6± 4.0	0.4	77.9± 24.2	83.0 ± 13.4	1.5	54.6 ± 30.0	99.2 ± 0.7	0.0	97.7 ± 2.4
21	74.8± 7.5	0.4	67.5± 6.2	80.4 ± 11.1	0.8	69.3 ± 30.0	99.3± 1.2	0.0	98.4 ± 2.9
22	89.7± 2.5	0.2	89.8± 10.7	69.1 ± 3.9	0.7	58.9 ± 12.4	96.9± 2.2	0.3	86.2 ± 13.2
23	87.6± 2.9	0.1	90.3± 5.0	86.3 ± 12.1	0.3	70.1 ± 24.1	97.0± 2.7	0.2	85.1 ± 20.9
24	97.0± 0.4	0.6	68.5 ± 24.3	83.0 ± 3.3	0.4	69.8 ± 12.6	98.5± 1.3	0.3	83.9 ± 16.0
25	95.6± 1.9	0.3	78.1± 4.5	65.3 ± 4.5	0.5	65.7 ± 11.0	98.9± 1.1	0.4	76.5 ± 8.4
26	93.3± 7.0	0.2	78.1± 20.8	94.2 ± 5.7	0.1	90.5 ± 1.7	98.2± 2.0	0.1	87.6 ± 5.8
27	83.4± 5.9	0.2	58.7± 13.5	66.7 ± 3.5	0.9	52.9 ± 23.1	93.7± 5.8	0.3	80.2 ± 18.6
28	97.3± 3.0	0.1	92.1± 13.7	50.3 ± 36.6	0.3	56.0 ± 11.8	97.9± 0.1	0.0	93.0 ± 6.2
29	95.4± 2.9	0.2	83.2± 18.1	62.1 ± 14.9	0.4	57.0 ± 13.2	93.9± 4.2	0.1	86.0 ± 10.2
30	68.2± 44.0	0.3	68.0± 13.1	35.3 ± 16.8	0.6	45.0 ± 11.5	91.6± 8.0	0.3	80.3 ± 13.7
6b	90.7± 6.3	0.2	72.3± 3.0	98.6 ± 1.2	0.1	88.2 ± 9.3	98.6± 1.2	0.1	89.0 ± 15.7
10b	75.7± 17.9	0.2	76.4± 16.7	91.9 ± 7.7	0.1	86.6 ± 9.3	97.5± 1.7	0.1	86.0 ± 10.8
26b	86.1± 10.5	0.2	75.6± 25.9	99.7 ± 0.5	0.1	92.3 ± 6.7	99.3± 1.2	0.1	92.8 ± 6.8
96	75.8± 33.9	0.2	87.3± 18.2	95.0 ± 4.0	0.0	98.9 ± 0.4	86.7± 20.0	0.2	90.5 ± 14.8
97	84.1± 27.1	0.2	88.5± 8.2	93.7 ± 2.5	0.4	84.2 ± 2.6	94.2± 9.5	0.2	91.1 ± 2.1
98	97.4± 2.2	0.0	95.9± 2.5	100 ± 0	0.1	95.1 ± 6.5	97.9± 2.7	0.0	99.6 ± 0.8
99	91.4± 1.9	5.8	96.7± 0.6	81.4 ± 1.3	12.6	91.0 ± 1.5	91.3± 2.0	2.8	98.4 ± 0.4
mean	88.9 ± 1.4	0.4	82.2 ± 2.4	76.7 ± 4.7	0.8	70.4 ± 5.8	**96.2 ± 0.7**	**0.2**	**86.9 ± 1.6**

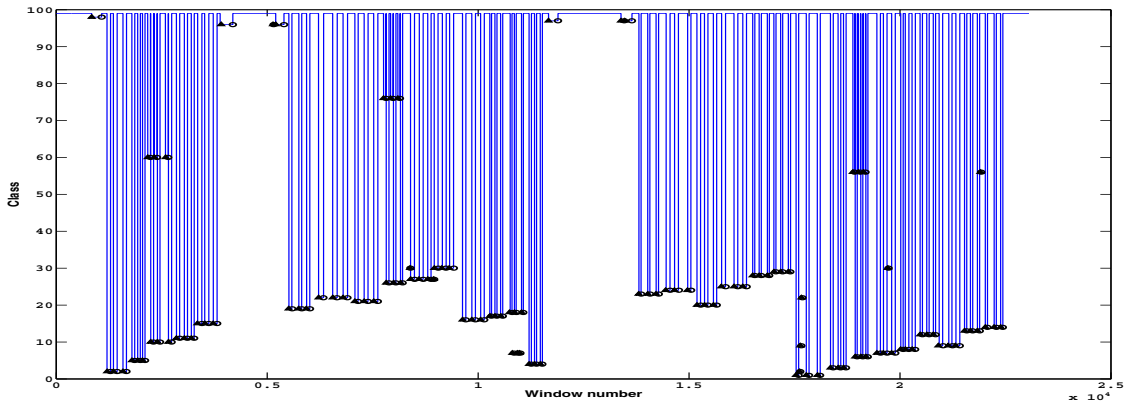

Fig. 2. Training data event recognition results. A beginning of an event is marked as triangle and the end as sphere. Thus if the blue line (actual class)and triangles/spheres intersect each other the event is recognized correctly.

B. Event-wise recognition

The results shown by now were window-wise accuracies, but when wanting to shift the interest more towards event recognition, the state-machine approach was utilized, while it can be assumed that the potential end-users are interested in automatic activity diaries showing reliably how many sets of which given exercise they have performed.

The estimated classes for every window were passed one at a time to state-machine introduced as Algorithm 1. The transitions of the state machine were decided utilizing a priori information about the duration of the tasks and the activity recognition accuracy. For example, in this study, the $W_s, W_e =$

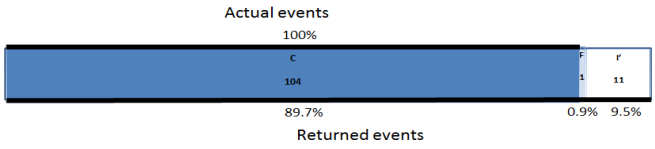

Fig. 3. Training data event analysis diagram (EAD) of all the exercises and sets of gym data: correctly found events (C), fragmented events (F) and inserted events (I).

8 for every class and the percentage was decided utilizing true positive and precision rates with their standard deviation so that $p(a_i) = (3 * (tpr(a_i) - std(a_i)) + pr(a_i) - std(a_i))/4$. The actual start and end moments of the events were then achieved using the algorithm.

The state machine results achieved are shown in Figure 2. The x-axis denotes the number of sequential windows (23, 000 windows). The y-axis denotes the activity classes for every window. The actual classes are drawn with blue lines. The beginnings of event estimation are marked as triangles while the sphere tells when the exercise has ended.

As it can be seen from Figures 2 and 3 (Event Analysis Diagram), the state machine approach has performed quite accurately. 100 percent of the actual events have been correctly founded and 89.7 percent of returned events have been correct ones. The inserted activity classes were from

- Bar Skullcrusher (2), an insertion
- Spider Curl (7), three insertions
- Incline Hammer Curl (9), an insertion
- Dumbbell Flyes (22), an insertion
- Wide-Grip Pulldown Behind The Neck (27), an insertion
- Wide-Grip Front Pulldown (30), two insertions
- Counter Arm Tricep Dumbbell Kickback (6b), an insertion
- Cross-trainer, an insertion.

When exploring the corresponding classes in Table III, it can be seen that all the classes have high variability in precision rate deviations. Nevertheless, there are several other classes having also high deviation in precision rates regardless of which events have been correctly classified. To overcome the problem, several different sets of the transition rules for Algorithm 1 were tested, but the decrease in insertions was effected by means of increasing deletions. Thus, the transition rules were kept unchanged.

V. Validation results

To validate the promising results achieved in the previous section, the data set from the second data collection was taken into consideration. Although, the data set was collected by the same person, it can be considered as more separate compared to the sequential sets from the same exercise occasion.

To be consistent with the previous section, the results are again shown separately as window- and event-wise.

A. Window-wise recognition

For window-wise classification, the k nearest neighbors classifier was utilized with k = 9 based on the results of the previous section. But instead of presenting classification accuracies only for the case with both sensors, which was found out to be the most accurate, the separate sensor rates are also introduced. These results can be found from Table IV. It has to be noted that while in this case the whole validation data is considered as a single test set, no deviations were achieved.

The first obvious observation is that there are significant decrease in the classification accuracies between Tables III and IV. The decrease is almost 20 percent with every sensor combination. A closer study reveals that the drops are clustered into specific exercise classes, and still in 24 classes, the true positive rate is over 90 percent. The specific classes with low accuracy (TPR $<$ 80%) includes ten classes (1, 3, 8, 12, 13, 14, 25, 26, 6b and 26b). In half of them, (8, 13, 14, 25 and 26), the mixing has occurred with the NULL-class meaning that the exercise performance has resembled some miscellaneous activities performed between the exercise sets. For them, it can be concluded that some of the actual movements during those exercises are not unambiguous even to be classified as exercises. Nevertheless, the remaining five classes give opposite examples.

The five exercise classes mixing most with other classes are Close-Grip Barbell Bench Press, Triceps Pushdown, Hammer Curl, and counter arms in Triceps Dumbbell Kickback and One-arm Dumbbell row (1, 3, 12, 6b and 26b). The mixing occurring between counter arm exercises is quite obvious while in both cases the exercises are done by leaning similarly on the same bench. Moreover, the exercises are concentrated on the active arm causing very little movement in torso area or to the counter arm. On the other hand, when considering Triceps Pushdown using rope (3) and Hammer Curl (12), the movements are quite identical although the force is in former case downwards and in the latter case upwards. In these, the paces of the movements have also been similar enough to cause mixing between classes. Moreover, the Hammer Curl which is mixed with Dumbbell Alternate Bicep Curl (8) is a part of the movement, while the difference between those is that in Hammer Curl, the movement is stopped in a half way. The mix from Close-Grip Barbell Bench Press to Bar SkullCrusher is not so obvious, but when looking at the true positive rates from Table IV, it can be assumed that the chest sensor features have effected at most and thus mixed the classes.

Another interesting aspect in the results of Table IV is that by combining the data from both sensors the accuracies have not always improved and in some cases the joint true positive rate is significantly lower than the sensor-wise accuracy. For example, with exercises Seated Cable Rows (25) and One-Arm Dumbbell Row (class 26), the classification accuracy is over 90 percent with the wrist sensor data, but the chest sensor data based accuracies are very low. When combining the data from both sensors, it would be assumed that the accuracy would be at least 90 percent, but the combined accuracy is in another case even zero. This kind of drop with the combined accuracy was not anticipated beforehand.

B. Event-wise recognition

By now, it has been shown that the true precision rates have dropped significantly with some classes in window-wise recognition. Now, the question is, if the same drop effect to the event-wise recognition rates.

TABLE IV. TRUE POSITIVE, FALSE POSITIVE AND PRECISION RATES FOR ALL THE ACTIVITIES USING THE SENSORS SEPARATELY AND JOINTLY BY USING SEPARATE VALIDATION DATA AS TEST SET.

Class	wrist			chest			both sensors		
	TPR	**FPR**	**PR**	**TPR**	**FPR**	**PR**	**TPR**	**FPR**	**PR**
1	5.4	0.4	12.9	48.3	0.1	89.1	51.2	0.1	81.3
2	96.5	0.1	89.9	97.8	2.5	38.3	97.4	0.4	71.8
3	5.3	0.1	44	21.6	0.1	77.4	15.5	0.1	71.1
4	81.8	0.1	82.3	80.4	0.1	93.3	96.6	0.0	96.0
5	96.9	0.0	95.0	48.7	0.8	46.7	97.4	0.0	95.0
6	82.5	0.0	94.4	52.6	0.3	61.3	92.0	0.0	96.2
7	80.2	0.5	68.0	95.8	0.2	89	97.3	0.0	96.6
8	50.4	1.0	38.0	34.1	0.6	48.6	68.1	0.7	56.4
9	32.9	0.2	61.9	89.5	0.7	70.6	98.8	0.2	86.8
10	97.2	0.6	63.7	69.1	0.1	93.4	94.9	0.1	94.4
11	46.5	0.0	97.3	45.4	0.7	52.6	90.9	0.1	94.6
12	22.4	0.2	47.4	29.5	0.5	39.7	58.4	0.5	48.2
13	23.4	0.1	72.7	28.7	0.1	70.5	52.6	0.1	82.6
14	68.5	0.0	92.4	28.4	0.1	71.1	56.7	0.1	87.8
15	85.5	0.4	73.8	81.6	1.3	56.9	96.6	0.4	79.7
16	39.8	0.6	40	58.3	0.6	57.9	97.2	0.0	96.8
17	92.9	0.1	92.4	55.9	0.2	75.0	93.5	0.1	90.3
18	80.7	0.0	95.1	92.1	1.1	46.2	91.0	0.2	81.0
19	23.6	0.1	70.7	42.5	0.1	86.0	84.5	0.1	86.0
20	69.8	0.2	81.0	52.4	0.1	93.8	94.2	0.0	99.6
21	64.3	1.5	33.8	65.9	0.6	65.7	98.4	0.1	94.6
22	94.2	0.5	69.9	50	0.3	74.9	100.0	0.2	84.5
23	41.7	0.1	77.4	77.2	0.2	84.3	76.9	0.1	83.9
24	99.0	0.1	94.6	51.6	0.4	61.8	99.0	0.1	91.8
25	95.7	0.4	75.8	28.0	0.4	56.1	55.7	0.3	70.6
26	93.8	0.2	76.9	0	0.1	0.0	0.0	0.1	0.0
27	83.4	0.5	67.4	43.4	0.9	43.6	86.0	0.3	75.7
28	96.4	0.1	82.5	65.0	0.2	77.7	97.8	0.0	94.4
29	96.3	0.0	93.5	77.6	0.5	61.2	97.0	0.2	76.5
30	96.2	0.3	78.8	61.0	0.5	67.6	93.2	0.4	71.2
6b	0	0.7	0	2.7	0.7	3.6	2.3	0.2	6.7
10b	87.2	0.4	67.5	87.2	0.5	68.5	98.9	0.2	81.9
26b	86.8	0.5	55.9	51.9	1.1	34.4	74.2	0.9	38.4
96	97.0	0.3	79.8	96.8	0.1	94.7	98.7	0.0	99.6
97	98.4	0.1	92.3	88.1	0.1	85.7	80.5	0.1	93.2
98	90.8	0.2	90.3	99.3	0.1	96.5	99.4	0.1	96.0
99	92.7	12.2	93.1	83.3	22.7	77.5	94.2	10.5	94.1
mean	70.2	0.6	71.4	59.0	1.1	65.2	80.5	0.5	79.6

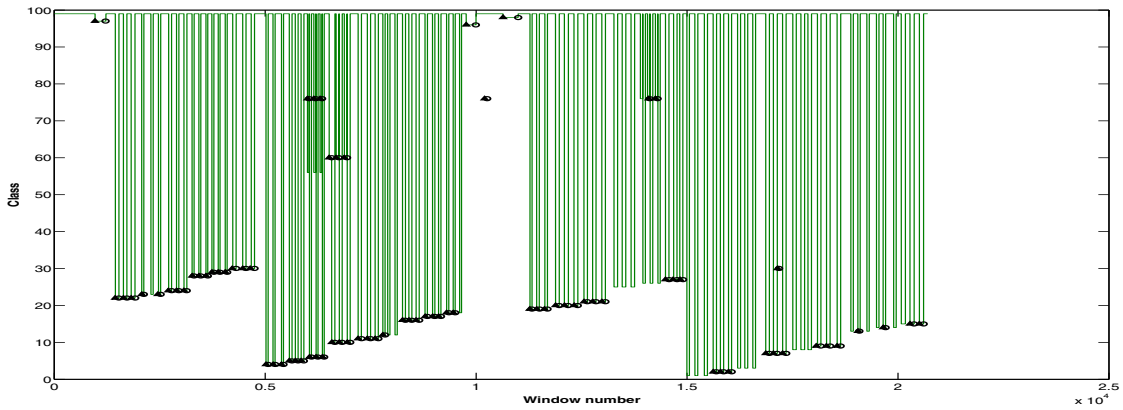

Fig. 4. Validation data event recognition results. A beginning of an event is marked as triangle and the end as sphere. Thus if the green line (actual class)and triangles/spheres intersect each other the event is recognized correctly.

The results from the state-machine are shown as Figures 4 and 5. It can be seen that the ten classes with low accuracy (1, 3, 8, 12, 13, 14, 25, 26, 6b and 26b) are the ones where events are not returned correctly. The state-machine has missed 24 sets from 102 actual events. Only with one case (counter arm exercise 6b) the state-machine has marked the set to belong in the other class. In other cases, it has missed the set completely.

Based on the validation results, it can be concluded that although in the most cases the approach gives a valid estimate of the exercise sets and their performance, there are classes where reliable results could not be achieved. Moreover, even with person-wise classification, the differences between exercise occasions (e.g. weights, pace, posture, exercise order effecting to muscle strength) can be in a significant role complicating the activity recognition.

On the other hand, it has to be noted that this study

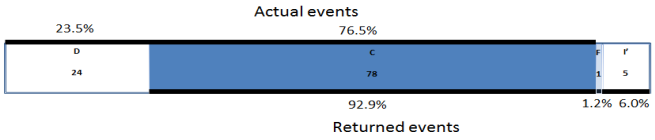

Fig. 5. Validation data event analysis diagram (EAD) of all the exercises and sets of gym data: not found events / deletions (D), correctly found events (C), fragmented events (F) and inserted events (I).

comprised real data from 30 upper body exercises with a few extra activities which is over three times more than in the previous gym recognition studies. With this approach, and suitable initial selection of activities, very high accuracies would have been achieved in both cases. For example, by selecting the most accurately classified 20 upper body exercises with no counter-arm movements, a data set twice the amount used in the previous studies, no obstacles would not even been found.

In a short, more data would be needed to test the system and to optimize the parameters and classifiers with gym approaches. The data collected during single exercise occasion is not versatile enough for training data.

VI. Conclusion

In this paper, a gym exercise recognition case study was shown by utilizing acceleration information from 30 different upper-body exercises. The results were shown separately as window- and event-wise and a separate test set was used to validate the methods. It was noticed, that although the results were surprisingly accurate within exercise sets from the same occasion, in the validation data set, there were classes not consistent with training results. Based on the study, it can be concluded that it is possible in most cases to recognize gym activities reliably, but more versatile data would be needed to improve the system further, because with the most difficult cases, the classification accuracies of nearly 100 percent in training studies do not give enough variability for incoming exercise occasion.

As a future work, a more versatile gym data set, including more exercises and more subjects, will be collected from the gym exercises to be able to develop more accurate methods, especially for the event recognition. Moreover, to release this more versatile data set to other researchers in the area will be a future goal of the author.

VII. Acknowledgments

Author would like to thank Academy of Finland for funding this work (Decision 257468, Postdoctoral Researcher project: Mobile Sensors and Behavior Recognition in Real-world).

References

[1] K. V. Laerhoven and O. Cakmakci, "What shall we teach our pants?" *The Fourth International Symposium on Wearable Computers*, pp. 77–83, 2000.

[2] M. Ermes, J. Pärkkä, J. Mäntyjärvi, and I. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 1, pp. 20–26, 2008.

[3] O. Banos, M. Damas, H. Pomares, A. Prieto, and I. Rojas, "Daily living activity recognition based on statistical feature quality group selection," *Expert Systems with Applications*, vol. 39, no. 9, pp. 8013–8021, 2012.

[4] M. Zhang and A. Sawchuk, "Human daily activity recognition with sparse representation using wearable sensors," *Biomedical and Health Informatics, IEEE Journal of*, vol. 17, no. 3, pp. 553–560, May 2013.

[5] H. Leutheuser, D. Schuldhaus, and B. M. Eskofier, "Hierarchical, multi-sensor based classification of daily life activities: Comparison with state-of-the-art algorithms using a benchmark dataset," *PLoS ONE*, vol. 8, 10 2013.

[6] Polar Loop, http://www.polarloop.com/.

[7] Nike+ FuelBand SE, http://www.nike.com/us/en_us/c/nikeplus-fuelband.

[8] Fitbit, http://www.fitbit.com/.

[9] P. Siirtola and J. Röning, "Ready-to-use activity recognition for smartphones," *IEEE Symposium on Computational Intelligence and Data Mining*, 2013.

[10] Location API, "https://developer.android.com/reference/com/google," /android/gms/location/package-summary.html.

[11] MotionActivity, "https://developer.apple.com/library/ios/," documentation/CoreMotion/Reference/CMMotionActivity, _class/Reference/Reference.html.

[12] P. Siirtola, H. Koskimäki, and J. Roning, "Periodic quick test for classifying long-term activities," in *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on.* IEEE, 2011, pp. 135–140.

[13] P. Siirtola, P. Laurinen, J. Röning, and H. Kinnunen, "Efficient accelerometer-based swimming exercise tracking," in *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on*. IEEE, 2011, pp. 156–161.

[14] K. Chang, M. Chen, and J. Canny, "Tracking free-weight exercises," *UbiComp 2007: Ubiquitous Computing*, pp. 19–37, 2007.

[15] C. Li, M. Fei, H. Hu, and Z. Qi, "Free weight exercises recognition based on dynamic time warping of acceleration data," *Intelligent Computing for Sustainable Energy and Environment Communications in Computer and Information Science*, pp. 178–185, 2013.

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

[17] J. A. Ward, P. Lukowicz, and H. W. Gellersen, "Performance metrics for activity recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 1, pp. 6:1–6:23, 2011.

[18] Bodybuilding, http://www.bodybuilding.com/exercises/.

[19] GeneActiv, http://www.geneactiv.co.uk/.

[20] L. Gao, A. K. Bourke, and J. Nelson, "Sensor positioning for activity recognition using multiple accelerometer-based sensors," *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2013.

[21] Data Analysis and Inference Research Group, http://www.oulu.fi/cse/node/23065.

[22] H. Koskimäki, V. Huikari, P. Siirtola, P. Laurinen, and J. Röning, "Activity recognition using a wrist-worn inertial measurement unit: a case study for industrial assembly lines," *The 17th Mediterranean Conference on Control and Automation*, pp. 401–405, 2009.

[23] D. J. Hand, H. Mannila, and P. Smyth, *Principles of data mining*. Cambridge, MA, USA: MIT Press, 2001.

[24] E. Fix and J. L. Hodges Jr., "Discriminatory analysis - nonparametric discrimination: Consistency properties," *Technical Report 4, U.S. Air Force, School of Aviation Medicine, Randolph Field, TX*, 1951.

[25] T. Mitchell, *Machine Learning*. The McGraw-Hill Companies, Inc., 1997.

[26] H. Koskimäki, V. Huikari, P. Siirtola, and J. Röning, "Behavior modeling in industrial assembly lines using a wrist-worn inertial measurement unit," *Journal of Ambient Intelligence and Humanized Computing*, vol. 4, no. 2, pp. 187–194, 2013.