Astrid Ustad

Validation of algorithms for physical activity type detection in children using raw acceleration data

Master's thesis in Human Movement Science

Department of Neuroscience

NTNU

Trondheim, Norway

Spring 2016

NTNU
Norwegian University of
Science and Technology

Abstract

Background: Accelerometry has become the objective method of choice to assess physical activity in children. However a number of limitations are related to how accelerometer data from children are analyzed. Valid algorithms to classify accelerometer data into physical activity types can enhance our understanding of children's physical behavior and provide useful information of different aspects of physical activity.

Aim: The aim of this study was to examine the validity of three algorithms for physical activity type detection in children using raw accelerometer data. The physical activity types of interest were the everyday activities walking, running, stair walking, cycling, standing, sitting and lying.

Methods: 15 typically developing children (7 boys and 8 girls) in the age range between 6 and 12 years conducted several repetitions of the everyday activities of interest while they wore accelerometers (Axivity AX3) on lower back and mid-thigh and were video recorded. The videos were labeled and used as gold standard for validation of the physical activity types identified by the algorithms. Three algorithms were evaluated: the Acti4 algorithm and the NTNU-adults algorithm, which were developed on data from adults, and the NTNU-children algorithm that was developed based on the children data from this study.

Results: The overall accuracy was 84.5%, 63.6% and 70.6% for the NTNU-children algorithm, the NTNU-adults algorithm and the Acti4 algorithm, respectively. The children algorithm showed consistently higher sensitivity, specificity and positive predictive values than the two adults algorithms. Sensitivity for the children algorithm was >0.89 for all activities of interest except stair walking (0.57-0.73). The NTNU-adults algorithm had very low sensitivity for walking, stair walking and cycling (<0.56) and high sensitivity for standing, sitting, lying and running (0.78-0.88). The Acti4 algorithm showed high sensitivity for walking, stair walking, running, cycling and lying (0.81-1.00), and lower sensitivity for standing (0.73) and sitting (0.66).

Conclusion: The children algorithm showed higher overall accuracy than the adults algorithms and detected the activities walking, running, cycling, standing, sitting and lying with very high precision in children. The results indicate that children-specific algorithms are necessary. This study showed that raw acceleration data from two monitors placed on lower back and mid-thigh can be used to detect and separate the static activities standing, sitting and lying, and detect the dynamic activities walking, running and cycling, with high precision in children.

Sammendrag

Bakgrunn: Akselerometre er mye brukt som en objektiv metode for å undersøke fysisk aktivitet hos barn. Det er imidlertid en rekke svakheter knyttet til hvordan akselerometerdata fra barn har blitt analysert. Valide algoritmer for å klassifisere akselerometerdata i ulike aktivitetstyper kan øke vår forståelse av barns fysiske atferd og gi nyttig informasjon om ulike aspekter ved fysisk aktivitet.

Mål: Målet med denne studien var å undersøke validiteten av tre algoritmer for å detektere aktivitetstype på barn ved å bruke råakselerometerdata. Aktivitetstypene som var av interesse var de hverdagslige aktivitetene gange, løping, trappegange, sykling, å stå, å sitte og å ligge.

Metode: 15 funksjonsfriske barn (7 gutter og 8 jenter) i alderen 6-12 år gjorde flere repetisjoner av de hverdagslige aktivitetene. De hadde et akselerometer (Axivity AX3) plassert på korsryggen og ett midt på låret og de ble hele tida filmet av et kamera. Videoene ble annotert og brukt som gullstandard for validering av aktivitetstypene som ble detektert av algoritmene. Tre algoritmer ble evaluert: Acti4 algoritmen og NTNU-adults algoritmen som var basert på data fra voksne og NTNU-children algoritmen som var utviklet basert på dataene fra barn samlet i denne studien.

Resultat: Overordna nøyaktighet (overall accuracy) var henholdsvis 84,5%, 63,6% og 70,6% for NTNU-children algoritmen, NTNU-adults algoritmen og Acti4 algoritmen. Barnealgoritmen viste konsistent høyere sensitivitet, spesifisitet and positive prediktive verdier enn de to voksenalgoritmene. Sensitivitet for NTNU-children var >0.89 for alle aktivitetene av interesse unntatt trappegange (0,57-0,73). NTNU-adults viste lav sensitivitet for gange, trappegange og sykling (<0,56) og høy sensitivitet for å stå, å sitte, å ligge og for løping (0,78-0,88). Acti4 algoritmen viste høy sensitivitet for gange, trappegange, løping, sykling og å ligge (0,81-1,00), og lavere sensitivitet for å stå (0,73) og å sitte (0,66).

Konklusjon: Barnealgoritmen var mer nøyaktig enn voksenalgoritmene og detekterte aktivitetstypene gange, løping, sykling, å stå, å sitte og å ligge med veldig høy presisjon hos barn. Disse resultatene indikerer at spesifikke algoritmer for barn er nødvendig. Denne studien viste at råakselerasjonsdata fra to akselerometre plassert på korsryggen og midt på låret kan brukes til å detektere og skille mellom de statiske aktivitetene å stå, å sitte og å ligge, samt å detektere de dynamiske aktivitetene gange, løping og sykling med høy presisjon hos barn.

Acknowledgement

First of all, I would like to thank my supervisor Karin Roeleveld for guidance and supervision through the year and in the writing process. I would also like to thank my co-supervisors Ellen Marie Bardal, Alan Bourke and Espen Ihlen for good advices and providing expertise for the analysis work.

I am grateful and want to thank the participants and their parents for their contribution.

I would like to thank the master students Astrid Tessem and Hans-Olav Hessen at the Department of Computer and Information Science for the collaboration and their work with the algorithms. I would also like to thank my fellow students Hilde Bremseth Bårdstu and Atle Melleby Kongsvold for interesting discussions and collaboration through the year.

Finally, I want to thank my fellow students Ingvild Maalen-Johansen and Ane Øvreness for the collaboration and all the social breaks. This year has not been the same without you.

Table of Contents

Introduction

It can be claimed that humans are born to be physically active. The human body is designed for movement with a muscular and skeletal system that enables a huge variety of degrees of freedom, and a physiology that enhances strength and endurance the more it is used. Today, it is well known that loading these systems through physical activity, which by Caspersen et al. are defined as "any bodily movement produced by skeletal muscles that results in energy expenditure"[1], provides fundamental health benefits[2]. Nevertheless, it has become an increasing problem that the population in modern communities is insufficient physical active to maintain good health[3]. Inadequate physical activity constitutes a major health risk and was in 2009 identified as the fourth leading risk factor for global mortality[4]. Although the serious consequences of an inactive life style seldom are seen before adulthood, there is evidence that more active children in general show healthier cardiovascular profiles, are leaner and develop higher peak bone masses than less active children[5]. It is also reasonable to assume that an active lifestyle in childhood has positive effects on health later in life and by the promotion of physical activity the appropriated habits can persist into adulthood. Although current research on the tracking of physical activity from childhood to adulthood is not unanimous[6], there is a wide agreement that promoting active lifestyles in children is important. According to the World Health Organization, children should accumulate at least 60 minutes of moderate- to vigorous- intensity physical activity every day, and additional health benefits are provided with amounts greater than the 60 minutes[2]. Valid and reliable methods for assessment of physical activity are necessary for investigating the extent to which children meet these recommendations, as well as for understanding the determinants of physical activity and for the evaluation of interventions aiming to increase physical activity. Traditionally, self-report methods as questionnaires have been used to assess children's physical activity. Because children tend to engage in brief sporadic bursts of intense activity and change activity frequently, the ability to recall physical activity can be especially challenging for children[7,8]. In this context, objective methods for monitoring physical activity have great potential because they are not subject to recall problems or the reporting bias associated with subjective methods[9-11].

Accelerometry has become the objective method of choice and is widely used to assess physical activity in children[10,12,13]. Accelerometers are small lightweight monitors that easily can be attached to the body and sample accelerations generated by body movement in one or more directions[12]. Continuous improvements in properties as battery and storage capacity

make these devices attractive for the assessment of physical activity in free-living children for prolonged periods. A substantial amount of studies have used accelerometer technology for monitoring physical activity in children. The majority of the studies have focused on energy expenditure and the analyses are based on what often is called activity counts. Activity counts are post-filtered accelerometer data that are summarized over specified time epochs, typically 1-minute, and processed using calibrated cut off points based on regression models for energy expenditure[9,14-16]. These cut off points are then used to estimate the amount of time spent in different intensities of physical activity[17]. There are a number of limitations associated with the use of activity counts for assessment of physical activity. First, there are no standards for conversion of raw accelerometer data to counts and activity counts are therefore derived from commercial "black box" software that is specific for brands and models[15]. Second, there is enormous variation in the use of cut off points and different studies have used different cut off values[10,16,18,19]. Third, the accuracy depends on the type of physical activity performed[20,21]. Therefore, the studies are not directly comparable and regression models taking activity type into account will have the potential to estimate energy expenditure in free-living situations more accurate.

Valid algorithms for detection of activity type can enhance our understanding of children's physical behavior and provide useful information of different aspects of physical activity. In addition to have the potential to improve energy expenditure estimations, methods providing valid information about activity type makes it possible to study specific activities, which in itself can be interesting. Pattern recognition is a method for classifying accelerometer data and algorithms using this approach are often called machine learning algorithms[17]. A small number of studies that have applied pattern recognition algorithms for detection of physical activity type in children are identified. De Vries at al.[22], Ruch et al.[23] and Trost et al.[24] classified activity types in children with an overall accuracy of 77%, 67% and 88%, respectively. These studies have all used the activity counts recorded over 1-second epochs for the activity classification. In general, the activities in these studies were conducted in unrealistic long sequences without the frequent shifts between activity types that characterize children's daily life.

Separating static activities as standing, sitting and lying from each other are shown problematic using the activity counts approach[22-24], and can simply be explained by the lack of generation of counts while doing no movements. An alternative approach is to use the unfiltered raw accelerometer data. The accelerometer signal does not only depend on

movement, but also on orientation due to gravitation[25]. Raw data from two accelerometers, one placed on the upper body and one placed on the lower body, make it possible to separate static activities based on the combined information about position from both accelerometers. Also dynamic activity types can be recognized based on the combined information about patterns in the raw acceleration signals from two monitors. The use of raw acceleration data for detection of activity types in children has the potential of improving accuracy in the study of children's physical activity, because a number of the limitations related to the activity counts approach are thereby avoided.

Skotte and co-workers have developed and validated an algorithm in adults called Acti4 that classify accelerometer data into activity types using standard deviation and angle as classification parameters[26,27]. A collaboration taking place at the Norwegian University of Science and Technology (NTNU) are working on developing a pattern recognition algorithm for activity type detection that will be used in the upcoming HUNT4 Study (The Nord-Trøndelag Health Study[28]). The NTNU-adults algorithm evaluated in this study was part of the preparation for the HUNT4, and the NTNU-children algorithm was based on the same computing methods.

The aim of this study was to examine the validity of the above mentioned algorithms for physical activity type detection in children using raw accelerometer data. The physical activity types of interest were the everyday activities walking, running, stair walking, cycling, standing, sitting and lying.

Methods

Participants

Healthy children in the age range between 6 and 12 years being without physical disabilities or medical conditions that affect normal daily activities, were invited to participate. Fifteen typically developing children from the mid region of Norway were recruited. There were 7 boys and 8 girls with a mean age of 9.5 years. Demographic data, including gender, age, height and weight, was collected from the participants. The children and their parents were informed of the aims of the study and the parents signed a written consent to participate. The study protocol was ethically approved by the Regional Ethical Committee for medical and health research in Middle Norway (REC Central).

Protocol and equipment

Accelerometers

Two activity monitors of the type Axivity AX3 (Axivity, Newcastle, United Kingdom) were used for the collection of accelerometer raw data. The Axivity AX3 is a data logger that record accelerations in 3 directions and that have an intern memory and a clock so that data can be recorded for prolonged periods. The weight of the monitor was 11 g and the dimensions were 23 x 32.5 x 7.6 mm.

The monitors were placed on the lower back (central at the lumbar vertebrae 3) and on the right thigh (on the front midline and midway between the patella and the anterior superior iliac spine), with the USB port pointing downwards and the fabric print towards the skin. The monitor was fixed to the skin with the following procedure: the monitor was wrapped in a finger cot, the toupee tape was attached to the print side of the monitor, a 5x5 cm piece of Fixomull (BSN Medical) was placed on the skin and the monitor was attached above and covered by FlexiFix (Smith & Nephew). When attached to the participants, the monitors were orientated so that the x-axis equaled to the vertical axis, the y-axis to the mediolateral axis, and the z-axis to the anteroposterior axis.

Figure 1. The orientation and direction of the axes of the Axivity AX3 monitor. Retrieved from axivity.com[29].

The AX3 software (Omgui version 1.0.0.28) was used to configure the devices and download the logged data. The monitors were set to sample at a frequency of 200Hz and a sample range of ±8g. This sampling range is sensible for moderate activities such as sprinting and jumping and suitable for most human movement studies[29]. A time interval for recording that covered the protocol was set during the configuration of the monitors. The AX3 monitors log raw data in a binary packed format named Continuous Wave Accelerometer (CWA) format.

Video recording

Video recordings were used as the gold standard for validation of the physical activity types identified by the algorithms. It was conducted using a GoPro HERO 3+ action camera. The camera was placed on a tripod (about 1.8 meters over ground) during the recording in the laboratory, and was hand held during the outdoors part of the protocol. The camera was held stable and not too far from the participant. The camera was set to a resolution of 720p and a frame rate of 30 frames per second, which according to GoPro is the best setting for handheld shots in low-light conditions[30].

The validation protocol

The validation protocol consisted of two parts: a standardized protocol conducted in a laboratory and a part that was conducted outdoors. The participants were instructed to perform daily activities while they were video recorded. The standardized protocol included several repetitions of the activities standing, sitting, lying, walking and stair walking. A detailed description of the protocol is shown in table 1. Some smaller deviations from the protocol did occur, and the number of repetitions stated in table 1 indicates the minimum number that was conducted by every participant. The laboratory was set up with a chair, a chair at a table, a bed and a 7 meters long pathway were the participants did the walking bout

of the protocol. The stair walking was conducted at stairs with 16 steps. The outdoors part of the protocol included several periods of running and a longer bout of cycling in addition to walking. No instructions were given whether the cycling should be performed in a seated or standing position. The duration of each sequence of the activities was not standardized and was in general very short. The typical duration of each repetition of sitting, lying and running was around 5 seconds, with a few repetitions of sitting lasting up to one minute. The walking bouts varied in duration from a few seconds up to one minute. Cycling was conducted in one sequence and lasted longer, typically around 45 seconds. Because of snow and slippery conditions, the outdoors part of the protocol had to be conducted in the basement under the St. Olavs Hospital in Trondheim for 5 of the participants. There was sufficient space there for running and cycling. The total duration of the protocol was between 25 and 35 minutes.

Table 1. The validation protocol in detail.

Instruction	**Repetitions**
stand - sit	1
sit - walk - stand	2
stand - walk - sit	2
sit - walk - sit at a table	1
sit at a table - stand	1
stand - sit at a table	1
sit at a table - walk - sit on the ground	1
sit on the ground - walk - sit at table	1
sit at a table - walk - lie (prone on the ground)	1
lie (prone on the ground) - walk - stand	1
stand - walk (normal) - stand	3
stand - walk (fast) - stand	3
stand - walk (slow) - stand	3
stand - walk - lie (supine in a bed) - turn to right/prone/left - stand	3
lie (in a bed) - sit (in a bed) - walk - stand	1
stand - walk - sit on the ground	1
sit on the ground - walk - sit	1
sit - stand	2
stand - sit	1
sit - stand - walk - stand	1
stand - descend stairs - stand	3
stand - ascend stairs - stand	2
stand - walk - stand	3
stand - run - stand	2
stand - run (fast) - stand	2
stand - walk (normal) - stand	2
stand - walk (fast) - stand	2
stand - walk (slow) - stand	2
stand - run - stand	2
stand - run (fast) - stand	2
stand - cycling - stand	1

Figure 2. The laboratory setup.

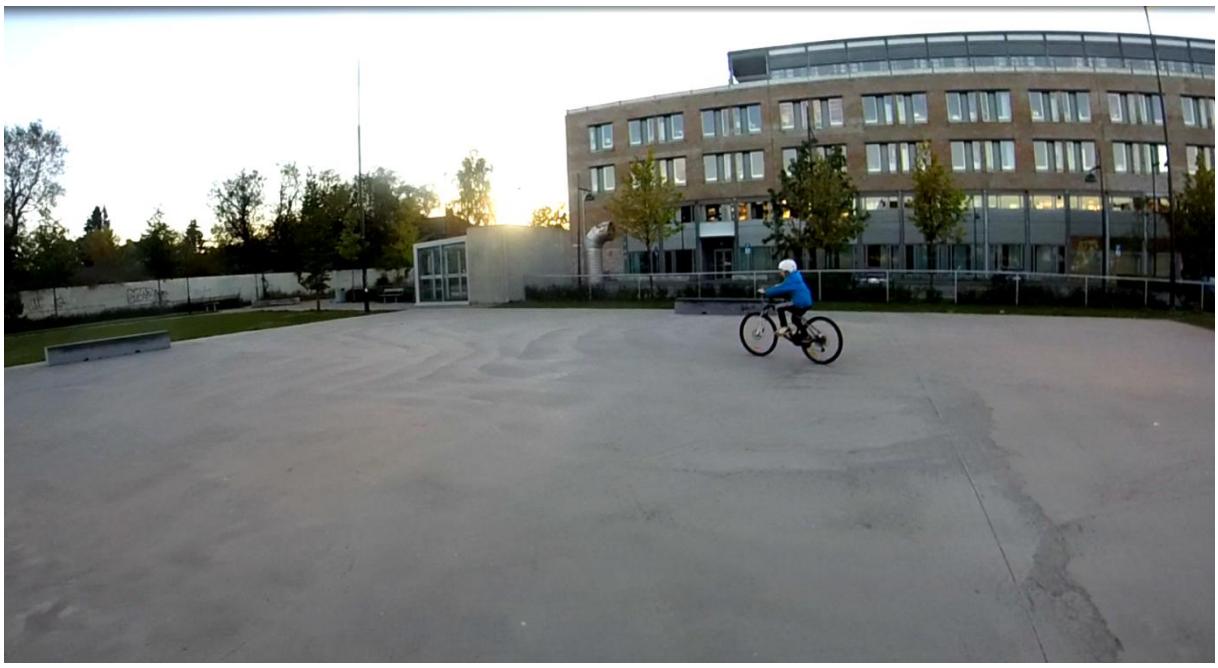

Figure 3. Picture from the outdoors part of the protocol with hand-held camera.

To be able to synchronize the accelerometer and video files afterwards, it was necessary to have a "hand-shake" that could be easily identified in the files. Heel drops were used for this purpose. The instruction for a heel drop was to stand still for at least 5 seconds, then rise up on toes and drop the heel quickly into the ground, and stand still for at least 5 seconds again. The time of the heel drops was noted, so that these hallmarks in the signal could be

recognized for the synchronization of the files. Three repetitions of the heel drop were conducted at the beginning and at the end of the protocol.

Definitions of activities

The video recordings were classified according to predetermined activity definitions. Because there exists no consensus of which or how activities should be defined in validation studies, definitions had to be developed for this study. In collaboration with two other master students (that did a similar validation study in adults and adolescents) and supervisors, definitions were worked out. Important principles were that it should be possible to classify all types of physical behavior according to these definitions, they were detailed with exact descriptions of onset and offset of activities, and they should be independent of technology. The defined activities were sitting, standing, walking, shuffling, stair walking (ascending/descending), lying, cycling (seated/standing), running, bending, picking, other vigorous activities and unclassified. The definitions of the activities, including description of onset/offset, are shown in Appendix 1. The participants were not instructed to perform the activities bending, picking or other vigorous activities as part of the protocol, but such physical behavior could occur anyway and it was important to have the possibility to label all physical activity precisely. However, bending, picking, other vigorous activities, as well as the periods labeled as unclassified, were not considered in this study.

Video analysis

The videos were classified using the ANVIL software (version 5.1.13) shown in figure 4. ANVIL is a video annotation tool where the coding schemes can be defined by the user[31]. The MP4 format was converted to a format with a frame size of 640 x 360 and a frame rate of 25Hz. Each frame of the videos was classified according to the definitions of activities. One rater labeled all the videos. The labeled video files were exported as text files and were the gold standard for the evaluation of the physical activity types detected by the algorithms.

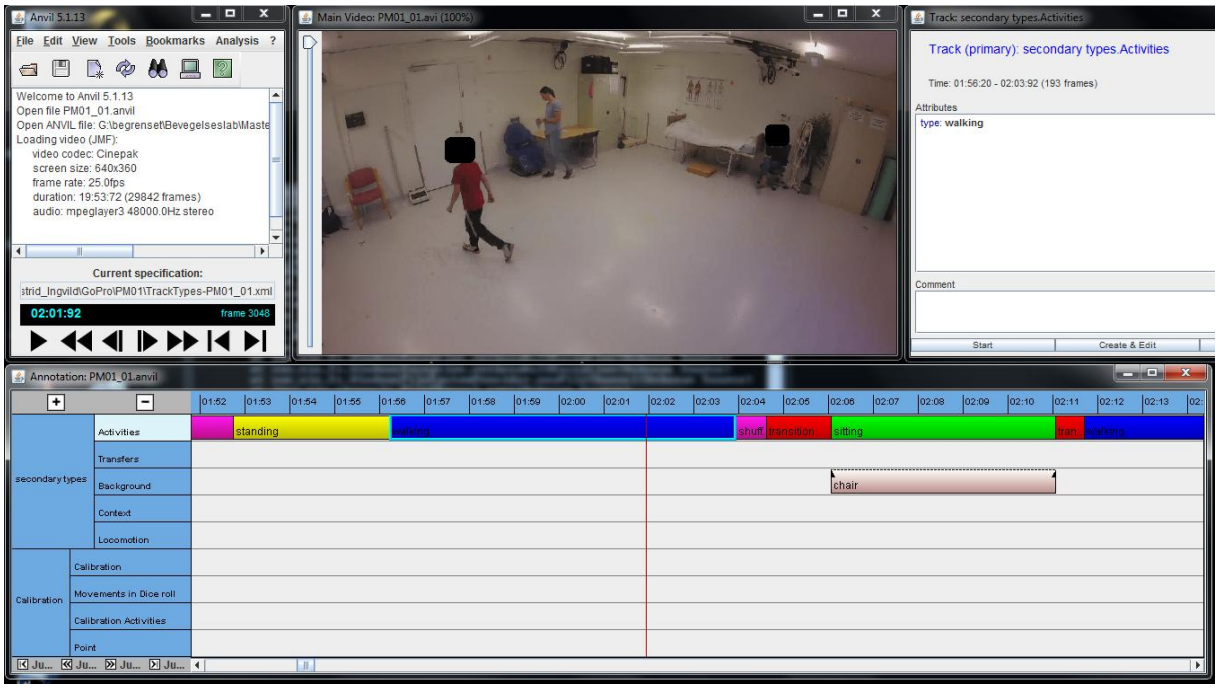

Figure 4. The ANVIL software.

Pre-processing

The accelerometer CWA files and the labeled video files were imported to Matlab for pre-processing. The accelerometer files were first resampled to 100Hz and then synchronized because of some deviation from the given sampling rate. The heel drops were used for the synchronization and identification of start and end point of the protocol. The two accelerometer files and the labeled video file for each participant resulted in labeled accelerometer data that was exported as mat files. The following figures show resampled and synchronized accelerometer data from the back monitor (figure 5) and mid-thigh monitor (figure 6), and how the heel drops looked for one of the participants (figure 7).

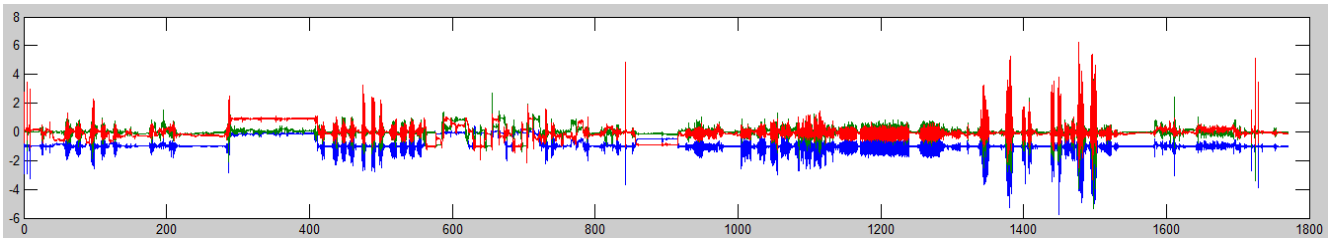

Figure 5. Accelerometer signal from the monitor on the back, with the time in seconds on x-axis and the sample range in g on y-axis. Blue curve: x-axis. Green curve: y-axis. Red curve: z-axis.

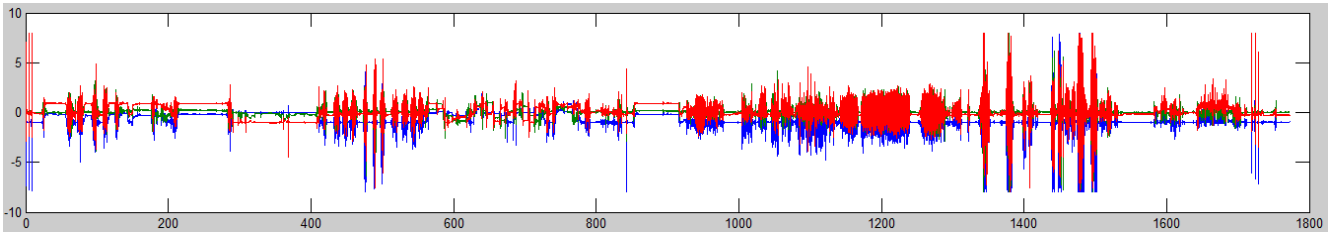

Figure 6. Accelerometer signal from the monitor on the thigh, with the time in seconds on x-axis and the sample range in g on y-axis. Blue curve: x-axis. Green curve: y-axis. Red curve: z-axis.

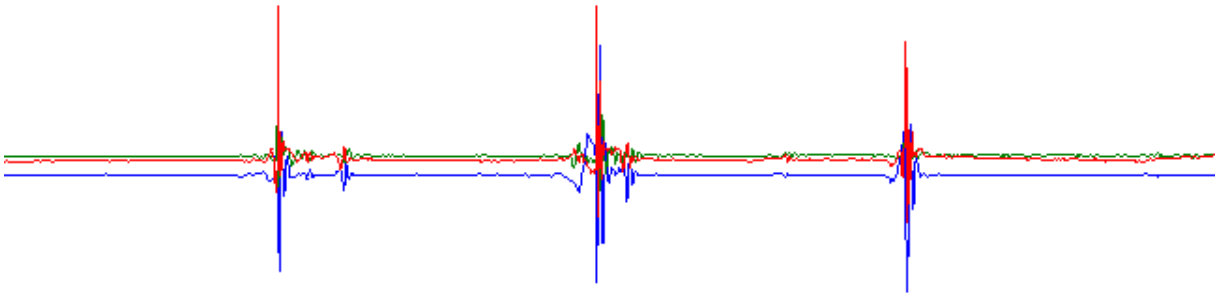

Figure 7. An extract from the thigh monitor signal showing the heel drops used for synchronization. Blue curve: x-axis. Green curve: y-axis. Red curve: z-axis.

Data analysis

Acti 4 algorithm

The Acti4 is a Matlab-based algorithm developed by Skotte and co-workers at the National Research Centre for the Working Environment in Copenhagen, Denmark. Details of the algorithm and how it was developed is provided in Skotte et. al[26]. The Acti4 classifies 30Hz sampled data in the activities lying, sitting, standing, moving, walking, running, stair walking and cycling. Moving was a left over category that normally corresponds to a standing posture neither detected as standing nor walking, a definition that matches with the definition of shuffling.

The AX3 raw data files (CWA) were converted to a Comma Separated Value (csv) format, resampled to 30Hz and then converted to the internal file format specially designed for use by Acti4 (called act4). A setup file with the time interval of the protocol and a reference period of standing for every participant were the input for Acti4. The output from Acti4 was in 1Hz, and was resampled up to 25Hz for the statistical analyses. The Acti4 classified the accelerometer data into activity types using standard deviation and angle as classification

parameters. For example, standing was detected if inclination of thigh was less than 45º and no thigh movement was detected (standard deviation in any direction below 0.1G). The complete list of definitions are shown in Appendix 2. The default settings of the algorithm were used. The window size was 2 seconds for walking, running, moving an standing, 5 seconds for sitting, lying and stair walking and 15 seconds for cycling. The thresholds for the activities standing/moving were 0.1 G, walking/running were 0.72 G, sitting/standing were 45º and stair walking/cycling were 40º.

Two different analyses with the acti4 algorithm were conducted, one with the entire raw data file, and one that was modified. In the modified Acti4 analysis, only activities that the algorithm actually classifies were included, meaning that periods of transitions, bending, picking, vigorous activities and unclassified activities were excluded. Cycling in seated and standing position were merged to cycling, and ascending/descending stairs were merged to stairs for both Acti4 analyses.

NTNU algorithms

The NTNU algorithms were developed as part of a master project by students at the Department of Computer and Information Science at NTNU. The labeled accelerometer data was used to generate the algorithms. The NTNU-adults algorithm was based on data from another master project that conducted a validation protocol on adults, while the NTNU-children algorithm was based on the data from this study. A window size of 1-second with 50% overlap between adjacent windows was used for all activities, and features were generated from each data window. A machine learning algorithm called Random Forest was used when training the classifier. A ten-fold cross validation was used to train the algorithms, meaning the data-set was split into ten equally sized sets and nine of the sets were used for training while the last one was used for testing. This process was repeated ten times so that each set was used for testing. This resulted in ten different classification models that were averaged for the final algorithm.

As the NTNU algorithms were developed based on the data from the labeling process, they classified exactly the activities that were defined in this study. It means that they, unlike the Acti4, separate the activities ascending/descending stairs and seated/standing cycling, and classify transitions, bending, picking, vigorous activities and unclassified.

Statistical analyses

Overall accuracy, sensitivity, specificity and positive predictive values

Agreement between the gold standard and the output from the algorithms were analyzed and confusion matrices were generated. The overall accuracy was calculated for each algorithm as the weighted average. Sensitivity, specificity and positive predictive values were calculated for each activity. The sensitivity is the proportion of instances correctly detected as the particular activity out of all instances of the particular activity. The specificity is the proportion of instances correctly detected as not the particular activity out of all the instances that are not the particular activity. The positive predictive value (PV+) is the proportion of detected instances of the particular activity that truly belong to the particular activity, and is therefore the probability that a detection of a particular activity is correct.

Inter-rater reliability of the video analysis

Inter-rater reliability is a term that refers to the extent of agreement among observers[32]. The video recording for one participant was labeled by two other trained raters, in addition to the rater that annotated all the videos. Inter-rater reliability was calculated with the Cohen´s kappa statistic[33], which is frequently used to test inter-rater reliability and developed to account for the agreement that occurs by chance[32].

Results

Participants

Descriptive characteristics of the 15 participants are presented in table 2. The sample contained approximately equal numbers of boys and girls and the two gender groups were almost identical according to the characteristics age, weight and height.

The labeled accelerometer data from one participant was excluded from the Acti4 analysis because the program of unknown reasons did not accept a reference period (see Methods;Acti4 algorithm) for this participant. This participant was a 7.3 year old girl (weight=22.6 kg, height=125.5 cm).

Table 2. Mean (SD) age, weight and height of the participants.

	Total (N=15)		Boys (N=7)		Girls (N=8)	
Age (years)	9.5	(1.5)	9.7	(1.7)	9.3	(1.3)
Weight (kg)	33.5	(8.4)	32.7	(7.4)	34.3	(9.6)
Height (cm)	138.1	(10.3)	139.8	(11.7)	136.6	(9.4)

Inter-rater reliability of the video analysis

The inter-rater reliability was calculated based on three different raters labeling of the video recording for one of the participants and the results are shown in table 3. The Cohen´s kappa was nearly identical between the three raters, with a mean close up to 0.95.

Table 3. Inter-rater reliability between the three raters calculated with the Cohen´s kappa statistic.

	Cohen´s kappa
Rater 1 vs. Rater 2	0.9482
Rater 2 vs. Rater 3	0.9466
Rater 3 vs. Rater 1	0.9439

Time spent in performing the activities

There was an average of 27 minutes of labeled accelerometer data per participant. That was equal to a total of 375 minutes of labeled accelerometer data for the 14 participants in the Acti4 analysis, and a total of 404 minutes for the 15 participants in the NTNU analyses.

Figure 8 shows the amount of time spent in the different activities of all the labeled accelerometer data. More time was spent in descending stairs (2.5%) than ascending stairs (1.2%), and more time was spent in cycling in a seated position (2.4%) versus cycling in a standing position (0.2%). Bending, picking, vigorous activity and unclassified activity were minor activities with respectively 0.1%, 0.1%, 0.2% and 1.3% of all the labeled accelerometer data and were not considered in this study.

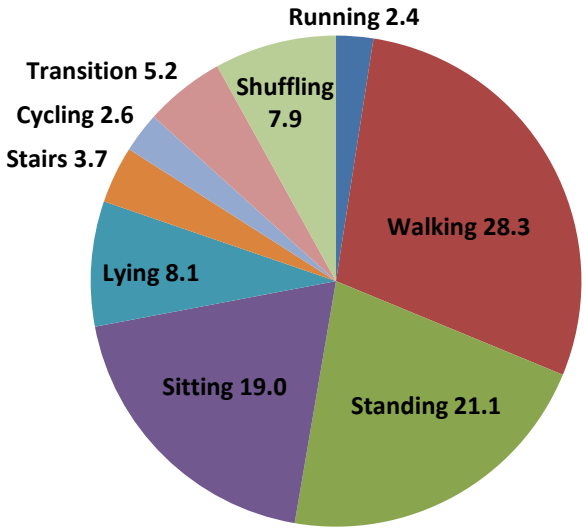

Figure 8. The amount of time spent in performing the different activities showed as a percentage out of all the labeled accelerometer data.

Confusion matrices

The confusion matrices shown in tables 4-7 show the results from the analysis with each algorithm. The number of instances in the confusion matrices varies between the algorithms because the methods differ. The output from the Acti4 algorithm was in 25 instances per second, while the NTNU algorithms was in 2 instances per second.

Table 4. Confusion matrix with instances of the activities for Acti4 with all activities. The rows represent the instances of labeled video data (gold standard) and the columns represent the instances of Acti4 detected data.

	Walking	Running	Shuffling	Stairs	Standing	Sitting	Lying	Transition	Bending	Picking	Cycling	VA	Unclassified	Total *
Walking	136584	6207	8265	1667	3336	3012	219	0	0	0	0	0	0	159290
Running	210	12243	872	0	232	110	0	0	0	0	0	0	0	13667
Shuffling	14653	1768	17592	2723	8828	695	206	0	0	0	0	0	0	46465
Stairs	546	0	2865	17046	579	0	0	0	0	0	0	0	0	21036
Standing	14633	613	14189	1673	86815	987	108	0	0	0	0	0	0	119018
Sitting	1663	42	1603	205	1943	68635	28852	0	0	0	1087	0	0	104030
Lying	22	0	24	0	0	139	45551	0	0	0	0	0	0	45736
Transition	3349	140	3762	48	1266	6452	14015	0	0	0	261	0	0	29293
Bending	16	0	319	37	229	115	21	0	0	0	0	0	0	737
Picking	2	0	185	104	128	35	0	0	0	0	0	0	0	454
Cycling	4	0	264	625	43	632	0	0	0	0	13136	0	0	14704
VA	504	116	476	47	0	0	5	0	0	0	0	0	0	1148
Unclassified	1337	69	1894	150	1913	1208	872	0	0	0	66	0	0	7509
Total**	173523	21198	52310	24325	105312	82020	89849	0	0	0	14550	0	0	563087

VA. Vigorous activity
* Total number of labeled instances
** Total number of detected instances

The Acti4 algorithm detected 397602 instances correctly (sum of the highlighted values in the main diagonal of the confusion matrix) of the total number of 563087 instances. The column to the right in table 4 includes the total number of instances labeled as each activity, and the bottom row includes the total number of instances detected as each activity by the algorithm. Walking was misclassified as running in 4% of cases (100*instances of walking detected as running/total number of walking instances, in this case 100*6207/159290=4%) and as shuffling in 5% of cases. Out of the instances that were detected as walking, 8% was actually shuffling (100*instances of shuffling detected as walking/total number of instances detected as walking, in this case 100*14653/173523=8%) and 8% standing. Running had some confusion with shuffling (6%), and out of the detected running, 29% was actually walking and 8% actually shuffling. Also stair walking was mainly confused with shuffling (14%). Out of the instances that were detected as stair walking, 11%, 7% and 7% was actually shuffling, walking and standing, respectively. Standing was misclassified with walking and shuffling in 12% of cases for both activities. 8% was actually shuffling and 3% actually walking of the

detected standing instances. As much as 28% of the sitting was wrongly detected as lying, and 8% of the detected sitting was actually transitions and 4% walking. Out of the detected lying, as much as 32% was actually sitting and 16% transitions. Cycling had some confusion with stair walking and sitting (both 4%), and 7% of the detected cycling was actually sitting.

Table 5. Confusion matrix with instances of the activities for modified Acti4. The rows represent the instances of labeled video data (gold standard) and the columns represent the instances of Acti4 detected data.

	Walking	Running	Shuffling	Stairs	Standing	Sitting	Lying	Cycling	Total*
Walking	136584	6207	8265	1667	3336	3012	219	0	159290
Running	210	12243	872	0	232	110	0	0	13667
Shuffling	14653	1768	17592	2723	8828	695	206	0	46465
Stairs	546	0	2865	17046	579	0	0	0	21036
Standing	14633	613	14189	1673	86815	987	108	0	119018
Sitting	1663	42	1603	205	1943	68635	28852	1087	104030
Lying	22	0	24	0	0	139	45551	0	45736
Cycling	4	0	264	625	43	632	0	13136	14704
Total**	168315	20873	45674	23939	101776	74210	74936	14223	523946

* Total number of labeled instances
** Total number of detected instances

The total number of instances tested in the modified Acti4 analysis were 523946, and 397602 instances were correctly detected. The activities that were excluded in this analysis corresponded to 7 % of all the labeled data, and consisted mainly of the transitions. The same confusions apply to this analysis compared to the Acti4 analysis with all activities.

Table 6. Confusion matrix with instances of the activities for NTNU-adults algorithm. The rows represent the instances of labeled video data (gold standard) and the columns represent the instances of NTNU-adults detected data.

	Walking	Running	Shuffling	Stairs ↑	Stairs ↓	Standing	Sitting	Lying	Transition	Bending	Picking	Cycling ↓	Cycling ↑	VA	Unclassified	Total*
Walking	7325	358	453	1566	2503	285	15	0	85	1	0	0	0	1146	1	13738
Running	23	1023	4	5	22	14	4	0	4	0	0	0	0	62	0	1161
Shuffling	1193	120	1156	138	179	769	53	0	75	0	0	0	0	143	1	3827
Stairs ↑	68	12	11	185	162	9	2	0	31	0	0	0	0	107	0	587
Stairs ↓	147	60	27	61	695	54	1	0	9	0	0	0	0	180	0	1234
Standing	627	20	1296	46	42	8003	119	0	15	0	0	0	1	36	14	10219
Sitting	15	1	59	8	2	344	7364	333	830	10	68	0	0	10	159	9203
Lying	0	0	0	0	0	0	7	3409	530	0	0	0	0	0	0	3946
Transition	194	4	108	60	45	41	227	189	1569	3	0	0	0	80	1	2521
Bending	9	0	4	2	0	16	7	0	18	0	0	0	0	4	0	60
Picking	1	0	1	3	0	2	25	0	1	0	0	2	0	0	0	35
Cycling ↓	7	0	14	8	2	14	555	0	517	2	0	64	0	0	0	1183
Cycling ↑	0	0	1	17	0	0	1	0	44	0	0	0	0	23	0	86
VA	7	28	2	11	9	3	0	0	6	0	0	0	0	26	0	92
Unclassified	80	6	92	8	2	147	64	0	163	1	0	1	1	10	47	622
Total**	9696	1632	3228	2115	3663	9701	8444	3931	3897	17	68	67	2	1827	223	48514

↑ ascending (stairs) / standing (cycling)
↓ descending (stairs) /seated (cycling)
VA. Vigorous activity
* Total number of labeled instances
** Total number of detected instances

The NTNU-adults algorithm detected 30866 instances correctly of the total number of 48514 instances. There was confusion with the vigorous activity class for all the dynamic activities in the NTNU-adults analysis. Walking showed substantially confusion with stair walking for both ascending (11%) and descending stairs (18%). Out of the detected walking, 12% was actually labeled as shuffling and 6% as standing. Out of the detected running, 22% was actually walking. Stair walking was seriously confused with walking (12%) for both ascending and descending stairs, and 28% of the ascending stairs was detected as descending stairs. Out of the detected stair walking, 74% and 68% was actually walking for ascending and descending stairs, respectively. Some standing was detected as walking (6%) and shuffling (13%). Out of the detected standing, 3% was actually walking, 8% was actually shuffling and 4% was actually sitting. 9% of the sitting was detected as transitions, 4% was detected as standing and 4% as lying. Out of the detected sitting, 3% was actually transitions

and 7% was actually cycling. Lying was wrongly detected as transitions in 13% of cases, and out of the detected lying, 8% was actually sitting and 5% was actually transitions. Cycling was detected totally wrong in the NTNU-adults analysis.

Table 7. Confusion matrix with instances of the activities for NTNU-children algorithm. The rows represent the instances of labeled video data (gold standard) and the columns represent the instances of NTNU-children detected data.

	Walking	Running	Shuffling	Stairs ↑	Stairs ↓	Standing	Sitting	Lying	Transition	Bending	Picking	Cycling ↓	Cycling ↑	VA	Unclassified	Total*
Walking	12788	42	327	27	34	426	13	0	71	0	0	9	0	0	1	13738
Running	77	1032	5	15	15	10	4	0	2	0	0	1	0	0	0	1161
Shuffling	1117	55	1240	14	69	1229	13	0	58	1	0	25	0	0	6	3827
Stairs ↑	185	14	7	336	20	13	0	0	10	0	0	2	0	0	0	587
Stairs ↓	267	2	38	1	902	21	0	0	3	0	0	0	0	0	0	1234
Standing	319	2	498	6	10	9343	7	0	9	0	0	11	5	1	8	10219
Sitting	11	0	35	6	3	67	8872	0	159	0	0	35	7	0	8	9203
Lying	0	0	0	0	0	0	8	3720	218	0	0	0	0	0	0	3946
Transition	284	4	119	13	9	44	270	303	1428	0	0	36	0	0	11	2521
Bending	11	0	5	0	0	10	5	0	18	4	7	0	0	0	0	60
Picking	1	0	4	0	0	1	4	0	2	4	19	0	0	0	0	35
Cycling ↓	4	0	11	3	0	3	39	0	13	0	0	1110	0	0	0	1183
Cycling ↑	4	0	0	4	0	2	6	0	20	0	0	0	50	0	0	86
VA	44	12	2	1	7	4	0	0	8	0	0	0	0	14	0	92
Unclassified	61	2	70	2	3	139	87	0	92	0	0	29	0	0	137	622
Total**	15173	1165	2361	428	1072	11312	9328	4023	2111	9	26	1258	62	15	171	48514

↑ ascending (stairs) / standing (cycling)
↓ descending (stairs) /seated (cycling)
VA. Vigorous activity
* Total number of labeled instances
** Total number of detected instances

The NTNU-children algorithm detected 40995 instances correctly of the total number of 48514 instances. Walking had some confusion with shuffling (2%) and standing (3%), and 7% of the detected walking was actually shuffling. 7% of the running instances were detected as walking. Out of the detected running, 4% and 5% was actually walking and shuffling, respectively. Stair walking showed substantially confusion with walking. As much as 32% of the ascending stairs was detected as walking and the same number for descending stairs was 22%. 3% and 5% of the standing were detected as walking and shuffling, respectively. Out of

the detected standing, 4% was actually walking and 11% was actually shuffling. Sitting was only slightly confused with the transitions (2%), and the same applied to lying (6%). Cycling (seated) showed some confusion with sitting (3%), and out of the detected cycling (seated), 3% was actually sitting and 3% was actually transitions.

Overall accuracy, sensitivity, specificity and positive predictive values

Overall accuracy for the algorithms and sensitivity, specificity and positive predictive values (PV+) for each activity are shown in table 8. Of the three algorithms, the NTNU-children algorithm showed the highest overall accuracy compared to the gold standard (84.5%). The NTNU-adults algorithm had lowest overall accuracy (63.6%) and the Acti4 algorithm showed an overall accuracy of 70.6%. The overall accuracy increased to 75.9% in the modified Acti4 analysis (shown in table 9).

Specificity was very high (>0.91) in all algorithms and in the children algorithm 8 out of 11 activities showed almost perfect specificity (>0.99). There was more variation in the sensitivity and PV+. The NTNU-adults algorithm had very low sensitivity for the activities walking, stair walking and cycling (<0.56), moderate sensitivity for the activities standing, sitting and lying (0.78-0.86), and high sensitivity for running (0.88). Although the three algorithms had almost equally high sensitivity for running, the NTNU-children algorithm showed significantly higher PV+ and was therefore more precise than the two adults algorithms. The Acti4 algorithm showed high sensitivity for the activities walking, stair walking, running, cycling and lying (0.81-1.00), and lower for the activities standing (0.73) and sitting (0.66). Acti4 detected lying with highest sensitivity (1.00), however the PV+ was very low (0.51). The NTNU-children algorithm was more precise than the two adults algorithms with consistently higher sensitivity, specificity and positive predictive values for all activities. The exception was stair walking where the children algorithm showed low sensitivity (0.57-0.73), and the Acti4 a sensitivity of 0.81. Compared to the Acti4 analysis with all activities, the modified Acti4 analysis showed higher PV+ for all activities, and substantially higher for the activities sitting and lying as a result of the exclusion of transitions.

Table 8. Overall accuracy and sensitivity, specificity, and positive predictive value for each physical behavior.

	Acti4			NTNU-adults			NTNU-children		
Activity	Sensitivity	Specificity	PV+	Sensitivity	Specificity	PV+	Sensitivity	Specificity	PV+
Walking	0.86	0.91	0.79	0.53	0.93	0.76	0.93	0.93	0.84
Running	0.90	0.98	0.58	0.88	0.99	0.63	0.89	1.00	0.89
Standing	0.73	0.96	0.82	0.78	0.96	0.82	0.91	0.95	0.83
Sitting	0.66	0.97	0.84	0.80	0.97	0.87	0.96	0.99	0.95
Lying	1.00	0.91	0.51	0.86	0.99	0.87	0.94	0.99	0.92
Stairs ↑	0.81	0.99	0.70	0.32	0.96	0.09	0.57	1.00	0.79
Stairs ↓	NA	NA	NA	0.56	0.94	0.19	0.73	1.00	0.84
Cycling ↓	0.89	1.00	0.90	0.05	1.00	0.96	0.94	1.00	0.88
Cycling ↑	NA	NA	NA	0.00	1.00	0.00	0.58	1.00	0.81
Shuffling	0.38	0.93	0.34	0.30	0.95	0.36	0.32	0.97	0.53
Transition	NA	NA	NA	0.62	0.95	0.40	0.57	0.99	0.68
OA	70.6%			63.6%			84.5%		

PV+. Positive predictive value
NA. Not applicable
↑ ascending (stairs) / standing (cycling)
↓ descending (stairs) /seated (cycling)
OA. Overall accuracy

Table 9. Overall accuracy and sensitivity, specificity and positive predictive values for the modified Acti4 algorithm.

	Acti4 - modified		
Activity	Sensitivity	Specificity	PV+
Walking	0.86	0.91	0.81
Running	0.90	0.98	0.59
Standing	0.73	0.96	0.85
Sitting	0.66	0.99	0.92
Lying	1.00	0.94	0.61
Stairs	0.81	0.99	0.71
Cycling	0.89	1.00	0.92
Shuffling	0.38	0.94	0.39
OA	75.9%		

OA. Overall accuracy
PV+. Positive predictive value

Discussion

The aim of this study was to examine the validity of algorithms for physical activity type detection in children using raw accelerometer data. For this purpose, 15 children conducted several repetitions of the everyday activities walking, running, stair walking, cycling, standing, sitting and lying while they wore accelerometers on lower back and mid-thigh and were video recorded. The main findings were that the children algorithm was much more accurate detecting activity types than the two adults algorithms, and detected the activities walking, running, cycling, standing, sitting and lying with very high precision. The validity of the Acti4 algorithm and the two NTNU algorithms for detection of these activity types in children will be discussed in the following. Also aspects related to shuffling, transitions between activities and window size will be pointed out and debated. Finally, inter-rater reliability, strengths, limitations and practical implications will be discussed.

Overall accuracy

The NTNU-children algorithm showed the highest overall accuracy compared to the gold standard for detection of the activities. With an overall accuracy of 84.5%, the NTNU-children algorithm had substantially higher accuracy than both the NTNU-adults algorithm (63.6%) and the Acti4 algorithm (70.6%). The two NTNU algorithms were based on the same computing methods. The fact that the results differed so much between the adults and children algorithm indicates that children´s behavior causing magnitude and patterns in the accelerometer signal that are different from adults. This finding supports that development and validation of children-specific algorithms are necessary. An overall accuracy of 84% is just as good and better than earlier studies in children[22-24], and an accuracy level of <80% has been reported as acceptable for activity classification[34]. The overall accuracy was the weighted average for all the activities included in the analysis. Except for stair walking, the activity types of interest achieved agreement with the gold standard that were higher than 84% with the NTNU-children algorithm. The results from the NTNU-children algorithm showed that it is possible to detect the everyday activities walking, running, cycling, standing, sitting and lying with high precision in children (sensitivity >0.89).

Walking, running and stair walking

Walking was detected with high accuracy by the Acti4 algorithm (sensitivity=0.86), with very high accuracy by the NTNU-children algorithm (sensitivity=0.93) and with low accuracy by

the NTNU-adults algorithm (sensitivity=0.53). All three algorithms showed high sensitivity for running (0.88-0.90) but the children algorithm showed significantly higher PV+ (0.89) than the two adults algorithms (0.58-0.63). In contrast to the good results for walking and running, the sensitivity for stair walking was low for the children algorithm (0.57-0.73). The PV+ was considerable higher (0.79-0.84), meaning that when stair walking is detected it is likely that it is correct. Descending stairs was detected with higher precision than ascending stairs. More time had been spent in descending stairs than ascending stairs (because the children descended more stairs as part of the protocol) which may explain this difference, because more data were then provided to training of the algorithm. It remains uncertain if ascending stairs is an activity that is more difficult to detect than descending stairs or if the difference can be explained by unequal time spent in performing these activities. In comparison, the Acti4 algorithm, merged the stair walking to one class and detected stair walking with higher sensitivity (0.81), but lower PV+ (0.70). What is most appropriate of separating or merging stair walking depend on the study question. If estimation of energy expenditure are of interest, a separation would be most appropriate because of quite different energy cost. Stair walking was mainly misclassified as walking by the children algorithm. Ascending stairs require more energy than level walking and a more precise detection of stair walking would therefore be desirable for evaluation of these physical activities in a health perspective.

Standing, sitting and lying

The children algorithm detected the static activities with very high accuracy (sensitivity >0.91), and sitting was in fact the activity showing highest precision of all activities with the children algorithm. Also the NTNU-adults algorithm detected the static activities quite well (sensitivity 0.78-0.86). It is understandable that features for the static activities are more similar between children and adults than they are for the dynamic activities. The results for the Acti4 algorithm were mixed. Looking solely on the sensitivity, lying seemed to be perfectly detected (1.00) and sitting poorly detected (0.66). However, as much as 28% of the sitting was wrongly detected as lying, meaning that the Acti4 algorithm severely misclassify these activities. The Acti4 algorithm classify sitting and lying based on different angles. The results showed that the angle that separate sitting and lying is not useable in children because lying was detected when they were sitting in a slightly reclined position. Standing was detected with moderately high sensitivity (0.73) by the Acti4 algorithm. Thus, the pattern recognition algorithms were more precise detecting static activities than the Acti4 algorithm,

but adjusting the angle as classification parameter in the Acti4 will probably increase the accuracy for these activities.

This study found that it is possible to detect and separate the static activities standing, sitting and lying with high precision in children. Separating static activities from each other have been shown problematic using the activity counts approach[22-24]. This finding is therefore important because it shows that using raw acceleration data from two monitors can provide this type of information.

Cycling

Cycling is for many children a common activity for transportation and thereby important to identify. It has, however, been reported to be a problematic activity type to measure using accelerometers due to underestimation of energy expenditure[9,16,19,23]. This is related to the use of activity counts because cycling is a nonweight-bearing activity that generates a small amount of activity counts relative to the energy cost. The children algorithm did indeed detect cycling (in a seated position) with high precision (sensitivity=0.94), and so did the Acti4 algorithm (sensitivity=0.89). These results show that using the raw acceleration data sampled from two monitors placed on lower back and mid-thigh enables valid detection of the physical activity type cycling. Detecting cycling does not solve the problem of energy expenditure estimation directly, but precise detection of the activity is an important first step.

The NTNU-adults algorithm did not detect the cycling correctly at all. This algorithm was trained with accelerometer data where the cycling had been conducted on a stationary bike. Although similar movement, the external forces will have large impact on the accelerometer signal when cycling outdoors and are probably of high importance for the detection of this activity. This finding demonstrates that algorithms have to separate cycling outdoors and stationary cycling. The ability of algorithms to detect cycling conducted on stationary bikes is probably not very important for children, but more important for adults.

Shuffling

The results showed shuffling to be a problematic activity to detect with low sensitivity for all algorithms. Although the activity type shuffling is not mentioned as one of the activities of interest in this study, it is an important activity to include in the discussion because of the relation with standing and walking. A high degree of misclassification between shuffling, walking and standing was seen for all three algorithms. This is understandable considering the

similarity between these activities. Typically, shuffling occurred as small feet movements while standing or in the beginning or end of a walking period. Thus, some parts of the labeled accelerometer data are very similar to walking and other parts are very similar to standing. The practical implication of this is that no features are distinctive for shuffling, which make accurate classification of this activity very difficult if not impossible. Finding alternative ways to classify this physical activity are therefore necessary. An option is to do a post-processing of the shuffling periods and redefine them into the activity classes walking and standing. To replace the physical activity type shuffling with expanded categories of standing and walking are probably more meaningful for most study purposes. A proposal for how this can be done, inspired by the Acti4 algorithm, is to apply a certain standard deviation as classification parameter for which periods of shuffling that belong to walking or standing. Another possibility is to post-process in relation to adjacent activities. For example, if a short period of shuffling is detected with standing before and afterwards, this should all be detected as standing. The most appropriate method for this needs to be investigated and may depend on the study goal.

Transitions between activities

In the modified Acti4 analysis, only the accelerometer data that were labeled as the activities the algorithm actually classify, were included. This analysis was primarily done with the objective to investigate how much the transitions influenced the accuracy of the algorithm. In other similar studies, transitions between activities are typically either excluded for the analysis and/or the activities are performed in long bouts resulting in few transitions. The studies by De Vries et al.[22], Trost et al.[24] and Stemland et al.[27] excluded the transition periods between activities and conducted the activities in long bouts, while Ruch et al.[23] labeled the transitions with the more strenuous activity before or afterward. The overall accuracy for the Acti4 analysis increased with more than 5% in the modified version, meaning that whether the transitions are included or excluded in the analysis have significant impact on the results. In free-living situations, transitions cannot be excluded, and validation studies that exclude or minimize the impact of transitions are therefore unrealistic. Algorithms for detection of activity type will be used outside laboratory and validation protocols should therefore represent children's physical activity habits with frequent activity changes[8]. It is a strength of this study that the protocol included a great number of shifts between activities and that the transitions are included in the analyses.

Algorithms for activity type detection can handle transitions in different ways. The Acti4 algorithm "assumes" that the person performs the previous activity until the definition for another activity is achieved, and in that way sort of hides the transitions in adjacent activities. In contrast, the NTNU algorithms detect the transitions as a separate activity, however with low sensitivity (0.57-0.62) and PV+ (0.40-0.68). Low results for transitions are not surprising because this activity class includes a variety of different movements. A classification into different types of transitions is probably necessary for enhancing the accuracy of detection of transitions. Most confusion of transitions was seen with the activities lying and sitting, which have the natural explanation that there is always a transition before and after these activities. With more than 5% of the data material, the transitions are not irrelevant for this study. However, the importance of detecting transitions when studying physical activity in free living can be questioned. We know that when there has been a shift between two postures, for example from standing to sitting, a transition has necessarily happened. For many study scopes, the Acti4 approach that sort of hides the transitions in the adjacent activities may be sufficiently precise and even most suitable.

Window size

The NTNU algorithms used 1-second data windows for all activities, while the Acti4 used windows of 2 seconds for walking, running, shuffling and standing, 5 seconds for sitting, lying and stair walking and 15 seconds for cycling. This is of importance for how sensitive the algorithm is to frequent changes in behavior, and may be more crucial in children than in adults because we expect more sporadic physical behavior[8]. Algorithms with smaller window size have of course a potential to detect rapid changes between activities more precise, but how crucial that is depends on the study question. If the objective is to investigate physical activity in daily living, activities with duration less than a couple of seconds may not be critical to "loose". However, in a validation study like this, window size can possibly have great impact on the results because activities with very short duration may not be detected. The protocol was designed with frequent changes between activities and the duration of each activity bout was in general short. Examples of periods with duration less than 2 seconds were frequently seen in the upright activities standing, walking and shuffling. There was more misclassification between these activities in the Acti4 results than in the NTNU-children results, and some of this difference may be attributed to the larger window size of the Acti4 algorithm.

Inter-rater reliability of the video analysis

The inter-rater reliability between three raters was calculated for evaluation of the video analysis. Because one rater annotated all the videos, the inter-rater reliability was perhaps not of very high importance in this study. However, it was found important because it is a measure of the quality of the gold standard and besides indicates how precise the definitions for activity classification developed for this study were in this study group. Precise definitions are important for avoiding subjective labeling. The Cohen´s kappa statistic showed an agreement of near 0.95 between the three raters. According to Landis and Koch´s interpretation of kappa values from 1977, a kappa agreement of 0.95 represents an "almost perfect agreement", which is in the range 0.81-0.99[35]. This result indicates that it is high quality of the video analysis in this study and that the activity definitions are precise for activity classification in this study group. Nevertheless, the agreement was not perfect and indicates some uncertainty of the video analysis. It was mainly within the activities shuffling, walking and standing that the raters disagreed (data not shown). This finding gives further support to the proposal that later studies should define these activities in another way or conduct some post-processing of shuffling. A possible limitation to the test of agreement was that the raters were the ones that developed the activity definitions. Independent raters would perhaps have given a more reliable test of the robustness of the definitions. Another possible limitation related to the video annotation was that it cannot be excluded that there exists some intra-rater variation. However, the high inter-rater reliability of the video analysis indicates high quality of the gold standard and is a strength of this study.

How representative are the validated activities for children´s daily living?

It is an important principle in validation studies that the participants included represent the target group[36], and likewise it is important that the validated physical activity types represent children's daily life. This study included everyday activities that probably represent much of children's physical activity, however some shortcomings do clearly exist. Young children do probably numerous activities in free-living that are not represented in the validated activities, for example rolling around, skipping, climbing and jumping. The protocol did only include running as a high-intensive activity. Many children participate in different sports activities and if they spend much time performing these activities, the validated activities are not representative for their daily life. Upcoming studies in children should therefore include more children-specific and sports-like activities.

Strengths and limitations

A limitation of this study is the small sample size. Validation studies need a certain amount of data to catch up the variation in the target group. More participants would probably also have led to better feature extraction for the development of the children algorithm and thereby a more robust algorithm. The activities that were conducted the least, for example ascending stairs, cycling and running, may have suffered from small amount of data for training of the algorithm. An alternative to include more participants that would have given a larger data set, had been a longer and more comprehensive protocol. However, the children also attended several tests for other studies the same day and this would have been a too large burden of the children. Therefore, data collection for validation studies should be conducted separately.

Although the sample size was small, the participant group had large variation in age. A heterogeneous participant group is a strength of this study because it ensures that there is variation, which is important in validation studies. The validation protocol did also ensure variation within the activities. Walking and running were performed in different self-selected paces, which contributes to variation in the data.

It is important to note that the children algorithm was both trained and evaluated based on the same participant group. The cross validation method used in this process ensured that training and testing data were separated and this shall give a valid estimate of the algorithm´s accuracy if applied to a population which it was not trained[34]. Nevertheless, direct comparisons between the results from the children algorithm with the adults algorithms do not appear to be fair. The adults algorithms were developed based on data from other participants and were therefore external validation for this study. While the NTNU-adults used the same monitors and computing methods as the children algorithm, the Acti4 was developed using another monitor type and technology. Considering this, an overall accuracy of 70.6% with the Acti4 algorithm is quite high. Somewhat lower accuracy for the adults algorithms compared to the children algorithm must be expected.

Practical implications

This study showed that algorithms can detect everyday physical activity types with high accuracy in children using raw acceleration data from two monitors placed on lower back and on mid-thigh. The small size of the Axivity AX3 monitor makes it suitable to be worn by children in many days without affecting daily living. A feasible method and valid algorithms

for activity type detection can give valuable information about how much time children spend in different types of physical activity and lead to more accurate estimations of energy expenditure in free-living situations.

Conclusion

Of the three algorithms for physical activity type detection evaluated in this study, the NTNU-children algorithm showed the highest overall accuracy and detected the activities walking, running, cycling, standing, sitting and lying with very high precision. The results indicate that children-specific algorithms are necessary. This study showed that raw acceleration data from two monitors placed on lower back and mid-thigh can be used to detect and separate the static activities standing, sitting and lying, and the everyday dynamic activities walking, running, stair walking and cycling with high precision in children. Upcoming studies in children should include more children-specific and sports-like activities, a larger participant group and a group for external validation of the algorithm.

References

1. Caspersen CJ, Powell KE, Christenson GM. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public health reports.* 1985;100(2):126.
2. WHO. Global recommendations on physical activity for health. *World Health Organization.* 2010.
3. Hallal PC, Andersen LB, Bull FC, et al. Global physical activity levels: surveillance progress, pitfalls, and prospects. *The lancet.* 2012;380(9838):247-257.
4. WHO. Global health risks: mortality and burden of disease attributable to selected major risks. *World Health Organization.* 2009.
5. Boreham C, Riddoch C. The physical activity, fitness and health of children. *Journal of sports sciences.* 2001;19(12):915-929.
6. Telama R. Tracking of physical activity from childhood to adulthood: a review. *Obesity facts.* 2009;2(3):187-195.
7. Trost SG. Measurement of physical activity in children and adolescents. *American Journal of Lifestyle Medicine.* 2007;1(4):299-314.
8. Bailey RC, Olson J, Pepper SL, Porszasz J, Barstow TJ, Cooper D. The level and tempo of children's physical activities: an observational study. *Medicine and science in sports and exercise.* 1995;27(7):1033-1041.
9. Trost SG, O'Neil M. Clinical use of objective measures of physical activity. *British journal of sports medicine.* 2014;48(3):178-181.
10. Reilly JJ, Penpraze V, Hislop J, Davies G, Grant S, Paton JY. Objective measurement of physical activity and sedentary behaviour: review with new data. *Archives of disease in childhood.* 2008;93(7):614-619.
11. Shephard RJ. Limits to the measurement of habitual physical activity by questionnaires. *British Journal of Sports Medicine.* 2003;37(3):197-206.
12. Rowlands AV. Accelerometer assessment of physical activity in children: an update. *Pediatric Exercise Science.* 2007;19(3):252-266.
13. Trost SG. Objective measurement of physical activity in youth: current issues, future directions. *Exercise and sport sciences reviews.* 2001;29(1):32-36.
14. McClain JJ, Tudor-Locke C. Objective monitoring of physical activity in children: considerations for instrument selection. *Journal of Science and Medicine in Sport.* 2009;12(5):526-533.
15. Butte NF, Ekelund U, Westerterp KR. Assessing physical activity using wearable monitors: measures of physical activity. *Medicine and science in sports and exercise.* 2012;44(1):5-12.
16. Chen KY, Bassett DR. The technology of accelerometry-based activity monitors: current and future. *Medicine and science in sports and exercise.* 2005;37(11):490.
17. Bassett Jr DR, Rowlands AV, Trost SG. Calibration and validation of wearable monitors. *Medicine and science in sports and exercise.* 2012;44(1):32.
18. Brazendale K, Beets MW, Bornstein DB, et al. Equating accelerometer estimates among youth: The Rosetta Stone 2. *Journal of Science and Medicine in Sport.* 2016;19(3):242-249.
19. Freedson P, Pober D, Janz KF. Calibration of accelerometer output for children. *Medicine and science in sports and exercise.* 2005;37(11):523.
20. Bassett DR, Ainsworth BE, Swartz AM, Strath SJ, O Brien WL, King GA. Validity of four motion sensors in measuring moderate intensity physical activity. *Medicine and science in sports and exercise.* 2000;32(9):471-480.
21. Crouter SE, Churilla JR, Bassett Jr DR. Estimating energy expenditure using accelerometers. *European Journal of Applied Physiology.* 2006;98(6):601-612.
22. De Vries SI, Engels M, Garre FG. Identification of Children's Activity Type with Accelerometer-Based Neural Networks. *Medicine and Science in Sports and Exercise.* 2011;43(10):1994-1999.

23. Ruch N, Rumo M, Mader U. Recognition of activities in children by two uniaxial accelerometers in free-living conditions. *European Journal of Applied Physiology.* 2011;111(8):1917-1927.
24. Trost SG, Wong W-K, Pfeiffer KA, Zheng Y. Artificial neural networks to predict activity type and energy expenditure in youth. *Medicine and science in sports and exercise.* 2012;44(9):1801.
25. Godfrey A, Conway R, Meagher D, OLaighin G. Direct measurement of human movement by accelerometry. *Medical engineering & physics.* 2008;30(10):1364-1386.
26. Skotte J, Korshøj M, Kristiansen J, Hanisch C, Holtermann A. Detection of physical activity types using triaxial accelerometers. *The Journal of Physical Activity and Health.* 2014;11(1):76-84.
27. Stemland I, Ingebrigtsen J, Christiansen CS, et al. Validity of the Acti4 method for detection of physical activity types in free-living settings: comparison with video analysis. *Ergonomics.* 2015;58(6):953-965.
28. NTNU. The HUNT Study - a longitudinal population health study in Norway. *http://www.ntnu.edu/hunt*. Accessed May, 2016.
29. Axivity. AX3 User Manual. *http://axivity.com/userguides/ax3/*. Accessed April, 2016.
30. GoPro. GoPro Product Manuals. *http://gopro.com/support/product-manuals-support*. Accessed April, 2016.
31. Kipp M. Anvil - A generic annotation tool for multimodal dialogue. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech), pp. 1367-1370.* 2001.
32. McHugh ML. Interrater reliability: the kappa statistic. *Biochemia medica.* 2012;22(3):276-282.
33. Cohen J. A coefficient of agreement for nominal scales. *Educational and psychological measurement.* 1960;20:37-46.
34. Staudenmayer J, Pober D, Crouter S, Bassett D, Freedson P. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *Journal of Applied Physiology.* 2009;107(4):1300-1307.
35. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977:159-174.
36. Lindemann U, Zijlstra W, Aminian K, et al. Recommendations for standardizing validation procedures assessing physical activity of older persons by monitoring body postures and movements. *Sensors.* 2013;14(1):1267-1277.

Appendix 1

DEFINITION OF ACTIVITIES	
Activity	**Description**
Sitting	When the person's buttocks is on the seat of the chair, bed or floor. Sitting can include some movement in the upper body and legs; this should not be tagged as a separate transition. Adjustment of sitting position is allowed.
Standing	Upright, feet supporting the person's body weight, with no feet movement, otherwise this could be shuffling/walking. Movement of upper body and arms is allowed until forward tilt and arm movement occurs below knee height. Then this should be inferred as bending. **For chest mounted camera**: If feet position is equal before and after upper body movement, standing can be inferred. Without being able to see the feet, if upper body and surroundings indicate no feet movement, standing can be inferred.
Walking	Locomotion towards a destination with one stride or more, (one step with both feet, where one foot is placed at the other side of the other). Walking could occur in all directions. Walking along a curved line is allowed.
Shuffling	Stepping in place by non-cyclical and non-directional movement of the feet. Includes turning on the spot with feet movement not as part of walking bout. **For chest mounted camera:** Without being able to see the feet, if movement of the upper body and surroundings indicate non-directional feet movement, shuffling can be inferred.
Stair ascending/descending	**Start:** Heel-off of the foot that will land on the first step of the stairs. **End:** When the heel-strike of the last foot is placed on flat ground. If both feet rests at the same step with no feet movement, standing should be inferred.
Lying down	The person lies down. Adjustment after lying down is allowed if it does not lead to a change between the prone, supine, right and left lying positons. Movement of arms and head is allowed. Movement of the feet is allowed as long as it does not lead to change in posture. **Prone:** On the stomach. **Supine:** On the back. **Right side:** On right shoulder. **Left side:** On left shoulder.
Sit cycling	Pedaling while the buttocks is placed at the seat. Cycling starts on first pedaling and finishes when pedaling ends. **For outdoor bicycling:** Cycling starts at first pedaling, or when both feet have left the ground. Cycling ends when the first foot is in contact with the ground. **Not pedaling:** Sitting without pedaling should be tagged separate as sitting.
Stand cycling	Pedaling while standing. Cycling starts on first pedaling and finishes when pedaling ends. Standing without pedaling should be tagged separate as standing.

Running	Locomotion towards a destination, with at least two steps where both feet leave the ground during each stride. **For chest mounted camera:** Running can be inferred when trunk moves forward is in a constant upward-downward motion with at least two steps. Running along a curved line is allowed.
Bending	While standing/sitting, bending towards an object placed below knee-height is bending.
Picking	This refers to picking/placing/touching an object from below knee height. Picking occurs when the trunk is at its lowest point and the person has touched/placed/picked an object. When the person starts to rise it's trunk, picking finishes, and bending begins.
Other vigorous activities	All non-cyclic rapid leg movements that do not classify as running. This includes sport like activities such as rapid change in direction and jumping. Can occur in all directions.
Unclassified	All non-cyclic movements that do not classify according to the definitions. Can occur in all directions. Can be crawling, rowing etc.
Undefined	Until all the sensors are attached, or final adjustment made to position the video can be tagged as undefined. All postures/movements that cannot be clearly identified should be tagged as undefined.
DEFINITON OF TRANSITIONS	
Transitions	**Description**
Bending to picking from standing/walking/sitting	As soon as forward/sideways trunk tilt occurs, bending has started. Bending finishes when the person has reached the lowest point of the movement and picking occurs. When the person starts to rise up, picking finishes and bending begins. When the trunk is in an upright and stable position, bending finishes. This should be tagged as "bending-picking-bending". Steps can occur during bending.
Walking to posture	Walking ends when both feet are at rest, or at first evident forward tilt of upper body. Steps can occur during the transition from walking to posture.
Upright to sitting	Can be from walking or standing, as soon as forward trunk tilt occurs, or a lowering of the trunk, the transition has started. Steps can occur during the transition for positioning. Transition ends when buttocks are in contact with the seat of the chair, bed or floor.
Sitting to upright	Transition starts when the person's buttocks leave the chair and ends when the trunk has reached its upright position. Steps and turning can occur during the transition from sitting to upright.
Standing/walking/sitting to lying	When the trunk flexion begins, or a lowering of the center of mass, the transition has started. Transition finishes when the person is lying flat with the trunk in a stable position.
Lying to standing/walking/sitting	While lying, the transition begins with an upward movement of the trunk or leg movement that leads to a stable upright position or continuous walking. The trunk angle should be in a steady posture for the transition to finish. Steps can occur during the transition.

Standing to walking	As soon as heel-off occurs, walking has started.
Standing to shuffling	As soon as one foot moves, shuffling has started.
Shuffling/walking to standing	As soon as the feet stop moving, walking/shuffling has finished and standing has started.
Shuffling to walking	As soon as walking direction is set and heel-off occurs, shuffling has ended and walking starts.
Walking to shuffling	When walking is interrupted by stepping in place, non-cyclical, non-directional movement of the feet or turning on the spot, this should be tagged as shuffling.
Sit cycling to stand cycling / stand cycling to sit cycling	When the buttocks leave the seat, stand cycling can be inferred. When the buttocks is placed at the seat, sit cycling can be inferred.

Appendix 2

Definition of Acti4 output parameters

lie
Length of periods (h) lying.
Lying is detected if the thigh inclination is above 45° and also the hip inclination is above 65° and trunk inclination is above 45° (default values). Also lying is detected if thigh inclination is above 45° and trunk is more than 45° backwards or sideways regardless of any hip inclination value (recordings by hip Actigraph may be missing). Lying also requires that no movement of the thigh in the direction of thigh's longitudinal axis is detected.

sit
Length of periods (h) sitting.
Sitting is detected if inclination of thigh is above 45° and lying is not detected (in previous versions (2013) it was also required that no movement of the thigh in the direction of thigh's longitudinal axis was detected).

stand (still)
Length of periods (h) standing still.
Standing still is detected if inclination of thigh is less than 45° and no movement of the thigh is detected (standard deviation in any direction of the thigh are below 0.1G).

move
Length of periods (h) moving (standing, neither still or walking).
This is a left over activity used if none of the activities lie, sit, stand, walk, run, stairs, cycle or row is detected. It will normally correspond to a standing posture that is neither detected as standing still nor walking.

walk
Length of periods (h) walking.
Walking is detected if the standard deviation in the thigh's longitudinal axis is between .1G and 0.72G (defaults values) and the mean forward/backward angle is less than the (individual) 'stair threshold' angle.

run
Length of periods (h) running.
Running is detected if standard deviation in the thigh's longitudinal axis is above 0.72G (default) (or below 0.72G and step frequency is above 2.5 Hz) and the mean forward/backward angle is less than the (individual) 'stair threshold' angle.

stairs
Length of periods (h) walking/running stairs.
Walking stairs is detected if the standard deviation in the thigh's longitudinal axis is between .1G and 0.72G and the mean forward/backward angle is between the (individual) 'stair threshold' angle and 40°.

cycle
Length of periods (h) cycling.
Cycling is detected if the standard deviation in the thigh's longitudinal axis is above .1G and the mean forward/backward angle is above 40° and the inclination is below 90°.

row

Length of periods (h) rowing.

Rowing is detected if the standard deviation in the thigh's longitudinal axis is above .1G and the mean forward/backward angle is above 40° and the inclination is above 90°.