A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors

ANDREAS BULLING, Max Planck Institute for Informatics, Germany
ULF BLANKE, Swiss Federal Institute of Technology (ETH) Zurich, Switzerland
BERNT SCHIELE, Max Planck Institute for Informatics, Germany

The last 20 years have seen ever-increasing research activity in the field of human activity recognition. With activity recognition having considerably matured, so has the number of challenges in designing, implementing, and evaluating activity recognition systems. This tutorial aims to provide a comprehensive hands-on introduction for newcomers to the field of human activity recognition. It specifically focuses on activity recognition using on-body inertial sensors. We first discuss the key research challenges that human activity recognition shares with general pattern recognition and identify those challenges that are specific to human activity recognition. We then describe the concept of an Activity Recognition Chain (ARC) as a general-purpose framework for designing and evaluating activity recognition systems. We detail each component of the framework, provide references to related research, and introduce the best practice methods developed by the activity recognition research community. We conclude with the educational example problem of recognizing different hand gestures from inertial sensors attached to the upper and lower arm. We illustrate how each component of this framework can be implemented for this specific activity recognition problem and demonstrate how different implementations compare and how they impact overall recognition performance.

Categories and Subject Descriptors: C.3 [**Special-Purpose and Application-Based Systems**]: Real-Time and Embedded Systems; C.3 [**Special-Purpose and Application-Based Systems**]: Signal Processing Systems; I.5.2 [**Pattern Recognition**]: Design Methodology; I.5.4 [**Pattern Recognition**]: Applications; I.5.5 [**Pattern Recognition**]: Implementation

General Terms: Algorithms, Design, Experimentation, Measurement, Standardisation

Additional Key Words and Phrases: Activity recognition, gesture recognition, on-body inertial sensors, Activity Recognition Chain (ARC)

ACM Reference Format:
Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. ACM Comput. Surv. 46, 3, Article 33 (January 2014), 33 pages.
DOI: http://dx.doi.org/10.1145/2499621

1. INTRODUCTION

Automatic recognition of physical activities—commonly referred to as Human Activity Recognition (HAR)—has emerged as a key research area in Human-Computer Interaction (HCI) and mobile and ubiquitous computing. One goal of activity recognition is to provide information on a user's behavior that allows computing systems to proactively assist users with their tasks [Abowd et al. 1998]. Traditionally, research in computer

Authors' addresses: A. Bulling and B. Schiele, Max Planck Institute for Informatics, Campus E1 4, 66123 Saarbrücken, Germany; email: andreas.bulling@acm.org, schiele@mpi-inf.mpg.de; U. Blanke, Swiss Federal Institute of Technology (ETH) Zurich, Gloriastrasse 35, 8092 Zurich, Switzerland; email: blankeu@ethz.ch.

DOI: http://dx.doi.org/10.1145/2499621

vision has been at the forefront of this work. A large number of researchers investigated machine recognition of gestures and activities from still images and video in constrained environments or stationary settings (see Mitra and Acharya [2007], Turaga et al. [2008], and Aggarwal and Ryoo [2011] for reviews). Efforts to recognize activities in unconstrained daily life settings caused a shift toward using inertial sensors worn on the body, such as accelerometers or gyroscopes. Advances in sensor technology now allow for form factors and battery lifetimes suitable for long-term recordings, computing, and continuous interaction on the move. On-body sensing extends the potential application areas of activity recognition beyond instrumented rooms and promises to provide smart assistance and interfaces virtually anywhere and at any time by observing activities from the user's perspective.

At the end of the 1990s, researchers performed the first feasibility studies on activity recognition using body-worn sensors, where the choice of activities seemed arbitrary and not always relevant to real-world applications. Still, the continuing success of activity recognition motivated steps toward more challenging and application-oriented scenarios. Several real-world domains were identified that would clearly benefit from activity recognition, such as the industrial sector [Maurtua et al. 2007; Stiefmeier et al. 2008], office scenarios, the sports and entertainment sector [Kunze et al. 2006; Minnen et al. 2006a; Ladha et al. 2013], and health care. Specifically, the Activities of Daily Living (ADLs) [Katz et al. 1970] attracted a great deal of interest (for examples see Bao and Intille [2004], Ravi et al. [2005], Logan et al. [2007], and Tapia et al. [2004]). Monitoring daily activity to support medical diagnosis, for rehabilitation, or to assist patients with chronic impairments was shown to provide key enhancements to traditional medical methods [Starner et al. 1997; Sung et al. 2005; Chen et al. 2005; Oliver and Flores-Mangas 2007; Bächlin et al. 2009; Tessendorf et al. 2011a; Plötz et al. 2012]. Early assistance to encourage humans to adopt a healthy lifestyle was regarded as another important goal. This led to a vast exploration of related human activities, for example, brushing teeth [Lester et al. 2006] or hand washing, food [Amft et al. 2007; Pirkl et al. 2008] and medication intake [Wan 1999; de Oliveira et al. 2010], or transportation routines [Krumm and Horvitz 2006].

Recently, activity recognition made its debut as a key component in several consumer products. For example, game consoles such as the Nintendo Wii and the Microsoft Kinect rely on the recognition of gestures or even full-body movements to fundamentally change the game experience. While originally developed for the entertainment sector, these systems have found additional applications, such as for personal fitness training and rehabilitation, and also stimulated new activity recognition research [Sung et al. 2011]. Finally, some sports products such as the Philips DirectLife or the Nike+ running shoes integrate motion sensors and offer both amateur and professional athletes feedback on their performance.

All of these examples underline the significance of human activity recognition in both academia and industry. Despite considerable advances in inferring activities from on-body inertial sensors and in prototyping and deploying activity recognition systems [Hartmann et al. 2007; Ashbrook and Starner 2010], developing HAR systems that meet application and user requirements remains a challenging task. This is the case even if HAR techniques that were successfully used for one recognition problem are to be adopted for a new problem domain.

Although activity recognition shares many methodological challenges with other fields, such as computer vision, natural language processing, or speech recognition, it also faces a number of unique challenges and requires a dedicated set of computational methods that extend on those developed in these fields. For example, computer vision and speech recognition can lend themselves to clear problem definitions, such as "detect object in image" or "detect a spoken word in a sentence," and focus on a

Table I. Main Characteristics of Human Activity Recognition Systems

Type	Characteristic	Description
Execution	Offline	The system records the sensor data first. The recognition is performed afterwards. Typically used for non-interactive applications such as health monitoring.
	Online	The system acquires sensor data and processes it in real time. Typically used for activity-based computing and interactive applications in human-computer interaction.
Generalisation	User independent	The system is optimised for working with a large number of users.
	User specific	The system is tailored to a specific user. Performance is usually higher than in the user-independent case, but does not generalise as well to other users.
	Temporal	The system should be robust to temporal variations caused by external conditions (sensor displacement, drifting sensor response such as barometers or gyroscopes)
Recognition	Continuous	The system automatically "spots" the occurrence of activities or gestures in the streaming sensor data.
	Isolated (Segmented)	The system assumes that the sensor data stream is segmented at the start and end of a gesture by an oracle. It only classifies the sensor data in each segment into one of the activity classes. The oracle can be an external system (e.g. cross-modality segmentation) or the experimenter when assessing classification performance in the design phase.
Activities	Periodic	Activities or gestures exhibiting periodicity, such as walking, running, rowing, biking, etc. Sliding window segmentation and frequency-domain features are generally used for classification.
	Sporadic	The activity or gesture occurs sporadically, interspersed with other activities or gestures. Segmentation plays a key role to isolate the subset of data containing the gesture.
	Static	The system deals with the detection of static postures or static pointing gestures.
System model	Stateless	The recognition system does not model the state of the world. Activities are recognised by spotting specific sensor signals. This is currently the dominant approach when dealing with the recognition of activity primitives (e.g. reach, grasp).
	Stateful	The system uses a model of the environment, such as the user's context or an environment map with location of objects. This enhances activity recognition performance, at the expense of more design-time knowledge and a more complex recognition system.

well-defined and fixed sensing system (i.e., a defined number and type of cameras or microphones). In contrast, HAR offers more degrees of freedom in terms of system design and implementation (see Table I for a description of the main characteristics of human activity recognition systems). First, there is no common definition, language, or structure of human activities that would allow us to formulate a clear and common problem statement (which activity has to be recognized, how a specific activity is characterized, etc.). For some applications, such as long-term behavioral monitoring, relevant activities can often not even be clearly defined up front. Second, human activity is highly diverse and its recognition therefore requires careful selection of several heterogeneous sensors that differ in their capabilities and characteristics. Sensor composition can also change as sensors may be added and removed opportunistically based on current application requirements [Roggen et al. 2009]. Finally, activity recognition

typically requires specific evaluation metrics to reflect the quality of the system for the intended application.

1.1. Article Scope and Contributions

To date, there is no single comprehensive tutorial on human activity recognition using on-body inertial sensors. There are several widely cited papers on the topic, such as Randell and Muller [2000], van Laerhoven et al. [2002], Bao and Intille [2004], and Lester et al. [2006], but these works do not present the design, implementation, and evaluation of HAR systems from a unified perspective. Given that they focus on specific activity recognition problems and typically present a single best solution to the problem under investigation, these works also can't provide the breadth of information expected from an educational tutorial. Only a few of them discuss and compare alternative design options, which we believe is crucial to educate and inform newcomers to the field of human activity recognition.

This article aims to fill this gap by providing the first tutorial on human activity recognition using on-body inertial sensors. It provides a comprehensive introduction to the standard procedures and best practices developed by the activity recognition community for designing, implementing, and evaluating HAR systems. Note that the presented methods are generic and are not limited to activity recognition using wearable sensors. For educational purposes, the article is complemented with a publicly available dataset and a feature-rich activity recognition framework implemented in MATLAB. More specifically, we first discuss the key research challenges that human activity recognition shares with general pattern recognition and identify those challenges that are specific to human activity recognition. We introduce the concept of an Activity Recognition Chain (ARC) as a general-purpose framework for designing and evaluating activity recognition systems. The framework comprises components for data acquisition and preprocessing, data segmentation, feature extraction and selection, training and classification, decision fusion, and performance evaluation. We detail each component of the framework, provide references to previous research, and introduce the best practice methods developed by the activity recognition research community. We conclude with the educational toy problem of recognizing different hand gestures from inertial sensors attached to the upper and lower arm. We describe how each component of this framework can be implemented for this specific activity recognition problem and demonstrate how different design decisions compare and how they impact overall recognition performance.

2. RESEARCH CHALLENGES IN ACTIVITY RECOGNITION

While human activity recognition shares a number of research challenges with the more general field of pattern recognition, it also faces a number of unique challenges.

2.1. Common Research Challenges

2.1.1. Intraclass Variability. The first challenge that HAR shares with general pattern recognition is to develop recognition systems that are robust to intraclass variability. Such variability occurs because the same activity may be performed differently by different individuals. Intraclass variability can also occur if an activity is performed by the same individual. Several factors can affect the performance of the activity, such as stress, fatigue, or the emotional or environmental state in which the activity is performed. For example, the walking style may be more dynamic in the morning after a good night's sleep than in the evening after a full day of work. If an HAR system is trained for a single person—so-called person-dependent training—robustness to intraperson variability in performing a specific activity can be increased by using a larger amount of training data that captures as much of the variability as possible.

For an HAR system that was trained for several people—so-called person-independent training—the system may additionally become subject to considerable interperson variability. To address this issue, one can either again increase the amount of training data or develop person-independent features that are robust to this variability (e.g., features derived from full-body models instead of low-level signals [Zinnen et al. 2009a]). In the latter case, the design of the HAR system is subject to a delicate tradeoff between using a highly specific and discriminative feature set and using a feature set that is more generic and therefore potentially less discriminative, but also more robust across different people.

2.1.2. Interclass Similarity. An inverse challenge is given by classes that are fundamentally different but that show very similar characteristics in the sensor data (so-called interclass similarity). For example, in automatic dietary monitoring [Amft et al. 2007], drinking coffee or water from a glass both involve similar arm movements but have different nutritional results. Such close similarity can often only be resolved by using additional cues captured by different sensor modalities [Stikic et al. 2008] or by analyzing co-occurring activities [Huynh et al. 2008], in this example the activities of using the coffee machine or opening the tap, respectively.

2.1.3. The NULL Class Problem. Typically, only a few parts of a continuous data stream are relevant for HAR systems. Given this imbalance of relevant versus irrelevant data, activities of interest can easily be confused with activities that have similar patterns but that are irrelevant to the application in question—the so called NULL class. The NULL class problem is an active area of research. Explicitly modeling the NULL class is difficult, if not impossible, since it represents a theoretically infinite space of arbitrary activities. In some cases, the NULL class can be identified implicitly if the corresponding signal characteristics, for example, the signal variance, differ considerably from those of the desired activities. The NULL class can then be identified by thresholding on either the raw feature values or the confidence scores calculated by the classifier. In most cases, the NULL class is just a large unknown space that can be ambiguous and that leads to confusion with the activities at hand. Recent methods, such as self-learning [Amft 2011], may allow one to make use of some of the NULL class for classifier training.

2.2. Challenges Specific to HAR

2.2.1. Definition and Diversity of Physical Activities. The first challenge specific to the design of HAR systems is to develop a clear understanding of the definition of the activities under investigation and their specific characteristics. This may seem trivial at first. But human activity is highly complex and diverse and an activity can be performed in many different ways, depending on different contexts, and for a multitude of reasons. Katz et al. [1970] developed the Activities of Daily Living (ADLs) index as a tool in elderly care. Providing a good initial taxonomy of activities, it served many researchers as an inspiration to recognize activities relevant to real-world applications. Other resources include the comprehensive compendium of physical activity [Ainsworth et al. 2011]. It groups physical activity in categories based on the metabolic equivalent. Another resource for activity definition is given by time use databases. These were assessed by the government to understand citizens' time use, and Partridge and Golle [2008] investigate the potential of this data repository for activity recognition systems. Besides providing prior probabilities for activities at a certain time of day or location, it provides a taxonomy that can serve as a good reference for activity recognition researchers.

While state-of-the-art systems achieve decent performance on many activity recognition tasks, research so far mainly focuses on recognizing "which" activity is being performed at a specific point in time. In contrast, only little work investigated means

to extract qualitative information from sensor data that allow us to infer additional activity characteristics, such as the quality or correctness of executing an activity. For instance, while recognizing the task of brushing one's teeth is itself relevant and part of the ADL index, it may be even more relevant for a specific application to recognize whether this task is performed correctly. It is easy to see that such qualitative assessments are more challenging to perform automatically and have so far only been demonstrated for constrained settings, such as in sports [Velloso et al. 2013a, 2013b; Tessendorf et al. 2011b; Kranz et al. 2013]. For general activities or physical behaviors, activity recognition research is still far from reaching a similar understanding. First, we have to learn what information about the activity is relevant for the potential application. Second, we need to identify the requirements to the recognition systems, to obtain the desired information about the activities. For example, for obtaining regularity of daily routines, it is not necessary to detect the activity, but using statistics based on clustering may be sufficient (e.g., by using topic models [Huynh et al. 2008]). ADLs, on the other hand, comprise several complex activities as well as subactivities that might be performed in an interleaved fashion, in changing order or at different speeds, and thereby with considerable variation in execution. Hierarchies become relevant that allow recognition on different levels in order to zoom in or out of specific activities or parts thereof, and the hierarchical structures are necessary for the system's recognition performance [Blanke and Schiele 2010].

2.2.2. Class Imbalance. A related challenge is that of modeling different activity classes in the face of considerable class imbalance. For many activity recognition problems, such as for long-term behavioral monitoring, only few activities occur often, such as sleeping or working, while most activities occur rather infrequently, such as taking a sip of a drink [Blanke et al. 2010]. In general pattern recognition, class imbalance can often be addressed rather easily by recording additional training data. Alternatively, generating artificial training data to extend a smaller class to equal another class's size can mitigate class imbalance. One technique is oversampling (i.e., duplicating) a smaller class size to equal the bigger class size [Bulling et al. 2013]. In activity recognition recording, additional training data is more challenging, particularly if experimental procedures are not to be constrained or fully scripted to ensure equal class distributions. It is important to note, however, that the problem of class imbalance also depends on level of activity (high-level physical behaviors vs. low-level gestures) to be recognized by the particular HAR system.

2.2.3. Ground Truth Annotation. Another challenge for supervised HAR recognition tasks is the collection of annotated or "ground truth labeled" training data. Ground truth annotation is an expensive and tedious task, as the annotator has to perform the annotation in real time [Bulling et al. 2012] or to skim through the raw sensor data and manually label all activity instances post hoc. In addition, motion data recorded from an accelerometer or gyroscope is often more difficult to interpret than data from other sensors, such as cameras. In stationary and laboratory settings, annotation can often be obtained by relying on post hoc labeling based on video footage [Roggen et al. 2009; Blanke and Schiele 2010]. In daily life settings, ground truth annotation is a far more difficult problem. Researchers have investigated different techniques to address this problem, including daily self-recall methods [van Laerhoven et al. 2008], experience sampling [Kapoor and Horvitz 2008], and reinforcement or active learning—all of which involve the user. If only a few labeled training samples are available, semisupervised [Stikic et al. 2011], unsupervised [Huynh et al. 2008], or knowledge transfer [Zheng et al. 2009; van Kasteren et al. 2010; Blanke and Schiele 2010] learning techniques can be used.

2.2.4. Data Collection and Experiment Design. Finally, there are also experimental challenges associated with data collection and the evaluation of HAR systems in real-world environments. One challenge is to collect datasets on which HAR systems can be evaluated. This challenge arises from the fact that, in contrast to other research fields such as speech recognition or computer vision, the research community in activity recognition has not yet started a joint effort to collect rich and thus more general-purpose datasets of human physical activity, nor has it agreed on the (scientific) value of collecting them. This challenge is intensified because data collection may focus on quite diverse requirements, such as high data quality, large numbers of modalities or sensors, long-term recordings, or large numbers of participants. Using standard datasets is crucial for reproducible research and is becoming increasingly important in HAR as a research discipline. Second, to properly design and conduct an HAR experiment is also more difficult than it may at first seem. Researchers are faced with a tradeoff between unobtrusiveness and ease of use of the sensors; the time required to prepare, conduct, and maintain the experiment; and the logistics and costs for participants, experimenters, and the equipment.

2.3. Application Challenges

2.3.1. Variability in Sensor Characteristics. A practical challenge for implementing HAR in real-world applications is caused by the sensing equipment, more specifically the variability in sensor characteristics. This variability may have internal and external causes. Internal causes are hardware errors or complete failures, as well as sensor drift. External causes may include changes in the operating temperature or loose straps [Kunze and Lukowicz 2008; Bayati et al. 2011]. Some sensors are particularly sensitive to the environment, such as a barometer that requires frequent recalibration or magnetometers that are sensitive to ferromagnetic influences. Finally, portable devices containing sensors, such as mobile phones, may be used in different ways or carried at different locations on the body [Blanke and Schiele 2008]. Sensor displacement and changes in sensor orientation can be detected if they cause obvious differences in the recorded signals [Kunze et al. 2005]. Subtle deviations over time, such as signal drift, are much more difficult to identify.

2.3.2. Tradeoffs in Human Activity Recognition System Design. Designers of HAR systems also face challenges associated with the tradeoff between accuracy, system latency, and processing power [Yan et al. 2012]. Depending on the available resources and the recognition problem, some of these challenges are related. For many real-world applications, such as gesture-based input, real-time signal processing and classification are required. For others, such as behavioral monitoring or trend analysis over longer periods of time, offline data analysis and classification may be sufficient [Van Laerhoven and Berlin 2009]. The same is true for the second design dimension, the requirements in terms of latency of adaptation. While for some HAR systems low-latency classification and immediate feedback may be required, for others this may be less critical. Highly miniaturized embedded sensors for data recording typically have only limited processing power. Increasing the processing power of the sensors typically decreases runtime. One solution to this problem is to introduce a central component in the experimental setup to aggregate, process, and fuse the information drawn from different sensors [Lu et al. 2010].

3. THE ACTIVITY RECOGNITION CHAIN

An Activity Recognition Chain (ARC) is a sequence of signal processing, pattern recognition, and machine learning techniques that implements a specific activity recognition system behavior (see Figure 1). As can be seen from the figure, an ARC bears strong

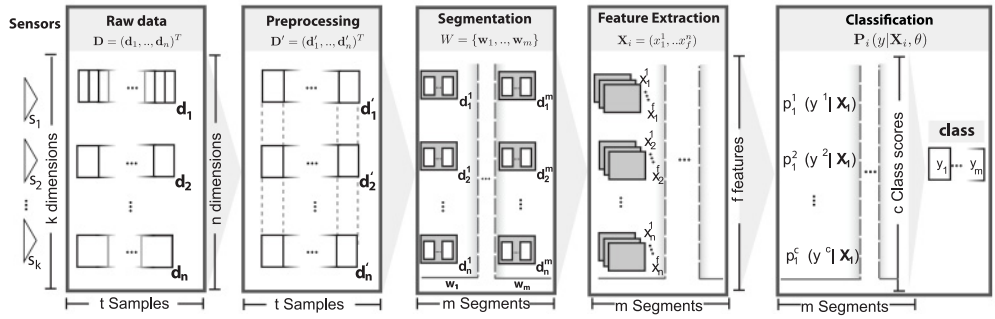

Fig. 1. Typical Activity Recognition Chain (ARC) to recognize activities from wearable sensors. An ARC comprises stages for data acquisition, signal preprocessing and segmentation, feature extraction and selection, training, and classification. Raw signals ($\mathbf{D}$) are first processed ($\mathbf{D}'$) and split into m segments ($\mathbf{W_i}$) from which feature vectors ($\mathbf{X_i}$) are extracted. Given features ($\mathbf{X_i}$), a model with parameters θ scores c activity classes $\mathbf{Y_i} = \{y^1, \ldots, y^c\}$ with a confidence vector $\mathbf{p}_i$.

similarity to general-purpose pattern recognition systems but, as we will detail in the following sections, also has a number of specific requirements and constraints. Also note that the chain can be executed in two different modes of operation if supervised classification algorithms are used, namely, training (modeling) and classification. Unsupervised classification doesn't require a dedicated training step but directly infers activities from the sensor data.

Input to the ARC consists of streams of sensor data acquired using multiple sensors worn on the body. The sensor data is first preprocessed to filter out signal variability or artifacts (see Section 3.1). The processed data is then segmented into sections of interest that are likely to contain an activity or gesture (see Section 3.2). Afterward, features that capture the activity characteristics are extracted from the signals within each segment (see Section 3.3). In training mode, the extracted features and corresponding ground truth class labels are used as input to train a classifier model in the training stage (see Section 3.4). In classification mode, the features and a previously trained model are used to calculate a score for each activity class and to map these scores into a single class label in the classification stage. If multiple sensors or classifiers are used, the output of several classifiers may subsequently be fused into a single decision (see Section 3.5). In addition, although typically only used during design time, a performance evaluation stage allows the assessment of the performance of the ARC (see Section 3.6).

3.1. Sensor Data Acquisition and Preprocessing

In the first stage of a typical ARC, raw data is acquired using several sensors attached to different locations on the body. In addition, advanced HAR systems may also include sensors placed in the environment. Such systems may capture additional data, for example, from objects in use or changes in the user's close surroundings (see Table III for an overview of common sensor modalities). Since some sensors can provide multiple values (e.g., an acceleration sensor provides a 3D acceleration typically referred to as x, y, and z direction), or multiple sensors are jointly sampled, vector notation is used to describe the sensor's output:

$$\mathbf{s}_i = (\mathbf{d}^1, \mathbf{d}^2, \mathbf{d}^3, \ldots, \mathbf{d}^t), \quad \text{for } i = 1, \ldots, k, \tag{1}$$

where k denotes the number of sensors and d^i the multiple values at a time t. Each of the sensors is sampled at regular intervals, which results in a multivariate time series. Often, however, the sampling rates of different types of sensors can differ. For example,

the typical sampling frequency for GPS is 5Hz, whereas acceleration is sampled at 25Hz or more. Sensors can also change their sampling frequency for other reasons, for example, for power saving or due to requirements of the operating system. In any case, t differs across $\mathbf{s}_i$, and synchronization across multimodal sensor data becomes a central technical issue. Moreover, raw sensor data can be corrupted by artifacts caused by a variety of sources (e.g., physical activity or sensor malfunction). AC power lines can cause electromagnetic interference with amplified electrical sensing techniques like EEG, EMG, EOG, and so forth. The function of the second stage of an ARC, the preprocessing stage, is to synchronize and to remove such artifacts and to prepare the acquired signals for feature extraction. It is important to note that this preprocessing is supposed to be generic; that is, it should not depend on anything but the data itself. It should not, for example, be specific to any particular person. The preprocessing stage transforms the raw multivariate and nonsynchronous time series data into a preprocessed time series $\mathbf{D}'$:

$$\mathbf{D}' = \begin{pmatrix} d_1'^1 & \cdots & d_1'^t \\ \vdots & \ldots & \vdots \\ d_n'^1 & \cdots & d_n'^t \end{pmatrix} = (\mathbf{d}_1', \ldots, \mathbf{d}_n')^T, \tag{2}$$

where $\mathbf{d}_i'$ corresponds to one dimension of the preprocessed time series, n to the number of total data dimensions, and t to the number of samples. The transformation aims to enhance the robustness of the extraction by applying signal processing algorithms that reduce noise or filter out artifacts. At the same time, these algorithms need to preserve those signal characteristics that carry relevant information about the activities of interest. Preprocessing of acceleration and gyroscope signals may involve calibration, unit conversion, normalization, resampling, synchronization, or signal-level fusion (see Figo et al. [2010] for a review). Physiological signals, such as electro-oculography (EOG), typically require preprocessing algorithms for denoising or baseline drift removal [Bulling et al. 2011].

3.2. Data Segmentation

The data segmentation stage identifies those segments of the preprocessed data streams that are likely to contain information about activities (also often referred to as activity detection or "spotting"). Information on activity segments not only is useful for classification but also can be used for other purposes, for example, to turn off the ARC to save power when no activity is sensed. Each data segment $\mathbf{w}_i = (t_1, t_2)$ is defined by its start time t_1 and end time t_2 within the time series. The segmentation stage yields a set of segments W containing a potential activity y:

$$W = \{\mathbf{w}_1, \ldots, \mathbf{w}_m\}. \tag{3}$$

Segmenting a continuous sensor stream is a difficult task. Humans perform activities fluently and consecutive activities blur into each other rather than being clearly separated by pauses. Another problem arises from the definition of an activity (see Section 2.2.1). Often, the exact boundaries of an activity are difficult to define. A drinking activity, for instance, might start with reaching for the cup *or* holding the cup and end after sipping *or* after putting the cup back on the table. In the literature, various methods exist to approach the problem of segmentation. To follow, we will explain in more detail the following methods specific to activity recognition: segmentation using a sliding window, energy-based segmentation, rest-position segmentation, the use of one sensor modality to segment data of a sensor of another modality, and the use of external context sources.

3.2.1. Sliding Window. In this approach, a window is moved over the time series data to "extract" a data segment that is then used in subsequent ARC processing stages. The window size directly influences the delay of the recognition system. The bigger the window size, the longer the ARC has to "wait" for a new segment to be available for processing. Also, the optimal (single) size is not clear a priori and can influence the recognition performance [Huynh and Schiele 2005]. The step size is subject to a tradeoff between segmentation precision and computational load. The larger the step size, the less frequently all subsequent stages of the ARC are executed, which reduces computational load, but also the less accurately the segmentation borders can be defined. Although commonly used, a fixed-size sliding window is agnostic about the type and structure of the underlying time series data.

3.2.2. Energy Based. Energy-based segmentation exploits the fact that for many HAR problems, different activities are performed with different intensities. These differences in intensity directly translate to different energy levels of the recorded sensor signals. The energy E of a signal s is calculated as $E = \int_{-\infty}^{\infty} |s(t)|^2 \, \mathrm{d}t$. By thresholding on E, data segments can be identified that are likely to belong to the same activity [Guenterberg et al. 2009]. A special case of energy-based segmentation is to require the user to assume a predefined rest position between each activity [Lee and Xu 1996; Amft et al. 2005]. Segmentation based on a rest position is particularly suited for gesture-based HCI and HAR problems that involve discrete activities or gestures. Whenever the rest position is detected by the HAR system, a segment border is assumed [Wilson and Bobick 2000]. For whole-body activity recognition, the rest position can be a certain posture; for the recognition of gestures, a defined hand position can be used. To allow for more natural movements, an adaptive sliding window technique has been proposed based on naturally occurring pauses, such as the turning point of arms [Zinnen et al. 2009b].

3.2.3. Additional Sensors and Contextual Sources. Sensor data recorded with one modality can also be segmented using information derived from additional modalities. For example, long-term acceleration data recorded on a mobile phone can be segmented using GPS traces [Ashbrook and Starner 2003] or sound recorded using the internal microphone [Lu et al. 2009]. Similarly, segmentation can be performed using external context sources (i.e., sensors external to the recording device), such as a diary or calendar that may hold information about the start and duration of activities such as meetings.

3.3. Feature Extraction and Selection

The feature extraction and selection stage reduces the signals into features that are discriminative for the activities at hand. Features may be calculated automatically (see Plötz et al. [2011] for an example) and/or derived based on expert knowledge. Features are extracted as feature vectors $\mathbf{X}_i$ on the set of segments W, with $\mathcal{F}$ being the feature extraction function:

$$\mathbf{X}_i = \mathcal{F}(\mathbf{D}', \mathbf{w}_i). \tag{4}$$

The total number of features extracted from the data form the so-called feature space. Generally speaking, the more clearly each activity can be separated in the feature space, the higher the achieved recognition performance. Ideally, features corresponding to the same activity should be clustered in the feature space, while features corresponding to different activities should be far apart. At the same time, "good" features need to be robust across different people as well as to intraclass variability of an activity. Depending on the type of activities, these features may be extracted on oversegmenting windows (for repetitive activities) or on windows covering the entire activity or gesture

(for nonrepetitive activities). Research in activity recognition has resulted in a wide range of features, for example:

—*Signal-based features*: these are mostly statistical features, such as the mean, variance, or kurtosis. These features are popular due to their simplicity as well as their high performance across a variety of activity recognition problems [Bao and Intille 2004; Ravi et al. 2005]. For physiological or audio signals, these can also be frequency-domain features, such as mel-frequency cepstral coefficients, or energy in specific frequency bands [Kang et al. 1995].
—*Body model features*: these are calculated from a 3D skeleton using multiple on-body sensors [Zinnen et al. 2009b]. Encoding prior knowledge increases robustness across persons and can lead to higher performance [Zinnen et al. 2009a]. Polynomial features that describe signal trends such as mean, slope, and curvature are used for trajectories of limbs [Blanke et al. 2010].
—*Event-based features*: for example, for eye movements, these are features extracted from saccades, fixations, or blinks, as well as features describing the characteristics of repetitive eye movement sequences [Bulling et al. 2011].
—*Multilevel features*: the data is first clustered, for example, using k-means on which occurrence statistics are calculated on a sliding window. Encoded duration, frequency, and co-occurrences of data provide expressive features [Huynh et al. 2008; Blanke and Schiele 2009; Zhang and Sawchuk 2012].

The higher the dimensionality of the feature space, the more training data is needed for model parameter estimation and the more computationally intensive the classification. Particularly for real-time processing on embedded systems, the objective is to minimize memory, computational power, and bandwidth requirements. It is therefore important to use a minimum number of features that still allow the ARC to achieve the desired target performance. Manual selection of such features is a difficult task. A large variety of methods for automatic *feature ranking and selection* has been developed (see Guyon and Elisseeff [2003] for an introduction). These can be categorized into wrapper [Kohavi and John 1997], filter [Peng et al. 2005], or hybrid [Somol et al. 2006] approaches, each with its specific properties. Modern machine learning approaches such as SVM or AdaBoost include a "built-in" feature selection mechanism. Relevant features are automatically selected while ensuring generalization at the same time.

3.4. Training and Classification

Research in machine learning and computational statistics developed a large variety of inference methods. Table II provides an overview of approaches used for different activities over the last 15 years. HAR researchers have successfully demonstrated template-based similarity metrics such as Dynamic Time Warping (DTW) [Blanke et al. 2011] or string matching [Stiefmeier et al. 2007; Bulling et al. 2008]. For more complex data exhibiting temporal dependencies, temporal probabilistic models such as Hidden Markov Models (HMMs) [Rabiner 1989; Bulling et al. 2008; Fink 2008], Conditional Random Fields (CRFs) [Liao et al. 2005; van Kasteren et al. 2008; Blanke and Schiele 2010], or dynamic Bayesian networks [Patterson et al. 2005] have been used. Discriminative approaches, for example, Support Vector Machines (SVMs) [Huynh et al. 2007; Bulling et al. 2011, 2012], C4.5 decision trees [Bao and Intille 2004], or (joint) boosting [Lester et al. 2005; Blanke and Schiele 2009], have been successfully applied to a variety of activities and sensor settings. Newcomers to the field may experience difficulties in interpreting the state of the art due to the numerous evaluation metrics used. However, we can still estimate certain tendencies (cf. Table II). For example, discriminative learning schemes showed higher recognition performance for multiple studies: *(2 vs. 3), (4), (10 vs. 11)*. Especially the ability to identify most contributing features

Table II. Examples of Activity Recognition Using On-Body Sensors to Illustrate the Diversity of Methods and Activities to be Recognised (Evaluation metrics are abbreviated: precision: "prec", recall: "rec", acccuracy: "acc", 1- equal error rate: "EER")

	Methods	Activities	# classes	# participants	Results	Reference
1	HMM	daily situations	12	1	85.8% - 99.7% acc	[Clarkson et al. 2000]
2	Topic models	daily routines	4	1	77% prec, 66% rec	[Huynh et al. 2008]
3	Joint boosting	daily routines	4	1	88% prec, 90% rec	[Blanke and Schiele 2009]
4	CRF/HMM	daily home activities	7	1	96%/95%	[van Kasteren et al. 2008]
5	Decision tree	selected daily activities	20	20	84% acc	[Bao and Intille 2004]
6	AdaBoost+HMM	selected daily activities	8	12	90%	[Lester et al. 2006]
7	HMM	eating and drinking arm gestures	5	2	87% acc	[Amft et al. 2005]
8	SVM	office activities from eye movements	6	8	76.1% prec, 70.5% rec	[Bulling et al. 2011]
9	String matching/SVM	reading from eye movements	2	8	88.9% prec, 72.3% rec / 87.7% prec, 87.9% rec	[Bulling et al. 2012]
10	HMM/LDA	assembly tasks	9	5	63% prec, 66% rec	[Ward et al. 2006]
11	CRF	composite and low-level DIY activities	10 and 6	6	75% EER and 88% EER	[Blanke and Schiele 2010]
12	String matching	bike maintenance tasks	5	3	82.7%	[Stiefmeier et al. 2007]
13	naive Bayes/kNN	car maintenance tasks (person dependent)	20	8	48% prec, 71% rec	[Ogris et al. 2008]
14	Joint Boosting	car maintenance tasks (person independent)	20	8	93% EER	[Zinnen et al. 2009b]
15	kNN	Tai Chi movements	3	4	85% acc	[Kunze et al. 2006]
16	HMM	American sign language	40	–	around 95%	[Starner et al. 1997]
17	–	walking styles	4	4	–	[Lukowicz et al. 2006]
18	HMM	self-stimulatory behaviour in autism	8	1	68.57%	[Westeyn et al. 2005]

Table III. Common Sensors and Example Applications in Human Activity Recognition (HAR). To Improve Recognition Performance, HAR Systems Make Use of Multiple Modalities, i.e., Sensors Integrated into the Environment (E), into Objects (O), or Wearable Sensors Attached to the Body (B)

Sensor	Location	Applications
Microphone	EOB	Speaker recognition, localisation by ambient sounds, activity detection, object self-localisation [Amft et al. 2005; Clarkson et al. 2000; Lu et al. 2009]
Accelerometers or gyroscopes	EOB	Detection of body movement patterns, object use, ambient infrastructure [Godfrey et al. 2008; Westeyn et al. 2005; Huynh and Schiele 2005; Blanke and Schiele 2009; Bächlin et al. 2009]
Magnetometer	—B	Orientation of the body [Lee and Mase 2002] or relative position sensing of body parts [Pirkl et al. 2008]
Inertial measurement units	-OB	Absolute orientation, multiple sensors for body model reconstruction [Blanke et al. 2011; Ogris et al. 2008; Stiefmeier et al. 2007; Zinnen et al. 2009a; Blanke and Schiele 2010; Bulling et al. 2012]
Capacitative sensing	—B	Breathing, fluid intake [Cheng et al. 2010]
Pressure sensor	—B	Vertical motion, e.g. in elevator or staircase [Lester et al. 2005]
Light sensor (visible, IR, UV)	—B	Localisation of windows, lamps, light tubes [Maurer et al. 2006; van Laerhoven and Cakmakci 2000]
Skin temperature	—B	Health state (e.g. fever) [Anliker et al. 2004]
Galvanic skin response	—B	Measure of skin conductivity to infer emotional states or levels of arousal [Pentland 2004]
Environment temperature	E—	Discrimination of outdoor vs. indoor settings
Oximetry	—B	Blood oxygen: Detection of sleep apnoea [Oliver and Flores-Mangas 2007]
ECG	—B	Electrocardiography: Monitoring of physical activity and health state
EOG	—B	Electrooculography: Analysis of eye movements and recognition of cognitive processes [Bulling et al. 2011, 2012; Bulling and Roggen 2011; Bulling et al. 2009]
EMG	—B	Electromyography: Detection of muscle activation [Kang et al. 1995]
EEG, fNIR	—B	Electroencephalography and functional near-infrared spectroscopy: Measure of brain activity
Strain, stress	—B	User's breathing (respiration belt), movement (strain sensors in clothes) [Lukowicz et al. 2006; Mattmann et al. 2007; Morris and Paradiso 2002]
UWB	E—	Ultra wide band: User localisation [Ogris et al. 2008]
GPS	E-B	Global positioning system: User localisation, activities at locations, prediction of future locations [Liao et al. 2005; Krumm and Horvitz 2006]
Camera	E-B	Localisation, body model reconstruction [Clarkson et al. 2000]
Reed switches	EO-	Use of objects and ambient infrastructure [van Kasteren et al. 2008]
RFID	EO-	Radio-frequency identification: Use of objects and ambient infrastructure [Philipose et al. 2004; Stikic et al. 2008; Wang et al. 2007; Buettner et al. 2009]
Proximity	E-B	motion detection, tracking, localisation [Schindler et al. 2006]; behaviour analysis [Wren et al. 2007]; obstacle avoidance [Cassinelli et al. 2006]

helps to discriminate well between activities (and to background class). This can allow better recognition for a person-independent case compared to a person-dependent case *(14 vs. 13)*. What we cannot observe is a general "best of breath" selection of machine learning algorithms. For example, (3) makes use of a rich feature representation (representing duration and co-occurrence) in combination with a simple algorithm, while (2) or (1) uses a more complex model (representing co-occurrences respective to temporal relationships). Furthermore, if characteristics become apparent in the feature space, even a kNN classification can suffice (15). The choice for a particular inference method is subject to a tradeoff between computational complexity and recognition performance. With a view to classification on embedded systems with limited resources, the goal is to minimize computational complexity and memory requirements while still achieving high recognition performance. Feature selection allows one to identify contributing features during training and thereby reduce computational complexity during classification [Blanke and Schiele 2009]. Therefore, inference methods are typically selected depending on the type of activity and the complexity of the feature space. They may also be selected based on other factors such as latency or online operation and adaptation. Depending on the mode of operation of the ARC, either the training or the classification stage is active to further process the extracted features.

3.4.1. Training. The models of supervised inference methods need to be trained before operation. Training is performed using training data $\mathcal{T} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N}$, with N pairs of feature vectors $\mathbf{X}_i$ and corresponding ground truth labels y_i. Model parameters θ can be learned to minimize the classification error on $\mathcal{T}$. For example, hidden Markov models are defined by parameters $\theta = (\pi, A, B)$, with matrix A corresponding to transitions between states, B to the output probabilities of each state, and π to the initial state probabilities. Given the training data $\mathcal{T}$ and an initial guess of the parameters θ, a separate model is trained for each class using expectation maximization [Rabiner 1989; Fink 2008]. Discriminative approaches minimize the error by gradient descend. In contrast, nonparametric classifiers such as kNN take as parameters the labeled training data $\theta = (\mathcal{T})$ without further training and match the label of the k-nearest neighbors to the test sample.

3.4.2. Classification. The classification stage performs two distinct steps. In the first step, using a trained model with parameters θ, each feature vector $\mathbf{X}_i$ is mapped to a set of class labels $\mathcal{Y} = \{y^1, \ldots, y^c\}$ with corresponding scores (or confidence values) $\mathcal{P}_i = \{p_i^1, \ldots, p_i^c\}$:

$$p_i(y|\mathbf{X}_i, \theta) = \mathcal{I}(\mathbf{X}_i, \theta), \quad \text{for } y \in \mathcal{Y}, \tag{5}$$

with the inference method $\mathcal{I}$. For Bayesian approaches, such as dynamic Bayesian networks or naïve Bayes classifiers, the scores correspond to probabilities. Many non-Bayesian classifiers can be calibrated to provide similar probabilistic outputs [Cohen and Goldszmidt 2004]. In a second step, the calculated scores $\mathcal{P}_i$ can then be used in different ways. One of the most common uses is to calculate the maximum score and to take the corresponding class label y_i as the classification output:

$$y_i = \underset{y \in \mathcal{Y},\ p \in \mathcal{P}_i}{\operatorname{argmax}}\ p(y|\mathbf{X}_i, \theta). \tag{6}$$

Alternatively, the scores can be used by the end application to decide whether to trust the system's output. In particular, if all scores fall below th_{rej}, the corresponding data sample is considered to belong to the NULL class—a mechanism typically referred to as NULL class rejection. The threshold th_{rej} directly influences the recognition system's performance. A large threshold (i.e., a low tolerance to activity outliers) may lead to a large number of activity instances not being detected by the system. In contrast, a high

tolerance may lead to a large number of falsely detected activity instances. The threshold th_{rej} is therefore typically trained at design time, using multiobjective optimization techniques jointly with the feature extraction, feature selection, or classifier training stages. Finally, the scores can directly be used as input to another inference method (so-called classifier stacking), for example, to find higher-level structure in the activity data (see Clarkson and Pentland [1999], Lester et al. [2005], Wang et al. [2007], and Blanke and Schiele [2010] for examples).

3.5. Decision Fusion

Multiple sensors or multiple classifiers (also known as ensemble classifiers or boosting) were shown to increase recognition performance [Ho et al. 1994; Kittler et al. 1998; Polikar 2006]. The decision fusion stage combines several intermediate (often weaker) classification results into a single decision. Fusion can happen either at the early stage (i.e., at the level of the features) or at a later stage (i.e., at the level of classifiers). Fusion rules commonly used in activity recognition research are summation, majority voting [Stikic et al. 2008], Borda count [Ward et al. 2006], and Bayesian fusion [Zappi et al. 2007]. Although Bayesian approaches have recently been gaining more widespread popularity, the limited resources of embedded systems often require limiting the complexity of the fusion approaches. Introduced in machine learning and computer vision [Friedman et al. 2000; Torralba et al. 2007], boosting as a variant of decision fusion has been successfully applied to activity recognition as well [Lester et al. 2005; Blanke and Schiele 2009; Zinnen et al. 2009b]. Besides increased recognition performance, sensor fusion has additional benefits for an HAR system, such as (1) increased robustness (e.g., to faults or variability in sensor characteristics); (2) reduced classification problem complexity through use of classifiers dedicated to specific activity subsets, selected according to another sensor modality (e.g., the user's position constrains the activities that may occur at that location); (3) derivation of confidence measures from the agreement between classifiers; (4) classification with missing features; and (5) discriminative training.

3.6. Performance Evaluation

Evaluating the recognition performance of an ARC is crucial and is usually done in the design phase. During operation, performance evaluation may allow optimization of the runtime behavior of an ARC. Generally speaking, activity recognition systems can miss, confuse, or falsely detect activities that did not occur. Besides correct classification in terms of *True Positives* (TPs) and *True Negatives* (TNs), classification can be wrong and lead to *False Negatives* (FNs) and *False Positives (FPs)*. The optimization objective may be to maximize a single performance metric or several at the same time. The choice of metric to be optimized depends on the application. Often it is favorable to reduce FNs at the price of FPs [Altakouri et al. 2010], for example, for prefiltering video data for human analysis [Patterson and Singh 2010]. In other cases, a high FP rate can make people ignore the system's notifications and eventually abandon the system.

Activity recognition has adopted several performance metrics that have proven to be beneficial in other fields, such as *confusion matrices*; related measures such as *accuracy*, *precision*, *recall*, and *F-scores*; or decision-independent *Precision-Recall (PR)*- or *Receiver Operating Characteristic (ROC)* curves. For further details on metrics specifically geared toward activity recognition, we point the reader to Minnen et al. [2006b] and Ward et al. [2011]. We now summarize some common metrics that are frequently used in activity recognition research.

3.6.1. Confusion Matrix. A confusion matrix summarizes how many instances of the different activity classes got confused (i.e., misclassified) by the system. Typically, the

rows of a confusion matrix show the number of instances in each actual activity class (defined by the ground truth), while the columns show the number of instances for each predicted activity class (given by the classifier's output). Each row of the matrix is filled by comparing all ground truth instances of the corresponding actual class with the class labels predicted by the system. From the matrix, *precision* ($\frac{TP}{TP+FP}$) and *recall* ($\frac{TP}{TP+FN}$) values as well as the overall *accuracy* ($\frac{TP+TN}{all}$) and the harmonic mean of precision and recall, the F1 score ($\frac{2*precision*recall}{precision+recall}$), can be calculated for each activity class. If a dataset is unbalanced (i.e., the number of ground truth instances of the activity classes vary significantly), the overall accuracy is not representative of the true performance of a classifier. The number can be strongly biased by dominant classes, usually the less relevant background class. To address this "class skew" problem, normalized confusion matrices should be used to allow for objective comparison between the different activity classes. Instead of absolute counts of instances, a normalized confusion matrix shows the confusion as a percentage of the total number of ground truth activity instances.

3.6.2. ROC and PR Curves. It is often difficult to set the optimal decision threshold on the classifier's score beforehand. Therefore, a common strategy is to sweep the threshold on the score for each individual class (one vs. all) and analyze the behavior in so-called Receiver Operating Characteristic (ROC) or Precision-Recall (PR) curves [Fawcett 2006]. ROC curves plot the true positive rate (recall) against False-Positive Rate (FPR) ($\frac{FP}{FP+TN}$). Typically, lowering the decision threshold increases the recall and respectively the FPR. Best-case results approach the top left corner, while worst-case (i.e., random) results follow the diagonal if class distributions are balanced. As ROC curves depend on TN counts, imbalanced class distributions (i.e., percentage of relevant activity vs. percentage of all other activities including NULL) may lead to "overoptimistic" ROC curves. PR curves do not depend on the true negative count. Therefore, they are suited to detection tasks, where activities of interest are "buried" in a large corpus of irrelevant data (NULL class). Similarly to ROC curves, lowering the decision threshold results in an increased recall and typically decreases the precision by increasing false positives.

Several metrics can be extracted from ROC and PR curves to summarize them into a single and thus more easily comparable number. *Equal Error Rate (EER)* represents the point in the PR curve where precision equals recall. The higher this value, the better. Another measure is *average precision*. Precision is measured at uniform steps (e.g., 10% steps) of the recall and subsequently averaged [Everingham and Winn 2007]. Finally, the *Area Under Curve (AUC)* can be calculated from ROC curves as a measure to describe the overall performance of a classifier [Ling et al. 2003]. The AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

3.6.3. Time-Based Evaluation. Activity recognition performance is typically evaluated with respect to time, that is, by performing a frame-by-frame comparison between the ground truth and the classifier's prediction. By understanding classification as a segmentation problem, further metrics were introduced that allow a more detailed performance analysis [Ward et al. 2006; Minnen et al. 2006b]:

(1) *Insertion* (an activity segment is detected where there is none in the ground truth) and *deletion* (failure to detect an activity segment).
(2) *Fragmentation* and *merge*: Fragmentation errors denote when activity segments in the ground truth correspond to several segments in the recognition system output.

Merge is the opposite; that is, several ground truth activity segments are combined into one segment.

(3) Timing errors: *Overfill* errors are where an activity segment in the output of the system extends into regions of *NULL*. The opposite of *overfill* is *underfill* (u): in this case, the segment recognized by the system fails to "cover" some parts of the ground truth segment.

3.6.4. Event-Based Evaluation. An alternative approach is to evaluate a system's performance in spotting activity *events* rather than detecting the exact start and end times of activity segments. To evaluate for such events, the evaluation criterion can be modified. A segment $\mathbf{w}_i$ is counted as a true positive if the annotated label $\mathbf{l}$ has the same activity label for $start(\mathbf{l}) \leq center(\mathbf{w}_i) \leq stop(\mathbf{l})$. For example, ensuring a 50% overlap of the event with the ground truth label adds a second criterion: $o = \frac{samples(w_i \cap l)}{samples(w_i \cup l)}$ for $o \geq 0.5$ [Everingham and Winn 2007] (see Ward et al. [2011] for more sophisticated event-based evaluation techniques).

3.6.5. Evaluation Schemes. Evaluation is typically conducted using leave-one-out cross-validation to assess how the recognition system generalizes to a new situation. To this end, the experimental dataset is partitioned into multiple folds. All folds except one are used to train the recognition system. The left-out fold is used for testing. The process is repeated rotating the left-out fold until all folds have been used once for testing. Folds are built differently to assess different aspects of generalization. Datasets may include recordings of multiple persons, on multiple days, and of multiple runs containing repetitions of a set of activities. Leave-one-person-out is used to assess generalization to an unseen user for a user-independent recognition system. Leave-one-run-out is used to assess a user-specific system. Since the user's movement trajectories or even strategies may change over time, leave-one-day-out is used to assess the robustness of the system over time.

4. CASE STUDY

We conducted a small user study on the example problem of recognizing hand gestures from body-worn accelerometers and gyroscopes. Hand gestures are commonly used in activities of daily living (such as in the kitchen) and gesture-based video game interfaces (e.g., for playing sports with characteristic movements, such as tennis or golf). As the focus of this tutorial is on providing an educational example, the goal of the study was to demonstrate how different design decisions in the ARC compare and how they impact overall recognition performance. The case study was therefore deliberately kept simple in terms of the number and type of sensors, the experimental setup and procedure, and the number of participants. It is important to note that activity recognition in real-world settings is much more challenging with respect to these aspects and will also typically include confounding "garbage" activity events that need to be taken care of.

4.1. Setup

We recorded arm movements of two people performing a continuous sequence of eight gestures of daily living, as listed in Figure 2 (right). To increase diversity, we also recorded typical arm movements performed while playing tennis. In addition, we included periods with no specific activity, the NULL class. For NULL class periods, no activity was required of the participants, but they were asked not to engage in any of the other activities. Taken together, this constitutes a 12-class recognition problem. The activities were performed in succession with a brief break between each activity. Each activity (including NULL) lasted between two and eight seconds and was repeated 26

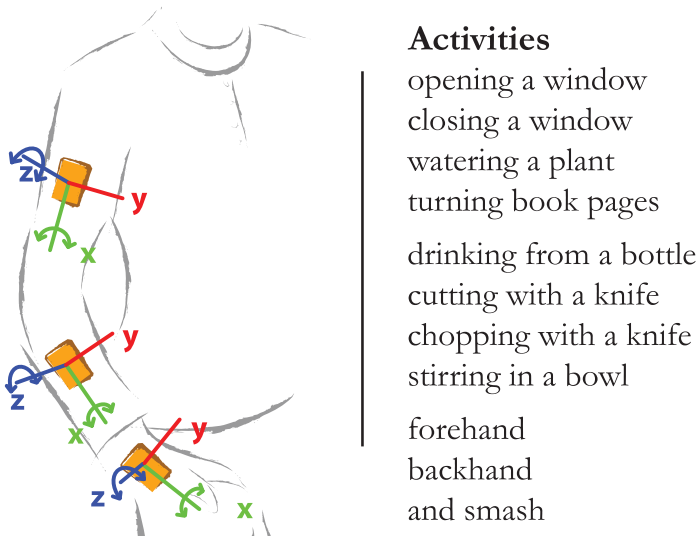

Fig. 2. Sensor setup and activities performed.

times (287 times for NULL) by each participant, resulting in a total dataset of about 70 minutes.

4.2. Apparatus

Arm movements were tracked using three custom Inertial Measurement Units (IMUs) placed on top of each participant's right hand, as well as on the outer side of the right lower and upper arm as depicted in Figure 2 (left). The IMUs comprise a three-axis accelerometer and a two-axis gyroscope recording timestamped motion data at a joint sampling rate of 32Hz. All recorded data was sent via Bluetooth to a laptop placed in close proximity to the participants. Data synchronization was handled offline using the SenseHub synchronization software (see Roggen et al. [2010] for details on SenseHub). Participants were observed by an assistant who instructed them and manually annotated their current gesture.

5. EVALUATIONS

Each stage of the ARC framework described in Section 3 can be implemented using a variety of methods (e.g., by choosing a specific set of features or a specific classifier). The parameters of each of these methods directly influence the overall recognition performance of the system. In addition, several stages of the chain depend on each other and need to be evaluated jointly to achieve high recognition performance. This poses an optimization problem that becomes even more challenging during operation if feedback from the user feeds into the ARC or if optimizations have to be performed in real time to allow for continuous adaptation of the ARC [Roggen et al. 2013]. Generally speaking, the optimal solution to this problem can only be found by using sophisticated, multidimensional optimization procedures.

For the sake of simplicity and intelligibility, in this tutorial we evaluate each stage of the ARC separately. We present a series of evaluations, each highlighting one stage of the ARC. It is important to note that these evaluations are not geared toward yielding the overall best recognition performance. Instead, they illustrate the key design decisions in implementing activity recognition systems by reducing the complexity of the problem. Specifically, we report on the following evaluations:

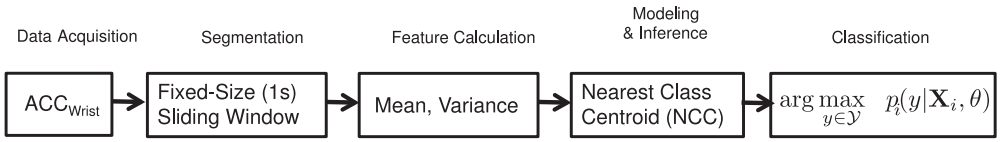

Fig. 3. Basic Activity Recognition Chain (ARC) for recognizing hand gestures from a wrist-worn accelerometer. The ARC consists of signal segmentation using a fixed-size sliding window, extraction of simple features such as mean and variance from the acceleration signal, and classification using a Nearest Class Centroid (NCC) classifier.

—*Basic ARC:* We first evaluate a basic ARC that comprises a single accelerometer attached to the right hand, simple features, and a lightweight classifier. This evaluation will serve as the baseline for all following evaluations (see Section 5.1).
—*Features:* We then analyze the influence of different types of features and combinations of these on the recognition performance (see Section 5.2).
—*Feature Extraction:* We evaluate different parameters of the feature extraction stage, such as the window size (see Section 5.3), sensor placement (see Section 5.4), and the type of sensors used (see Section 5.5).
—*Classifiers:* We compare different classifiers that are commonly used in HAR research with respect to their recognition performance (see Section 5.6).
—*Feature Selection:* Finally, we show how to optimize recognition performance by using feature selection mechanisms (see Section 5.7).

We perform each of these evaluations along two dimensions: the evaluation scheme and the sensor configuration. We compare two evaluation schemes, person-dependent vs. person-independent leave-one-repetition-out cross-validation. For person-dependent evaluation, for each participant, we leave out one repetition for testing and train on all remaining repetitions of the same participant. For the person-independent case, we train on all repetitions of one participant and test on all repetitions of the second. In both cases, overall recognition performance is calculated as the average performance across all cross-validation rounds. In addition, we compare two different sensor configurations, namely, using only one accelerometer attached to the right hand vs. using all sensors. The evaluation of different numbers of sensors is motivated by the diversity of application areas for a typical activity recognition system. Implementing a recognition system for long-term use (e.g., a step counter in a watch) requires only a small number of simple, low-power sensors, such as a single accelerometer. In contrast, a wearable system for tracking full-body movements (as, for example, those commonly used in the film industry to animate virtual characters) requires a network of powerful inertial measurement units spread over the whole body. For performance evaluation, we opted to use a time-based evaluation, because it is commonly used in HAR research.

5.1. Basic Activity Recognition Chain

Figure 3 shows a basic ARC that addresses the specific requirements of this gesture recognition problem. The chain uses simple features and a lightweight classifier to minimize computational complexity. The first stage of the ARC segments the signals using a sliding window with a fixed window of length $W_s = 1s$ and a fixed step size of $S_s = 1s$. In each step, two common features are extracted individually for each sensor dimension: the mean and the variance of the signal in the current window. The features are fed into an NCC classifier, which is equivalent to a k-NN classifier with $k = 1$. The NCC classifier is well suited for embedded implementation and real-time recognition, as it is lightweight and has only low computational complexity. Each instance is then assigned to one of the defined activity classes.

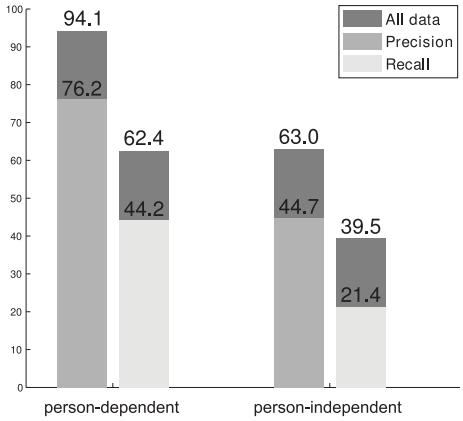

Fig. 4. Precision and recall for person-dependent and person-independent evaluation using a single accelerometer attached to the right hand (blue and red bars) and using all available sensors (green bars). Results are averaged over both participants using the ARC shown in Figure 3.

groundtruth \ classification	NULL	Open window	Drink	Water plant	Close window	Cut	Chop	Stir	Book	Forehand	Backhand	Smash	recall
NULL	24267	216	444	3228	48	24	60	75	45		3		85.42
Open window	3849	1938	453	291	48	12	9		24				29.26
Drink	3984	927	3780	321	3	9							41.89
Water plant	3984	726	774	3735	21	57	15						40.11
Close window	3891	381	1173	945	1533								19.35
Cut	2940		264	450		6585	456		3				61.55
Chop	2895	168	435	153		909	5742		126				55.06
Stir	4947	39	135	42	21	474	561	4392	207				40.60
Book	4560	27	144	951		354	1725	60	6687				46.09
Forehand	3195	330		144	609	9	66		3	969	6	3	18.17
Backhand	3003	207	21		21	3	6	24	33		1302		28.18
Smash	1860	57		78	185		42	45		1567	137	230	5.47
precision	38.29	38.64	49.59	36.13	61.59	78.06	66.14	95.56	93.81	38.21	89.92	98.71	

Fig. 5. Confusion matrix for person-independent evaluation and all data for participant one, fold 1.

Results. Figure 4 shows precision and recall for the two dimensions (evaluation scheme and amount of data) of the basic ARC given in Figure 3 averaged over both participants. As can be seen from the figure, training and testing on the same person results in 76.2% precision (44.2% recall) using only the accelerometer attached to the right hand. When using all sensors, precision increases to 94.1% (62.4% recall). For the person-independent case, results are lower: 44.7% precision and 21.4% recall. Using data from all sensors attached to three different positions on the body improves the recognition performance to 63% precision (39.5% recall).

Figure 5 shows the confusion matrix for person-independent evaluation. Overall, the household activities *opening window*, *drinking*, *watering the plant*, and *closing window* are mostly confused. *Forehand* is confused with *closing window* as well as with *smash*. Since the NULL class is overrepresented and the boundaries of activities are not always precisely detected, nearly all activities lose recall to the NULL class. Overall lower performance in the person-independent case is expected, as different users tend to perform activities differently. For person-independent evaluation, the model is trained on one user and used to classify activities of the other. The variability

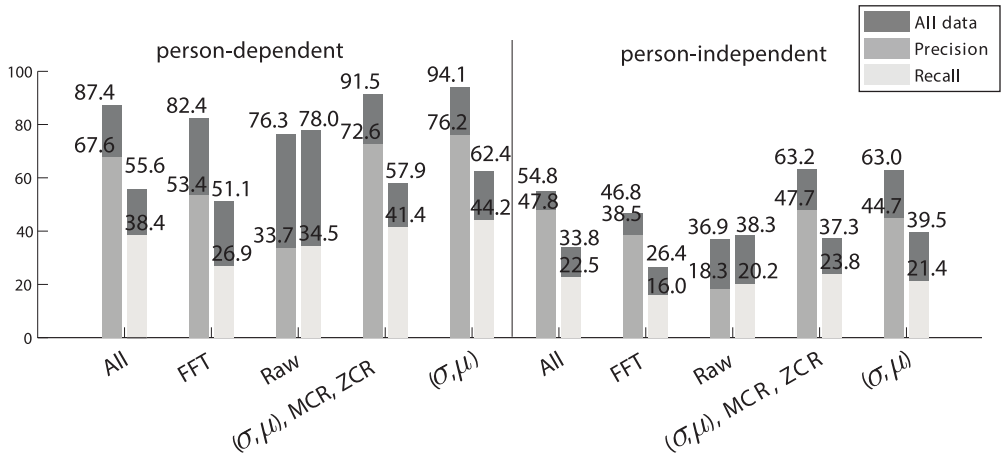

Fig. 6. Recognition performance for typical feature types used in activity recognition: using the raw signals (Raw), mean and variance, Zero Crossing Rate (ZCR), Mean Crossing Rate (MCR), features based on FFT, and combinations of these.

in executing the activities across both users reduces generalization. Given the strong differences in execution by both subjects, even a 1-nn classifier suffers from overfitting. To address this problem, more participants could be recorded to cover more of this variability, hence improving generalization ability.

5.2. Feature Types

Based on the basic ARC, we first analyze the influence of different feature types on the recognition performance. To this end, we compare five different cases:

(1) No feature extraction (raw signals)
(2) Mean and variance of the signal
(3) Mean, variance, Zero Crossing Rate (ZCR), and Mean Crossing Rate (MCR)
(4) Features based on fast Fourier transform: coefficients grouped in four logarithmic bands, 10 cepstral coefficients, spectral entropy, and overall energy [Lester et al. 2005].
(5) Combination of all features from (2) to (4).

Results. Figure 6 shows the results for different feature types. The best performance is achieved by using mean and variance as features. This result is consistent with previous activity recognition studies described in the literature and illustrates the popularity of these features in the HAR community. Combining mean and variance with other features (FFT and zero crossings) leads to a small decrease of performance. This may seem counterintuitive, as one might expect that additional features always lead to improved recognition performance. The k-NN classifier, however, is very sensitive to the feature quality, and adding low-quality features to the feature set can have a negative impact on the performance. Typically it is not clear beforehand which features to choose. Feature selection techniques can be used to identify the most relevant features (see Section 5.7). One can see that choosing a specific feature type can have significant impact on the recognition. A second observation is the performance difference using all data versus using one sensor only. This becomes particularly evident when using raw data, which results in a surprisingly high recognition rate when using all available

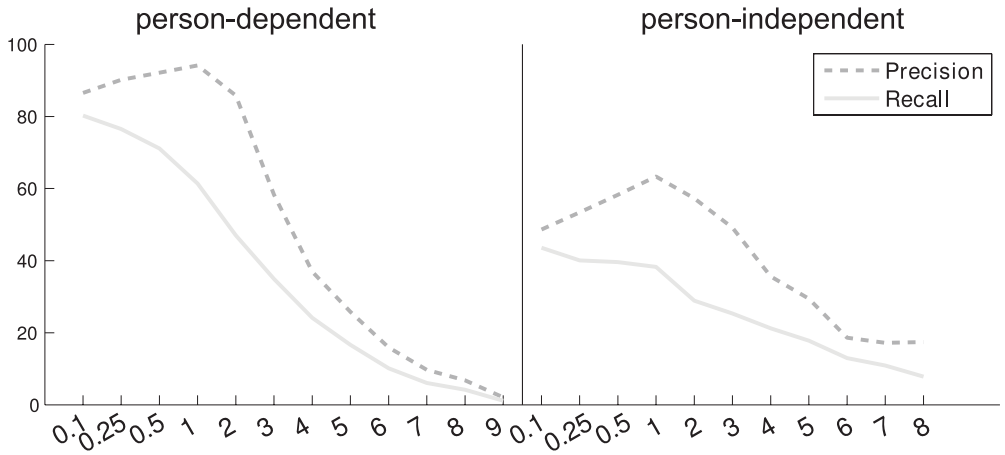

Fig. 7. Recognition performance for different feature extraction window sizes using data from all sensors.

data. This suggests that sensor type and placement might play an important role for the activities addressed. In Sections 5.4 and 5.5, we evaluate both of these settings.

5.3. Window Size Used During Feature Extraction

A parameter closely related to the feature type is the size of the window used during feature extraction. To investigate the tradeoff between window size and the performance of the recognition system, we sweeped $W_s = 0.1, 0.25, 0.5, 1, 2, 3, 4, 5, 6, 7, 8s$.

Results. Figure 7 shows precision and recall for different window sizes W_s. Note that we used equal window size for all activities. We can see that precision reaches a maximum $W_s = 1s$ for both the person-dependent and the person-independent case. For all evaluations described in this article, we therefore fixed $W_s = 1s$. At the same time, however, increasing W_s leads to a decrease of recall. This is also visible in the experiments from the previous section. Using the raw signal (i.e., each frame instead of a window) led to higher recall at the cost of precision.

5.4. Sensor Placement

As the findings from Section 5.2 suggest, the number and type of sensors play an important role for this activity recognition problem. In this section, we therefore analyze the influence of sensor placement on the recognition performance. In a second evaluation, we then look at different types of sensors (see Section 5.5).

Results. Figure 8 shows the results for different sensor placements using both accelerometers and gyroscopes. As can be seen from the figure, results for the person-dependent case do not vary as much as might have been expected for different placements. The best result for individual placement is obtained at the hand at 87.2% precision and 55.1% recall. Combining several sensors, either in pairs or all together, allows us to increase precision beyond that of individual sensor modalities up to precision and recall of 94.1% and 62.4%, respectively. For the person-independent case, the best recognition performance is obtained by the combination of all sensors (precision: 63%, recall: 39.5%). The worst performance is obtained at the upper arm (30.2% precision and 11.4% recall). This is as expected, given that all of the investigated gestures involve hand movements and show increasing motion levels the farther down the sensors are placed on the arm.

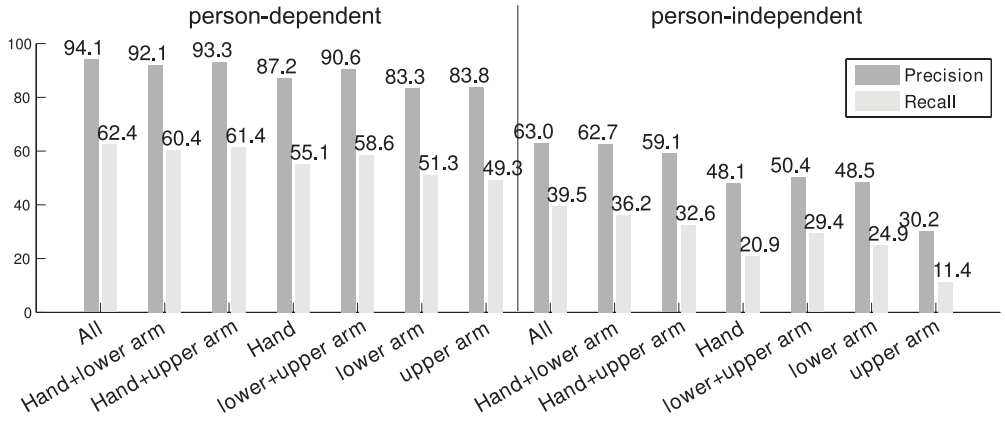

Fig. 8. Recognition performance for different sensor placements.

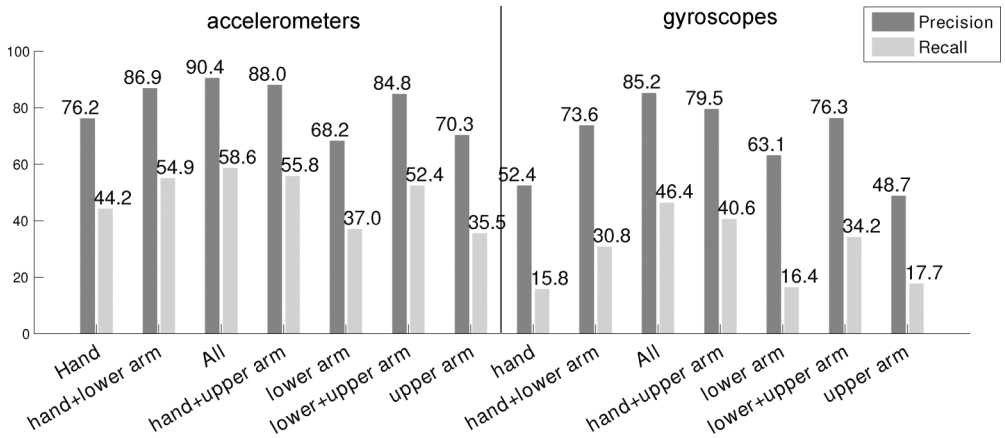

Fig. 9. Recognition performance for different sensor types (accelerometers and gyroscopes) using person-dependent evaluation.

5.5. Sensor Modality

All evaluations so far were based on both accelerometers and gyroscopes. We will now evaluate each modality separately.

Results. Given results in Figures 9 and 10, we can observe a strong influence of the sensor type at different placements. Overall classification using acceleration sensors performs significantly better than using gyroscope sensors. The best results for the person-dependent case are achieved by combining all three acceleration sensors ($p = 90.4\%$, $r = 58.6\%$). Using gyroscopes only, the best performance is a precision of 85.2% (recall 46.4%). The best performance is achieved by combining all placements. This ranking of sensor types is confirmed in the person-independent case. In this section, we analyzed sensor modalities at different placements. Interestingly, the combination led to the best results. This is not in line with the results of Section 5.4, where the best result was on par with or better than the combination with other placements. This result can be interpreted as follows. The combination of sensor modalities at the hand might preserve more information about movement of the entire arm than a single modality. In order to leverage information from both modalities, as well as from

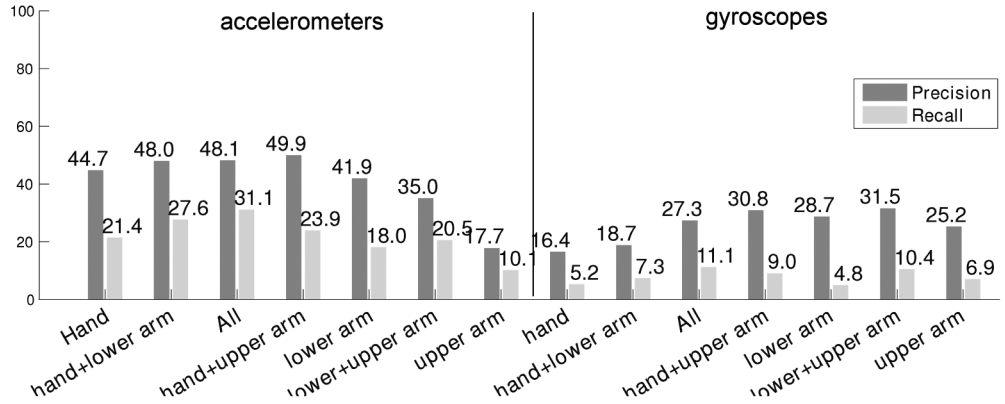

Fig. 10. Recognition performance for different sensor types (accelerometers and gyroscopes) using person-independent evaluation.

multiple placements, we face the same problem as with the feature definition. While some parts of modalities do contribute (e.g., a single axis of a modality), other parts might introduce noise. Feature selection can extract and leverage multiple modalities and placements automatically (see Section 5.7).

5.6. Classifier

So far, we have investigated different settings based on a simple k-NN classifier. Another important source of influence on the recognition performance is the classifier itself. For this reason, it is quite common in activity recognition research to evaluate and compare different classifiers for a specific recognition problem. The decision for or against a specific classifier can be made for several reasons, including but not limited to lower computational complexity or simply superior performance. In this evaluation, we investigate how recognition performance is influenced by several common classifiers used in HAR research and provide an intuition for a potential choice of classifiers. We evaluate the following classification techniques: Discriminative Analysis (DA), Naive Bayes (NB), Support Vector Machine (SVM), Hidden Markov Models (HMM), Joint Boosting (JB), and k-NN with $k = 1$. For the HMM, we used a left-right model with three states, each with a unimodal Gaussian.

Results. Figure 11 summarizes the recognition performance achieved using these different classifiers. As can be seen from the figure, the best results for all available data are achieved by SVM ($p = 96\%$, $r = 84.8\%$). The worst performance is exhibited by naive Bayes ($p = 78.2\%$, $r = 69.2\%$); k-NN suffers from lowest recall ($p = 94.1\%$, $r = 62.4\%$). In the person-independent case, results are less conclusive. The k-NN and HMM classifiers, however, lead to significantly higher precision.

5.7. Feature Selection

A large number of features may improve recognition performance but also increases computational complexity (see Section 5.2). For low-power sensors with limited processing power, a small feature set is desired. Automatic feature selection techniques can be used to reduce the feature set to the most relevant features for a given classification problem. To investigate the tradeoff between feature set size and recognition performance, we evaluated minimum-Redundancy Maximum-Relevance (mRMR) feature selection [Peng et al. 2005]. The mRMR algorithm selects a feature subset of arbitrary size S that best characterizes the statistical properties of the target classes given the

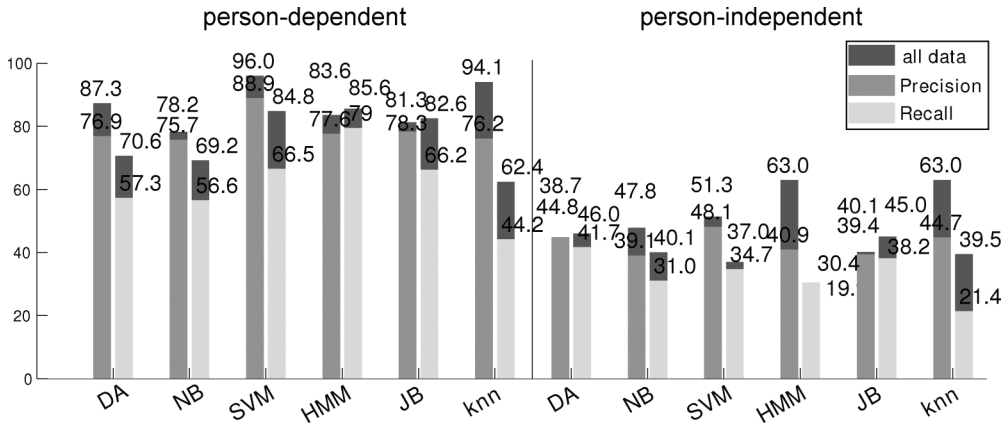

Fig. 11. Recognition performance for different classifiers using parameters as in Figure 3. Results are for using all data (all sensor placements) as well as only using the sensor attached to the hand.

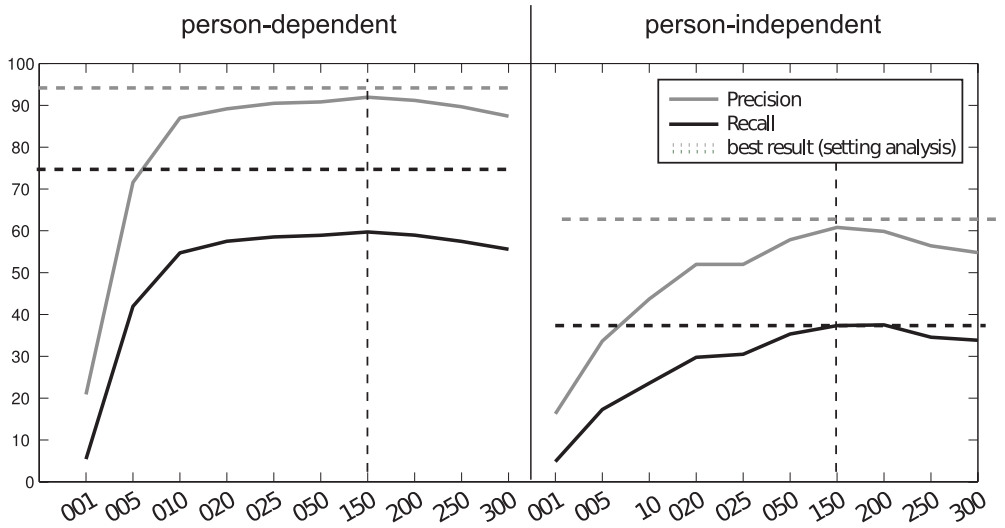

Fig. 12. Recognition performance for different subsets of the feature ranking averaged over both participants. The dashed line indicates the recognition performance only using mean and variance.

ground truth. In contrast to other methods such as the F-test, mRMR also considers relationships between features during the selection. Among the possible underlying statistical measures described in the literature, a mutual information difference was shown to yield the most promising results and was thus selected in this work.

Results. Figure 12 shows precision and recall curves for person-dependent and person-independent evaluation averaged over both participants for $S = 1, 5, 10, 20, 25, 50, 150, 200, 250, 300$. In both cases, the best recognition performance is achieved for a feature set size of $S = 150$ (person-dependent evaluation: precision: 91.9%, recall: 59.7%; person-independent: precision: 60.8%, recall: 37.3%). It is important to note that feature selection operates not only on the sensor level but also on each dimension from that sensor. This allows a far more detailed selection. It is likely that we introduced, together with highly relevant features, a few less relevant features that reduced

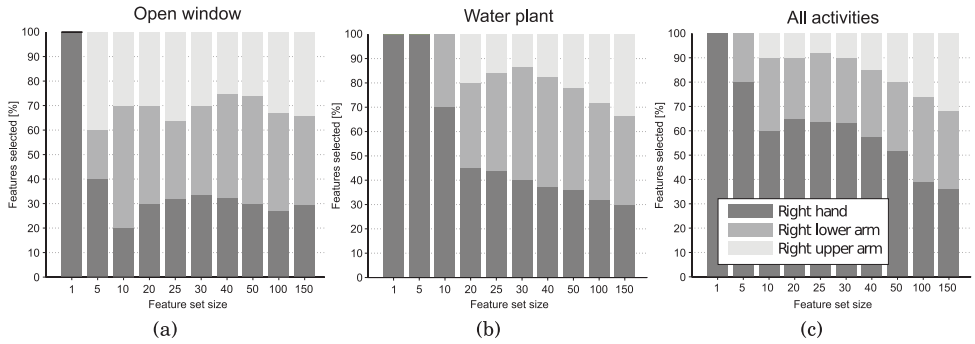

Fig. 13. Feature distribution for opening the window (a), watering the plant (b), and all activities (c) for different feature set sizes for participant 1. Each bar shows the percentage of features selected by mRMR for placements on the right hand, lower arm, and upper arm.

the overall recognition rate using the k-NN classifier. The modality can be accounted similarly. Feature selection allows us to analyze each axis of the sensor. Consequently, the fine interplay between selected features leads to better performance than using all features without preselection. The figure, however, also shows that using only the features' mean and variance yields higher recognition performance than using mRMR (precision: 94.1, recall: 76.2 for person-dependent evaluation). It is surprising that neither the larger feature set size nor a different combination of features resulted in increased performance.

We then analyzed the feature set sizes up to $S = 150$ in more detail. First, we were interested to see which sensor placements contributed most to the overall recognition performance. Figure 13 shows the feature distribution for *open window*, *water plant*, and all activities for participant 1. As can be seen from the figure, the feature distributions for the two selected activities differ quite considerably from the distribution across all activities. The top 10 features selected for *water plant* contain only features extracted from the sensor at the right hand. In contrast, for the same feature set size, *open window* is best characterized by a mixture of features derived from all three sensor placements. These distributions nicely reflect the characteristics of the activities at hand. Watering the plant mainly involves hand movements to lower the watering can toward the flower pot; opening the window requires the whole arm to reach the window handle and the hand to rotate it and swing open the sash.

Finally, we analyzed how mRMR ranked the features on each of the leave-one-repetition-out folds for the *water plant* activity. The rank of a feature is the position at which mRMR selected it within a set. The position corresponds to the importance with which mRMR assesses the feature's ability to discriminate between classes in combination with the features ranked before it. Figure 14 shows the top 15 features according to the median rank over all sets. Each vertical bar represents the spread of mRMR ranks: for each feature, there is one rank per training set. The most useful features are those found with the highest rank (close to one) for most training sets, indicated by shorter bars. Sometimes a useful feature that is ranked low by mRMR might be the one that improves a classification; for example, F7 (*ZCR_gyr_1_y*) is spread between rank four and 40 but is included in all 26 folds. This analysis confirms that the top four features (*Mean_acc_1_y*, *Mean_gyr_1_x*, *ceptrCoeff4_gyr_1_x*, and *MCR_acc_1_z*) are based on the sensors attached to the right hand, as judged by high ranks for all folds. F5 (*ceptrCoeff6_gyr_1_x*) belongs to the same placement but is ranked high for only 19 of the 26 folds. The most useful features for the other placements are F6 (*ZCR_gyr_2_y*) for the lower arm and F11 (*energy_gyr_3_y*) for the upper arm. The feature with

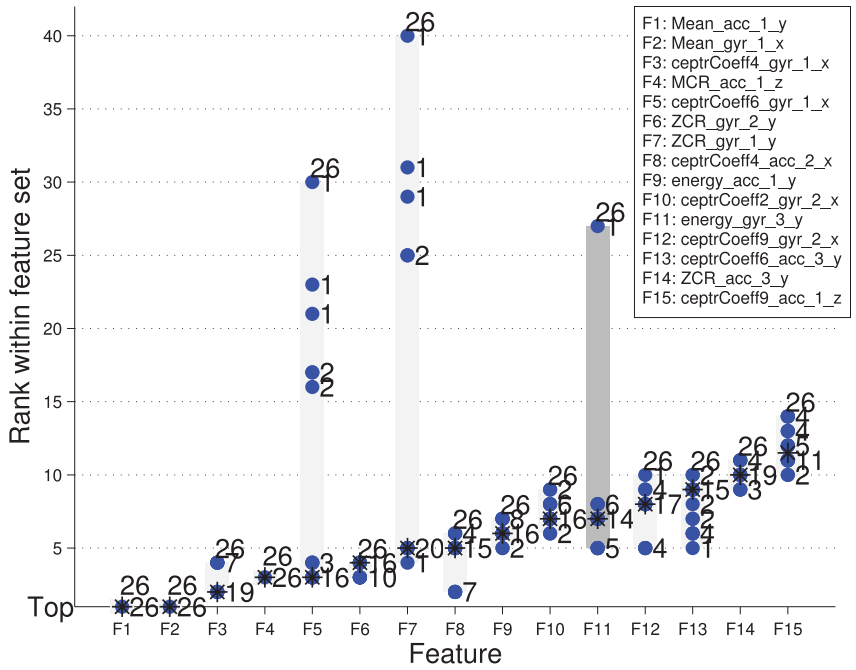

Fig. 14. The top 15 features selected by mRMR for "water plant" averaged over all folds of participant 1. The x-axis shows feature number and group; the key on the right shows the corresponding feature names; the y-axis shows the rank (top = 1). For each feature, the bars show the total number of folds for which the feature was chosen (bold number at the top), the rank of the feature within each set (dots, with a number representing the set count), and the median rank over all sets (black star). For example, a useful feature is F1—the mean acceleration signal in y direction at the hand—selected for all folds, in all of which it is ranked top; less useful is F7—the zero crossing rate of the gyroscope signal in y direction at the hand—used in all folds but only ranked between 4 and 40.

the most rank variations is F7 (*ZCR_gyr_1_y*), which is spread between ranks four and 40.

5.8. Discussion

It is challenging to evaluate an ARC according to a waterfall model (i.e., by selecting the best-performing method for a particular stage based on the methods chosen for the previous stage). This is because design decisions in different stages depend on each other and require a joint evaluation and optimization. The closest to best approach, which we also followed in this tutorial article, is to evaluate each stage separately, to identify and discuss dependencies, and to show how different design decisions impact overall recognition performance. Generally speaking, a person-dependent system achieves higher accuracy than a person-independent one (see Figure 4). In the latter case, recognition performance can be increased either with training on more data of multiple users to obtain a better generalizing model or by using more robust features (i.e., by introducing human knowledge into the process). Figure 8 shows that using information from multiple body locations achieved higher performance, as did the use of accelerometers compared to gyroscopes (see Figures 9 and 10). While accelerometers are able to capture rotation changes (through gravity) and linear motion, gyroscopes are limited to rotation. As can be seen from the same figures, for example, for the upper arm, combining both sensor types can still improve recognition performance by 14%. Finally, as could have been expected given the set of activities, the sensor at the

wrist achieved the overall best performance, which can also be seen from the ranking of features (see Figure 14).

6. CONCLUSION

This tutorial is specifically geared toward newcomers to the field of activity recognition using on-body inertial sensors. We first discussed the key research challenges that researchers in human activity recognition face. We then described in detail the activity recognition chain as a general-purpose framework for designing and evaluating activity recognition systems and provided an overview of best practice methods developed by the activity recognition research community. To illustrate an actual implementation of the framework, we concluded with the educational example problem of recognizing different hand gestures from inertial sensors. The deliberately low complexity of the example allowed us to compare different algorithms with respect to overall recognition performance, which we hope will prove helpful to newcomers also for designing more complex activity recognition systems.

REFERENCES

Gregory D. Abowd, Anind K. Dey, R. Orr, and J. Brotherton. 1998. Context-awareness in wearable and ubiquitous computing. *Virtual Reality* 3, 3 (1998), 200–211.

J. K. Aggarwal and M. S. Ryoo. 2011. Human activity analysis: A review. *Comput. Surveys* 43, 3 (2011), 16:1–16:43. DOI: http://dx.doi.org/10.1145/1922649.1922653

Barbara E. Ainsworth, William L. Haskell, Stephen D. Herrmann, Nathanael Meckes, David R. Bassett, Catrine Tudor-Locke, Jennifer L. Greer, Jesse Vezina, Melicia C. Whitt-Glover, and Arthur S. Leon. 2011. 2011 compendium of physical activities: A second update of codes and MET values. *Medicine and Science in Sports and Exercise* 43, 8 (2011), 1575–1581.

Bashar Altakouri, Gerd Kortuem, Agnes Grunerbl, Kunze Kai, and Paul Lukowicz. 2010. The benefit of activity recognition for mobile phone based nursing documentation: A Wizard-of-Oz study. In *Proceedings of ISWC*. 1–4.

Oliver Amft. 2011. Self-taught learning for activity spotting in on-body motion sensor data. In *Proceedings of ISWC* 0 (2011), 83–86. DOI: http://dx.doi.org/10.1109/ISWC.2011.37

Oliver Amft, Holger Junker, and Gerhard Tröster. 2005. Detection of eating and drinking arm gestures using inertial body-worn sensors. In *Proceedings of the IEEE International Symposium on Wearable Computing*. 160–163.

Oliver Amft, Martin Kusserow, and Gerhard Tröster. 2007. Probabilistic parsing of dietary activity events. In *Proceedings of BSN*. Springer, 242–247.

Urs Anliker, Jamie A. Ward, Paul Lukowicz, Gerhard Tröster, François Dolveck, Michel Baer, Fatou Keita, Eran B. Schenker, Fabrizio Catarsi, Luca Coluccini, Andrea Belardinelli, Dror Shklarski, Menachem Alon, Etienne Hirt, Rolf Schmid, and Milica Vuskovic. 2004. AMON: A wearable multiparameter medical monitoring and alert system. *IEEE Trans. Inf. Technol. Biomed.* 8, 4 (2004), 415–427.

Daniel Ashbrook and Thad Starner. 2003. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing* 7, 5 (2003), 275–286.

Daniel Ashbrook and Thad Starner. 2010. MAGIC: A motion gesture design tool. In *Proceedings of CHI*. 2159–2168.

Marc Bächlin, Daniel Roggen, Meir Plotnik, Noit Inbar, Inbal Meidan, Talia Herman, Marina Brozgol, Eliya Shaviv, Nir Giladi, Jeffrey M Hausdorff, and Gerhard Tröster. 2009. Potentials of enhanced context awareness in wearable assistants for Parkinson's disease patients with freezing of gait syndrome. In *Proceedings of ISWC*. 123–130.

Ling Bao and Stephen S. Intille. 2004. Activity recognition from user-annotated acceleration data. In *Proceedings of Pervasive*. 1–17.

H. Bayati, J. d. R. Millán, and R. Chavarriaga. 2011. Unsupervised adaptation to on-body sensor displacement in acceleration-based activity recognition. In *Proceedings of ISWC*.

Ulf Blanke, Robert Rehner, and Bernt Schiele. 2011. South by South-East or sitting at the desk. Can orientation be a place? In *Proceedings of ISWC*.

Ulf Blanke and Bernt Schiele. 2008. Sensing location in the Pocket. In *Adj. Proceedings of UbiComp*.

Ulf Blanke and Bernt Schiele. 2009. Daily routine recognition through activity spotting. In *Proceedings of LoCa*. 192–206.

Ulf Blanke and Bernt Schiele. 2010. Remember and transfer what you have learned—recognizing composite activities based on activity spotting. In *Proceedings of ISWC*. 1–8.

Ulf Blanke, Bernt Schiele, Matthias Kreil, Paul Lukowicz, Bernard Sick, and Thiemo Gruber. 2010. All for one or one for all? Combining heterogeneous features for activity spotting. In *Proceedings of the IEEE PerCom Workshop on Context Modeling and Reasoning*. 18–24.

M. Buettner, R. Prasad, M. Philipose, and D. Wetherall. 2009. Recognizing daily activities with RFID-based sensors. In *Proceedings of UbiComp*. 51–60.

Andreas Bulling and Daniel Roggen. 2011. Recognition of visual memory recall processes using eye movement analysis. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp'11)*. ACM, 455–464. DOI:http://dx.doi.org/10.1145/2030112.2030172

Andreas Bulling, Jamie A. Ward, and Hans Gellersen. 2012. Multimodal recognition of reading activity in transit using body-worn sensors. *ACM Transactions on Applied Perception* 9, 1 (2012), 2:1–2:21. DOI:http://dx.doi.org/10.1145/2134203.2134205

Andreas Bulling, Jamie A. Ward, Hans Gellersen, and Gerhard Tröster. 2008. Robust recognition of reading activity in transit using wearable electrooculography. In *Proceedings of the 6th International Conference on Pervasive Computing (Pervasive'08)*. Springer, 19–37. DOI:http://dx.doi.org/10.1007/978-3-540-79576-6_2

Andreas Bulling, Jamie A. Ward, Hans Gellersen, and Gerhard Tröster. 2009. Eye movement analysis for activity recognition. In *Proceedings of the 11th International Conference on Ubiquitous Computing (UbiComp'09)*. ACM, 41–50. DOI:http://dx.doi.org/10.1145/1620545.1620552

Andreas Bulling, Jamie A. Ward, Hans Gellersen, and Gerhard Tröster. 2011. Eye movement analysis for activity recognition using electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 4 (April 2011), 741–753. DOI:http://dx.doi.org/10.1109/TPAMI.2010.86

Andreas Bulling, Christian Weichel, and Hans Gellersen. 2013. EyeContext: Recognition of high-level contextual cues from human visual behaviour. In *Proceedings of the 31st SIGCHI International Conference on Human Factors in Computing Systems*. 305–308. DOI:http://dx.doi.org/10.1145/2470654.2470697

A. Cassinelli, C. Reynolds, and M. Ishikawa. 2006. Augmenting spatial awareness with haptic radar. In *Proceedings of ISWC*. 61–64.

J. Chen, K. Kwong, D. Chang, J. Luk, and R. Bajcsy. 2005. Wearable sensors for reliable fall detection. In *Proceedings of the 27th IEEE International Conference of Engineering in Medicine and Biology*. 3551–3554.

Jingyuan Cheng, Oliver Amft, and Paul Lukowicz. 2010. Active capacitive sensing: Exploring a new wearable sensing modality for activity recognition. In *Proceedings of Pervasive*. 319–336.

B. Clarkson, K. Mase, and A. Pentland. 2000. Recognizing user's context from wearable sensors: Baseline system. In *Proceedings of ISWC*. 69–76.

B. Clarkson and A. Pentland. 1999. Unsupervised clustering of ambulatory audio and video. In *Proceedings of ASSP*. 3037–3040.

I. Cohen and M. Goldszmidt. 2004. Properties and benefits of calibrated classifiers. In *Proceedings of the International Conference on Knowledge Discovery in Databases*. 125–136.

R. de Oliveira, M. Cherubini, and N. Oliver. 2010. MoviPill: Improving medication compliance for elders using a mobile persuasive social game. In *Proceedings of UbiComp*, Vol. 1001. 36.

M. Everingham and J. Winn. 2007. *The PASCAL Visual Object Classes Challenge 2007 Development Kit*. Technical Report.

Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 8 (2006), 861–874. DOI:http://dx.doi.org/DOI: 10.1016/j.patrec.2005.10.010

Davide Figo, Pedro Diniz, Diogo Ferreira, and João Cardoso. 2010. Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing* 14, 7 (2010), 645–662.

Gernot A. Fink. 2008. *Markov Models for Pattern Recognition: From Theory to Applications*. Springer.

J. Friedman, T. Hastie, and R. Tibshirani. 2000. Additive logistic regression: a statistical view of boosting. *Annals of Statistics* 28, 2 (2000), 337–407.

A. Godfrey, R. Conway, D. Meagher, and G. ÓLaighin. 2008. Direct measurement of human movement by accelerometry. *Medical Engineering and Physics* 30, 10 (2008), 1364–1386.

Eric Guenterberg, Sarah Ostadabbas, Hassan Ghasemzadeh, and Roozbeh Jafari. 2009. An automatic segmentation technique in body sensor networks based on signal energy. In *Proceedings of BAN*. 21:1–21:7. DOI:http://dx.doi.org/10.4108/ICST.BODYNETS2009.6036

Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3 (2003), 1157–1182.

Björn Hartmann, Leith Abdulla, Manas Mittal, and Scott R. Klemmer. 2007. Authoring sensor-based interactions by demonstration with direct manipulation and pattern recognition. In *Proceedings of CHI*. 145–154.

T. Ho, J. Hull, and S. Srihari. 1994. Decision combination in multiple classifier systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 16 (1994), 66–75.

T. Huynh, U. Blanke, and B. Schiele. 2007. Scalable recognition of daily activities with wearable sensors. In *Proceedings of LoCa*. 50–67.

Tâm Huynh, Mario Fritz, and Bernt Schiele. 2008. Discovery of activity patterns using topic models. In *Proceedings of UbiComp*. 10–19.

Tam Huynh and Bernt Schiele. 2005. Analyzing features for activity recognition. In *Proceedings of the Joint Conference on Smart Objects and Ambient Intelligence*. 159–163. DOI:http://dx.doi.org/10.1145/1107548.1107591

Wen-Juh Kang, Jiue-Rou Shiu, Cheng-Kung Cheng, Jin-Shin Lai, Hen-Wai Tsao, and Te-Son Kuo. 1995. The application of cepstral coefficients and maximum likelihood method in EMG pattern recognition. *IEEE Trans. on Biomedical Engineering* 42, 8 (1995), 777–785.

Ashish Kapoor and Eric Horvitz. 2008. Experience sampling for building predictive user models: A comparative study. In *Proceedings of CHI*. 657–666. DOI:http://dx.doi.org/10.1145/1357054.1357159

S. Katz, T. D. Downs, H. R. Cash, and R. C. Grotz. 1970. Progress in development of the index of ADL. *The Gerontologist* 10, 1 Part 1 (1970), 20.

J. Kittler et al. 1998. On combining classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 20, 3 (1998), 226–239.

Ron Kohavi and George H. John. 1997. Wrappers for feature subset selection. *Artificial Intelligence* 97, 1–2 (1997), 273–324.

Matthias Kranz, Andreas Möller, Nils Hammerla, Stefan Diewald, Thomas Plötz, Patrick Olivier, and Luis Roalter. 2013. The mobile fitness coach: Towards individualized skill assessment using personalized mobile devices. *Pervasive and Mobile Computing* 9, 2 (2013), 203–215.

J. Krumm and E. Horvitz. 2006. Predestination: Inferring destinations from partial trajectories. In *Proceedings of UbiComp*. 243–260.

K. Kunze, M. Barry, E.A. Heinz, P. Lukowicz, D. Majoe, and J. Gutknecht. 2006. Towards recognizing tai chi—An initial experiment using wearable sensors. *Proceedings of FAWC* (2006), 1–6.

Kai Kunze and Paul Lukowicz. 2008. Dealing with sensor displacement in motion-based onbody activity recognition systems. In *Proceedings of UbiComp*. 20–29.

K. Kunze, P. Lukowicz, H. Junker, and G. Tröster. 2005. Where am I: Recognizing on-body positions of wearable sensors. In *Proceedings of the International Workshop on Location and Context-Awareness*. 257–268.

Cassim Ladha, Nils Hammerla, Patrick Olivier, and Thomas Plötz. 2013. ClimbAX: Skill assessment for climbing enthusiasts. In *Proceedings of the Int. Conf. Ubiquitous Comp. (UbiComp)*. to appear.

C. Lee and Y. Xu. 1996. Online, interactive learning of gestures for human/robot interfaces. In *Proceedings of the IEEE International Conference on Robotics and Automation*. 2982–2987.

S. W. Lee and K. Mase. 2002. Activity and location recognition using wearable sensors. *IEEE Pervasive Computing* 1, 3 (2002), 24–32.

Jonathan Lester, Tanzeem Choudhury, and Gaetano Borriello. 2006. A practical approach to recognizing physical activities. In *Proceedings of the International Conference on Pervasive Computing*. 1–16.

Jonathan Lester, Tanzeem Choudhury, Nicky Kern, Gaetano Borriello, and Blake Hannaford. 2005. A hybrid discriminative/generative approach for modeling human activities. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. 766–772.

Lin Liao, Dieter Fox, and Henry Kautz. 2005. Location-based activity recognition using relational Markov networks. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. 773–778.

Charles X. Ling, Jin Huang, and Harry Zhang. 2003. AUC: A statistically consistent and more discriminating measure than accuracy. In *Proceedings of the 18th International Conference on Artificial Intelligence*. 329–341.

B. Logan, J. Healey, M. Philipose, E.M. Tapia, and S. Intille. 2007. A long-term evaluation of sensing modalities for activity recognition. In *Proceedings of UbiComp*. Springer-Verlag, 483–500.

Hong Lu, Wei Pan, Nicholas D. Lane, Tanzeem Choudhury, and Andrew T. Campbell. 2009. SoundSense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services*. 165–178. DOI:http://dx.doi.org/10.1145/1555816.1555834

Hong Lu, Jun Yang, Zhigang Liu, Nicholas D. Lane, Tanzeem Choudhury, and Andrew T. Campbell. 2010. The Jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of ENSS*. ACM, 71–84.

P. Lukowicz, F. Hanser, C. Szubski, and W. Schobersberger. 2006. Detecting and interpreting muscle activity with wearable force sensors. In *Proceedings of Pervasive*. 101–116.

C. Mattmann, O. Amft, H. Harms, G. Tröster, and F. Clemens. 2007. Recognizing upper body postures using textile strain sensors. In *Proceedings of ISWC*. 29–36.

U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher. 2006. Activity recognition and monitoring using multiple sensors on different body positions. In *Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks*. 113–116.

I. Maurtua, P. T. Kirisci, T. Stiefmeier, M. L. Sbodio, and H. Witt. 2007. A wearable computing prototype for supporting training activities in automative production. In *Proceedings of the 4th International Forum on Applied Wearable Computing*. 1–12.

David Minnen, Thad Starner, Irfan Essa, and Charles Isbell. 2006a. Discovering characteristic sctions from on-body sensor data. In *Proceedings of the 10th IEEE International Symposium on Wearable Computers (ISWC)*.

D. Minnen, T. Westeyn, T. Starner, J. Ward, and P. Lukowicz. 2006b. Performance metrics and evaluation issues for continuous activity recognition. In *Performance Metrics for Intelligent Systems*.

Sushmita Mitra and Tinku Acharya. 2007. Gesture recognition: A survey. *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 37, 3 (2007), 311–324.

S. J. Morris and J. Paradiso. 2002. Shoe-integrated sensor system for wireless gait analysis and real-time feedback. In *Proceedings of the 2nd Joint IEEE EMBS and BMES Conference*. 2468–2469.

G. Ogris, T. Stiefmeier, P. Lukowicz, and G. Tröster. 2008. Using a complex multi-modal on-body sensor system for activity spotting. In *Proceedings of ISWC*. 55–62.

N. Oliver and F. Flores-Mangas. 2007. HealthGear: Automatic sleep apnea detection and monitoring with a mobile phone. *Journal of Communications* 2, 2 (2007), 1–9.

K. Partridge and P. Golle. 2008. On using existing time-use study data for ubiquitous computing applications. In *Proceedings of UbiComp*. ACM, 144–153.

D. Patterson, D. Fox, H. Kautz, and M. Philipose. 2005. Fine-grained activity recognition by aggregating abstract object usage. In *Proceedings of ISWC*. 44–51.

Donald Patterson and Mohan Singh. 2010. Involuntary gesture recognition for predicting cerebral palsy in high-risk infants. In *Proceedings of ISWC*.

Hanchuan Peng, Fuhui Long, and C. Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 27, 8 (2005), 1226–1238.

A. S. Pentland. 2004. Healthwear: Medical technology becomes wearable. *Computer* (2004), 42–49.

M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Fox, H. Kautz, and D. Hahnel. 2004. Inferring activities from interactions with objects. *IEEE Pervasive Computing* (2004), 50–57.

G. Pirkl, K. Stockinger, K. Kunze, and P. Lukowicz. 2008. Adapting magnetic resonant coupling based relative positioning technology for wearable activity recognition. In *Proceedings of ISWC*. 47–54.

Thomas Plötz, Nils Y Hammerla, and Patrick Olivier. 2011. Feature learning for activity recognition in ubiquitous computing. In *Proceedings of the 22nd international Joint Conference on Artificial Intelligence*. AAAI Press, 1729–1734.

Thomas Plötz, Nils Y. Hammerla, Agata Rozga, Andrea Reavis, Nathan Call, and Gregory D. Abowd. 2012. Automatic assessment of problem behavior in individuals with developmental disabilities. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 391–400. DOI:http://dx.doi.org/10.1145/2370216.2370276

R. Polikar. 2006. Ensemble Based Systems in Decision Making. *IEEE Circuits and Systems Magazine* 6, 3 (2006), 21–45.

L. R. Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of IEEE* 77, 2 (1989), 257–285.

C. Randell and H. Muller. 2000. Context awareness by analysing accelerometer data. In *Proceedings of ISWC*. 175–176.

Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L. Littman. 2005. Activity recognition from accelerometer data. In *Proceedings of the 17th International Conference on Innovative Applications of Artificial Intelligence*. 1541–1546.

D. Roggen, M. Baechlin, J. Schumm, T. Holleczek, C. Lombriser, G. Tröster, L. Widmer, D. Majoe, and J. Gutknecht. 2010. An educational and research kit for activity and context recognition from on-body sensors. In *Proceedings of BSN*. 277–282.

D. Roggen, K. Förster, A. Calatroni, T. Holleczek, Yu Fang, G. Tröster, P. Lukowicz, G. Pirkl, D. Bannach, K. Kunze, A. Ferscha, C. Holzmann, A. Riener, R. Chavarriaga, and J. del R. Millan. 2009. OPPORTUNITY: Towards opportunistic activity and context recognition systems. In *Proceedings of Wowmom*. 1–6.

Daniel Roggen, Kilian Förster, Alberto Calatroni, and Gerhard Tröster. 2013. The adARC pattern analysis architecture for adaptive human activity recognition systems. *Journal of Ambient Intelligence and Humanized Computing* 4, 2 (2013), 169–186. DOI:http://dx.doi.org/10.1007/s12652-011-0064-0

G. Schindler, C. Metzger, and T. Starner. 2006. A wearable interface for topological mapping and localization in indoor environments. *Proceedings of LoCa* (2006), 64–73.

Petr Somol, Jana Novovičová, and Pavel Pudil. 2006. Flexible-Hybrid Sequential Floating Search in Statistical Feature Selection. In *Structural, Syntactic, and Statistical Pattern Recognition*. 632–639.

T. Starner, J. Weaver, and A. Pentland. 1997. A wearable computer-based American sign language recogniser. *Personal and Ubiquitous Computing* 1, 4 (1997), 241–250.

Thomas Stiefmeier, Daniel Roggen, Georg Ogris, Paul Lukowicz, and Gerhard Tröster. 2008. Wearable activity tracking in car manufacturing. *IEEE Pervasive Computing* 7, 2 (2008), 42–50.

Thomas Stiefmeier, Daniel Roggen, and Gerhard Tröster. 2007. Gestures are strings: efficient online gesture spotting and classification using string matching. In *Proceedings of the 2nd International Conference on Body Area Networks*. 1–8.

M. Stikic, T. Huynh, K. van Laerhoven, and B. Schiele. 2008. ADL recognition based on the combination of RFID and accelerometer sensing. In *Proceedings of PervasiveHealth*. 258–263.

Maja Stikic, Diane Larlus, Sandra Ebert, and Bernt Schiele. 2011. Weakly supervised recognition of daily life activities with wearable sensors. *IEEE Trans. on Pattern Analysis and Machine Intelligence* (2011).

J. Sung, C. Ponce, B. Selman, and A. Saxena. 2011. Human activity detection from RGBD images. In *Proceedings of the AAAI Workshop on Plan, Activity, and Intent Recognition*.

M. Sung, C. Marci, and A. Pentland. 2005. Wearable feedback systems for rehabilitation. *Journal of Neuro-Engineering and Rehabilitation* 2, 1 (2005).

E. Munguia Tapia, S. S. Intille, and K. Larson. 2004. Activity recognition in the home using simple and ubiquitous sensors. In *Proceedings of PERVASIVE*. 158–175.

Bernd Tessendorf, Andreas Bulling, Daniel Roggen, Thomas Stiefmeier, Manuela Feilner, Peter Derleth, and Gerhard Tröster. 2011a. Recognition of hearing needs from body and eye movements to improve hearing instruments. In *Proceedings of the 9th International Conference on Pervasive Computing*. Springer, 314–331. DOI:http://dx.doi.org/10.1007/978-3-642-21726-5_20

Bernd Tessendorf, Franz Gravenhorst, Bert Arnrich, and Gerhard Tröster. 2011b. An IMU-based sensor network to continuously monitor rowing technique on the water. In *Proceedings of ISSNIP*.

Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2007. Sharing visual features for multiclass and multiview object detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 29, 5 (2007), 854–869.

P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. 2008. Machine recognition of human activities: A survey. *IEEE Trans. on Circuits and Systems for Video Technology* 18, 11 (2008), 1473–1488.

T. L. M. van Kasteren, G. Englebienne, and B. J. A. Kröse. 2010. Transferring knowledge of activity recognition across sensor networks. *Proceedings of the Pervasive* (2010), 283–300.

T. van Kasteren, A. Noulas, G. Englebienne, and B. Kröse. 2008. Accurate activity recognition in a home setting. In *Proceedings of UbiComp*. 1–9.

Kristof Van Laerhoven and Eugen Berlin. 2009. When else did this happen? Efficient subsequence representation and matching for wearable activity data. In *Proceedings of ISWC*. IEEE Press, 69–77.

K. van Laerhoven and O. Cakmakci. 2000. What shall we teach our pants. In *Proceedings of ISWC*. 77–83.

Kristof van Laerhoven, David Kilian, and Bernt Schiele. 2008. Using rhythm awareness in long-term activity recognition. In *Proceedings of ISWC*. 63–68.

K. van Laerhoven, A. Schmidt, and H.-W. Gellersen. 2002. Multi-sensor context aware clothing. In *Proceedings of ISWC*. 49–56.

Eduardo Velloso, Andreas Bulling, and Hans Gellersen. 2013a. MotionMA: Motion modelling and analysis by demonstration. In *Proceedings of the 31st SIGCHI International Conference on Human Factors in Computing Systems*. 1309–1318. DOI:http://dx.doi.org/10.1145/2470654.2466171

Eduardo Velloso, Andreas Bulling, Hans Gellersen, Wallace Ugulino, and Hugo Fuks. 2013b. Qualitative activity recognition of weight lifting exercises. In *Proceedings of the 4th Augmented Human International Conference (AugmentedHuman 2013)*. 116–123. DOI: http://dx.doi.org/10.1145/2459236.2459256

D. Wan. 1999. Magic medicine cabinet: A situated portal for consumer healthcare. In *Handheld and Ubiquitous Computing*. Springer, 352–355.

S. Wang, W. Pentney, A.M. Popescu, T. Choudhury, and M. Philipose. 2007. Common sense based joint training of human activity recognizers. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. 2237–2242.

Jamie A. Ward, Paul Lukowicz, and Hans W. Gellersen. 2011. Performance metrics for activity recognition. *ACM Trans. on Intelligent Systems and Technology* 2, 1 (2011), 6:1–6:23. DOI: http://dx.doi.org/10.1145/1889681.1889687

Jamie A. Ward, Paul Lukowicz, Gerhard Tröster, and Thad E. Starner. 2006. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 28, 10 (2006), 1553–1567.

Tracy Westeyn, Kristin Vadas, Xuehai Bian, Thad Starner, and Gregory D. Abowd. 2005. Recognizing mimicked autistic self-stimulatory behaviors using HMMs. In *Proceedings of ISWC*. 164–169.

Andrew D. Wilson and Aaron F. Bobick. 2000. Realtime online adaptive gesture recognition. In *Proceedings of the 15th International Conference on Pattern Recognition*. 270–275. DOI: http://dx.doi.org/10.1109/ICPR.2000.905317

C. Wren, Y. Ivanov, I. Kaur, D. Leigh, and J. Westhues. 2007. Socialmotion: Measuring the hidden social life of a building. *Proceedings of LoCa* (2007), 85–102.

Zhixian Yan, Vigneshwaran Subbaraju, Dipanjan Chakraborty, Archan Misra, and Karl Aberer. 2012. Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach. In *Proceedings of ISWC*. IEEE, 17–24.

P. Zappi, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, and Tröster. 2007. Activity recognition from on-body sensors by classifier fusion: Sensor scalability and robustness. In *Proceedings of ISSNIP*. 281–286.

Mi Zhang and Alexander A. Sawchuk. 2012. Motion primitive-based human activity recognition using a bag-of-features approach. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. ACM, 631–640.

V. W. Zheng, D. H. Hu, and Q. Yang. 2009. Cross-domain activity recognition. In *Proceedings of UbiComp*. 61–70.

Andreas Zinnen, Ulf Blanke, and Bernt Schiele. 2009a. An analysis of sensor-oriented vs. model-based activity recognition. In *Proceedings of ISWC*.

Andreas Zinnen, Christian Wojek, and Bernt Schiele. 2009b. Multi activity recognition based on body model-derived primitives. In *Proceedings of LoCa*. DOI: http://dx.doi.org/10.1007/978-3-642-01721-6_1

Received October 2011; revised April 2013; accepted June 2013