UNIVERSIDAD CARLOS III DE MADRID

Departamento de Teoría de la Señal y Comunicaciones

DOCTORAL THESIS

MARKOV MODELLING ON HUMAN ACTIVITY RECOGNITION

Author: ALFREDO NAZÁBAL RENTERÍA

Supervised by: ANTONIO ARTÉS RODRÍGUEZ

July 18, 2017

Tesis Doctoral: MARKOV MODELLING
ON HUMAN ACTIVITY RECOGNITION

Autor: Alfredo Nazábal Rentería

Director: D. Antonio Artés Rodríguez

Fecha:

Tribunal

Presidente:

Vocal:

Secretario:

A mis padres, Alfredo y Pilar,
porque son ellos quienes realmente
han hecho posible esta Tesis.

Agradecimientos

To be done...

Abstract

Human Activity Recognition (HAR) is a research topic with a relevant interest in the machine learning community. Understanding the activities that a person is performing and the context where they perform them has a huge importance in multiple applications, including medical research, security or patient monitoring. The improvement of the smart-phones and inertial sensors technologies has lead to the implementation of activity recognition systems based on these devices, either by themselves or combining their information with other sensors. Since humans perform their daily activities sequentially in a specific order, there exist some temporal information in the physical activities that characterize the different human behaviour patterns. However, the most popular approach in HAR is to assume that the data is conditionally independent, segmenting the data in different windows and extracting the most relevant features from each segment.

In this thesis we employ the temporal information explicitly, where the raw data provided by the wearable sensors is fed to the training models. Thus, we study how to perform a Markov modelling implementation of a long-term monitoring HAR system with wearable sensors, and we address the existing open problems arising while processing and training the data, combining different sensors and performing the long-term monitoring with battery powered devices.

Employing directly the signals from the sensors to perform the recognition can lead to problems due to misplacements of the sensors on the body. We propose an orientation correction algorithm based on quaternions to process the signals and find a common frame reference for all of them independently on the position of the sensors or their orientation. This algorithm allows for a better activity recognition when feed to the classification algorithm when compared with similar approaches, and the quaternion transformations allow for a faster implementation.

One of the most popular algorithms to model time series data are Hidden Markov Models (HMMs) and the training of the parameters of the model is performed using the Baum-Welch algorithm. However, this algorithm converges to

local maxima and the multiple initializations needed to avoid them makes it computationally expensive for large datasets. We propose employing the theory of spectral learning to develop a discriminative HMM that avoids the problems of the Baum-Welch algorithm, outperforming it in both complexity and computational cost.

When we implement a HAR system with several sensors, we need to consider how to perform the combination of the information provided by them. Data fusion can be performed either at signal level or at classification level. When performed at classification level, the usual approach is to combine the decisions of multiple classifiers on the body to obtain the performed activities. However, in the simple case with two classifiers, which can be a practical implementation of a HAR system, the combination reduces to selecting the most discriminative sensor, and no performance improvement is obtained against the single sensor implementation. In this thesis, we propose to employ the soft-outputs of the classifiers in the combination and we develop a method that considers the Markovian structure of the ground truth to capture the dynamics of the activities. We will show that this method improves the recognition of the activities with respect to other combination methods and with respect to the signal fusion case.

Finally, in long-term monitoring HAR systems with wearable sensors we need to address the energy efficiency problem that is inherent to battery powered devices. The most common approach to improve the energy efficiency of such devices is to reduce the amount of data acquired by the wearable sensors. In that sense, we introduce a general framework for the energy efficiency of a system with multiple sensors under several energy restrictions. We propose a sensing strategy to optimize the temporal data acquisition based on computing the uncertainty of the activities given the data and adapt the acquisition actively. Furthermore, we develop a sensor selection algorithm based on Bayesian Experimental Design to obtain the best configuration of sensors that performs the activity recognition accurately, allowing for a further improvement on the energy efficiency by limiting the number of sensors employed in the acquisition.

Resumen

El reconocimiento de actividades humanas (HAR) es un tema de investigación con una gran relevancia para la comunidad de aprendizaje máquina. Comprender las actividades que una persona está realizando y el contexto en el que las realiza es de gran importancia en multitud de aplicaciones, entre las que se incluyen investigación médica, seguridad o monitorización de pacientes. La mejora en los smart-phones y en las tecnologías de sensores inerciales han dado lugar a la implementación de sistemas de reconocimiento de actividades basado en dichos dispositivos, ya sea por si mismos o combinándolos con otro tipo de sensores. Ya que los seres humanos realizan sus actividades diarias de manera secuencial en un orden específico, existe una cierta información temporal en las actividades físicas que caracterizan los diferentes patrones de comportamiento, Sin embargo, los algoritmos más comunes asumen que los datos son condicionalmente independientes, segmentandolos en diferentes ventanas y extrayendo las características más relevantes de cada segmento.

En esta tesis utilizamos la información temporal de manera explícita, usando los datos crudos de los sensores como entrada de los modelos de entrenamiento. Por ello, analizamos como implementar modelos Markovianos para el reconocimiento de actividades en monitorizaciones de larga duración con sensores *wearable*, y tratamos los problemas existentes al procesar y entrenar los datos, al combinar diferentes sensores y al realizar adquisiciones de larga duración con dispositivos alimentados por baterías.

Emplear directamente las señales de los sensores paa realizar el reconocimiento de actividades puede dar lugar a problemas debido a la incorrecta colocación de los sensores en el cuerpo. Proponemos un algoritmo de corrección de la orientación basado en quaterniones para procesar las señales y encontrar un marco de referencia común independiente de la posición de los sensores y su orientación. Este algoritmo permite obtener un mejor reconocimiento de actividades al emplearlo en conjunto con un algoritmo de clasificación, cuando se compara con modelos

similares. Además, la transformación de la orientación basada en quaterniones da lugar a una implementación más rápida.

Uno de los algoritmos más populares para modelar series temporales son los modelos ocultos de Markov, donde los parámetros del modelo se entrenan usando el algoritmo de Baum-Welch. Sin embargo, este algoritmo converge en general a máximos locales, y las múltiples inicializaciones que se necesitan en su implementación lo convierten en un algoritmo de gran carga computacional cuando se emplea con bases de datos de un volumen considerable. Proponemos emplear la teoría de aprendizaje espectral para desarrollar un HMM discriminativo que evita los problemas del algoritmo de Baum-Welch, superándolo tanto en complejidad como en coste computacional.

Cuando se implementa un sistema de reconocimiento de actividades con multiples sensores, necesitamos considerar cómo realizar la combinación de la información que proporcionan. La fusión de los datos, se puede realizar tanto a nivel de señal como a nivel de clasificación. Cuando se realiza a nivel de clasificación, lo normal es combinar las decisiones de multiples clasificadores colocados en el cuerpo para obtener las actividades que se están realizando. Sin embargo, en un caso simple donde únicamente se emplean dos sensores, qué podría ser una implentación habitual de un sistema de reconocimiento de actividades, la combinación se reduce a seleccionar el sensor más discriminativo, y no se obtiene mejora con respecto a emplear un único sensor. En esta tesis proponemos emplear salidas blandas de los clasificadores para la combinación, desarrollando un modelo que considera la estructura Markoviana de los datos reales para capturar la dinámica de las actividades. Mostraremos como este método mejora el reconocimiento de actividades con respecto a otros métodos de combinación de clasificadores y con respecto a la fusión de los datos a nivel de señal.

Por último, abordamos el problema de la eficiencia energética de dispositivos alimentados por baterías en sistemas de reconocimiento de actividades de larga duración. La aproximación más habitual para mejorar la eficiencia energética consiste en reducir el volumen de dats que adquieren los sensores. En ese sentido,

introducimos un marco general para tratar el problema de la eficiencia energética en un sistema con multiples sensores bajo ciertas restricciones de energía. Proponemos una estrategia de adquisición activa para optimizar el sistema temporal de recogida de datos, basándonos en la incertidumbre de las actividades dados los datos que conocemos. Además, desarrollamos un algoritmo de selección de sensores basado diseño experimental Bayesiano y así obtener la mejor configuración para realizar el reconocimiento de actividades limitando el número de sensores empleados y al mismo tiempo reduciendo su consumo energético.

Contents

List of Acronyms

A2RMS	Double Adaptive Rejection Metropolis Sampling
AAR	Acceleration Angular Rate
AQ	Acceleration Quaternion
CNN	Convolutional Neural Network
CRF	Conditional Random Field
DL	Deep Learning
ECG	Electrocardiogram
EM	Expectation Maximization
FB	Forward-Backward
FFBS	Forward-Filtering Backward-Sampling
FHMM	Factorial Hidden Markov Model
GMM	Gaussian Mixture Model
GPS	Global Positioning System
HAR	Human Activity Recognition
HDM	Hierarchical Dynamic Model

HHMM	Hierarchical Hidden Markov Model
HMM	Hidden Markov Model
IBCC	Independent Bayesian Classifier Combination
IFHMM	Infinite Factorial Hidden Markov Model
IHHMM	Infinite Hierarchical Hidden Markov Model
IHMM	Infinite Hidden Markov Model
IMU	Inertial Measurement Unit
kNN	k-Nearest Neighbours
LDA	Linear Discriminant Analysis
MAP	*Maximum a Posteriori*
MARG	Magnetic, Angular Rate and Gravity
MC	Monte Carlo
MSCC	Markov Soft-output Classifier Combination
PCA	Principal Component Analysis
PPC	Posterior Probability Combination
RNN	Recurrent Neural Network
SCC	Soft-output Classifier Combination
SVD	Singular Value Decomposition
SVM	Support Vector Machine

1 Introduction

1.1 Human Activity Recognition

Human behavior analysis is a popular interdisciplinary research topic. Understanding the reasons behind human actions, or in another perspective what are people doing and why, has an increasing interest in a wide range of research fields, from medical research to marketing and finance. The main problem of human behavior analysis can be summarized with a simple question: What constitutes human behavior? It seems that there exist different possible answer to this question, since experts from different fields have their own definitions of human behavior [15, 116]. The activities that the people perform during the day, their social interactions or their habits are integral aspects of the essence of human behavior.

We can define the activities of the daily life as the basic unit that describes the structure of the different human behavior patterns. Indeed, the sequence of

activities that a person performs during a day, week or month essentially define their behavior. The nature of human activities constitutes a complex topic in itself. We can distinguish between the different activities attending to different criteria. For example, some of the activities are static in a physical sense, like sitting or lying while others are dynamic, like walking. Some activities are considered simple, like moving an arm or a leg, while others are complex, since they involve the combinations of different simple activities, e.g., brushing your teeth or sweeping the floor. Furthermore, people can perform more than one activity at the same time, adding even more possibilities to the different activity patterns. Consequently, the automatic recognition of physical activities is an arduous problem gaining a considerable attention from multiple research fields, including medical, security or military [20, 61].

Modern approaches in HAR date from the late nineties [31, 36]. The main idea behind the activity recognition is to provide information about the behavior of the users, allowing the implementation of computing systems that helps them with their daily life tasks [3]. For example, in ambulatory monitoring of elderly patients, knowing the activities that the patients are performing is vital to understand the context in which the patients are being monitored. This context awareness can help to overcome the limitations associated with the use of self-reporting in medical assessment, consequently improving the patients quality of life by reducing the frequency of visits to medical centres and reducing medical costs [11]. Another example is the employment of actigraphy for the treatment of chronic insomnia [102]. The authors showed that actigraphy data were more accurate than sleep-diary data when compared to polysomnography and it should be used as a complement to sleep-diary evaluation.

The inherent complexity of the daily life activities affect directly to the definition of the HAR problem. There is no specific set of activities that is considered universal in the implementation of a HAR system, leading to a large amount of possible design and implementation options. As a simple example, we will compare the differences in the implementation of a HAR system employed to monitor

the progress of injured athletes during rehabilitation and a HAR system employed during the monitoring of elderly patients in their daily life. In the first case, the activities that needs to be analyzed are dynamic activities, like walking and running, to find differences with common walking or running patterns. In the second case, the number of activities to be considered can be as small or large as we want, increasing the complexity of the system with the number of activities while at the same time improving its ability to characterize the daily behavior patterns of the patients. Furthermore, while in the first case the number of activities is fixed in advance, in the second case we can generalize the system for the appearance of unexpected new activities that only appear in real life. Another aspect that conditions the implementation of a HAR systen is the context of the monitoring. In the rehabilitation system, we want to monitor the activity patterns of an athlete under a controlled environment. In such systems, it is common to employ a large number of devices to track the patients, from video recordings, to accelerometers or Electrocardiogram (ECG). The number of devices and its nature it is not restricted, since the athletes only use these devices during the rehabilitation sessions. However, in the case of an ambulatory monitoring system, the number of possibilities is much more restricted, since the HAR system needs to be employed under naturalistic conditions, and privacy and comfortability of the patients must be addressed. Ultimately, the definition of the HAR problem is completely dependent on the application, and we refer to the review of HAR presented in [20] for an extensive description of the different HAR problem statements.

In human behavior analysis, there exist the necessity of huge amounts of data to perform the HAR and ultimately detect the different behavior patterns. Data recordings range from days to months or years, and long-term monitoring systems are essential under these conditions. The acquisition devices that can be employed in the implementation of a long-term monitoring system depend completely on the definition of the HAR problem. The number of available devices is huge and is increasing with the new advances in the sensing technologies [113].

One of such sensing technologies are infrastructure sensors. These devices are

fixed in different positions in a closed environment, and the recognition of the activities depend completely on the interaction of the users with such devices. The most popular infrastructure system and one of the first technologies employed in HAR is the use of video recordings [5, 86]. In such systems, different people are recorded performing several physical activities and image processing algorithms [4] are implemented in the core of the HAR system to perform different tasks, including automatic recognition of the physical activities [44] or detection of abnormal activities [47]. One of the first applications with a huge interest in HAR employing video cameras were video surveillance systems [79]. Such systems use real-world video data to automatically distinguish normal behaviors from suspicious ones in different indoor and outdoor scenarios, like parking lots or bank branches. Other examples include medical applications, like the life-logging activity recognition system for elderly care in smart home environments detailed in [52]. Even though the amount of information retrieved by such systems is huge, privacy, installation cost and restricted sensing area limit the implementation of such systems.

In addition to video recordings, another popular infrastructure technologies are environment sensors in home controlled scenarios [54]. These devices are attached to several objects in a house and detect the interaction of the users with them. It is the main approach in the implementation of smart home houses, where privacy issues limit the use of video recordings. The data recorded by these sensors are binary signals that indicate whether the object is being used or not. For example, when the sensors are attached to the doors inside the house, they can identify the location of the user. Combining this information with other sensors they can restrict the set of possible activities that the users are performing. One of the main problems of such systems is the identification of the best positions in the house to attach the sensors and the number of sensors to be employed. Depending on the set of activities to be recognized the design can vary significantly. Furthermore, maintenance and installation costs of such systems are huge, and they are not useful for ambulatory purposes, since the monitoring is restricted to the home environment.

Though infrastructure sensors are shown to be an interesting design choice for the implementation of HAR sytems, they lack the ability to analyze the activities of the daily life of different people under naturalistic conditions during long periods of time. The devices to be employed must be able to perform long-term monitoring independently of the places where the people are located (home, job, street, etc.), guaranteeing their privacy. With the development of the wireless protocols like Bluetooth Smart (also known as Bluetooth Low Energy) [27, 28] wireless wearable sensors are increasingly becoming an attractive alternative in the implementation of HAR systems [74]. Their unobtrusiveness and handiness make them appropriate for applications which require long-term continuous monitoring [17]. The first devices employed were 3-axis accelerometers attached to the body, where they record the acceleration of the person wearing these sensors to recognize the different activities. Later, with the advancement of the sensing technologies, the number of sensors included in the wearable sensors increased. One of such technologies are Inertial Measurement Units (IMUs) [6]. These devices consist on a 3-axis accelerometer and a 3-axis gyroscope enabling the measuring of acceleration and angular velocity, respectively. In addition, Magnetic, Angular Rate and Gravity (MARG) sensors extend the capabilities of IMUs by integrating a 3-axis magnetometer. More sophisticated devices include smart-phones, whose sensing power and capabilities are increasing with the advance of the technology [98], and make them and interesting alternative in the implementation of wearable sensors HAR systems. A smart-phone acquires data from the inertial sensors while additionally including information from many other sources, like the location of the patient (GPS) [80], the measurement of the air pressure to detect elevation (barometer) [63] the accessibility of the patients to wireless connections or even their personal environment from message applications and social media [115].

However, the implementation of HAR systems with wearable sensors present several limitations. First of all, the location of the sensors on the body directly affects to the recognition of the activities. In a single sensor system, the position of the sensor on the body greatly limits the recognition performance of the differ-

ent activities [13, 10]. Depending on the activities considered, the optimal position varies significantly. For example, while performing dynamic activities like walking, running or cycling, a sensor placed on one of the ankles provides the most information to distinguish them. However, this sensor is not optimal to discriminate between static activities, like standing, sitting or lying, which rely in the position of the body to identify them. Furthermore, under naturalistic conditions where the people are going to manipulate the sensors, we must expect misplacements on the position of the sensors. We need to find some mechanisms to correct the orientation of the sensors automatically, since the position of the sensors can vary significantly. As an example, a smart-phone is not a device that is attached to the body, and the data provided by the inertial sensors changes dramatically depending on the location. The location of the sensors and the presence of misplacements are still open problems in the implementation of HAR systems.

The energy efficiency of wearable sensors and smart-phones is another important limitation [84], and it is directly related with the long-term monitoring property. HAR systems implemented with a smart-phone need to share the battery resources between all of the applications, but the embedded sensors are one of the main sources of battery consumption. Furthermore, acquiring synchronous data from all the inertial sensors at high sampling frequencies, decreases dramatically the battery of the devices. Existing approaches to improve the energy efficiency include the work in [65], where the authors study the effect of the sampling frequency of the mobile phones in a HAR system. Previous work in the topic claims that the sampling frequency needed to recognize physical activities should be no less than 20Hz [19]. However, the authors demonstrate in [65] that an activity recognition based on low sampling frequencies is feasible for long-term activity monitoring. Furthermore, in relation with the sensor position problem, not all the devices are informative while recognizing some of the activities, and most of the data acquired is either useless or redundant. In that respect, the authors in [109] develop a HAR system where they dynamically adapt the sampling frequency and the classification features employed on the system depending on the

performed activities. They select a priori the parameters needed in terms of the activities and change the working operation of the mobile phones accordingly while tracking continuously the ongoing activities. Both approaches consider a continuous monitoring, adapting the sampling frequency of the sensors depending on the specifications. However, sampling frequency reduction based methods can be still inefficient with some long-term activities, e.g. sleeping, where data acquisition can be halted completely without losing information. Though all of the previous works present interesting alternatives, energy efficiency of wearable sensors remains an open problem in the literature.

A HAR system can be implemented with a single sensor or with the combination of multiple sensors. Although the implementation of a single sensor system is feasible for a simple recognition problem, employing multiple sensors increases the recognition performance in general [85]. For example, the detection of epilepsy seizures can be improved substantially combining the information provided by a continuous video/EEG monitoring system with the data provided by a wearable sensor located at the wrist [67]. In a simple system, several IMUs can be attached to different positions on the body and additional sensors can be attached as well. These sensors could present different sampling frequencies and synchronisation across the different sensors becomes a technical problem. When combining the raw data from different sensors directly, the main technique consist of transforming the potentially non-synchronous data from the sensors in a multivariate time series matrix D

$$D = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_T]$$

where $\mathbf{d}_t \in \mathbb{R}^N$ is the N-dimensional data point at time instant t. This processing step only depend on the devices employed and is independent of the people being monitored. However, the differences in sampling frequencies and the presence of missing data during the acquisition makes the combination of the data at signal level problematic in many cases. How to perform the data fusion of several sensors is a main topic in the implementation of HAR systems, since nowadays most of them employ several sensors jointly. An alternative to the combination at

signal level consists of performing the combination at classifier level, i.e., after a classification of the different activities has been already performed. This method will be explained later in more detail.

Modeling the activities in wearable sensor based systems is not a simple task. In this setting, the activities are just physical signals provided by the sensors, changing constantly depending on the activity. For example, a ECG measures the heart rate of the people monitored. It is expected that when people are running the heart rate increases, returning to a neutral state when people are relaxed. With an accelerometer, during static activities, the gravity component will provide information about the relative position of the sensor, and thus, of the position of the people's body. In a long-term monitoring, due to the nature of the human behavior, it is expected that there exist a temporal relation between the activities that dictates the dynamics of the physical signals. One of the main approaches while modeling wearable sensors based systems is to consider that the activity sequences can be divided in several segments, each of them with a unique label or activity. In such cases, the temporal relation between the activities is lost, and the data can be considered conditionally independent. Another approach consists of benefiting from the temporal dynamics of the physical activities, employing the raw signals provided by the sensors. I this case, there exists an activity label for every time instant, and the activities that a person is performing at one time instant depend on the activities that they were performing the previous instant. This property makes Markov modelling an interesting approach for the HAR setting with wearable sensors.

Modelling of HAR when the data is conditionally independent usually follows the same structure. First, the data provided by the sensors is grouped in reduced information units called windows or segments, which contain relevant information about the activities. Each window is defined by an initial t_i and ending t_e time instant in the time series, $\mathbf{w}_i = (t_i, t_e)$. This data segmentation process obtains a set of windows W from the data, each of them assigned to a label or activity

$$W = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_M]$$

The basic approach to obtain these windows consists of fixing a window of a specific size and divide all the data in fixed length windows, assigning a single label to every window. The size of the window is a parameter of the system and the performance of the HAR system is directly related to this window size [49]. However, a fixed sized window does not consider the structure of the time series data. Other schemes consider dynamic window sizes, where the size of the window is not fixed in advance, and a probabilistic approach is employed to derive the window size for every sensor event, depending mainly on the information contained in the signals [58]. Another approach is the implementation of a sliding window algorithm, where the size of the window can be fixed or adjusted dynamically and the data is grouped in each window sequentially, with the data being overlapped in consecutive windows [60]. The amount of overlapping constitutes a trade-off between the precision of the segmentation and the computational cost of the system [49].

When the data is segmented, we need to extract relevant information or features from this segments, allowing the system to improve the recognition of the activities. This feature extraction step involves the application of several transformations on every data segment $\mathbf{w}_i$ to obtain a vector of F features $\mathbf{x}_i \in \mathbb{R}^F$ that are relevant to discriminate between the different activities

$$\mathbf{x}_i = f(D, \mathbf{w}_i).$$

Each vector of features constitutes a data point, and a label is assigned to every data point. Feature extraction allows the HAR system to reduce the length of the data sequences by grouping every data window in vectors containing all the relevant information of the raw signals inside these windows. There are multiple types of features that have been demonstrated to work well during the implementation of HAR systems in the literature [90, 13], though the most frequent ones are time domain features, which include the mean, standard deviation, variance, covariance, entropy and kurtosis among others, and frequency domain features, including energy, Fast Fourier Transform, Discrete Fourier Transform and many others [61].

Depending on the features employed, feature extraction can be computation-

ally expensive. Furthermore, the more highly dimensional is the feature space the more data is needed in training to obtain the parameters and the more computationally expensive becomes the classification. Working with high dimensionality data is usually complicated. There are two common approaches in HAR systems: feature reduction and feature selection. Feature reduction or dimensionality reduction algorithms involve obtaining a transformation of the data that reduces the number of features needed to separate the data in different groups, with each group corresponding to a single label. Some common approaches include Principal Component Analysis (PCA) [7] and Linear Discriminant Analysis (LDA) [23]. PCA performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. However, PCA does not take into account the different labels. On the contrary, LDA finds a linear combination of features that characterizes or separates two or more labels. LDA explicitly attempts to model the difference between the labels in the data. In feature selection approaches, no transformation is made on the data, and a reduced subset of the features extracted is selected instead. Feature selection method are usually divided in filter methods and wrapper methods and we refer to [106] for more details.

Once the data has been collected and processed, we need to discover automatically patterns in the data corresponding to the different activities. Research in machine learning and computational statistics has developed many different algorithms to explain the contents of the data in applications like biology [99], marketing [25] or geology [70] among other research fields. In general, models in machine learning can be divided in supervised and unsupervised [16]. In supervised learning, we try to infer the parameters of a function given a labelled dataset. On the contrary, in unsupervised learning we try to infer a function that describes hidden structures in unlabelled data. In the context of HAR, supervised learning is the most common approach, since the pattern discovery process is performed employing some example instances, or training set, to find the parameters that characterize the different classification algorithms and evaluate their performance

in a different data set, the test set [61].

More specifically, given a labelled training dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{T}$, with T pairs of processed data points $\mathbf{x}_i$ and its label or activity y_i, we need to infer the parameters defining a model $\mathcal{M}$ such as we minimize the classification error on $\mathcal{D}$. After the model $\mathcal{M}$ is trained, every data point in the test set is mapped to the set of possible labels Y and we compute the classification performance of the algorithm. One of the most common approaches is to compute the probability of every possible activity given the data and the model $p(y|\mathbf{x}_i, \mathcal{M})$ and assign the estimated label y_i^* as the one that maximizes these probabilities

$$y_i^* = \arg\max_{y \in Y} p(y|\mathbf{x}_i, \mathcal{M}).$$

Depending on the correspondence between the ground truth labels of the test set and the estimated labels, we can determine the performance of the algorithm. The most common performance metrics employed in classification algorithms with multiple classes are precision and recall, though other metrics are also possible. We refer to [20] for a more extensive review in the topic.

In the context of conditionally independent data points, multiple classification algorithms have been employed for the recognition of the activities in the literature. Some of these approaches include decision trees [82], Support Vector Machines (SVMs) [8], logistic regression [59] or k-nearest neighbours [57] among others. We refer to these HAR reviews [20, 61] for a more extensive description of the different classification approaches. A recent research line in HAR that is becoming popular is employing Deep Learning (DL) for the implementation of the activity recognition system. The first approaches were proposed for video recording HAR systems [12], where Convolutional Neural Networks (CNNs) are employed to extract the relevant information from the recordings and automatically classify different sets of activities. Additionally, in the last years DL is being employed extensively in HAR systems with wearable sensors [81]. The main approach of these systems is to perform the feature extraction and feature selection steps automatically employing CNNs and to model the temporal structure of the data employing Recurrent Neural Networks (RNNs).

Once the classification of the different activities is obtained, in HAR systems implemented with several wearable sensors, it is possible to perform a combination of the classification results. Data fusion can be implemented at signal or feature level and at classification level [20], as previously stated. Since combination at signal level imposes high bandwidth and synchronization requirements in wireless transmission, an estimation of the performed activity can be transmitted instead, leading to a decision fusion configuration (also named as classifiers combination). Existing approaches to the classifier combination problem include the seminal work of Dawid and Skene [26]. The authors propose a model where each classifier is characterized by a confusion matrix and they use the Expectation Maximization (EM) algorithm to estimate the most likely values of both the ground truth and the parameters governing the behaviour of each classifier. Another common approach for the combination of classifiers are ensemble methods [112]. Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone [92]. Classifiers based on ensembles are more computationally expensive, since they require several models to be trained and evaluated.

In general, the classifier combination problem consists of finding a mapping between the classification results from every single classifier and the final estimated activity

$$y_t^* = g(y_t^1, y_t^2, \cdots, y_t^K)$$

where y_t^k is the estimated activity of the classifier k at time instant t, and $g(\cdot)$ is the mapping function. This general approach allows for the employment of different classifiers for each sensor, as only the combination of the estimated activities is needed. Furthermore, decision fusion offers the advantage of robustness against sensor failures, since only a single available sensor is needed to have an estimation of the performed activity. However, HAR is a setting that employs a reduced number of wearable sensors, and in such cases, the combination reduces to always selecting the best sensor.

In general, the conditionally independent approach performs a mapping be-

tween the data provided by the sensors D and the set of features X that are later feed to the classification algorithms. One of the main drawbacks of this structure is that the feature extraction and selection on the data can be computationally expensive depending on the features selected and the number of features. Although these computations can be performed offline, when online HAR systems are considered, this could become problematic. Furthermore, humans commonly perform their daily activities sequentially in a specific order, existing some temporal relationship between the physical activities. For example, when a person is sitting in a sofa, the probability that this person suddenly starts running is much less than lying down. A different approach to the HAR problem is employing Markov models to consider the temporal dynamics of the activities. Employing the signals provided by the inertial sensors, without performing any data segmentation or feature extraction processing, the computational cost of the data processing becomes negligible, allowing for online implementations of HAR systems. Additionally, the raw physical signals maintain the Markovian structure, since the sampling frequency of the wearable sensors typically ranges between tens and hundred of Hertz. However, the raw signals are particularly sensitive to the placement of the sensors on the body in terms of location and orientation. There exist several techniques in the literature dealing with the sensor orientation problem [29, 110]. In particular, in [35] the authors transform the accelerometer and gyroscope data of the sensors to a virtual reference system where all the raw signals are invariant to sensor orientation. However, this work only deals with the orientation of a sensor on the waist, and is not extended to other locations. A computationally efficient algorithm that corrects the orientation of the sensors independently of their position is to be desired in this approach.

The most common temporal probabilistic models employed in HAR systems are HMMs [55] and conditional random fields [101]. These methods have been employed extensively in combination with the feature extraction and selection methods described before for the recognition of human activities, but they are more suitable when the data retains the temporal structure of activities. In particular,

a HMM is described by a sequence of observations $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$ of length N and a sequence of unobserved hidden states $S = [s_1, s_2, \cdots, s_N]$ explaining the data. A first order Markov process models the hidden states, and the observations are conditionally independent given the states (Figure 5.1).

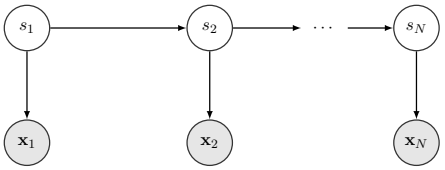

Figure 1.1: Graphical model of a HMM.

A HMM is characterized by three parameters $\mathcal{M} = (\boldsymbol{\pi}, \Psi, \boldsymbol{\theta})$, the initial probabilities distribution $\boldsymbol{\pi}$, with $\pi_i = p(s_1 = i)$, representing the probability distribution of the first state of the sequence, the transition probabilities matrix Ψ, with $\Psi_{ij} = p(s_{n+1} = j | s_n = i)$, which describes the probability of transitioning between the different states of the chain, and the parameters $\boldsymbol{\theta}$ of the conditional likelihood of the observations given the states of the HMM. Depending on the data employed in the system, the observation model can be a discrete model, e.g., Bernoulli distribution, or a continuous probability distribution, e.g., Gaussian distribution.

Inference over a HMM to obtain the sequence of states S that explains the data X is performed using the Forward-Backward (FB) algorithms. To compute the parameters of the model $\mathcal{M} = (\boldsymbol{\pi}, \Psi, \boldsymbol{\theta})$ an EM algorithm is employed. In particular, in the literature of HMMs this EM algorithm is known as the Baum-Welch algorithm [89]. For a more extensive review on HMMs we refer to [89].

However, the Baum-Welch algorithm exhibits two main problems: 1) the likelihood of the observations $p(X|S)$ is in general multi-modal so the EM is guaranteed to converge only to a local maxima, and 2) although the complexity of the algorithm growths linearly with the length of the training sequence, the multiple initializations required for minimizing the effects of the local convergence and the more than quadratic growth with the number of hidden states makes EM computationally expensive with large training datasets. Other possible inference methods for HMMs include Gibbs sampling [91] and variational inference [68],

but are even more computationally expensive and global convergence is still not guaranteed. Additionally, HMMs are unsupervised learning algorithms, meaning that the information of the activities is not considered while performing the inference. A mapping between the hidden states and the different activities is needed to perform classification.

HMMs are not the only existing model that preserves the temporal dynamics of the data. As stated before, Conditional Random Field (CRF) is another common machine learning approach employed while performing structure prediction. The main difference with HMMs is that CRF is an undirected graphical model whose nodes can be divided into the observed and output variables, respectively. Furthermore, exact inference in CRF is intractable in general, and the approximate solutions available are computationally expensive when employed with long duration data sequences. Other alternatives in the same family of HMMs include the Hierarchical Hidden Markov Models (HHMMs) [32], which consider a hierarchical structure on the hidden states and the Factorial Hidden Markov Models (FHMMs) [39], which consider several parallel hidden chains. Furthermore, non-parametric versions of all this algorithms also exist in the literature, the Infinite Hidden Markov Models (IHMMs) [38], the Infinite Factorial Hidden Markov Model (IFHMM) [38] and the Infinite Hierarchical Hidden Markov Model (IHHMM) [43]. In non-parametric algorithms the number of hidden states is not set in advance, and though theoretically it can increase infinitely, in practice, only a finite number of states appear during the inference. However, again, the inference in these methods is computationally prohibitive in long-term monitoring scenarios, where the activity sequences of one day can perfectly surpass one million of observations.

Markov modelling can be employed in during the classifier combination algorithms. There are many approaches to the classifier combination problem that consider the Markovian structure of the data. In [111], the authors train a discriminative HMM classifier for each individual sensor, and then a Naive Bayes classifier fuses these individual classification results. A more elaborated approach to classifier combination in a general framework, that can be seen as a Bayesian ex-

tension of [26], is the Independent Bayesian Classifier Combination (IBCC) method proposed in [56]. In particular, the authors develop models for dependent and independent classifiers that simultaneously infer both the ground truth and the model parameters from the individual classifier outputs. However, in the extreme case of using only two sensors, these methods are equivalent to selecting the most discriminative sensor and, therefore, the combination does not yield better precision than the best single classifier. Furthermore, in [96] the authors derive a variational inference algorithm for the IBCC and, more interestingly, a time-varying classifier combination model. While all of these methods consider hard-output classifiers, i.e., the output of the classifier is the activity (running, sitting, etc.) an alternative that needs to be explores is employing soft-output classifiers, where the outputs are the probability of performing every possible activity defined in the HAR problem instead of the actual activity.

1.2 Contributions and organization

In this thesis we are addressing the main limitations in the Markov modelling implementation of a HAR system with wearable sensors, detailed in the previous section.

In Chapter 2 we consider the problem of the wearable sensors orientation and presence of misplacements. We present an orientation correction method using quaternions that estimates the orientation of a person with respect to the Earth frame, whose results are summarized in [30]. This estimation is performed automatically without any extra information about where the sensor is placed on the body of the person, only detecting the walking patterns to estimate the orientation of the sensor with respect to the person. Furthermore, we show that this method is robust to displacements of the sensor with respect to the body. This novel data processing technique is used to feed a classification algorithm showing excellent results that outperform those obtained by an existing state-of-the-art feature extraction techniques.

In Chapter 3 we address the limitations of the Baum-Welch algorithm while

training HMMs with large training databases. To overcome its convergence problems and reduce the computational cost related with the multiple initializations required, we extend the spectral learning method of HMMs developed in [46] to train a discriminative HMM and we detail the algorithm in [77]. The resulting method provides the posterior probabilities of the activities without explicitly determining the parameters of the HMM, being able to deal with missing labels and avoiding the optimization problems of the Baum-Welch algorithms. We apply and evaluate this method in two different settings: a wearable inertial sensors setting, and a wireless binary sensor network implemented in a home environment setting. This method outperform the standard Baum-Welch algorithm in both complexity and computational cost.

The classifier combination problem is treated in Chapter 4. We propose new Bayesian models to combine the outputs of a small number of sensors in a HAR setting and we perform the inference in [78]. The models are based on a soft outputs combination of individual classifiers, more appropriate for a small number of sensors setting, instead of the most common approach of hard outputs for the combination. We also incorporate the dynamic nature of human activities as a first order homogeneous Markov chain. We develop both inductive and transductive inference methods for each model to be employed in supervised and semi-supervised situations, respectively. Using different real HAR databases, we compare our classifiers combination models against a single classifier that employs all the signals from the sensors and to previous classifier combination models. Additionally, we will demonstrate how our models exhibit an increase of robustness against sensor failures.

In Chapter 5 we study the optimization of the energy efficiency of wearable sensors in a HAR system. We present an active sensing strategy to maximize the performance of the system while increasing at the same time the efficiency of the devices, whose main results are described in [76]. Under this strategy the sensors decide when they must acquire data samples based on the entropy of the posterior probability distribution of the activities. Furthermore, we develop a

general framework for the sensor configuration employed when a HAR system is implemented combining several sensors. The sensor selection problem is treated as a Bayesian Experimental Design problem, where the maximization of the mutual information between the posterior of the activities and the observation model of the sensors results in the optimal sensor selection strategy. We evaluate this framework in a publicly available HAR database, demonstrating that employing an optimal data acquisition strategy allows the system to increase significantly the energy efficiency of the devices, consequently extending the duration of the monitoring, while maintaining the performance of the HAR system.

To conclude this thesis, in Chapter 6 we summarize the conclusions extracted from our contributions, partially published in [30, 77, 78, 76], and propose possible future lines that can further improve HAR systems.

2

Data Processing

2.1 Introduction

Inertial based sensory systems (see [62] for a review) are the most popular approach for the implementation of a Human Activity Recognition (HAR) system with wearable sensors. A basic Inertial Measurement Unit (IMU) consists of a 3-axis accelerometer and a 3-axis gyroscope, enabling the measuring of acceleration and angular velocity respectively. A Magnetic, Angular Rate and Gravity (MARG) sensor is an extended IMU that also integrates a 3-axis magnetometer. Most of the work dealing with data processing in a HAR system is based in the extraction of features from the data provided by the wearable sensors. The basic approach for the data processing in such systems consists of dividing the data acquired from these sensors in different segments or windows and obtaining a feature vector that groups all the information of the windows. Common feature extrac-

tion schemes include time domain features and frequency domain features among others.

A different approach is considered in this chapter, where the raw signals provided by the wearable sensors are employed directly to perform the activity recognition. Deleting the feature extraction and selection steps from the HAR system makes the wearable sensors more energy efficient and computationally efficient and allows performing the data processing in the devices. The only transformation applied to the raw signals is an orientation algorithm that returns the raw signals from the sensors in the same reference system.

We focus on the processing of the signals acquired by a MARG sensor, since it is the one that provides more information in indoor scenarios (i.e., without Global Positioning System (GPS) signal). A MARG sensor provides the measurements referenced to the sensor frame. However, these raw signals are sensitive to the placement of the sensor on the body of the person, in terms of position and orientation. Most of the classification algorithms for HAR proposed in the literature are fed with raw or mildly processed signals [62, 87]. Few of them try to extract the orientation of the sensor or the person in order to feed the classification algorithms [35]. In this chapter, we propose to use as inputs of the classification algorithms the orientation of the person w.r.t. the earth frame, and the acceleration in the person frame. To that end, a novel scheme of data processing for HAR systems is presented, including an efficient algorithm based on quaternion representation that computes the orientation of the person from the measurements of the MARG sensor.

2.1.1 Coordinate Systems and Notation

The orientation methods developed in this chapter are described in terms of three different three-dimensional frames that define the orientation of the person w.r.t. the earth. First, the sensor frame (S) is defined along the orthonormal axes of the physical devices, $\{{}^S\mathbf{x},{}^S\mathbf{y},{}^S\mathbf{z}\}$. The recorded signals are referred to this frame. Secondly, the earth frame (E) is defined by the orthonormal set of vectors

$\{^E\mathbf{x}, ^E\mathbf{y}, ^E\mathbf{z}\}$={North, West, Up}. Finally, we describe the person frame (P), defined by an orthonormal set of vectors whose directions when the person is standing are aligned as $\{^P\mathbf{x}, ^P\mathbf{y}, ^P\mathbf{z}\}$={Forward, Left, Up}.

We use a notation system of leading superscripts and subscripts to describe relative frame orientations and vector representations adopted from [69]. A leading subscript denotes the frame being described, and a leading superscript denotes the frame this is with reference to. For example, ${}^A_B\hat{\mathbf{q}}$ describes the orientation of frame B relative to frame A while ${}^A\mathbf{v}$ represents a vector described in frame A.

Throughout the rest of this chapter we use quaternions to represent three-dimensional orientations and rotations. Quaternions retain several advantages compared to Euler rotation matrices: they do not suffer from problematic singularities such as gimbal lock [94], and they are more compact, computationally efficient, and numerically stable.

Quaternions constitute a four-dimensional space over the real numbers. They are composed by the real axis and three imaginary orthogonal axes. Here we list some relevant quaternion properties, where the $\otimes$ operator denotes the Hamilton product, and $\hat{\mathbf{v}}$ denotes the normalised vector $\mathbf{v}$:

1. A rotation through an angle of α around a unit vector $\hat{\mathbf{u}}$ is represented by the unit quaternion

$$\hat{\mathbf{q}} = \cos\left(\frac{\alpha}{2}\right) + \sin\left(\frac{\alpha}{2}\right)\left(u_x\mathbf{i} + u_y\mathbf{j} + u_z\mathbf{k}\right). \tag{2.1}$$

2. Two rotation quaternions $\hat{\mathbf{q}}_1$ and $\hat{\mathbf{q}}_2$ can be combined into one equivalent quaternion, $\hat{\mathbf{q}} = \hat{\mathbf{q}}_2 \otimes \hat{\mathbf{q}}_1$ that represents a rotation given by $\hat{\mathbf{q}}_1$ followed by a rotation given by $\hat{\mathbf{q}}_2$.[1]

3. For any unit quaternion $\hat{\mathbf{q}}$, its inverse is equal to its conjugate $\hat{\mathbf{q}}^{-\mathbf{1}} = \hat{\mathbf{q}}^*$.

4. If a quaternion $\hat{\mathbf{q}}$ represents a rotation, and $\mathbf{v}$ is a three-dimensional vector, the rotated vector $\mathbf{v}'$ can be computed as $\mathbf{p}' = \hat{\mathbf{q}} \otimes \mathbf{p} \otimes \hat{\mathbf{q}}^*$, where $\mathbf{p} = p_x\mathbf{i} + p_y\mathbf{j} + p_z\mathbf{k}$ and $\mathbf{p}' = p'_x\mathbf{i} + p'_y\mathbf{j} + p'_z\mathbf{k}$.

[1] Note that quaternion multiplication is not commutative

2.2 Acceleration Angular Rate method

In [35] the authors propose a data processing method for a HAR system based entirely on raw data from the sensors instead of traditional feature extraction methods. The inertial sensors they used were equipped with 3-axis accelerometers and 3-axis gyroscopes. The accelerometers measure in m/s^2 the total inertial force applied on the sensor. The gyroscopes measure the angular velocity of the sensors in rad/s.

They compute the frame transformation between the sensor frame and a virtual frame F, where the gravitational component of the force of acceleration is constant. To do this, they compute the angle of roll θ_x (angle between the ${}^S y$-axis and the ${}^F xy$ plane) and angle of pitch (angle between the ${}^S x$-axis and the ${}^F xy$ plane) using the acceleration component of each sensor:

$$\begin{aligned} \theta_x &= \arctan\left(\frac{{}^S a_y}{{}^S a_z}\right) \\ \theta_y &= \arcsin\left(-\frac{{}^S a_x}{||{}^S \mathbf{a}||}\right) \end{aligned}$$

They compute the roll and pitch when the person is in standing position, and compute the rotation matrix that is needed to apply to transform the signals from the sensor frame to the person frame:

$$R(\theta_x, \theta_y) = \begin{pmatrix} \cos(\theta_y) & \sin(\theta_x)\sin(\theta_y) & \cos(\theta_x)\sin(\theta_y) \\ 0 & \cos(\theta_x) & -\sin(\theta_x) \\ -\sin(\theta_y) & \sin(\theta_x)\cos(\theta_y) & \cos(\theta_x)\cos(\theta_y) \end{pmatrix}$$

Using the previous matrix, the measured signals are all transformed by the same constant rotation at each time instant, such that it appears that all measurements have been recorded from the virtual frame F. As an example, the transformation of the acceleration at time instant t is:

$${}^P \mathbf{a}_t = R(\theta_x, \theta_y) \cdot {}^S \mathbf{a}_t$$

2.3 Acceleration Quaternion method

The proposed data processing scheme processes the magnetic, angular rate, and accelerometer signals provided by the MARG sensors in order to excerpt

1. the orientation of the person w.r.t. the earth frame, and
2. the acceleration in the person frame, ${}^{P}\mathbf{a}$.

In contrast to the Acceleration Angular Rate (AAR) method, we consider that angular rate measurements provided by the gyroscopes are not valuable signals any longer for the classification algorithms, since their information is incorporated to the orientation of the person.

Therefore, the main goal consists of computing ${}^{P}_{E}\hat{\mathbf{q}}$, i.e., the orientation of the earth frame (E) relative to the person frame (P). The proposed algorithm uses the quaternion property 2., decomposing the estimation of ${}^{P}_{E}\hat{\mathbf{q}}$ as a concatenation of the estimation of the orientation of ${}^{E}\mathbf{z}$ w.r.t. to ${}^{P}\mathbf{z}$, ${}^{P}_{E}\hat{\mathbf{q}}_{\mathbf{z}}$, followed by the estimation of the orientation of the plane ${}^{E}xy$ w.r.t. the plane ${}^{P}xy$, ${}^{P}_{E}\hat{\mathbf{q}}_{\mathbf{xy}}$, i.e.,

$$ {}^{P}_{E}\hat{\mathbf{q}} = {}^{P}_{E}\hat{\mathbf{q}}_{\mathbf{xy}} \otimes {}^{P}_{E}\hat{\mathbf{q}}_{\mathbf{z}}, \tag{2.2} $$

where ${}^{P}_{E}\hat{\mathbf{q}}_{\mathbf{z}}$ is also decomposed as

$$ {}^{P}_{E}\hat{\mathbf{q}}_{\mathbf{z}} = {}^{S}_{E}\hat{\mathbf{q}} \otimes {}^{P}_{S}\hat{\mathbf{q}}_{\mathbf{z}}. \tag{2.3} $$

The orientation of the earth frame relative to the sensor frame, ${}^{S}_{E}\hat{\mathbf{q}}$, is computed by means of the gradient descent algorithm proposed in [69]. This algorithm has shown an accurate performance close to a Kalman-based algorithm [93], while remaining computationally very efficient. The algorithm updates the current orientation via integration of the provided angular rate, and corrects the gyroscope drift with accelerometer and magnetometer measurements. This correction is driven by a parameter, β, that represents the correction rate of the gyroscope drift (see [69] for more details). The authors prove that, if the sampling rate is large enough, the algorithm performs accurately just computing one gradient descent iteration

per sample, which implies a very low computational cost. The convergence of the algorithm can be tuned by increasing the parameter β (see Section 2.4 for more details).

The second term of equation (2.3), ${}^{P}_{S}\hat{\mathbf{q}}_{\mathbf{z}}$, corresponds to the orientation of the ${}^{S}\mathbf{z}$ axis w.r.t. the ${}^{P}\mathbf{z}$ axis. Note that, if the sensor is strongly attached to the body of the person, this orientation should remain constant. Nevertheless, considering that the sensor is fixed to the clothes (for instance bounded by a belt at the waist), ${}^{P}_{S}\hat{\mathbf{q}}_{\mathbf{z}}$ may suffer from small variations. Although knowing ${}^{S}_{E}\hat{\mathbf{q}}$ during the standing position would be enough to find this orientation, with unlabelled data it is not possible to determine a priori when the person is standing. Nonetheless, walking sequences are easier to detect automatically, and while walking, the person is in average also upright; i.e., the ${}^{P}\mathbf{z}$ axis is aligned to the ${}^{E}\mathbf{z}$ axis in average. For this purpose, we have used a walking detection algorithm similar to that proposed in [35]. Therefore, ${}^{S}_{P}\hat{\mathbf{q}}_{\mathbf{z}}$ can be computed by averaging ${}^{S}_{E}\hat{\mathbf{q}}$ during the walking period. Due to quaternion property 3., we obtain the second term of equation (2.3) as ${}^{P}_{S}\hat{\mathbf{q}}_{\mathbf{z}} = {}^{S}_{P}\hat{\mathbf{q}}^{*}_{\mathbf{z}}$. Note that although there exist several ways to average a quaternion [71], we use an unweighted mean of ${}^{S}_{E}\hat{\mathbf{q}}$ during the walking period since it provides good results while being computationally efficient. In this way, ${}^{P}_{S}\hat{\mathbf{q}}_{\mathbf{z}}$ is updated every time a walking period is detected.

Finally, we compute the first term of equation (2.2), ${}^{P}_{E}\hat{\mathbf{q}}_{\mathbf{xy}}$, by estimating the direction of the velocity in ${}^{E}xy$ plane when the person is walking. For that purpose, we integrate the acceleration in the earth frame to get the velocity [18], we remove the velocity drift [105], and we compute the angle γ of the projection of the velocity vector onto the ${}^{E}xy$ plane w.r.t. ${}^{E}\mathbf{x}$. Let ϕ be the angle between ${}^{E}\mathbf{x}$ and the projection of the vector ${}^{P}_{E}\hat{\mathbf{q}}_{\mathbf{z}} \otimes \mathbf{i}$ onto the ${}^{E}xy$ plane. Then, defining $\theta = \gamma - \phi$, and according to quaternion properties 3. and 4., ${}^{P}_{E}\hat{\mathbf{q}}_{\mathbf{xy}} = \cos(\theta/2) + \mathrm{sen}(\theta/2)\mathbf{k}$.

Algorithm 1 summarises the process to compute ${}^{P}_{E}\hat{\mathbf{q}}[n]$, the orientation of the earth frame w.r.t. the person frame. The calculation is performed for the N available samples of magnetic field, angular rate, and acceleration measurements acquired by the MARG sensor. Note that β, the key parameter of the sensor

orientation algorithm [69] must be selected at the beginning, and it plays a key role in the performance of the classification algorithm, as it can be seen in Section 2.4.

Algorithm 1 Pseudocode of person orientation algorithm

Select β

for $n = 1 : N$ **do**

Compute ${}^{S}_{E}\hat{\mathbf{q}}[n]$ with the algorithm of [69] and β

Detect whether the person is *walking*

if *walking* **then**

Update ${}^{P}_{S}\hat{\mathbf{q}}_{\mathbf{z}}[n]$

Update ${}^{P}_{E}\hat{\mathbf{q}}_{\mathbf{xy}}[n]$

else

${}^{P}_{S}\hat{\mathbf{q}}_{\mathbf{z}}[n] = {}^{P}_{S}\hat{\mathbf{q}}_{\mathbf{z}}[n-1]$

${}^{P}_{E}\hat{\mathbf{q}}_{\mathbf{xy}}[n] = {}^{P}_{E}\hat{\mathbf{q}}_{\mathbf{xy}}[n-1]$

end if

${}^{P}_{E}\hat{\mathbf{q}}[n] = {}^{P}_{E}\hat{\mathbf{q}}_{\mathbf{xy}}[n] \otimes {}^{S}_{E}\hat{\mathbf{q}}[n] \otimes {}^{P}_{S}\hat{\mathbf{q}}_{\mathbf{z}}[n]$

end for

2.4 Experiments

2.4.1 Experimental Setting

The evaluation of the proposed method is performed using real data acquired by APDM OPAL miniature sensors [50]. These sensors provide 3-axis acceleration, 3-axis gyroscope, and 3-axis magnetometer data. 18 data sequences have been collected, each one from a different person. A single sensor has been placed at the waist of each subject, and they have been asked to perform some of activities in no particular order. These sequences are combinations of five different activities: running, walking, standing, sitting, and lying. This data acquisition procedure has provided us with 6 hours and 21 minutes of real data samples acquired at a

sampling rate of 128 Hz.

In order to randomize the testing process, we have built 25 sets of sequences. For each set, we have randomly selected 12 sequences for training from the database, and the 6 ones left have been used for testing. The 25 sets have been used to test all feature extraction algorithms, in order to maintain the consistency. The data have been processed both with the Acceleration Quaternion (AQ) method presented in this chapter (using different values of β) and with the AAR method.

For sake of simplicity and a fair comparison in terms of computational complexity, we have not made use of ${}^{P}_{E}\hat{\mathbf{q}}_{\mathbf{xy}}$ in equation (2). Thus, we have provided the classification algorithm with the orientation of the ${}^{P}\mathbf{z}$ w.r.t. the earth.[2] We have visually checked that the processed acceleration and quaternion signals are consistent with the dynamics of the activities performed.

2.4.2 Training description

Although the proposed feature extraction technique is not restricted to any classification algorithm, in this paper, we evaluate its performance by applying it to an state-of-the-art Hierarchical Dynamic Model (HDM) based on Hidden Markov Models (HMMs). We train a different HMM for each activity independently using the Baum-Welch algorithm, following the scheme of [33] that will be explained in detail in Chapter 3.

Each HMM is modelled using five states per activity, i.e., having a global model with 25 identifiable states, and a Gaussian Mixture Model (GMM) observation probability distribution with three mixture components. We use the Forward-Backward (FB) algorithm to obtain the *Maximum a Posteriori* (MAP) estimate of the test sequences.

[2] We believe that most of the useful information residing in the orientation of the subject must rely on the inclination of its z-axis w.r.t. the earth. Nevertheless, further investigations will be performed.

2.4.3 Results

We compare the performance of the proposed AQ algorithm (with three different values of β) with the AAR algorithm. Table 2.1 shows the probability of error of both methods broken down by activity. The proposed AQ algorithm exhibits a lower error rate for all tested β, largely outperforming the AAR algorithm in some activities, and remaining very close in the others. Note that decreasing from 0.16 to 0.11 in probability of error is a remarkable reduction, since the bottleneck must presumably lie on the classification algorithm.

Activity	AAR	AQ $\beta = 1$	AQ $\beta = 3$	AQ $\beta = 5$
Running	0.38	0.18	0.19	0.20
Walking	0.02	0.05	0.02	0.05
Standing	0.03	0.06	0.05	0.05
Sitting	0.15	0.12	0.06	0.07
Lying	0.21	0.23	0.23	0.23
Mean	**0.16**	0.13	**0.11**	0.12

Table 2.1: Probability of error comparison of the AAR method and the proposed AQ method.

In Table 2.2, the feature extraction algorithms have also been compared in terms of the F-measure [104]. For all different values of β, the classification with the proposed algorithm outperforms that obtained with the AAR method. Again, the AQ method with $\beta = 3$ obtains the best results.

Finally, Figure 2.1 shows the F-measure range of accumulating the 6 test sequences of the 25 different sets, i.e., 150 different test sequences in total. For each method, the horizontal red line inside every box shows the median value, the upper and lower edges of the blue boxes are the 25th and 75th percentiles respectively, and the vertical black dashed lines extend to the extreme cases. It can be seen that most of the test sequences for all three values of β fall around a F-measure of

Activity	AAR	AQ $\beta = 1$	AQ $\beta = 3$	AQ $\beta = 5$
Running	0.75	0.88	0.87	0.86
Walking	0.92	0.95	0.95	0.94
Standing	0.98	0.96	0.97	0.97
Sitting	0.81	0.76	0.81	0.79
Lying	0.82	0.84	0.86	0.85
Mean	**0.86**	0.88	**0.89**	0.88

Table 2.2: F-measure of the AAR method and the proposed AQ method.

0.9 whereas for the AAR method they are around 0.85. The worst sequence with the proposed AQ method with $\beta = 3$ obtains a F-measure = 0.8 while the worst one with AAR remains at 0.75.

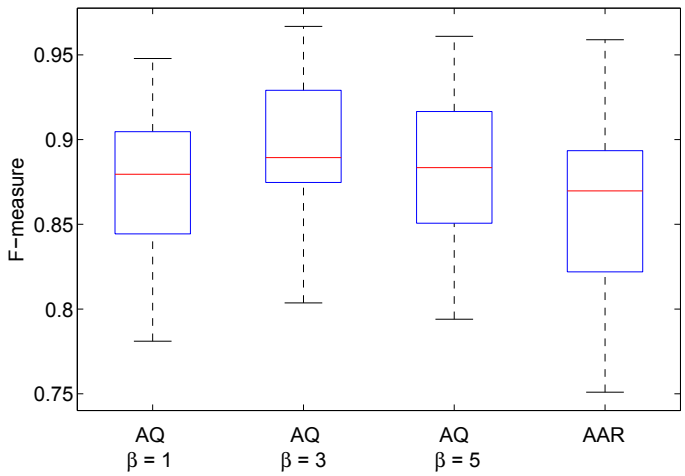

Figure 2.1: F-measure results for all test sequences using the AAR method and the proposed AQ method.

2.5 Conclusions

We have presented a novel data processing technique for human activity recognition based on quaternion representation. The proposed algorithm computes the acceleration referred to the person frame, and the orientation of the person frame with respect to the earth frame. Numerical results show a substantial improvement in the results of the classification algorithm when the feature extraction is performed with the proposed method. The computational cost of the proposed algorithm is linear with the length of the sequence and extremely low for each sample, requiring only few quaternion multiplications and additions. Moreover, the simplicity of the algorithm would also allow, with a slight adjustment, an online estimation of the person orientation.

3

Discriminative Learning

3.1 Introduction

Hidden Markov Models (HMMs) have been applied to a wide range of sequence modelling problems like speech recognition, Human Activity Recognition (HAR) or time series analysis [88]. The Expectation Maximization (EM) algorithm is the classical method used to learn the parameters of HMMs [14]. However, it exhibits two main problems: 1) the likelihood is multi-modal so the EM is guaranteed to converge only to a local maxima, and 2) although the complexity of the algorithm growths linearly with the length of the training sequences, the multiple initializations required for minimizing the effects of the local convergence and the more than quadratic growth with the number of hidden states makes EM computationally expensive with large training datasets. Bayesian inference methods including Gibbs sampling [91], variational optimization [68], or Bayesian non-parametric methods

[100] are even more computationally expensive and global convergence is still not guaranteed.

The authors in [46] propose a spectral algorithm for learning HMMs with discrete observations. Basically, the method adjusts the model by moment matching instead of maximizing the likelihood, and it relies on the use of the observable operators view of the HMM [51]. They use this approach to solve the prediction and filtering problems. Although the authors focus on HMMs with discrete observations, there exists several extensions for continuous observations using kernels [95, 97], solving the prediction and filtering problems too.

The application of HMMs to the HAR problem follows two main approaches. The first one [114] consists of learning a unique HMM, modelling only the temporal dependencies between different classes and assigning the same number of hidden states and classes. This is a very simple model and it usually must be combined with supervised learning algorithms. The second one [41] consists of learning one HMM for each possible class and then choosing the model with the maximum likelihood for each test case. The main problem of this approach is the need of defining a sequence size to learn each model and to infer the test sequences. A similar approach has been used in speech recognition where they initially train a different HMM for each class (phoneme or word) using the EM, and then they fine tune them discriminatively [53]. A more direct approach is based on discriminative training of the HMM, computing directly the posterior probabilities of the classes [108].

In this chapter, we propose a spectral algorithm for learning a discriminative HMM with discrete observations. We extend the work in [46] obtaining a recursive algorithm for estimating the labels of an observation sequence. We will introduce the hierarchical dynamic model [34] employed to implement the discriminative HMM and we will compare this approach with the spectral learning algorithm, both considering discrete observations and a Gaussian Mixture Model (GMM) for modelling the observations. The comparison between all the models will be evaluated in two different HAR settings. We will show than under the same conditions,

our algorithm outperforms the EM algorithm in computation time increasing the accuracy performance. When comparing with continuous observations, the accuracy error is slightly worse, but again, the computational time duration of our algorithm dramatically outperforms the EM.

3.2 Hierarchical Dynamical Model

The hierarchical dynamical model of a HMM was introduced in [34]. Under this model, every activity is modelled independently using the raw signals from the sensors and then the temporal dependencies between the different activities are included in the model. The resulting model is a global HMM, $\mathcal{M} = (\boldsymbol{\pi}, \Psi, \boldsymbol{\theta})$ built from several sub-HMMs, one for each activity, which are joined to yield the final HMM. This model incorporates two different temporal dynamics, the intra-activity dynamics, explaining the significant events happening in between the activities and the inter-activity dynamics, explaining the behavior of the system when we are transitioning between different activities. The observation model depends on the data and the specifications of the problem.

3.2.1 Intra-activity dynamics

The intra-activity dynamics are modelled as a HMM for every activity $j = \{1, \cdots, J\}$, $\mathcal{M}^j = (\boldsymbol{\pi}^j, \Psi^j, \boldsymbol{\theta}^j)$. The number of states used to train the different activities N_j is a parameter that needs to be fixed in advance. We can model the joint distribution of the activities and the states corresponding to a particular activity employing the Markov Property as

$$p(X, S|j) = \prod_{t=1}^{T} p(\mathbf{x}_t|s_t, j)p(s_t|s_{t-1}, j) \tag{3.1}$$

In [34], the authors study different topologies for the transition matrix depending on the nature of the activities. For this thesis however, we are considering a fully connected transition matrix with one initial state, one ending state and $N_j - 2$

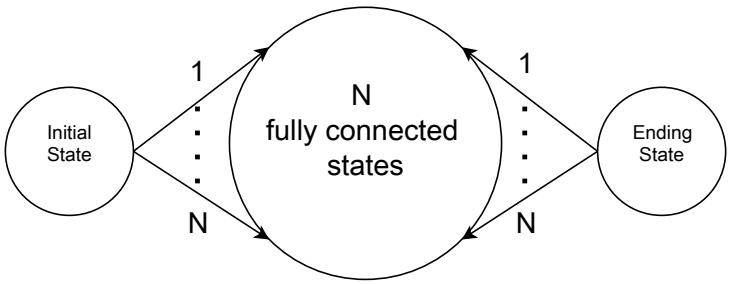

Figure 3.1: Topology of the transition matrix for every independent activity.

fully connected intermediate states to explain the temporal dependencies of the Markov Chain (Figure 3.1). The initial probability distribution for each intra-activity model $\boldsymbol{\pi}^j$ is forced to start always with the first state of the HMM, that is $\boldsymbol{\pi}^j = [1, 0, \cdots, 0]$.

3.2.2 Inter-activity dynamics

Once the intra-activity models have been trained, we can obtain the global HMM defining the transition probabilities between the different activities. This global HMM is expressed as $\mathcal{M} = (\boldsymbol{\pi}, \Psi, \boldsymbol{\theta})$, where the initial probabilities and the observations matrices are built by concatenating the corresponding parameters of the different intra-activity models, and the transition matrix Ψ is built with two steps:

1. Setting the transition matrices of the intra-activity models Ψ^j in the block diagonal matrix

$$\Psi = \begin{pmatrix} \Psi^1 & 0 & \cdots & 0 \\ 0 & \Psi^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Psi^J \end{pmatrix} \tag{3.2}$$

2. Connecting the sub-HMMs, setting the last state of the sub-HMMs to the initial state of all of the rest.

3. Resetting the self-transition probabilities corresponding to the last state of each activity, such as all the rows in the transition matrix sum to 1.

Finally, the initial probabilities distribution of the global HMM, $\boldsymbol{\pi}$, is a uniform distribution over the number of activities, where we set to $\frac{1}{J}$ the probability of the first state of every sub-HMM, and the rest to 0.

3.3 Spectral Algorithm for Learning HMMs

In this section, we briefly explain the spectral algorithm for learning HMMs with discrete observations presented in [46]. Let $\mathcal{S} \in \{1, \cdots, N\}$ the set of hidden states of a HMM and $\mathcal{X} \in \{1, \cdots, M\}$ the alphabet of the discrete observations. The probability transition matrix is $\Psi \in \mathbb{R}^{N \times N}$ where $\Psi_{ij} = p(s_{t+1} = i | s_t = j)$, the observation probability matrix is $\mathbf{O} \in \mathbb{R}^{M \times N}$ where $O_{ij} = p(x_t = i | s_t = j)$, the initial probabilities distribution is $\boldsymbol{\pi} \in \mathbb{R}^N$ with $\pi_i = p(s_1 = i)$ and t is any time instant. We can compute the probability of a sequence of observations in terms of observable operators [51] for each observation in $\mathcal{X}$:

$$\begin{aligned} \mathbf{A}_x &= \Psi \cdot \operatorname{diag}(\mathbf{O}_{x,:}) \\ \mathbf{O}_{x,:} &= (O_{x,1}, \cdots, O_{x,M}) \end{aligned}$$

where $\operatorname{diag}(\mathbf{O}_{x,:})$ is a diagonal matrix with elements $\mathbf{O}_{x,:}$ and $\underline{\mathbf{A}}$ is the tensor of observable operators.

For any t, the probability of any sequence of observations $x_{1:t} = [x_1, \cdots, x_t]$ can be written as the following product of matrices:

$$p(x_{1:t}) = \mathbf{1}_N^T \mathbf{A}_{x_t} \cdots \mathbf{A}_{x_1} \boldsymbol{\pi} = \mathbf{1}_M^T \mathbf{A}_{x_{t:1}} \boldsymbol{\pi} \tag{3.3}$$

where $\mathbf{1}_N$ is a column vector of N ones. This expression depends on the transition matrix Ψ and the observation matrix $\mathbf{O}$ of the model, but neither are known at the training stage.

If we are only interested in the value of (3.3) and not in the parameters of the model, we can use an invertible transformation of this equation that only depends on observable quantities:

$$\begin{aligned} \mathbf{1}_N^T \mathbf{A}_{x_{t:1}} \boldsymbol{\pi} &= \mathbf{1}_N^T \mathbf{T}^{-1} \mathbf{T} \mathbf{A}_{x_t} \mathbf{S}^{-1} \cdots \mathbf{T} \mathbf{A}_{x_1} \mathbf{T}^{-1} \mathbf{T} \boldsymbol{\pi} \\ &= \mathbf{b}_\infty^T \mathbf{B}_{x_{t:1}} \mathbf{b}_1 \end{aligned} \tag{3.4}$$

where $\mathbf{T} \in \mathbb{R}^{N \times N}$ is the invertible transformation matrix and

$$\mathbf{b}_\infty^T = \mathbf{1}_N^T \mathbf{T}^{-1}, \quad \mathbf{B}_x = \mathbf{T}\mathbf{A}_x\mathbf{T}^{-1}, \quad \mathbf{b}_1 = \mathbf{T}\boldsymbol{\pi}$$

The idea of this transformation is to avoid the identification problem of the hidden structure of the model by expressing (3.4) in terms of the marginal probabilities of the vector of singletons $\mathbf{p}_1 \in \mathbb{R}^M$, matrix of pairs $\mathbf{P}_{21} \in \mathbb{R}^{M \times M}$ and tensor of triples $\underline{\mathbf{P}_{31}^x} \in \mathbb{R}^{M \times M}, \forall x$. These quantities can be estimated from the data, and they are related to the parameters of the HMM

$$\begin{aligned}
\mathbf{p}_1 &= p(x_1) = \mathbf{O}\boldsymbol{\pi} \\
\mathbf{P}_{21} &= p(x_2, x_1) = \mathbf{O}\Psi\mathrm{diag}\left(\boldsymbol{\pi}\right)\mathbf{O}^T \\
\underline{\mathbf{P}_{31}^x} &= p(x_3, x_2 = x, x_1) = \\
&= \mathbf{O}\mathbf{A}_x\Psi\mathrm{diag}\left(\boldsymbol{\pi}\right)\mathbf{O}^T, \forall x
\end{aligned}$$

In [46], the authors proof that $\mathbf{T} = \mathbf{U}^T\mathbf{O}$ is a valid invertible transformation for the HMM model, where $\mathbf{U} \in \mathbb{R}^{M \times N}$ is the matrix of N left singular vectors of the joint probability matrix $\mathbf{P}_{21}$. With this transformation, we can express the new parameters of the model as:

$$\begin{aligned}
\mathbf{b}_1 &= \mathbf{U}^T\mathbf{p}_1 \\
\mathbf{b}_\infty &= \left(\mathbf{P}_{21}^T\mathbf{U}\right)^\dagger \mathbf{p}_1 \\
\mathbf{B}_x &= \left(\mathbf{U}^T\underline{\mathbf{P}_{31}^x}\right)\left(\mathbf{U}^T\mathbf{P}_{21}\right)^\dagger, \forall x
\end{aligned} \tag{3.5}$$

where $(\cdot)^\dagger$ is the Moore-Penrose pseudo-inverse operation.

To compute the spectral algorithm for learning a HMM, first we take m *i.i.d.* triples $[x_1, x_2, x_3]$ from the training observations to obtain an estimation of $\mathbf{p}_1$, $\mathbf{P}_{21}$ and $\underline{\mathbf{P}_{31}^x}$. Then, we compute the Singular Value Decomposition (SVD) of $\mathbf{P}_{21}$, obtaining the matrix of N left singular vectors $\mathbf{U}$. Finally, we compute the HMM transformed parameters in (3.5).

3.4 Spectral Algorithm for Learning Discriminative HMMs

The algorithm in Section 3.3 can be used to obtain the probability of a sequence of observations or the probability of the last observation given all the previous ones. However, the focus of this thesis is the classification of a sequence of discrete labels (or activities) given the observations, using the discriminative HMM model in Figure 4.1. We want to solve 1) the problem of predicting the probability of a sequence of labels given the observations and 2) the problem of predicting the conditional probability of a label given all the previous labels and observations.

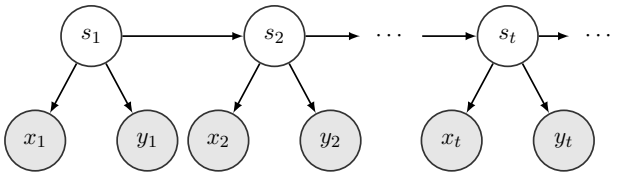

Figure 3.2: Discriminative HMM graphical model of a sequence of t observations $x_{1:t}$ and labels $y_{1:t}$

Let $\mathcal{L} \in \{1, \cdots, L\}$ be the alphabet of labels of the model and $\mathbf{D} \in \mathbb{R}^{L \times N}$ the label probability matrix where $D_{ij} = p(y_t = i | s_t = j)$. We define the joint probability of a sequence of labels $y_{1:t}$ and observations $x_{1:t}$ as in (3.3)

$$p(y_{1:t}, x_{1:t}) = \mathbf{1}_N^T \mathbf{A}_{y_t x_t} \cdots \mathbf{A}_{y_1 x_1} \boldsymbol{\pi} \tag{3.6}$$

where the label and observation operators are

$$\mathbf{A}_{y_t x_t} = \Psi \text{diag}(\mathbf{D}_{y_t,:}) \text{diag}(\mathbf{O}_{x_t,:}) \tag{3.7}$$

We have ML label and observation operators, i.e., the combination of all its possible values. In this discriminative HMM model, the labels and the observations are independent given the hidden states, so we can define for simplicity a joint label-observation matrix $\mathbf{F} \in \mathbb{R}^{ML \times N}$ where

$$F_{ijk} \quad = \quad P(y_t = i | s_t = k) P(x_t = j | s_t = k) = D_{ik} O_{jk}$$

and from (3.7), our new operators are $\mathbf{A}_{y_t x_t} = \Psi \text{diag}(\mathbf{F}_{y_t x_t,:})$.

To solve the first problem, we compute the probability of a sequence of labels given the observations applying the Bayes theorem and combining the expressions (3.3) for the denominator and (3.6) for the numerator:

$$p(y_{1:t}|x_{1:t}) = \frac{p(y_{1:t}, x_{1:t})}{p(x_{1:t})} = \frac{\mathbf{1}_N^T \mathbf{A}_{y_t x_t} \cdots \mathbf{A}_{y_1 x_1} \boldsymbol{\pi}}{\mathbf{1}_N^T \mathbf{A}_{x_t} \cdots \mathbf{A}_{x_1} \boldsymbol{\pi}} \tag{3.8}$$

We are again only interested in the value of (3.8), and not in the values of the matrices Ψ, $\mathbf{O}$ and $\mathbf{D}$. We can apply an invertible transformation in the numerator and the denominator of this expression so it only depends on the training observations and labels. We follow Section 3.3 to obtain the denominator, and we use an invertible transformation matrix $\mathbf{R} \in \mathbb{R}^{ML \times N}$ as in (3.4) to obtain the numerator:

$$\mathbf{1}_N^T \mathbf{A}_{y_{t:1} x_{t:1}} \boldsymbol{\pi} = \mathbf{c}_\infty^T \mathbf{C}_{y_{t:1} x_{t:1}} \mathbf{c}_1$$

where $\mathbf{A}_{y_{t:1} x_{t:1}}$ is the product $\mathbf{A}_{y_t x_t} \cdots \mathbf{A}_{y_1 x_1}$ and

$$\mathbf{c}_\infty^T = \mathbf{1}_N^T \mathbf{R}^{-1}, \quad \mathbf{C}_{yx} = \mathbf{R} \mathbf{A}_{yx} \mathbf{R}^{-1}, \quad \mathbf{c}_1 = \mathbf{R}\boldsymbol{\pi} \tag{3.9}$$

We want to represent the numerator in (3.8) in terms of the observable vector $\mathbf{q}_1 \in \mathbb{R}^{ML}$, matrix $\mathbf{Q}_{21} \in \mathbb{R}^{ML \times ML}$ and tensor $\underline{\mathbf{Q}_{31}^{yx}} \in \mathbb{R}^{ML \times ML}, \forall y, x$. Also, these observable quantities can be expressed in terms of the hidden parameters of the model.

$$\begin{aligned}
\mathbf{q}_1 &= p(y_1, x_1) = \mathbf{F}\boldsymbol{\pi} \\
\mathbf{Q}_{21} &= p(y_2, x_2, y_1, x_1) = \mathbf{F}\Psi \mathrm{diag}\left(\boldsymbol{\pi}\right) \mathbf{F}^T \\
\underline{\mathbf{Q}_{31}^{yx}} &= p(y_3, x_3, y_2 = y, x_2 = x, y_1, x_1) = \\
&= \mathbf{F}\mathbf{A}_{yx}\Psi \mathrm{diag}\left(\boldsymbol{\pi}\right) \mathbf{F}^T
\end{aligned}$$

From [46] it is straightforward to proof that $\mathbf{R} = \mathbf{V}^T\mathbf{F}$ is a valid invertible transformation matrix for the model, where $\mathbf{V}$ is the matrix of left singular vectors of $\mathbf{Q}_{21}$. We can calculate the transformed parameters of the model in (3.9) in terms of $\mathbf{V}$ and these observable quantities:

$$\begin{aligned}
\mathbf{c}_1 &= \mathbf{V}^T \mathbf{q}_1 \\
\mathbf{c}_\infty &= (\mathbf{Q}_{21}^T \mathbf{V})^\dagger \mathbf{q}_1 \\
\mathbf{C}_{lx} &= (\mathbf{V}^T \underline{\mathbf{Q}_{31}^{yx}})(\mathbf{V}^T \mathbf{Q}_{21})^\dagger, \forall y, x
\end{aligned}$$

To solve the second problem, we use the Bayes Theorem to express the conditional probability of a label y_t given the all the previous labels $y_{1:t-1}$ and all the observations $x_{1:t}$

$$p(y_t|y_{1:t-1}, x_{1:t}) = \frac{\mathbf{c}_\infty^T \mathbf{C}_{y_{t:1} x_{t:1}} \mathbf{c}_1}{\mathbf{b}_\infty^T \mathbf{B}_{x_t} \mathbf{Z} \mathbf{C}_{y_{t-1:1} x_{t-1:1}} \mathbf{c}_1} \tag{3.10}$$

where $\mathbf{Z} = (\mathbf{U}^T\mathbf{O})(\mathbf{V}^T\mathbf{F})$ is the transformation term between $\mathbf{B}_{x_t}$ and $\mathbf{C}_{y_{t-1}x_{t-1}}$. We can further express $\mathbf{Z}$ in terms of an additional observable operator $\mathbf{W}_{21} \in \mathbb{R}^{M \times ML}$

$$\begin{aligned} \mathbf{W}_{21} &= P(x_2, y_1, x_1) = \mathbf{O}\Psi \mathrm{diag}\,(\boldsymbol{\pi})\, \mathbf{F}^T \\ \mathbf{Z} &= (\mathbf{U}^T\mathbf{W}_{21})(\mathbf{V}^T\mathbf{Q}_{21})^\dagger \end{aligned}$$

Finally, equation (3.10) can also be represented recursively as follows:

$$\begin{aligned} P(y_t|y_{1:t-1}, x_{1:t}) &= \frac{\mathbf{c}_\infty^T \mathbf{C}_{y_t x_t} \mathbf{c}_t}{\sum_y \mathbf{c}_\infty^T \mathbf{C}_{y x_t} \mathbf{c}_t} \\ \mathbf{c}_{t+1} &= \frac{\mathbf{C}_{y_t x_t} \mathbf{c}_t}{\mathbf{b}_\infty^T \mathbf{B}_{x_{t+1}} \mathbf{Z} \mathbf{C}_{y_t x_t} \mathbf{c}_t} \end{aligned} \tag{3.11}$$

We describe the spectral algorithm to solve both classification problems for discriminative HMMs in Algorithm 2. It is interesting to notice that this algorithm can also handle missing labels in the training. If we take a triple $\{x_1, x_2, x_3\}$ with missing labels, we can just actualize $\mathbf{p}_1$, $\mathbf{P}_{21}$ and $\underline{\mathbf{P}_{31}^x}$ without changing the other operators and proceed normally.

3.5 Results

3.5.1 Inertial sensors database

In the first experiment, we use a HAR database obtained using miniature inertial sensors from APDM [50], which provide acceleration, gyroscope and magnetometer data. 18 different people were used as subjects for the experiments and only one sensor was placed at the waist of each of them. Then, they were asked to perform a sequence of activities, that were combinations of running, walking, standing,

Algorithm 2 Spectral Algorithm for discriminative HMMs

Input:

Number of hidden states N and labels L.

Sample size M

m *i.i.d.* groups of triples $\{x_1, x_2, x_3\}$ and $\{y_1, y_2, y_3\}$,

Test sequence $x_{1:t}$

Output: $p(y_t|y_{1:t-1}, x_{1:t})$

1. Compute the empirical estimates of $\hat{\mathbf{p}}_1$, $\hat{\mathbf{P}}_{21}$, $\underline{\hat{\mathbf{P}}^x_{31}}$, $\hat{\mathbf{q}}_1$, $\hat{\mathbf{Q}}_{21}$, $\underline{\hat{\mathbf{Q}}^{yx}_{31}}$ and $\hat{\mathbf{W}}_{21}$.
2. Compute both the SVD of $\hat{\mathbf{P}}_{21}$ to get the matrix of N left singular vectors $\hat{\mathbf{U}}$ and the SVD of $\hat{\mathbf{Q}}_{21}$ to get the matrix of N left singular vectors $\hat{\mathbf{V}}$.
3. Compute the transformed HMM model quantities $\hat{\mathbf{b}}_1$, $\hat{\mathbf{b}}_\infty$, $\hat{\mathbf{B}}_x$, $\hat{\mathbf{c}}_1$, $\hat{\mathbf{c}}_\infty$, $\hat{\mathbf{C}}_{yx}$ and $\hat{\mathbf{Z}}$.
4. Compute sequentially for all t the probabilities of the labels $y_{1:t}$ using (3.11).

sitting and lying, in no particular order. Also, the data was processed to make it invariant to sensor orientation, following the work presented in Chapter 2. The bottleneck of the spectral algorithm is the computation of the SVD. However, we can overcome this problem by substituting the SVD by a random normalized matrix and a regularization step, following the work in [73].

To evaluate our algorithm, we get 1000 random unique observations as centroids from the training data and assign each observation to one of these centroids according to the minimum euclidean distance. We re-estimate the centroids 100 times getting the mean accuracy error for each case[1]. Also, the number of centroids for each activity is proportional to the number of training observations of each activity. We perform a leave-one-out strategy, where we use 17 sequences of activities as training and one sequence as test to evaluate the performance of the algorithms. The number of hidden states is set to $N = 25$. For the EM algorithm, we perform 5 iterations of the algorithm with k-means initialization. We consider two different EMs, one with continuous observations, assuming that the probability of the observations is a GMM and the other one with the same discrete observations as with the spectral case. We include the results from performing clustering on the observations using the minimum euclidean distance as a baseline performance. In Figure 3.3 we show the comparison of all the methods for each leave-one-out case. We observe that the accuracy error for the GMM EM is in mean 3.39% better than in the spectral learning. Also, the discrete EM algorithm performs a 1.40% worse than the spectral learning algorithm. We can conclude that the reason for the loss in accuracy is due to the use of discrete observations. In fact, under the same assumptions, the absence of local minima in the spectral algorithm results in a better performance.

The most important point of this comparison is the drastic difference in computation time. The EM algorithm requires multiple initializations to avoid the local minima. Furthermore, the EM algorithm is a recursive algorithm with the number of observations, meanwhile this restriction is not present in the spectral algorithm.

[1]The standard deviation for all the cases is less than 0.04.

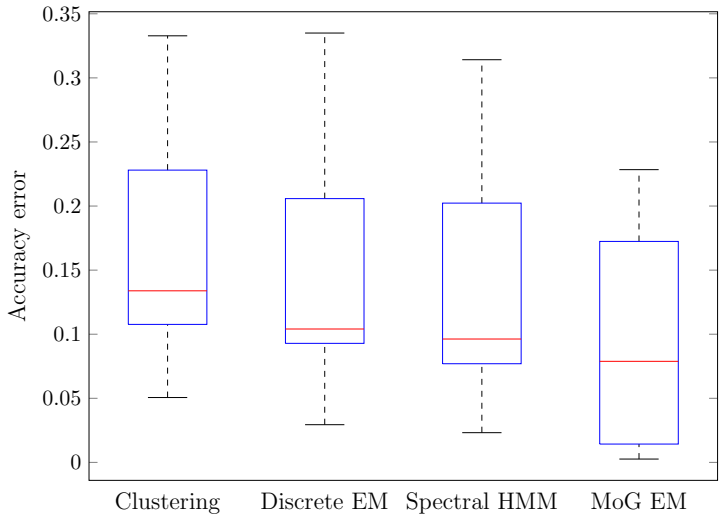

Figure 3.3: Accuracy error of the activity classification using a clustering classification, the discriminative HMM with the EM algorithm, both with continuous and discrete observations, and the discriminative HMM with spectral learning and discrete observations.

In particular, in a 2.5 Ghz Intel Core i5 processor with Matlab and C for the recursive section of the EM, 30.9 minutes are needed in the training step for one of the sequences of the leave-one-out case, while in the spectral algorithm case implemented exclusively in Matlab, only 1.37 seconds are needed.

3.5.2 Binary sensors database

In the second experiment, we have used the databases with discrete observations generated by the authors in [103]. They implemented a binary sensor network in three different home environments to detect several events, like movement of objects or motion in a specific area. They placed 14, 23 and 21 sensors in each house respectively and they tagged between 10 and 16 activities. We use the last-fired feature representation of the database, where a sensor remains activated as long as another sensor is not activated. We compare the results of the HMM using the EM algorithm with the spectral algorithm for each of the three settings in terms of the accuracy error in Table 3.1. We can see that our spectral algorithm obtains better results in the first two houses while performing worse in the other one.

House	HMM+EM	Spectral HMM
A	0.105	0.053
B	0.516	0.332
C	0.161	0.227

Table 3.1: Accuracy error of the activity classification using both the discriminative HMM with the EM algorithm and the discriminative HMM with spectral learning and discrete observations.

3.6 Conclusions

In this chapter, we have proposed a spectral algorithm for learning a discriminative HMM with discrete observations. Our algorithm exhibits a similar performance as the continuous observations HMM while dramatically outperforming it in terms of time computation. Also, it performs better in general than the discrete HMM, leading to the hypothesis that the loss of accuracy is due to the discreteness of the observations, not to the spectral algorithm implementation. The low computation time and complexity and the possibility of obtaining directly the probability of the labels make it useful for on-line tracking problems, e.g. human on-line activity recognition applications. In the future, we will try to extend this spectral algorithm for discriminative HMMs with continuous observations and for other models where the large amount of training data make iterative methods prohibitive.

4

Classifier Combination

4.1 Introduction

In multi-sensor systems we have to address the problem of information fusion. The most common approach is to send the sensors' raw signals to a central unit which estimates the performed activity. This configuration, referred as signal fusion, imposes high bandwidth and synchronization requirements for wireless transmission. Instead, a summary of the signals (features), or an estimation of the performed activity can be transmitted, leading to feature fusion and decision fusion (also named as classifiers combination) configurations respectively. Decision fusion also offers the advantage of robustness against sensor failures with respect to feature fusion, because only a single available sensor is needed to have an estimation of the performed activity.

In ambulatory Human Activity Recognition (HAR) systems, the number of

sensors employed in [83] and [111] are unrealistic. In some experiments a single sensor system is used [42] and the placement of the sensor is of critical importance, i.e., depending on the activity some positions are preferred over others. More sensible numbers of sensors are 2 (e.g., waist and ankle, or waist and wrist, or wrist and chest), 3 or 4. A system with this number of sensors can be practically implemented using smartphones, wristbands, smartwatches, and some additional low-consumption wearable sensors [40], being some of them commercially available [1].

In this work we are considering a classifier combination approach. The same approach is followed in [111], where a discriminative Hidden Markov Model (HMM) classifier is trained for each individual sensor, and a Naive Bayes classifier fuses these individual classification results. A more elaborated approach to Bayesian classifier combination in a general framework is proposed in [56]. In particular, the authors develop models for dependent and independent classifiers that simultaneously infer both the ground truth and the model parameters from the individual classifier outputs. However, in the extreme case of using only two sensors, these methods are equivalent to select the most discriminative sensor and, therefore, the combination does not yield better precision than the best single classifier. To obtain advantages even in the two sensors case, we propose to directly combine the soft outputs of the classifiers.

Another important question in the HAR problem is the dynamic structure of the human activities. Most HAR state-of-the-art methods (see [22] for a review not restricted to inertial sensors) make use of some kind of dynamic Bayesian network (mostly HMMs) to capture the dynamics of human activities. However, to the extent of our knowledge, none of the proposed classifier combination methods consider the dynamic structure on the combination result, i.e., no classifier combination method considers that the result of the combination depends on previous classifier combination results.

Our work extends the Independent Bayesian Classifier Combination (IBCC) model proposed in [56] in two directions. First, we substitute the categorical (hard)

output of the individual classifiers with a vector that contains the confidence on being performing each of the possible activities (soft output), which allows dealing with a low number of sensors. The confidence can be the posterior probability of the performed activity, but can be also a categorical value (for example: not likely, perhaps, almost sure), or any other probabilistic or non probabilistic confidence measure. The only restrictions to the confidence measure are to be positive and normalized (the sum of all confidence measures must be one). Second, we include a first-order Markov chain approximation to model the classifier combination output, i.e., the ground truth, to capture the dynamics of the human activities. We call Soft-output Classifier Combination (SCC) the model that includes the soft output extension, and Markov Soft-output Classifier Combination (MSCC) the model that includes both the soft output and the Markovian ground truth extensions. MSCC provides good results in comparison with existing classifier combination methods for the HAR problem in accuracy, speed of the inference process, and robustness against sensor failures.

4.2 Human Activity Recognition by Soft Output Classifier Combination

4.2.1 Problem Formulation

We observe sequences from different people. Each sequence is composed of two parts, the output of the soft classifiers, C, and the ground truth, T. We assume that each pair (C, T) is a realization of the same random process. Each sequence C is composed by I instances of K soft output classifiers, $C = \{\mathbf{c}_i^k : i = 1, \ldots, I,\ k = 1, \ldots, K,\ \mathbf{c}_i^k \in \mathcal{S}^J\}$, where J is the number of classes (activities) and $\mathcal{S}^J$ is the J-dimensional probability simplex, i.e., $c_{ij}^k \in \mathbb{R}_+, \sum_j c_{ij}^k = 1$. If $c_{ij}^k > c_{i\ell}^k$ then $p(t_i = j|\mathbf{x}_i^k, \theta^k) > p(t_i = \ell|\mathbf{x}_i^k, \theta^k)$, where $\mathbf{x}_i^k$ is the i-instance input to classifier k, that is characterized by a set of parameters θ^k. No other assumptions are made on $\mathbf{c}_i^k$. The ground truth sequence is $T = \{t_i : i = 1, \ldots, I,\ t_i \in \{1, \ldots, J\}\}$, and it can be unknown for some sequences.

We consider two inference problems. The first and most computationally demanding is a transductive semi-supervised estimation problem. The data consist of a set of labelled sequences, $\{(C_n, T_n) : n = 1, \ldots, N\} \equiv (C_{1:N}, T_{1:N})$, and a set of unlabelled sequences, $\{C_n^u : n = 1, \ldots, N^u\} \equiv C_{1:N^u}^u$, and the aim is to estimate the unknown ground truth sequences, $T_{1:N^u}$. In the second problem, the aim is to design an inductive classifier that avoids retraining the combination algorithm every time we want to estimate the ground truth T^* from a new sequence C^*.

4.2.2 Soft Output Combination of Classifiers model

We assume that the outputs $\{\mathbf{c}_i^k : i = 1, \ldots, I\}$ of the classifier k are conditionally independent of the ground truth. We propose a Dirichlet prior distribution on $\mathbf{c}_i^k$ with mean $\mathbf{h}_j^k$ and strength α_j^k as a natural choice for variables that lie in the $\mathcal{S}^J$ simplex, contrary to the multinomial distribution used as a prior for discrete observations in [56]. The observation model is the following

$$\mathbf{c}_i^k | t_i = j, \boldsymbol{\alpha}^k, \mathbf{H}^k \sim \mathrm{Dir}(\alpha_j^k \mathbf{h}_j^k), \qquad t_i | \mathbf{p} \sim \mathrm{Cat}(\mathbf{p}),$$

where $\mathbf{H}^k = \{\mathbf{h}_j^k : j = 1, \cdots, J\}$ is the confusion matrix of the k classifier, and $\underline{\mathbf{H}}$ is the tensor containing the confusion matrices of the K classifiers. Like in [56], the model of the ground truth t_i is a categorical distribution with parameters $\mathbf{p} \in \mathcal{S}^J$, where $p_j = P(t_i = j)$.

We propose a Gamma prior on the strength α_j^k and Dirichlet priors on $\mathbf{h}_j^k$ and $\mathbf{p}$,

$$\alpha_j^k | a, b \sim \mathrm{Gamma}(a, b), \qquad \mathbf{h}_j^k | \beta, \boldsymbol{\lambda}_j \sim \mathrm{Dir}(\beta \boldsymbol{\lambda}_j),$$
$$\mathbf{p} | \epsilon, \boldsymbol{\gamma} \sim \mathrm{Dir}(\epsilon \boldsymbol{\gamma}),$$

where $a, b, \beta, \epsilon \in \mathbb{R}_+$ and $\boldsymbol{\lambda}_j, \boldsymbol{\gamma} \in \mathcal{S}^J$. We further parametrize $\lambda_{j\ell} = \kappa\delta(j, \ell) + (1 - \kappa)/J$, where $\kappa \in [0, 1]$, and $\delta(j, \ell) = 1$ when $j = \ell$ and 0 otherwise. This prior reflects our belief that the classifiers perform better than random, and κ controls how close we believe the classifiers are to the ideal one, i.e., $\kappa = 1$. We call this model SCC, and its graphical representation is shown in Figure 4.1.

The joint probability distribution of the observations and the model parameters

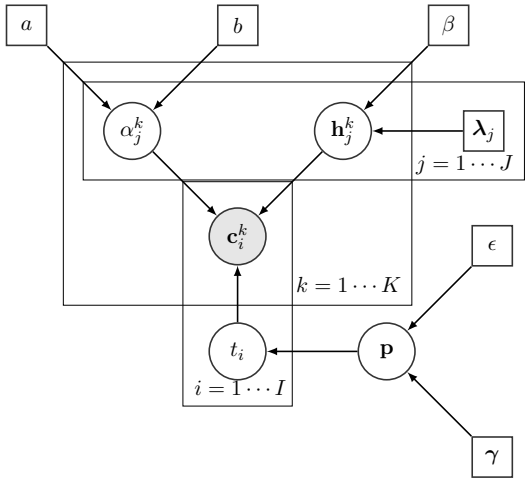

Figure 4.1: Graphical representation of the SCC model.

is

$$p(C, T, \boldsymbol{\alpha}, \underline{\mathbf{H}}) = p(C|T, \boldsymbol{\alpha}, \underline{\mathbf{H}})p(T|\epsilon, \boldsymbol{\gamma})p(\boldsymbol{\alpha}|a, b)p(\underline{\mathbf{H}}|\beta, \boldsymbol{\lambda}),$$

where the result of integrating out $\mathbf{p}$, $p(T|\epsilon, \boldsymbol{\gamma})$, is the Dirichlet compound multinomial distribution.

The probability of C given the rest of the parameters is

$$\begin{aligned} p(C|T, \boldsymbol{\alpha}, \underline{\mathbf{H}}) &= \prod_{j=1}^{J} \prod_{k=1}^{K} \prod_{i:t_i=j} \mathrm{Dir}\left(\mathbf{c}_i^k | \alpha_j^k, \mathbf{h}_j^k\right) = \\ &= \prod_{j=1}^{J} \prod_{k=1}^{K} \left(\frac{\Gamma(\alpha_j^k)}{\prod_{\ell=1}^{J} \Gamma(\alpha_j^k h_{j\ell}^k)} \prod_{\ell=1}^{J} (\bar{c}_{j\ell}^k)^{\alpha_j^k h_{j\ell}^k - 1} \right)^{n_j}, \end{aligned} \tag{4.1}$$

where $\bar{c}_{j\ell}^k = \left(\prod_{i:t_i=j} c_{i\ell}^k\right)^{\frac{1}{n_j}}$ is the geometric mean of the observations whose $t_i = j$, $n_j = \sum_{i=1}^{I} \mathbb{I}(t_i = j)$ is the number of instances with ground truth j, $\mathbb{I}(\cdot)$ is the indicator function and $\Gamma(\cdot)$ is the gamma function. To infer the hidden variables from the observations, we use a Gibbs sampling algorithm where the conditional

distribution of each parameter given the rest is

$$p(\alpha_j^k|\text{rest}) \propto \prod_{i:t_i=j} \text{Dir}\left(\mathbf{c}_i^k|\alpha_j^k, \mathbf{h}_j^k\right) \text{Gamma}(\alpha_j^k|a,b), \tag{4.2}$$

$$p(\mathbf{h}_j^k|\text{rest}) \propto \prod_{i:t_i=j} \text{Dir}\left(\mathbf{c}_i^k|\alpha_j^k, \mathbf{h}_j^k\right) \text{Dir}(\mathbf{h}_j^k|\beta, \boldsymbol{\lambda}_j), \tag{4.3}$$

$$p(t_i = j|\text{rest}) \propto \prod_{k=1}^{K} \text{Dir}\left(\mathbf{c}_i^k|\alpha_j^k, \mathbf{h}_j^k\right) (\epsilon\gamma_j + n_j^{\neg i}), \tag{4.4}$$

where $n_j^{\neg i}$ is the same as n_j without the i-th element.

In the case of the conditionals (4.2) and (4.3) there is no analytic expression for the normalization constant, so we cannot directly sample from them. Sampling from (4.2) is done using the Double Adaptive Rejection Metropolis Sampling (A2RMS) algorithm proposed in [72]. Sampling from (4.3) is done with a Metropolis-Hastings algorithm, using as proposal a uniform distribution in the simplex of $\mathcal{S}^J$. To draw samples uniformly from the simplex, we use the method described in [45]. As an alternative, we also considered the Dirichlet prior of the conditional distribution of $\mathbf{h}_j^k$ as proposal, obtaining similar results in both cases.

For the semi-supervised inference problem, the initial sampling from (4.4) is done using the values of $\mathbf{h}_j^k$ and α_j^k estimated from the labelled sequences. This strategy has shown to be effective against the peaky nature of the Dirichlet distributions due to the low values of α_j^k obtained in the HAR problem.

To build the inductive classifier we use the predictive distribution which is approximated by averaging M samples after the Gibbs algorithm reaches the stationary regime

$$\begin{aligned} p(T^*|C^*, C_{1:N}, T_{1:N}) &\propto p(C^*|T^*, C_{1:N}, T_{1:N}) p(T^*|C_{1:N}, T_{1:N}) \approx \\ &\approx \frac{1}{M} \sum_m p(C^*|T^*, \boldsymbol{\alpha}^{(m)}, \underline{\mathbf{H}}^{(m)}) p(T^*|T_{1:N}). \end{aligned}$$

We compare this expression with a point estimate inference using the average values of $\boldsymbol{\alpha}$ and $\underline{\mathbf{H}}$ in the Gibbs sampler, $\overline{\boldsymbol{\alpha}}$ and $\overline{\underline{\mathbf{H}}}$,

$$t^* = \arg\max_{1\le j\le J} \log p(t^*|\mathbf{c}^*, \overline{\boldsymbol{\alpha}}, \overline{\underline{\mathbf{H}}}) = \arg\max_{1\le j\le J} \left\{ \sum_{k=1}^{K} \mathbf{w}_j^{k^T} \log \mathbf{c}^{*k} + w_{0_j}^k \right\},$$

where

$$\begin{aligned} \mathbf{w}_j^k &= \bar{\alpha}_j^k \bar{\mathbf{h}}_j^k - 1, \\ w_{0j}^k &= \log\left(\epsilon\gamma_j + n_j\right) + \sum_{k=1}^{K}\left(\log\Gamma(\bar{\alpha}_j^k) - \sum_{\ell=1}^{J}\log\Gamma(\bar{\alpha}_j^k \bar{h}_{j\ell}^k)\right). \end{aligned}$$

We can see that the soft outputs of each classifier are weighted differently depending on the ground truth.

4.2.3 SCC With Markovian Ground Truth

In the HAR problem, the dynamic structure of the activities performed by a person can be modelled using a Markov model. For example, the probability that a person is running at a given instant is different depending on whether that person was running or sitting at the previous instant. In its simplest form, the ground truth can be modelled as a first order homogeneous Markov chain defined by the initial state probability distribution $\boldsymbol{\pi}$ and the transition probability matrix $\boldsymbol{\Psi}$. We will consider Dirichlet priors for both $\boldsymbol{\pi}$ and the rows of $\boldsymbol{\Psi}$, $\boldsymbol{\psi}_j$, with independence between the rows ($1 \leq j \leq J$), and favouring self transitions. In summary,

$$p(T|\boldsymbol{\pi}, \boldsymbol{\Psi}) = p(t_1|\boldsymbol{\pi}) \prod_{i=2}^{I} p(t_{i+1}|t_i, \boldsymbol{\Psi}),$$

with priors

$$\begin{aligned} t_1|\boldsymbol{\pi} &\sim \text{Cat}(\boldsymbol{\pi}), \\ t_{i+1}|t_i = j, \boldsymbol{\Psi} &\sim \text{Cat}(\boldsymbol{\psi}_j), \\ \boldsymbol{\pi}|\eta &\sim \text{Dir}(\eta \mathbf{u}_J), \\ \boldsymbol{\psi}_j|\mu, \boldsymbol{\phi}_j &\sim \text{Dir}(\mu \boldsymbol{\phi}_j), \end{aligned}$$

where $\eta, \mu \in \mathbb{R}_+$ are the strength parameters, $\mathbf{u}_J$ is the J-dimensional vector $[1/J, \ldots, 1/J]$ and $\boldsymbol{\phi}_j \in \mathcal{S}^J$ with $\phi_{j\ell} = \phi\delta(j, \ell) + (1 - \phi)/J$, where $\phi \in [0, 1]$. $\mathbf{u}_J$ can be replaced with a more informative hyper parameter if further information is available. The observation model is the same as in the SCC model (4.1). We call this model MSCC, and its graphical representation is shown in Figure 4.2.

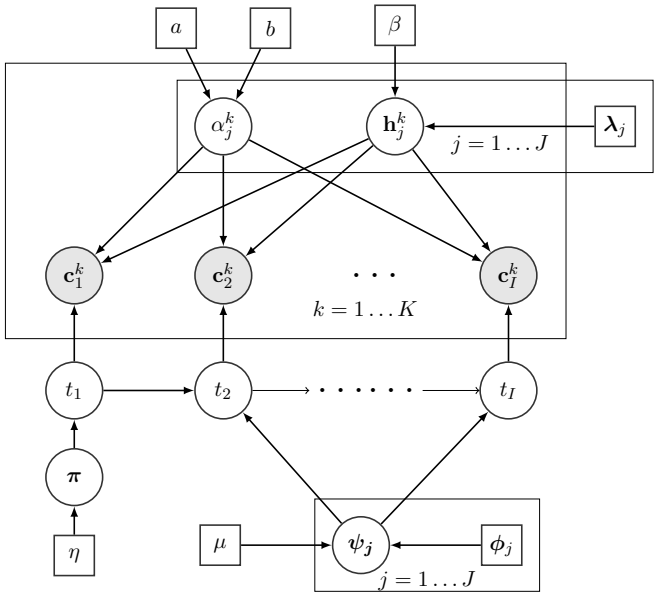

Figure 4.2: Graphical representation of the MSCC model.

The joint probability distribution of the observations and the model parameters is now

$$p(C, T, \boldsymbol{\alpha}, \underline{\mathbf{H}}, \boldsymbol{\Psi}, \boldsymbol{\pi}) = p(C|T, \boldsymbol{\alpha}, \underline{\mathbf{H}})p(T|\boldsymbol{\pi}, \boldsymbol{\Psi})p(\boldsymbol{\alpha}|a, b)p(\underline{\mathbf{H}}|\beta, \boldsymbol{\lambda})p(\boldsymbol{\pi}|\eta)p(\boldsymbol{\Psi}|\mu, \boldsymbol{\phi}).$$

In the semi-supervised inference problem, inference for the unknown ground truth sequences is done using the Forward-Filtering Backward-Sampling (FFBS) algorithm [37]. The forward step of the FFBS for each unlabelled sequence is defined by the recursion

$$\begin{aligned} a_1(j) &= p(t_1 = j|\mathbf{c}_1^{1:K}, \boldsymbol{\alpha}, \underline{\mathbf{H}}, \boldsymbol{\pi}) = \pi_j \prod_{k=1}^{K} \mathrm{Dir}(\mathbf{c}_1^k|\alpha_j^k, \mathbf{h}_j^k), \\ a_i(j) &= p(t_i = j|\mathbf{c}_{1:(i-1)}^{1:K}, \boldsymbol{\alpha}, \underline{\mathbf{H}}, \boldsymbol{\Psi}) \propto \prod_{k=1}^{K} \mathrm{Dir}(\mathbf{c}_i^k|\alpha_j^k, \mathbf{h}_j^k) \sum_{\ell=1}^{J} \psi_{\ell j} a_{i-1}(\ell), \end{aligned}$$

and the backward step is done by sampling recursively from

$$\begin{aligned} t_i &\sim p(t_i = j|t_{i+1} = \ell, \mathbf{c}_{1:(i+1)}^{1:K}, \boldsymbol{\alpha}, \mathbf{H}, \boldsymbol{\Psi}) \propto \\ &\propto \frac{\prod_{k=1}^{K} p(\mathbf{c}_{i+1}^k|t_{i+1} = \ell, \alpha_\ell^k, \mathbf{h}_\ell^k)\psi_{j\ell} a_i(j)}{a_{i+1}(\ell)}, \end{aligned}$$

initializing by sampling from $t_I \sim \text{Cat}(a_I)$. The Gibbs sampling updates for $\boldsymbol{\alpha}$ and $\mathbf{H}$ are the same as in the SCC model. The conditional distributions of $\boldsymbol{\pi}$ and $\boldsymbol{\Psi}$ are easily determined due to conjugacy

$$\begin{aligned} p(\boldsymbol{\pi}|\text{rest}) &= \text{Dir}(\boldsymbol{\pi}|q_1,\ldots,q_J), \quad q_j = \frac{\eta}{J} + \mathbb{I}(t_i = j), \\ p(\boldsymbol{\psi}_j|\text{rest}) &= \text{Dir}(\boldsymbol{\psi}_j|\mu\boldsymbol{\phi}_j + \mathbf{m}_j), \end{aligned}$$

where $\mathbf{m}_j$ is the count of transitions from state j to the rest of the states.
The derivation of the sampled based approximation of the predictive distribution, as well as the point estimate method in the second inference problem are straightforward.

4.3 Performance evaluation

4.3.1 Databases

Two HAR databases were created using APDM Opal [2] wearable inertial sensors. These sensors provided synchronized measurements from three axis accelerometers, gyroscopes and magnetometers. The first database contains the measurements from eight different people with two sensors placed on the waist and ankle, whereas the second database contains the measurements from seven different people with four sensors placed on the waist, ankle, chest and wrist. People with ages between 24 and 33 were chosen as subjects for the data acquisition system. Each of them was asked to perform a combination of five different activities: running, walking, standing, sitting and lying (in no particular order) under semi-naturalistic conditions in an indoor environment during a minimum of 20 minutes.[1] In Table 4.1 we show the empirical probabilities of each activity in both databases, that we consider typical from an active person during the day. The activities were labelled with the help of a synchronous video recording and the original data was downsampled to 16 Hz. The orientation correction of the database was performed using the algorithm in [30].

[1]The datasets are available at `http://www.tsc.uc3m.es/dataproy/har/databases.zip`

Table 4.1: Empirical activities distribution in the databases.

Database	Running	Walking	Standing	Sitting	Lying
Two sensors	0.034	0.208	0.296	0.287	0.175
Four sensors	0.026	0.156	0.305	0.300	0.213

We also employ the publicly available database described in [64] and named DaLiAc. This database contains the synchronized measurements from three axis accelerometers and gyroscopes of four wearable inertial sensors placed on the wrist, chest, waist and ankle, similar to our second database. 19 people performed a sequence of 13 different activities (standing, sitting, lying, washing dishes, vacuuming, sweeping, walking, ascending and descending stairs, treadmill running, bicycling (50 watt), bicycling (100 watt) and rope jumping). We merged the two bicycling activities because this distinction is the main source of errors of the system and obscures the comparison among methods. This modification does not change the conclusions of the rest of this work. We downsampled the data to 16 Hz to make a fair comparison between all the databases. As the DaLiAc database does not contain data from magnetometers, we employ the orientation correction algorithm in [35].

Among the multiple classification techniques that provide a posterior probability or soft output (see [61] for a review), we apply HMMs directly on the observations, so no feature selection is required. HMMs offer the advantage of estimating the posterior probability of the activities using the Forward-Backward (FB) algorithm. In particular, we trained a HMM classifier with a Gaussian mixture observation model for each sensor using the standard Baum-Welch algorithm [75]. We configure our HMMs using the structure described in [34], assigning three states per activity. The outputs of the classifiers are obtained following a leave-one-person-out methodology, i.e. the soft outputs C for each person's record correspond to the posterior distribution obtained using the forward-backward algorithm with the HMM that is trained with the rest of the people in the database. This methodology allows us to evaluate the inter-person variability of the database.

The error results from each person and sensor when performing *Maximum a Posteriori* (MAP) estimation over the individual classifiers are shown in Table 4.2.

4.3.2 Baseline models

For the validation of the whole approach, we apply a single classifier trained with all the sensor signals, labelled as joint sensors in Table 4.2. As a baseline measure for classifier combination we use a model that assumes i.i.d. classifiers. Under this model, given a true posterior probability output $c_{ij}^k = p(t_i = j|\mathbf{x}_i^k, \theta^k)$, we directly calculate the posterior probability of the ground truth as the product of the outputs of the soft classifiers

$$p(T|\mathbf{c}^{1:K}) = \prod_{k=1}^{K} \mathbf{c}^k.$$

This model assumes that we have an i.i.d. equiprobable ground truth, $p(T) = \prod_{i=1}^{I} p(t_i)$, $p(t_i) = 1/J$ and i.i.d. classifiers $p(\mathbf{c}_i^{1:K}) = \prod_{k=1}^{K} p(\mathbf{c}_i^k)$. In Table 4.2 we call this baseline model Posterior Probability Combination (PPC) and we also include the error results of the joint sensors model and the error results of performing MAP estimation on the PPC model.

We observe that there exist significant differences in the error results between different sequences and sensors. A significant inter-person variability is present in the databases due to the nature of the experiments. This problem naturally arises in all HAR databases where data from more than one subject is collected. One method to reduce this variability is to employ an increasing number of sensors. Another issue is that the location of the sensors changes between individuals. As an example, right and left ankles were equally used. This obviously leads again to important differences in error performance.

In general, PPC performs better than the joint sensors baseline model. The high dimensionality of the data makes the posterior probability highly multimodal, and finding the optimal solution becomes challenging. Moreover, as we increase the number of sensors, this approach obtains worse error results than the

Table 4.2

Leave-one-person-out test error results of all the databases, with a comparison between each independent classifier, the PPC method and the joint sensors method.

(a) Two sensors database

Person	Waist sensor	Ankle sensor	PPC	Joint sensors
1	0.139	0.349	**0.081**	0.166
2	0.112	0.320	0.086	**0.060**
3	0.046	0.239	**0.027**	0.136
4	**0.140**	0.497	0.299	0.183
5	0.057	0.443	0.094	**0.048**
6	**0.122**	0.374	0.177	0.127
7	0.101	0.185	0.126	**0.012**
8	0.017	0.423	**0.013**	0.191
Average	**0.092**	0.354	0.113	0.116

(b) Four sensors database

Person	Wrist sensor	Chest sensor	Waist sensor	Ankle sensor	PPC	Joint sensors
1	0.390	0.164	0.228	**0.018**	0.056	0.084
2	0.340	0.304	0.249	0.265	**0.174**	0.296
3	0.385	0.286	0.141	0.154	**0.120**	0.256
4	0.248	0.249	0.096	0.143	**0.086**	0.140
5	0.171	0.109	0.200	0.129	**0.016**	**0.016**
6	0.283	**0.024**	0.110	0.093	0.028	0.114
7	0.406	0.182	0.307	**0.059**	0.097	0.279
Average	0.317	0.188	0.190	0.123	**0.083**	0.169

(c) DaLiAc database

Person	Wrist sensor	Chest sensor	Waist sensor	Ankle sensor	PPC	Joint sensors
1	0.099	0.183	0.048	0.117	**0.042**	0.263
2	0.330	0.110	0.339	0.189	0.095	**0.052**
3	0.615	**0.071**	0.273	0.201	0.125	0.176
4	0.162	0.362	0.349	0.798	**0.140**	0.285
5	0.572	0.359	0.275	0.192	0.246	**0.146**
6	0.326	0.086	**0.041**	0.181	0.216	0.103
7	0.605	0.188	0.306	0.188	**0.098**	0.147
8	0.534	0.349	0.386	0.176	0.111	**0.081**
9	0.244	**0.047**	0.162	0.153	0.091	0.171
10	0.494	**0.144**	0.270	0.224	0.152	0.149
11	0.182	**0.054**	0.179	0.139	0.081	0.139
12	0.297	**0.096**	0.454	0.343	0.113	0.322
13	0.367	0.116	0.240	0.114	**0.071**	0.083
14	0.451	0.224	**0.124**	0.176	0.139	0.295
15	0.336	0.248	0.156	0.178	0.102	**0.038**
16	0.272	0.175	0.205	0.261	**0.112**	0.342
17	0.324	**0.084**	0.554	0.321	0.299	0.309
18	0.199	0.155	**0.119**	0.338	0.125	0.232
19	0.109	**0.032**	0.075	0.298	0.147	0.071
Average	0.343	0.162	0.240	0.241	**0.132**	0.179

PPC method.

PPC performs better with four than two sensors due to averaging. We should expect an increase of performance in this baseline model as we increase the number of classifiers. Although in all databases we could choose one of the sensors to classify the different activities, obtaining a similar performance, the high inter-person variability of HAR systems makes this approach intractable, as we would over-fit the solution to our data.

4.3.3 Basic set of activities experiment

We use the first two databases to test the SCC and MSCC models in a low number of activities setting. First we describe some details about the inference, then we compare SCC and MSCC and finally we evaluate their performances in comparison to PPC and the IBCC as proposed in [56]. The IBCC results have been obtained after performing a MAP classification in each of the $\mathbf{c}_i^k$ values in the databases.

As hyper parameters in the SCC model we set $a = 20$ and $b = 10$, $\beta = 20$ and $\kappa = 0.6$ (this prior enforces our belief that our classifiers perform better than random and have most of their mass located at the true class), $\epsilon = 20$ and $\gamma_j = \frac{1}{J}$. In the MSCC model we additionally set $\mu = 20$ and $\phi = 0.8$ (we assume a diagonal dominant transition matrix because of the nature of human activities and the sampling rate in the database), and $\eta = 20$. We have tried several combinations of values for the hyper parameters without any significant differences in the results. This independence on the hyper parameters is a consequence of the large databases used in the experiments, that make the likelihood term in (4.2) dominant.

In the semi-supervised inference setting, we first run 10 iterations of the Gibbs sampler using the labelled sequences to obtain the initial values $\underline{\mathbf{H}}$ and $\boldsymbol{\alpha}$. Then, with both the labelled and the unlabelled sequences we gather 500 samples to estimate the unknown T after a burn-in period of 100 iterations. To build the inductive classifier, we run the Gibbs sampler using exclusively the labelled data. Likewise, after a burn-in period of 100 samples, we collect 500 samples from the posterior probability of $\underline{\mathbf{H}}$ and $\boldsymbol{\alpha}$ to compute the predictive distribution and the

point estimate approximation.

Although we choose the same number of iterations in the Gibbs sampler, the behaviour of both the SCC and MSCC methods differ significantly. In Figure 4.3 we represent the error rate convergence of both algorithms after the 100 burn-in iterations in one sequence of the two sensors database, and Figure 4.4 represents the percentage of the estimated activity samples that varies between consecutive iterations. We can see how averaging is essential for SCC but both averaging and the 500 iterations were unnecessary for the MSCC. This is corroborated in Figures 4.5-4.6, where we represent the estimated activity in the last Gibbs iteration. It becomes evident that the iterations in the MSCC model are more stable than in the SCC model, leading to a faster practical convergence of the algorithm.

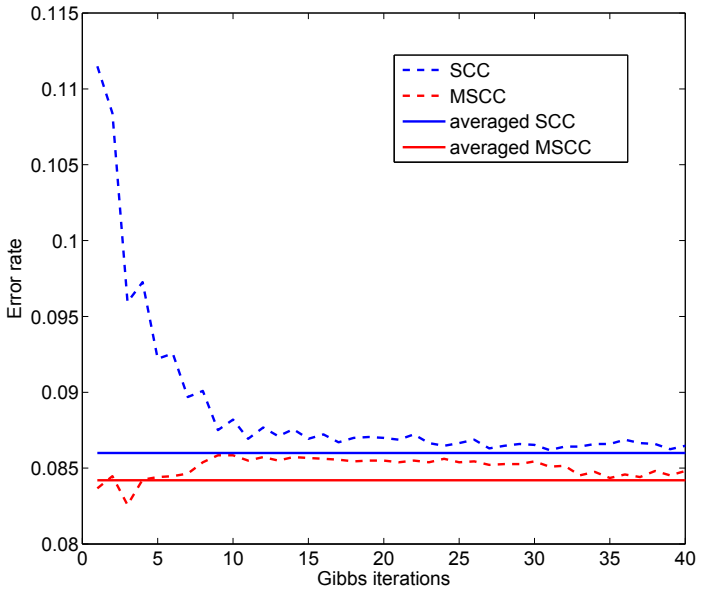

Figure 4.3: Cumulative mean error rate of SCC and MSCC models.

Tables 4.3 to 4.6 show the error rates obtained with PPC, IBCC, and both SCC and MSCC using the three inference methods. The standard deviations of the error rates for all cases are of order 10^{-3} and were not included in the results. As expected, both SCC and MSCC achieve a lower error rate than IBCC, due to the use of soft classifiers instead of hard ones. Also, SCC and MSCC perform

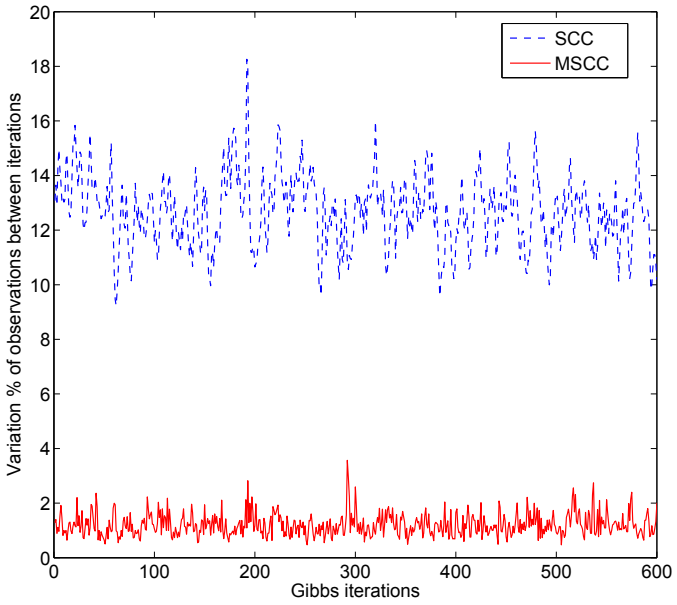

Figure 4.4: Variation percentage of ground truth samples between consecutive Gibbs sampling iterations.

better than PPC, due to the classifier combination structure of both models. The differences between the SCC and MSCC with semi-supervised inference are negligible but, as mentioned above, the run-time complexity of the MSCC can be significantly lower. In the two sensors database there are no differences between the semi-supervised and the inductive methods, but in the four sensors database semi-supervised inference methods perform better. The difference between the semi-supervised and the Gibbs averaging inference methods is due to the change in the paradigm, from transductive to inductive learning. The difference between the Gibbs averaging and the point estimate of the predictive distribution is simply due to the averaging effect. The later differences are smaller in the MSCC due to the stability of the estimated ground truth.

4.3.4 Rich set of activities experiment

We have replicated the experiments performed by the authors in [64] using the DaLiAc database. We used a sliding window of 5s with 50% overlap in the data,

Table 4.3

Leave-one-person-out mean error results of SCC in the two sensors database.

			SCC		
Person	PPC	IBCC	Semi-supervised	Gibbs averaging	Point estimate
1	**0.081**	0.135	0.115	0.115	0.115
2	0.088	**0.066**	**0.066**	**0.066**	**0.066**
3	0.027	0.016	**0.014**	**0.014**	**0.014**
4	0.299	0.140	0.139	**0.138**	**0.138**
5	0.094	0.065	**0.046**	**0.046**	**0.046**
6	0.177	0.112	**0.086**	0.087	0.087
7	0.126	0.126	**0.096**	0.097	**0.096**
8	0.013	0.079	**0.012**	0.013	**0.012**
Average	0.113	0.092	**0.072**	**0.072**	**0.072**

Table 4.4

Leave-one-person-out mean error results of MSCC in the two sensors database.

			MSCC		
Person	PPC	IBCC	Semi-supervised	Gibbs averaging	Point estimate
1	**0.081**	0.135	0.115	0.113	0.113
2	0.088	**0.066**	**0.066**	**0.066**	**0.066**
3	0.027	0.016	**0.013**	**0.013**	**0.013**
4	0.299	0.140	0.137	0.137	**0.136**
5	0.094	0.065	0.041	**0.040**	**0.040**
6	0.177	0.112	**0.084**	0.085	0.085
7	0.126	0.126	**0.095**	**0.095**	**0.095**
8	0.013	0.079	0.011	**0.010**	**0.010**
Average	0.113	0.092	**0.070**	**0.070**	**0.070**

Table 4.5

Leave-one-person-out mean error results of SCC in the four sensors database.

			SCC		
Person	PPC	IBCC	Semi-supervised	Gibbs averaging	Point estimate
1	0.056	**0.008**	0.011	0.010	0.010
2	0.174	0.058	**0.039**	0.052	0.128
3	0.120	0.149	**0.119**	0.120	0.120
4	0.086	**0.013**	**0.013**	0.070	0.065
5	**0.016**	0.017	0.019	0.026	0.026
6	0.028	0.037	**0.018**	0.063	0.087
7	0.097	0.117	0.018	**0.017**	**0.017**
Average	0.083	0.057	**0.034**	0.055	0.065

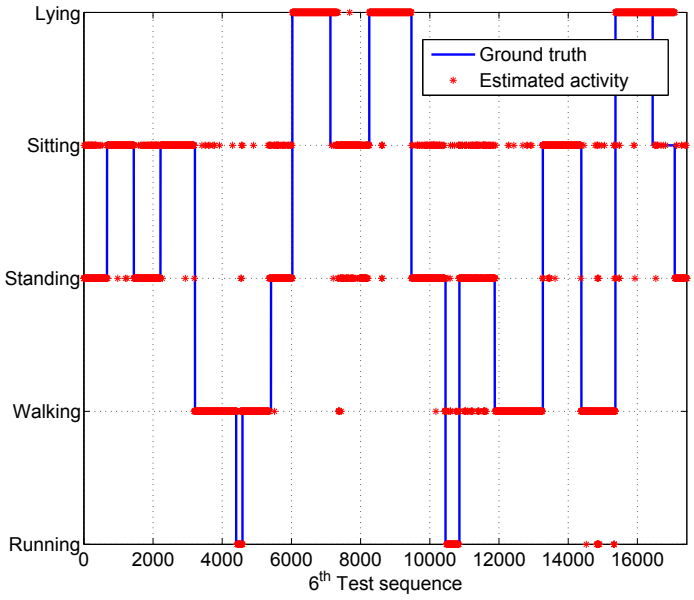

Figure 4.5: Estimated activity in the last Gibbs iteration of the 6^{th} sequence using the SCC.

then we extracted the 152 specified features and we replicated the hierarchical classification algorithm based on different classifiers depending on the group of activities.

For comparison, we used Support Vector Machines (SVMs) for the whole system, as we obtained better performance than using K-Nearest Neighbourss (kNNs) for the walk classifier and AdaBoost for the house classifier.[2] Considering one bicycling activity, the overall error is 5%, estimated from the confusion matrix reported in [64].

[2]The results were obtained with Matlab®2012b version for all the experiments. To compute the SVM we employed the functions *svmtrain* and *svmclassify* from the Bioinformatics Toolbox using a radial basis kernel. We set the scaling factor to 1 and gamma to 1 divided by the number of features, as proposed in [64]. To compute the KNNs we employed the function *knnclassify*, also from the Bioinformatics Toolbox, setting $k = 5$ and using an euclidean distance. To compute the AdaBoost we employed the *adaboost* function from the publicly available spider toolbox for Matlab® (http://people.kyb.tuebingen.mpg.de/spider/main.html). We used KNNs as weak classifiers setting $k = 5$ and performing a 5-folds cross validation.

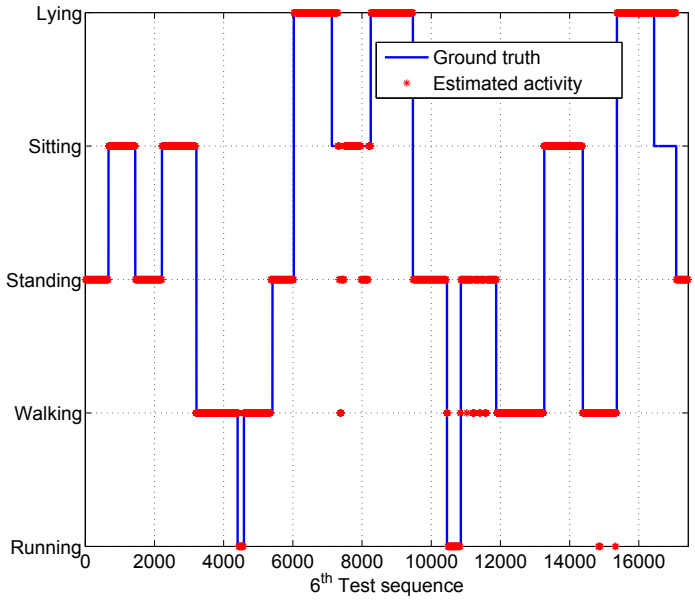

Figure 4.6: Estimated activity in the last Gibbs iteration of the 6^{th} sequence using the MSCC.

For our classifier combination algorithms, we obtain the classifiers as in the previous section. The FB algorithm provides 16 samples of the posterior probability of the activities per second. We calculate the average of the posterior probabilities every 5s with a sliding window with 50% overlap to make a fair comparison between the different methods. The rest of the settings of the experiment is unaltered; i.e., the hyper parameters and the semi-supervised inference conditions. We also compare the MSCC results to the PPC and IBCC models in the same conditions.

In Table 4.7 we show the error results of all the algorithms. We only present the results of the best case algorithm on the previous experiment, the MSCC method with the semi-supervised inference setting.

We observe that MSCC is also the best method employing a rich set of activities database. We obtain an increase of performance of around 3% with respect to the IBCC model and the algorithm in [64], which is consistent with the basic activities experiment results. Replicating the experiments in the aforementioned conditions,

Table 4.6

Leave-one-person-out mean error results of MSCC in the four sensors database.

			MSCC		
Person	PPC	IBCC	Semi-supervised	Gibbs averaging	Point estimate
1	0.056	**0.008**	0.009	0.009	0.009
2	0.174	0.058	**0.036**	0.068	0.070
3	**0.120**	0.149	**0.120**	0.121	0.121
4	0.086	0.013	**0.012**	0.037	0.038
5	**0.016**	0.017	0.018	0.017	0.017
6	0.028	0.037	**0.018**	0.021	0.022
7	0.097	0.117	0.017	**0.016**	**0.016**
Average	0.083	0.057	**0.033**	0.041	0.042

Table 4.7

Leave-one-person-out error performance results of PPC, IBCC, MSCC and hierarchical classifiers method for the DaLiAc database with four sensors and 19 different sequences

Person	Joint	PPC	IBCC	MSCC	[64]
1	0.263	0.042	0.025	**0.023**	0.071
2	0.052	0.095	0.037	**0.010**	0.044
3	0.176	0.125	0.042	**0.018**	0.054
4	0.285	0.140	0.256	**0.012**	0.068
5	0.146	0.246	0.323	0.258	**0.111**
6	0.103	0.216	0.028	**0.023**	**0.023**
7	0.147	0.098	0.137	0.047	**0.043**
8	0.081	0.111	0.053	**0.011**	0.104
9	0.171	0.091	**0.012**	0.051	0.031
10	0.149	0.152	0.089	**0.059**	0.106
11	0.139	0.081	0.026	**0.015**	0.062
12	0.322	0.113	0.024	**0.006**	0.066
13	0.083	0.071	0.033	**0.008**	0.073
14	0.295	0.139	**0.029**	0.038	0.061
15	0.038	0.102	0.034	**0.023**	0.093
16	0.342	0.112	0.137	**0.088**	0.133
17	0.309	0.299	0.060	**0.008**	0.118
18	0.232	0.125	**0.058**	0.165	0.118
19	0.071	0.147	0.009	**0.004**	0.082
Mean	0.179	0.132	0.074	**0.046**	0.077

Table 4.8

Confusion matrix of the MSCC algorithm for the DaLiAc database. The code of the activities is: ST = standing, SI = sitting, LY = lying, WD = washing dishes, VC = vacuuming, SW = sweeping, WK = walking, AS = ascending stairs, DS = descending stairs, RU = running, BC = bicycle and RJ = rope jumping

	ST	SI	LY	WD	VC	SW	WK	AS	DS	RU	BC	RJ
ST	446	0	2	3	0	0	0	0	0	0	0	0
SI	0	430	5	20	0	0	0	0	0	0	0	0
LY	4	26	426	0	0	0	0	0	0	0	0	0
WD	4	0	0	931	3	0	0	0	0	0	0	0
VC	0	0	0	5	426	29	0	0	0	0	0	0
SW	0	0	0	0	41	698	0	0	0	0	0	0
WK	0	0	0	0	0	21	1951	64	7	0	0	0
AS	0	0	0	0	0	0	22	295	2	0	0	0
DS	0	0	0	0	0	0	2	18	255	0	0	0
RU	0	0	0	0	0	0	6	0	0	897	7	0
BC	0	96	0	0	2	19	0	0	0	0	1727	0
RJ	0	0	0	0	0	0	0	0	0	0	10	245

we obtained an error of 0.077 instead of the 0.050 reported in [64]. Nonetheless, we still improve their original results with our algorithm.

In Table 4.8 we show the confusion matrix of the MSCC algorithm for this experiment. We observe that all activities are mostly classified correctly, with less than an 8% error in any activity. This shows the robustness of our algorithms with respect to the different activities of the database, where our worst case activity is better by a 7% than the worst case activity reported by [64].

Table 4.9

Mean error results of the Daliac database when one of the sensors stops working at test.

Sensor	PPC	IBCC	MSCC	[64]	SVM + IBCC
Wrist	0.134	0.082	**0.051**	0.271	0.126
Chest	0.186	0.111	**0.052**	0.681	0.138
Waist	0.161	0.087	**0.042**	0.816	0.136
Ankle	0.138	0.117	**0.079**	0.735	0.115
No failure	0.132	0.074	**0.042**	0.077	0.131

4.3.5 Robustness

We have also tested the robustness of the system against sensor failures. In both SCC and MSCC, we only need to include an indicator variable a_i^k in the observation models (4.1). This variable is one if the classifier is active or zero if the classifier is inactive for a given time instant. The probability of the classifiers C given the rest of the parameters is now

$$p(C|T, \boldsymbol{\alpha}, \underline{\mathbf{H}}) = \prod_{j=1}^{J} \prod_{k=1}^{K} \prod_{i:t_i=j} \mathrm{Dir}\left(\mathbf{c}_i^k | \alpha_j^k, \mathbf{h}_j^k\right)^{\mathbb{I}(a_i^k=1)}.$$

According to the previous equation, when a sensor becomes inactive we simply do not consider its data in the observation model.

We have performed an experiment where we simulate the malfunction of one of the sensors during the test stage by deleting its data only in the test sequences. The results are shown in Table 4.9 where we consider all cases with one sensor broken.

The loss in performance in the PPC and IBCC is similar to the MSCC algorithm, as both methods are based on the combination of classifiers. Losing one of the sensors in the combination algorithms implies a loss in performance of around 4% whereas in the method based on hierarchical classifiers it is much larger. As expected, all schemes based on sensor combination are more robust against failures than single classifier schemes. In fact, if we compare the robustness of using the model in [64] against training one SVM per sensor and then combining the

classification using the IBCC model, we can see that there exists almost no loss in performance.

4.4 Conclusions

We have proposed new Bayesian techniques to combine soft outputs classifiers for the person-independent HAR problem. Our work extends the IBCC model proposed in [56] by using soft output classifiers to deal with a low number of sensors and a first-order Markov ground truth to capture the dynamics of the human activities. Our methods exhibit consistent error rate reduction and higher robustness against sensor failures when compared with a single classifier that employs all the sensor signals using different real HAR databases. These results show the advantages of classifier combination models over single classifier designs. When compared with the IBCC, both the SCC and the MSCC models take advantage of the soft output model and lower the error rate of the IBCC, but the real advantage here is the low run-time complexity of the MSCC due to algorithm stability and the use of the FFBS algorithm.

The contribution of this work is not a new classifier, but a classifier combination model for the HAR problem. Although we used HMMs as individual classifiers, the model can be employed without any modification for combining homogeneous or heterogeneous individual classifiers of any kind.

The remaining challenge is to increase the robustness of our methods against inter-person variability. Multi-task learning schemes may help to address this problem. Another future extension consists of considering more robust classifiers, which can be easily incorporated in the combination algorithms due to their ability to work with non-probabilistic soft or hard-output classifiers.

5

Energy efficiency in HAR Systems

5.1 Introduction

The energy efficiency of smart-phones is an important topic in many sensing related applications [84]. The HAR system needs to share the battery resources between all the applications of the smart-phones. However, the embedded sensors are one of the main sources of battery consumption, decreasing dramatically the battery of the devices during synchronous data acquisition at high sampling frequencies. Many different approaches exist to reduce the data acquisition of the wearable sensors. The authors in [65] study the effect of the sampling frequency of the smart-phones in a HAR system. They demonstrate that an activity recognition based on low sampling frequencies is feasible for long-term activity monitoring. In [109] the authors study the influence of the performed activities over the energy efficiency of the sensors. They develop a HAR system where they dynamically adapt the sam-

pling frequency and the classification features employed on the system depending on the performed activities. They obtain a priori the configuration parameters for every activity and select them accordingly while tracking continuously the ongoing activities. Both approaches consider a continuous monitoring, adapting the sampling frequency of the sensors depending on the specifications. However, with some long-term activities, e.g. sleeping, data acquisition can be halted completely without losing information, further enhancing the energy efficiency of the sensors. In this work, an active sensing solution based on the acquisition of intermittent data windows when the uncertainty over the performed activities exceeds a critical value is proposed.

Determining the optimal position of the sensors on the body is also vital to increase the energy efficiency of the sensors. Not all the body positions are informative while recognizing some of the activities, and most of the data acquired becomes either useless or redundant. In that respect, the authors in [10] analyze the recognition performance of IMUs located on several body areas, choosing the sensor that maximizes the accuracy of the system while employing many different sensor position configurations. Furthermore, they demonstrate that the performance of each device strongly depends on the measured human activities. A common alternative to optimizing the position of a unique sensor on the body is the combination of the information provided by multiple sensors. In [111] the authors implement a Naive Bayes classifier that fuses the individual classification results of each sensor. This approach selects the best sensor for the recognition of every activity independently. A more elaborate approach is the Bayesian Classifier Combination problem studied in [56], where the authors infer simultaneously the ground truth and the model parameters of the individual classifiers. In [78], a soft output combination of individual classifiers is considered, where the posterior probability of the activities provided by each individual classifier is employed to infer the parameters of the combination model. In these approaches, all the sensors are employed in the recognition and no consideration over the energy consumption of the sensors is addressed. A different approach is considered in [107], where

the authors propose a hierarchical sensor management strategy to recognize user activities and to detect activity transitions, deciding which sensors to use at any given time. Instead of the ad hoc approximation considered in [107], a systematic approach to decide when to perform the data acquisition with a HAR system based on Markov processes is followed.

The main purpose of this paper is to develop a general framework that achieves the joint optimization of the energy efficiency of the sensors, the number of sensors employed and its location and the performance of the HAR system. In a single sensor long-term monitoring HAR system, when the sensor acquires data with a sampling frequency of tens of Hertz, the mass of the posterior probability distribution of the activities given the data is located on a single activity during most of the time. Accordingly, the uncertainty (or entropy) of the performed activity is low in general. Modelling the data with a Markov process, the posterior probability distribution of the activities when data is not available approaches asymptotically to the stationary distribution of the process and the entropy of this posterior increases. A novel active sensing strategy exploiting this property is proposed, i.e., to stop acquiring observations when this entropy is low, and to estimate the next time instant when the entropy reaches a certain threshold and new data needs to be acquired. This is a reasonable assumption in a long-term activity recognition system, where some activities are performed continuously during extended periods of time and only a few observations are needed to recognize these activities.

Additionally, an optimization problem for the multiple sensors framework with energy constraints is proposed. The maximization of the mutual information between the activities and the models of the sensor observations provides the optimal sensor configuration to be employed [21]. Depending on the activities distribution and the energy constraints of the devices, the sensor configuration is adapted dynamically to provide the maximum amount of information to the HAR system. There exists several approaches to numerically characterize this mutual information. In [9] the authors find an expression of the lower and upper bound of the mutual information by employing an arbitrary auxiliary distribution that approx-

imates the conditional probability distribution of the mutual information. Alternatively, the authors in [48] find an expression of the lower and upper bound of the entropy of a random variable with Gaussian probability distribution. However, as reflected in this work, finding a tractable expression of the bounds for a general observation model is difficult. An approximation of the mutual information using Monte Carlo (MC) sampling [66] is considered instead, since the model of the observations is known in the energy efficiency setting.

5.2 Problem Statement

The implementation of a HAR system with K classifiers to recognize J different activities is considered. The sequence of observations $X^k = \{\mathbf{x}_n^k\}_{n=1}^N$ of length N, indicates the confidence of the k^{th} classifier over the sequence of performed activities $S = \{s_n\}_{n=1}^N$, with $s_n \in \{1, 2, \cdots, J\}$ and $\mathbf{x}_n^k$ belonging to the $J-1$ dimensional probability simplex. All the classifiers X are conditionally independent given S, and the conditional likelihood of the observation $\mathbf{x}_n$ for all the classifiers given the performed activity s_n is

$$p(\mathbf{x}_n|s_n) = \prod_{k=1}^{K} p(\mathbf{x}_n^k|s_n)^{z_n^k}, \tag{5.1}$$

where $z_n^k \in \{0, 1\}$ is an indicator variable representing whether the k^{th} classifier is active or inactive at time instant n. The observation model proposed for the classifiers is a Dirichlet distribution with parameters $\boldsymbol{\gamma}$, $p(\mathbf{x}_n^k|s_n) = \mathrm{Dir}(\mathbf{x}_n^k|\boldsymbol{\gamma}_{s_n}^k)$, since the observations of the classifiers are defined over a probability simplex. A HAR system modelled as a Hidden Markov Model (HMM) with observations X, hidden variables S and indicator variables Z over the observations is proposed (Fig. 5.1). This HMM is characterized by an initial probabilities distribution, a transition matrix $\Psi \in \mathbb{R}^{J \times J}$ and the model of the observations $p(\mathbf{x}_n^k|s_n)$ [89]. The parameters of the model are trained in advance, and the remaining problem is how to perform the data acquisition for a new sequence of observations.

The energy efficiency problem consists of selecting the optimal sensor configuration, $Z = \{z_n^k\}_{n=1}^N$ for $k = \{1, \cdots, K\}$ that minimizes the data acquisition of the

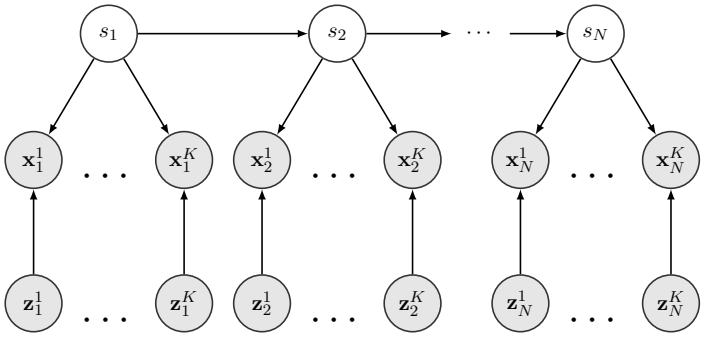

Figure 5.1: Graphical model of the HMM employed in the HAR system.

HAR system while maximizing its performance. No further assumptions are made over Z, i.e., data is acquired from a combination of the available sensors during some periods of time while no data is acquired during other periods. This configuration needs to be updated online, adapting the strategy employed depending on the observations provided by the classifiers, the posterior of the activities given these observations and the energy constraints of the system.

The energy efficiency method is divided in two separate optimization problems: 1) Selecting the temporal scheme for the data acquisition, 2) Selecting the sensor configuration to be employed during each data acquisition given our knowledge over the posterior of the activities, the sensor observation models and the energy constraints. The algorithm optimizes both problems iteratively to perform the activity recognition until the long-term monitoring concludes.

5.3 Active Sensing Strategy

The objective of the active sensing strategy consists of selecting an optimal temporal acquisition policy that reduces the data acquired by the sensors while maximizing the performance of the HAR system. Under this strategy the system decides when to acquire a new window of data observations based on the previous data acquisitions and the posterior of the activities given this data. When the system is modelled using HMMs, the posterior probability of the activity s_n given all the observations until time instant n, $\mathbf{x}_{1:n}$, and the sensor configuration employed,

$\mathbf{z}_{1:n}$, is defined by the forward step of the Forward-Backward algorithm [89]. The posterior probability distribution of any activity in the future given the current known data is computed by marginalizing all the activities between the current time instant n and the future time instant $n + n_0$. This marginalization results in an expression of the posterior distribution $p(s_{n+n_0}|\mathbf{x}_{1:n}, \mathbf{z}_{1:n})$ given by the n_0 power of the transition matrix multiplied by the posterior distribution of s_n

$$p(s_{n+n_0}|\mathbf{x}_{1:n}, \mathbf{z}_{1:n}) = \Psi^{n_0} p(s_n|\mathbf{x}_{1:n}, \mathbf{z}_{1:n}). \tag{5.2}$$

Equation (5.2) represents the evolution of the system's knowledge over the activities when data is not available. As n_0 increases, $p(s_{n+n_0}|\mathbf{x}_{1:n}, \mathbf{z}_{1:n})$ becomes less informative and the uncertainty over the activities increases. This uncertainty can be measured as the entropy of a random variable S_{n+n_0} with probability distribution $p(s_{n+n_0}|\mathbf{x}_{1:n}, \mathbf{z}_{1:n})$[1].

The active sensing strategy reduces to finding the next time instant n_0 where data needs to be acquired by restricting $H(S_{n+n_0})$ to a certain threshold H_0

$$H(S_{n+n_0}) < H_0. \tag{5.3}$$

When the entropy exceeds this threshold, the posterior distribution becomes unreliable and new data is acquired to reduce the entropy. Unfortunately, (5.3) is intractable in general, so three different numerical approximations are proposed to obtain n_0 [76].

5.3.1 Activity independent approximation

The transition matrix of a HMM is a stochastic matrix with a limiting distribution or stationary state $\mathbf{p}$. The existence of this limiting distribution implies that $H(S_{n+n_0})$ must converge

$$\lim_{n_0 \to \infty} H(S_{n+n_0}) = -\mathbf{p}^T \log(\mathbf{p}) \triangleq H_p.$$

[1]The definition of the entropy of a random variable X with probabilities $p(\mathbf{x})$ is $H(X) \triangleq -\sum_{i=1}^{|x|} p(x_i) \log p(x_i)$, where $|x|$ is the dimensionality of the vector $\mathbf{x}$.

The entropy of S_{n+n_0} is asymptotically independent of $p(s_n|\mathbf{x}_{1:n}, \mathbf{z}_{1:n})$, only depending on the structure of the transition matrix. Furthermore, every stochastic matrix contains at least one eigenvalue that is equal to 1 and the largest absolute value of all its eigenvalues is also 1. The limiting distribution $\mathbf{p}$ corresponds to the eigenvector with eigenvalue $\lambda_1 = 1$. Computing the eigendecomposition of the n-power of the transition matrix in (5.2),

$$\Psi^n = U\Lambda^n U^{-1},$$

where U is the matrix of eigenvectors of Ψ and Λ is a diagonal matrix containing its eigenvalues $\{\lambda_1, \cdots, \lambda_J\}$, with $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_J|$, it is observed that the n-power only affects Λ.

A naive solution of the active sensing method involves finding the value of n_0 where $p(s_{n+n_0}|\mathbf{x}_{1:n}, \mathbf{z}_{1:n}) = \mathbf{p}$ and the system reaches its stationary point. This approach, called activity independent method, consists of finding the minimum value of n_0 such as $|\lambda_2|^{n_0} < \epsilon$, where ϵ controls the precision of the approximation,

$$n_0^p = \left\lfloor \frac{\log(\epsilon)}{\log(|\lambda_2|)} \right\rfloor. \tag{5.4}$$

Under (5.4), for $j > 1$, all $|\lambda_j|$ are negligible, and the only contribution of $\Psi^{n_0^p}$ is the limiting distribution $\mathbf{p}$ with eigenvalue $\lambda_1 = 1$.

Fig. 5.2 shows an example of $H(S_{n+n_0})$ as a function of n_0 for a transition matrix with 10 states. For any transition matrix, this function is not monotonically increasing in general, with several local maxima in the interval $[0, n_0^p]$ that are larger than the value of the entropy in the limiting distribution. When $n_0 > n_0^p$, $H(S_{n+n_0})$ converges to the entropy of the limiting distribution H_p. When $n_0 = 0$, $H(S_{n+n_0})$ is just the entropy of the posterior of s_n, $H(S_n)$. Choosing n_0^p as the time instant when new data must be acquired is a naive approximation, since there exists in general an interval of values in $[0, n_0^p]$ where the uncertainty is larger than H_p, and n_0 must be chosen in the first interval where (5.3) holds.

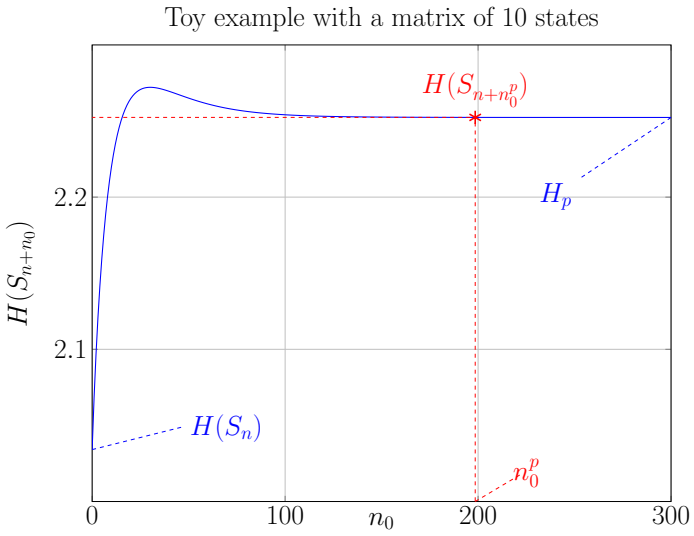

Figure 5.2: Representation of $H(S_{n+n_0})$ for a transition matrix with 10 states.

5.3.2 Threshold method

A direct approach to solve the active sensing problem consists of finding numerically the value of n_0 that satisfies (5.3), given that a suitable value of H_0 is chosen. When a small value of H_0 is considered the time between data acquisitions decreases and the energy consumption of the devices increases. When H_0 is too large, this value is not reached since the entropy is bounded above and data is never acquired again. In the threshold method, H_0 is chosen in terms of the entropy of the limiting distribution, $H_0 = cH_p$, where $c \in [0, 1]$ is a parameter that controls the distance to H_p. Consequently, the maximum value of n_0 is found such as (5.3) holds,

$$H(S_{n+n_0}) < cH_p.$$

When $c = 1$, the threshold is equal to H_p. Depending on the structure of the transition matrix, this could be problematic. Particularly, if $H(S_{n+n_0})$ is monotonically increasing in terms of n_0, this solution reduces to the activity independent approximation, i.e., $n_0 = n_0^p$. Choosing $c < 1$ a new data window is acquired faster, and the posterior distribution of the activities can be updated properly. For small values of c, data is acquired too quickly, and the data acquisition reduction becomes negligible. Increasing the value of c, the distance between time intervals when new

data is acquired becomes larger, reducing the number of observations employed and thus increasing the energy efficiency of the devices.

5.3.3 Line intersection method

Fixing a constant threshold H_0 is not always the best approach as it does not consider the curvature of the function $H(S_{n+n_0})$. An alternative to directly finding a solution for (5.3) involves selecting n_0 as the value of the intersection between two lines instead, a constant line defined by the entropy of the limiting distribution $y_1(n) = H_p$ and the tangent line of the entropy of the activities at some $n_k \in [0, n_0^p]$

$$y_2(n) = y_k + m(n - n_k),$$

where y_k is the entropy of the posterior at n_k, $H(S_{n+n_k})$, and m is the slope of the tangent line, corresponding to the derivative of the entropy at n_k, $H'(S_{n+n_k})$

$$\begin{aligned} y_k &= -(\Psi^{n_k}\boldsymbol{\alpha}_n)^T \log(\Psi^{n_k}\boldsymbol{\alpha}_n), \\ m &= -(U\log(\Lambda)\Lambda^{n_k}U^{-1}\boldsymbol{\alpha}_n)^T \log(\Psi^{n_k}\boldsymbol{\alpha}_n), \end{aligned}$$

with $\boldsymbol{\alpha}_n \triangleq p(s_n|\mathbf{x}_{1:n}, \mathbf{z}_{1:n})$.

Computing the intersection between both lines and replacing y_k and m by its expressions, the value of the intersection point is obtained

$$n = n_k + \frac{\mathbf{p}^T \log(\mathbf{p}) - (\Psi^{n_k}\boldsymbol{\alpha}_n)\log(\Psi^{n_k}\boldsymbol{\alpha}_n)}{(U\log(\Lambda)\Lambda^{n_k}U^{-1}\boldsymbol{\alpha}_n)^T \log(\Psi^{n_k}\boldsymbol{\alpha}_n)},$$

constantly increasing n_k as long as the value of the slope is greater than a certain value. With this approach, the limitation of the entropy is given by the slope of the tangent line $y_2(n)$, and the stopping value changes depending on the activities being performed. When n approaches the location of the maximum of the entropy, the slope will decrease and the intersections of both lines will provide a good estimate of n_0.

5.4 Sensor selection framework

Choosing an optimal sensor configuration during the data acquisition allows the HAR system to increase the energy efficiency of the devices. For every time window

obtained by the active sensing strategy the system needs to find the optimal sensor configuration to perform the data acquisition, based on the posterior distribution of the activities computed at a given time instant $n' = n + n_0$, $\boldsymbol{\alpha}_{n'} \triangleq p(s_{n'}|\mathbf{x}_{1:n}, \mathbf{z}_{1:n})$. This decision must be made before acquiring new data, so the observations $\mathbf{x}_{n'}$ are considered unknown.

The sensor selection problem can be formulated as a decision problem using the Bayesian Experimental Design theory [21], where an utility function reflecting the objective of the experiment is selected. The utility function considered in this problem is the mutual information between the known posterior probabilities of the activities $S_{n'}$ and the unknown sensor observations $X_{n'}$

$$I(S_{n'}; X_{n'}) \triangleq \int \sum_{s_{n'}} p(\mathbf{x}_{n'}, s_{n'}) \log \frac{p(\mathbf{x}_{n'}|s_{n'})}{p(\mathbf{x}_{n'})} d\mathbf{x}_{n'}, \tag{5.5}$$

with $p(\mathbf{x}_{n'}|s_{n'})$ defined in (5.1). The optimal sensor configuration depends on the energy constraints imposed on the data acquisition for the different sensors. The maximization of the mutual information in (5.5) results in the most informative sensor configuration $\mathbf{z}^*_{n'}$ for a particular posterior distribution $\boldsymbol{\alpha}_{n'}$

$$\begin{aligned} \underset{\mathbf{z}_{n'}}{\text{maximize}} \quad & I(S_{n'}; X_{n'}) \\ \text{subject to} \quad & \mathbf{c}^T \mathbf{z}_{n'} \leq E, \forall n' \end{aligned}$$

with $\mathbf{c} = \{c_1, \cdots, c_K\}$ being the data acquisition costs related to each classifier and E the maximum amount of energy consumption per time instant allowed by the system. The simplest solution to the sensor selection problem with energy constraints consists of choosing beforehand all the possible combinations of the sensor configurations that fulfil the energy constraints $\mathbf{z}^0_{n'} = \left\{ z^k_{n'} | \mathbf{c}^T \mathbf{z}_{n'} \leq E \right\}$, and then obtain $\mathbf{z}^*_{n'}$ as the strategy that maximizes the mutual information of all the available sensor combinations

$$\mathbf{z}^*_{n'} = \underset{\mathbf{z}^0_{n'}}{\arg\max}\ I(S_{n'}; X_{n'}). \tag{5.6}$$

Unfortunately, integral (5.5) cannot be solved analytically, since the likelihood of the observations of a sensor $p(\mathbf{x}_{n'})$ is a mixture distribution model depending on

the activities and the number of classifiers, and its logarithm is not tractable. Two different approaches are considered to compute numerically the mutual information: 1) Computing the bounds of the mutual information, 2) Approximating the mutual information by MC sampling.

5.4.1 Mutual Information Bounds

The mutual information can be expressed as the difference between two entropies

$$I(S_{n'}; X_{n'}) = H(S_{n'}) - H(S_{n'}|X_{n'}).$$

Finding the bounds of the entropies is an equivalent problem to finding the bounds of the mutual information. In [48] the authors obtain an expression of the lower and upper bounds on the entropy of a random variable with Gaussian probability distribution. In this section, the same approach is considered for the case where the conditional likelihood of the observations is a Dirichlet distribution.

Upper bound (UB)

An expression of the upper bound is obtained using the properties of the mutual information [24], i.e., the mutual information between two random variables must be positive

$$I(S_{n'}; X_{n'}) = H(S_{n'}) - H(S_{n'}|X_{n'}) \geq 0.$$

$H(S_{n'}|X_{n'})$ cannot be computed in this problem. However, $S_{n'}$ is discrete, and the entropy of a discrete random variable is always positive, so an expression of the upper bound is given by

$$I(S_{n'}; X_{n'}) \leq H(S_{n'}) = I_{UB}(S_{n'}; X_{n'}).$$

The upper bound only depends on the posterior distribution of the activities and is independent of the likelihood of the observations. As the posterior becomes less informative, the upper bound increases until reaching its maximum value when $\boldsymbol{\alpha}_{n'}$ is the uniform distribution over the J activities, $I_{UB}(S_{n'}; X_{n'}) = \log J$. This bound always exists and it is independent of the observation likelihood model, so a solution of (5.6) always exists.

Lower bound (LB)

In the expression of the mutual information as a difference of entropies

$$I(S_{n'}; X_{n'}) = H(X_{n'}) - H(X_{n'}|S_{n'}),$$

the conditional entropy $H(X_{n'}|S_{n'})$ is given by the entropy of a Dirichlet distribution, so only an expression of the lower bound of $H(X_{n'})$ is needed. Applying Jensen's inequality in $H(X_{n'})$

$$\begin{aligned} H(X_{n'}) &\geq -\sum_{s_{n'}} \alpha_{n'}(s_{n'}) \log(f), \\ f &\triangleq \int p(\mathbf{x}_{n'}|s_{n'}) p(\mathbf{x}_{n'}) d\mathbf{x}_{n'}. \end{aligned} \tag{5.7}$$

where $\alpha_{n'}(s_{n'})$ represents the element $s_{n'}$ of the posterior $\boldsymbol{\alpha}_{\boldsymbol{n'}}$. Although this lower bound can be computed for several distributions, e.g., Gaussian distributions [48], this integral is intractable for many others. In (5.7), the integral f is the product of Dirichlet distributions, which is also a Dirichlet distribution

$$f \propto \sum_{j=1}^{J} \alpha_{n'}(j) \prod_{k=1}^{K} \int \mathrm{Dir}(\mathbf{x}_{n'}^k | \boldsymbol{\gamma}_\ell^k)^{z_{n'}^k} d\mathbf{x}_{n'}^k,$$

with $\boldsymbol{\gamma}_\ell^k = \boldsymbol{\gamma}_{s_{n'}}^k + \boldsymbol{\gamma}_j^k - 1$, and $\boldsymbol{\gamma}_j^k$ the parameters of the sensor k for a particular activity j. This integral is only defined when $\boldsymbol{\gamma}_{s_{n'}}^k + \boldsymbol{\gamma}_j^k > 1$, since the support of the parameters of the Dirichlet distribution is the positive real numbers. This is a critical restriction in the sensor selection problem, since the classifiers are usually confident in this setting and the parameters of the observation model range between 0 and 1. For $\boldsymbol{\gamma}_\ell^k > 0$ a closed form solution of the lower bound of $H(X_{n'})$, and consequently of the lower bound of the mutual information is obtained

$$\begin{aligned} f &= \sum_{j=1}^{J} \alpha_{n'}(j) \left(\prod_{k=1}^{K} \frac{B(\boldsymbol{\gamma}_\ell^k)}{B(\boldsymbol{\gamma}_{s_{n'}}^k) B(\boldsymbol{\gamma}_j^k)} \right)^{z_{n'}^k}, \\ H_{LB}(X_{n'}) &= -\sum_{s_{n'}} \alpha_{n'}(s_{n'}) \log(f), \\ I_{LB}(S_{n'}; X_{n'}) &= H_{LB}(X_{n'}) - H(X_{n'}|S_{n'}), \end{aligned}$$

where the beta function of a vector $\mathbf{v} \in \mathbb{R}^D$ is defined as

$$B(\mathbf{v}) \triangleq \frac{\prod_{d=1}^{D} \Gamma(v_d)}{\Gamma(\sum_{d=1}^{D} v_d)},$$

with $\Gamma(\cdot)$ being the gamma function.

There exist no guarantees in the previous results implying that $I_{LB}(S_{n'}; X_{n'}) \geq 0$. In many cases, the best lower bound is $I_{LB}(S_{n'}; X_{n'}) = 0$, which is not an informative solution, since the observation model of the sensor is irrelevant when computing the bounds of the mutual information.

5.4.2 Monte Carlo approximation

The computation of the analytical solution of the mutual information bounds results in an uninformative solution for the sensor selection problem. An alternative involving an approximation of its value for any $p(s_{n'}|\mathbf{x}_{1:n}, \mathbf{z}_{1:n})$ and $p(\mathbf{x}_{n'})$ is considered. The mutual information in (5.5) can be expressed as the expectation with respect to the joint distribution of the variables

$$I(S_{n'}; X_{n'}) = \mathbb{E}_{p(s_{n'}, \mathbf{x}_{n'})} \left[\log \frac{p(\mathbf{x}_{n'}|s_{n'})}{p(\mathbf{x}_{n'})} \right]. \tag{5.8}$$

In the sensor selection problem the parameters of the likelihood model are trained in advance, and the posterior distribution of the activities is known. This implies that the joint distribution model is completely characterized, and the expectation in (5.8) can be approximated by MC sampling,

$$I(S_{n'}; X_{n'}) \approx \frac{1}{M} \sum_{m=1}^{M} \log \frac{p(\mathbf{x}_{n'm}|s_{n'm})}{p(\mathbf{x}_{n'm})}, \tag{5.9}$$

where M groups of samples $(s_{n'm}, \mathbf{x}_{n'm})$ are obtained from the joint probability distribution $p(s_{n'}, \mathbf{x}_{n'})$. As the number of samples increases, the MC approximation approaches the true value of the mutual information asymptotically. Furthermore, this approximation is independent of the observation likelihood model. As long as the models of $p(\mathbf{x}_{n'})$ and $p(\mathbf{x}_{n'}|s_{n'})$ are fully characterized, (5.9) holds.

To compute this approximation, M instances of $s_{n'm}$ are sampled from the posterior distribution $p(s_{n'}|\mathbf{x}_{1:n}, \mathbf{z}_{1:n})$. Then M instances of $\mathbf{x}_{n'm}$ are sampled from the corresponding mixture component $p(\mathbf{x}_{n'}|s_{n'm})$. Finally, an approximation of the mutual information is obtained evaluating (5.9) for every group of samples $(s_{n'm}, \mathbf{x}_{n'm})$. The solution of the sensor selection problem (5.6) with the MC

approach consists of choosing the sensor configuration $\mathbf{z}_{n'}^*$ that maximizes the approximation of the mutual information (5.9).

5.5 Experiments

The evaluation of the energy efficiency framework is conducted using the publicly available database described in [64] and named DaLiAc. This database contains the synchronized data of four wearable sensors (placed on the wrist, chest, waist and ankle) from 19 different people while performing a sequence of 12 activities (standing, sitting, lying, washing dishes, vacuuming, sweeping, walking, ascending and descending stairs, treadmill running, bicycling and rope jumping). The sensors provide data from three axis accelerometers and gyroscopes, and is processed using the sensor orientation correction technique described in [30]. This data is modelled using HMMs, and the soft outputs of the classifiers X are obtained estimating the posterior probability of the activities using the Forward-Backward algorithm (FB). In particular, a HMM classifier with a Gaussian mixture observation model for each sensor is trained using the Baum-Welch algorithm [89]. Three states and two mixture components are assigned per activity, following the configuration described in [34]. The parameters $\boldsymbol{\gamma}$ of the Dirichlet conditional observation model (5.1) and the transition matrix Ψ are trained using the MSCC method described in [78].

The duration of the window when data is acquired by the sensors is a fixed parameter for all the algorithms. Three different window sizes are considered in the experiments, $W = \{5, 10, 20\}$ seconds. When a window of data is obtained, the data acquisition ceases, and the active sensing algorithm decides the next time instant n_0 when the sensors need to acquire a new window of observations. In the threshold algorithm, three different values of $c = \{0.7, 0.8, 0.9\}$ are used. In the line intersection algorithm, the algorithm stops when the value of the slope of the line y_2, is less than 0.1, 0.01 and 0.001 respectively. In the activity independent algorithm, the values $\epsilon = \{0.1, 0.01, 0.001\}$ are employed.

The MC expression obtained in (5.9) for the sensor selection is evaluated using the Mean Square Error (MSE) between its true value and the approximation

for every sensor. The true value of the mutual information is reached when the number of samples considered is extremely large. In these experiments, 10^9 samples is assumed to be a large enough value. Sampling from the joint distribution $p(S, X)$ is performed with a sample size ranging from 10^1 to 10^7 samples. 1000 independent simulations are performed, and the average of the MSE of the set of simulations is computed and represented in Fig. 5.3. The MSE decreases

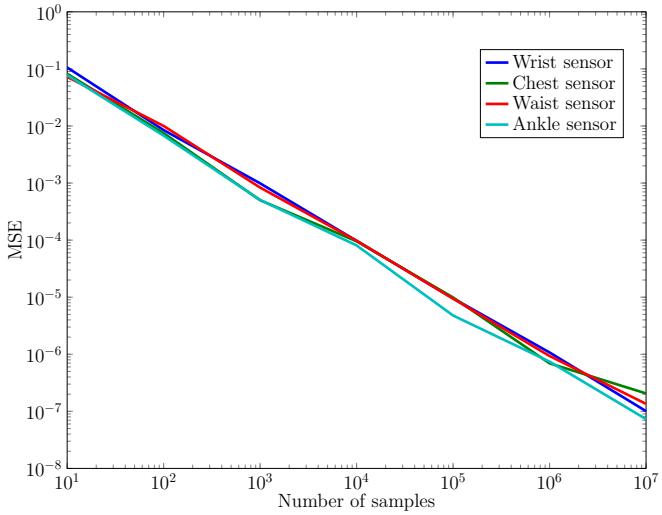

Figure 5.3: Logarithmic representation of the MSE between the true values of the mutual information for each sensor and the MC approximation.

logarithmically with the logarithm of the number of samples considered in the approximation for all the sensors. In this work, the number of samples considered for the rest of the experiments is 10^5, since the MSE between the true and approximated mutual informations is less than 10^{-5} in this case. Other alternatives to sample from (5.8) have been implemented, like Rejection Sampling or Importance Sampling [66]. However, neither of these methods seem to improve the MSE with the number of samples, so sampling directly from the distribution is the chosen method. While other sampling methods could find a better approximation with less samples, employing an efficient MC sampling method is out of the scope of this work.

Fig. 5.4 shows a comparison of the precision loss between all the methods in

terms of the number of observations employed in the system. Decreasing the number of samples acquired reduces the performance of the system. However, under the same conditions, the loss in precision is not heavily influenced by the reduction in window size. It is more important to update the model with new observations when the entropy increases than to acquire large windows of observations, since the entropy is practically zero during these windows.

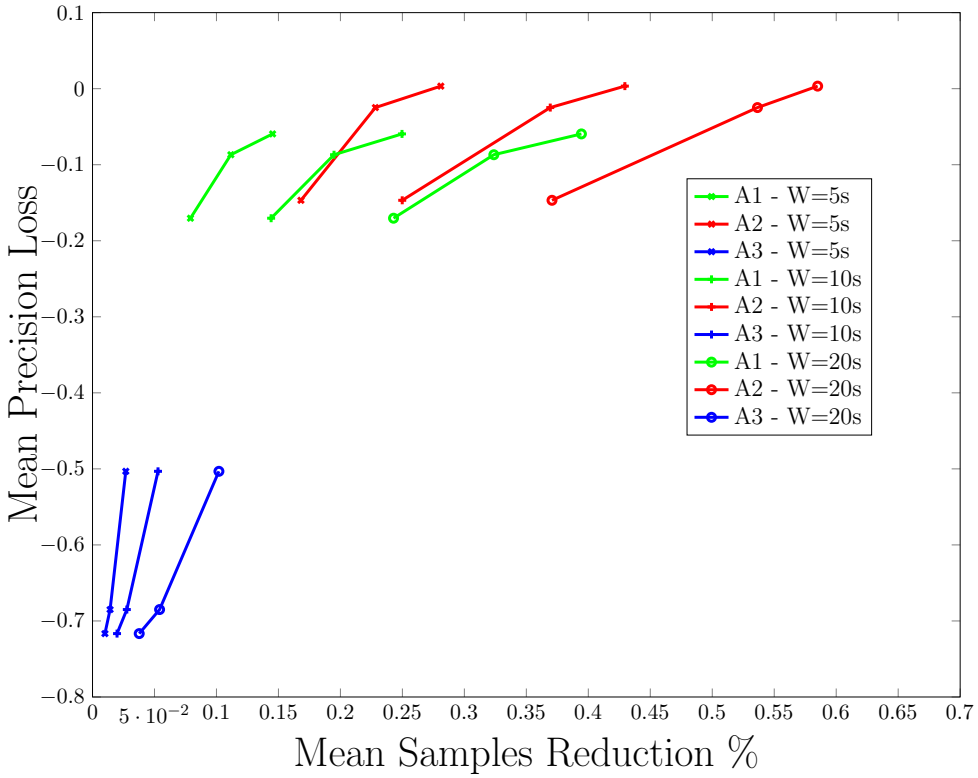

Figure 5.4: Comparison of all the algorithms in Section 5.3 using three different window sizes. A1 is the threshold algorithm, A2 is line intersection algorithm and A3 is the activity independent algorithm.

Table 5.1 shows in more detail the evaluation of the energy efficiency framework for all the different acquisition algorithms when no energy restrictions are assumed on the sensors. The estimation of the performed activities is compared with the case where all the data and sensors are used in the activity recognition, using the

MAP of the posterior probabilities combination (PPC)

$$p(S|X) = \prod_{k=1}^{K} p(S|X^k).$$

Table 5.1

Average data reduction and accuracy error obtained with all the active sensing algorithms. A1 corresponds to the threshold algorithm, A2 to the Line intersection algorithm and A3 to the activity independent algorithm.

Algorithms	Window = 5s		Window = 10s		Window = 20s	
	Reduction of data %	Accuracy error	Reduction of data %	Accuracy error	Reduction of data %	Accuracy error
A1 - 0.7	85.5	0.177	75.0	0.169	60.6	0.155
A1 - 0.8	88.8	0.204	80.5	0.220	67.6	0.164
A1 - 0.9	92.1	0.288	85.6	0.242	75.7	0.223
A2 - 1e-1	**71.9**	**0.114**	57.0	0.123	41.5	0.100
A2 - 1e-2	77.2	0.142	63.1	0.133	46.4	0.128
A2 - 1e-3	83.2	0.264	75.0	0.284	62.9	0.268
A3 - 1e-1	97.3	0.620	94.7	0.608	89.8	0.580
A3 - 1e-2	98.6	0.802	97.2	0.816	94.6	0.789
A3 - 1e-3	99.0	0.834	98.0	0.859	96.2	0.865
PPC	0.0	0.132	0.0	0.132	0.0	0.132

The results show that the best model in terms of precision loss is the line intersection algorithm, maintaining or even increasing the performance, though the number of samples employed is in general larger than in the other models. The activity independent algorithm performs much worse than the others, since the posterior of the activities is not considered while computing the next time instant when new data must be acquired.

The evolution of the entropy of the posterior of the activities strongly depends on the transition matrix of the system. With the MSCC method employed to obtain it, the transition matrix becomes more confident with the amount of data employed to train each of the activities. This implies that the evolution of the entropy becomes slower for a specific posterior when the data for that activity increases. This effect can be observed in Fig. 5.5 for the threshold algorithm. It is observed that the entropy of activities that are overrepresented present a slower

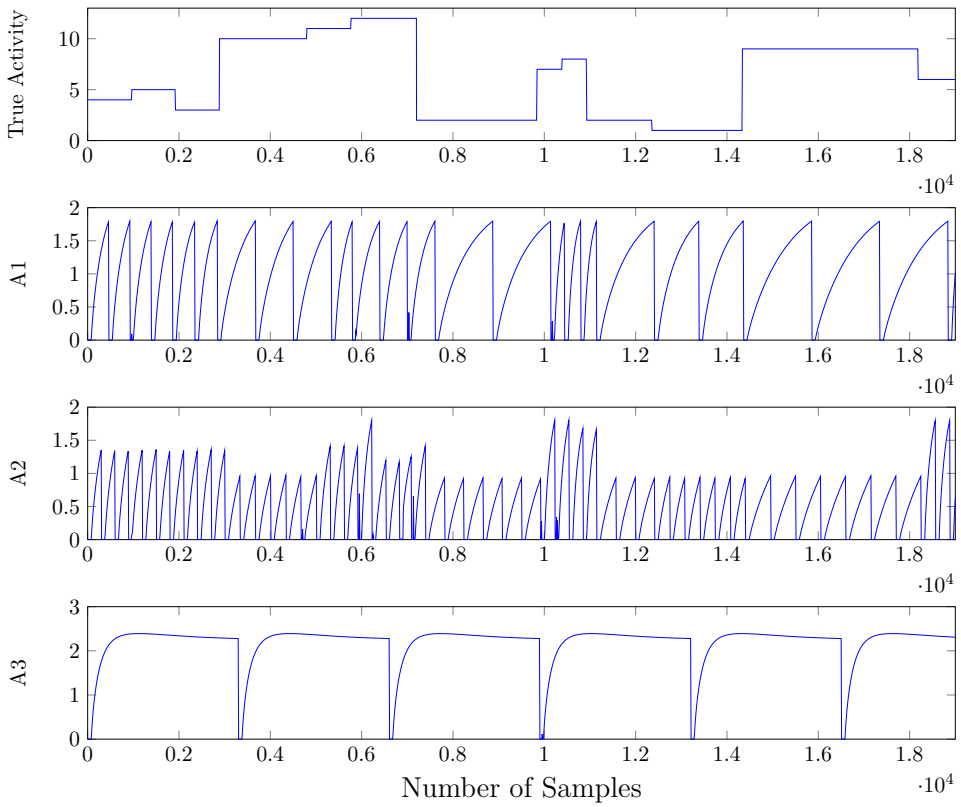

Figure 5.5: Entropy evolution of the posterior of the activities for each data acquisition algorithm (A1 = Threshold algorithm, A2 = Line intersection algorithm, A3 = Activity independent algorithm) employing the first sequence of the database. The true activity labels, from 1 to 12 are: 1-running, 2-walking, 3-standing, 4-sitting, 5-lying, 6-rope jumping, 7-ascending stairs, 8-descending stairs, 9-bicycling, 10-washing dishes, 11-vacuuming, 12-sweeping.

evolution, while the under-represented activities evolve faster.

There exists several differences in the entropy evolution between the different active sensing strategies. In the threshold algorithm the entropy increases until a certain fixed value, and a new window of data is acquired. The number of windows employed for each activity differs considerably depending on the performed activities. In the line intersection algorithm, the next time window is chosen when the derivative of the entropy, i.e. the slope, is less than 0.01. Depending on the shape of the entropy function, this method stops at different values. The number of windows acquired in this model is larger in general, leading to shorter periods

of time where there is no data acquisition and consequently to reduce the loss in precision of the system. The activity independent method reduces to a sampling method where the distance between data acquisitions is constant, since it is independent on the activities. Discarding the information of the activities in the data acquisition leads to poor results in the HAR system performance.

Table 5.2 shows an evaluation of the effects of the energy restrictions imposed on the sensors. When all the sensors present the same energy restrictions but only a given number of sensors can be used per data acquisition, the sensor selection algorithm finds the optimal configuration depending on the performed activities. It can be observed that the performance of the HAR system degrades with the reduction on the number of sensors employed. However, since the number of sensors employed decreases, the amount of data employed decreases too. Depending on the specifications of the problem it is possible to find an optimal working point for the HAR system in terms of the data employed and the performance.

Table 5.2

Average data reduction and accuracy error for the best data acquisition algorithm (Line intersection algorithm with $m < 0.1$ and a window of 5 seconds) for all the energy restriction cases.

# Sensors	Reduction of data %	Accuracy Error
1	92.6	0.239
2	85.8	0.160
3	79.1	0.136
4	71.9	0.114

Table 5.3 shows the probability of choosing each sensor depending on the energy constraints of the problem. When there exist no restrictions and all sensors can be selected, the algorithm combines the information of all the sensors. When some decisions over the number of sensors must be performed, the algorithm prefers some of the sensors (waist) among others (wrist) and the performance of the system degrades since less data is employed in the recognition task.

Table 5.3
Probability of choosing each sensor for all the energy restrictions considered and all sequences.

# Sensors	Wrist	Chest	Waist	Ankle
1	0	0.362	0.549	0.090
2	0.041	0.399	0.428	0.132
3	0.143	0.326	0.326	0.205
4	0.250	0.250	0.250	0.250

5.6 Conclusions

A general framework for the joint optimization of the energy efficiency of wearable sensors, the number of sensors and their location and the performance of a HAR system has been proposed. This work shows that employing the maximum entropy of the posterior of the activities, the data acquisition can be reduced while maintaining the performance of the recognition system. An active sensing algorithm with three different approaches has been implemented and evaluated, emphasizing the importance of updating the model with new data when the entropy increases. In addition, an optimization problem for the multiple sensors case with energy constraints was implemented using a MC approximation of the mutual information between the observations and the posterior probability of the activities. The sensor configuration that maximizes this mutual information provides the maximum amount of information to HAR system. The data acquisition employing the proposed methodology is reduced by a 72% while maintaining the performance of the system when no energy restrictions are considered and all the sensors are employed. When the energy constraints force the HAR system to choose other sensor configurations, the performance decreases by a 10% in the limit where only one of the sensors is employed.

6

Conclusions

6.1 Summary

The main objective of this thesis is the development of novel techniques to address several open problems in the Markov modelling implementation of a Human Activity Recognition (HAR) system employing the raw signals provided by wearable sensors. In this chapter, we summarize the contributions of this thesis, and we also describe some possible lines for future research.

- We have presented an orientation correction algorithm for the sensor misplacement on the body when using raw signals from Inertial Measurement Units (IMUs). This algorithm was implemented with quaternion operations allowing for a much faster computation of the rotations of the signals with respect to the reference system. Furthermore, numerical results in Chapter 2

show a substantial improvement in the activity recognition in comparison with other orientation correction methods.

- In Chapter 3, we have developed a discriminative spectral learning of Hidden Markov Models (HMMs) with discrete observations as an alternative training method. Our work is an extension of the spectral learning of HMM presented in [46], where we develop an algorithm to perform activity recognition with discriminative HMMs. We show how the low computation time and complexity and the possibility of obtaining directly the probability of the activities make it specially useful for the implementation of online HAR systems and an alternative to the Baum-Welch algorithm while training HMMs.

- We perform the inference of the Bayesian combination of soft-output classifiers in a HAR system. Our work extends the Independent Bayesian Classifier Combination (IBCC) model proposed in [56] by using soft output classifiers to deal with a low number of sensors and a first-order Markov ground truth to capture the dynamics of the human activities. The results in Chapter 4 show consistent error rate reduction and higher robustness against sensor failures when compared with a single classifier that employs all the sensor signals using different publicly available HAR databases.

- Finally, a general framework for the joint optimization of the energy efficiency of wearable sensors, the number of sensors and their location and the performance of a HAR system has been proposed. In Chapter 5 we show that employing the maximum entropy of the posterior of the activities, the data acquisition can be reduced while maintaining the performance of the recognition system. An active sensing algorithm [76] with three different approaches has been implemented and evaluated, emphasizing the importance of updating the model with new data when the entropy increases. In addition, we implemented an optimization problem for the multiple sensors case with energy constraints using a Monte Carlo (MC) approximation of the mutual information between the observations and the posterior proba-

bility of the activities. The sensor configuration that maximizes this mutual information provides the maximum amount of information to HAR system. The data acquisition employing the proposed methodology is reduced by a 72% while maintaining the performance of the system when no energy restrictions are considered and all the sensors are employed. When the energy constraints force the HAR system to choose other sensor configurations, the performance decreases by a 10% in the limit where only one of the sensors is employed.

6.2 Future Lines

Our work also suggests several paths for further research in the Markov modelling of a HAR system with wearable sensors. We provide below a list with what we consider are some of the main potential future research lines.

Complete orientation correction. The data processing algorithm proposed in Chapter 2 computes the acceleration referred to the person frame, and the orientation of the person frame with respect to the earth frame. However, the complete orientation of the person where we obtain the direction of movement is not computed. This is an interesting future line, since it allows for tracking methods based completely on IMUs and correction applications of simple Global Positioning System (GPS) systems.

Discriminative spectral learning with continuous observations. The algorithm described in Chapter 3 is implemented to work with discrete data, where several approximations were performed to employ the data provided by the wearable sensors, which are continuous observations in nature. There exists several extensions of the spectral learning for HMMs for continuous observations using kernels [95, 97]. However, to our best knowledge, an algorithm working with a discriminative version of the HMM is not yet developed, and it would be an important improvement in the implementation of fast learning methods for HAR systems.

Hybrid of features and classifiers combination. Most of the approaches of

the literature in data fusion either combine the data provided by the sensors or by the classifiers. However, a combination of classifiers and signals is not considered. As an example, in a smart-phone it can be implemented a HAR system of basic activities and combine the output with raw signals from other devices, like location or phone usage among others. This is a topic related with data processing from heterogeneous sources, and performing the inference for an hybrid combination of features and classifiers could improve the activity recognition or even extend the number and complexity of the activities to be recognized in the HAR system.

Adaptive HAR systems. The models developed in this thesis consider a general model of the activities for every user. However, there exists several differences between the physical activities performed by young people, elder people or people with physical diseases. Developing an adaptive HAR system based on a general model that adapts to every specific human group would allow for a better activity recognition.

References

[1] Amiigo multiple sensor system. `https://amiigo.com/`. Accessed: 2015-05-22.

[2] Apdm inc. opal technical specification. `http://www.apdm.com/`.

[3] D Abowd, Anind K Dey, Robert Orr, and Jason Brotherton. Context-awareness in wearable and ubiquitous computing. *Virtual Reality*, 3(3):200–211, 1998.

[4] Jake K Aggarwal and Quin Cai. Human motion analysis: A review. In *Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE*, pages 90–102. IEEE, 1997.

[5] Jake K Aggarwal and Michael S Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011.

[6] Norhafizan Ahmad, Raja Ariffin Raja Ghazilla, Nazirah M Khairi, and Vijayabaskar Kasi. Reviews on various inertial measurement unit (imu) sensor applications. *International Journal of Signal Processing Systems*, 1(2):256–262, 2013.

[7] Kerem Altun and Billur Barshan. Human activity recognition using inertial/magnetic sensor units. In *International Workshop on Human Behavior Understanding*, pages 38–51. Springer, 2010.

[8] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L Reyes-Ortiz. Human activity recognition on smartphones using a multiclass

hardware-friendly support vector machine. In *International Workshop on Ambient Assisted Living*, pages 216–223. Springer, 2012.

[9] Dieter-Michael Arnold, H-A Loeliger, Pascal O Vontobel, Aleksandar Kavcic, and Wei Zeng. Simulation-based computation of information rates for channels with memory. *IEEE Transactions on Information Theory*, 52(8):3498–3508, 2006.

[10] Louis Atallah, Benny Lo, Rachel King, and Guang-Zhong Yang. Sensor positioning for activity recognition using wearable accelerometers. *IEEE transactions on biomedical circuits and systems*, 5(4):320–329, 2011.

[11] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga. Activity Recognition Using Inertial Sensing for Healthcare, Wellbeing and Sports Applications: A Survey. In *International Conference on Architecture of Computing Systems*, pages 1–10, 2010.

[12] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In *International Workshop on Human Behavior Understanding*, pages 29–39. Springer, 2011.

[13] Ling Bao and Stephen S Intille. Activity recognition from user-annotated acceleration data. In *International Conference on Pervasive Computing*, pages 1–17. Springer, 2004.

[14] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Statist.*, pages 164–171, 1970.

[15] Bernard Berelson and Gary A Steiner. Human behavior: An inventory of scientific findings. 1964.

[16] Christopher M.. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[17] P. Bonato. Wearable Sensors and Systems. *IEEE Engineering in Medicine and Biology Magazine*, 29(3):25–36, 2010.

[18] J.E. Bortz. A new mathematical formulation for strapdown inertial navigation. *Aerospace and Electronic Systems, IEEE Transactions on*, (1):61–66, 1971.

[19] Carlijn VC Bouten, Karel TM Koekkoek, Maarten Verduin, Rens Kodde, and Jan D Janssen. A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity. *IEEE Transactions on Biomedical Engineering*, 44(3):136–147, 1997.

[20] A. Bulling, U. Blanke, and B. Schiele. A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors. *ACM Computing Surveys*, 46(3), January 2014.

[21] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[22] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu. Sensor-Based Activity Recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 42(6):790–808, 2012.

[23] Yen-Ping Chen, Jhun-Ying Yang, Shun-Nan Liou, Gwo-Yun Lee, and Jeen-Shing Wang. Online classifier construction algorithm for human activity detection using a tri-axial accelerometer. *Applied Mathematics and Computation*, 205(2):849–860, 2008.

[24] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[25] Geng Cui, Man Leung Wong, and Hon-Kwong Lui. Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science*, 52(4):597–612, 2006.

[26] A. P. Dawid and A. M. Skene. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society*, 28(1):pp. 20–28, 1979.

[27] Joseph Decuir. Introducing Bluetooth Smart: Part 1: A look at both classic and new technologies. *Consumer Electronics Magazine, IEEE*, 3(1):12–18, 2014.

[28] Joseph Decuir. Introducing Bluetooth Smart: Part II: Applications and updates. *Consumer Electronics Magazine, IEEE*, 3(2):25–29, 2014.

[29] Wan-Yu Deng, Qing-Hua Zheng, and Zhong-Min Wang. Cross-person activity recognition using reduced kernel extreme learning machine. *Neural Networks*, 53:1–7, 2014.

[30] V. Elvira, A. Nazabal-Renteria, and A. Artes-Rodriguez. A novel feature extraction technique for human activity recognition. In *2014 IEEE Workshop on Statistical Signal Processing (SSP)*, pages 177–180. IEEE, 2014.

[31] Jochen Fahrenberg, Friedrich Foerster, Manfred Smeja, and WOLFGANG MÜLLER. Assessment of posture and motion by multichannel piezoresistive accelerometer recordings. *Psychophysiology*, 34(5):607–612, 1997.

[32] Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine learning*, 32(1):41–62, 1998.

[33] B. Florentino, N. O'Mahony, and A. Artés-Rodríguez. Hierarchical dynamic model for human daily activity recognition. *Proceedings of BIOSIGNALS*, 2012.

[34] B. Florentino-Liano, N. O'Mahony, and A. Artés-Rodríguez. Hierarchical dynamic model for human daily activity recognition. In *BIOSIGNALS*, pages 61–68, 2012.

[35] B. Florentino-Liano, N. O'Mahony, and A. Artes-Rodriguez. Long term human activity recognition with automatic orientation estimation. In *2012*

IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pages 1–6. IEEE, 2012.

[36] F Foerster, M Smeja, and J Fahrenberg. Detection of posture and motion by accelerometry: a validation study in ambulatory monitoring. *Computers in Human Behavior*, 15(5):571–583, 1999.

[37] S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer, New York, 2006.

[38] Jurgen V Gael, Yee W Teh, and Zoubin Ghahramani. The infinite factorial hidden markov model. In *Advances in Neural Information Processing Systems*, pages 1697–1704, 2009.

[39] Zoubin Ghahramani and Michael I Jordan. Factorial hidden markov models. In *Advances in Neural Information Processing Systems*, pages 472–478, 1996.

[40] John J Guiry, Pepijn van de Ven, John Nelson, Lisanne Warmerdam, and Heleen Riper. Activity recognition with smartphone support. *Medical Engineering & Physics*, 36(6):670–675, 2014.

[41] C. W. Han, S. J. Kang, and N. S. Kim. Implementation of hmm-based human activity recognition using single triaxial accelerometer. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 93(7):1379–1383, 2010.

[42] Zhenyu He and Lianwen Jin. Activity recognition from acceleration data based on discrete consine transform and svm. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 5041–5044. IEEE, 2009.

[43] Katherine A Heller, Yee W Teh, and Dilan Görür. Infinite hierarchical hidden markov models. In *International Conference on Artificial Intelligence and Statistics*, pages 224–231, 2009.

[44] Javier Hernández, Raúl Cabido, Antonio S Montemayor, and Juan José

Pantrigo. Human activity recognition based on kinematic features. *Expert Systems*, 31(4):345–353, 2014.

[45] W. Hörmann, J. Leydold, and G. Derflinger. *Automatic Nonuniform Random Variate Generation*. Springer, 2003.

[46] D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

[47] Derek Hao Hu, Xian-Xing Zhang, Jie Yin, Vincent Wenchen Zheng, and Qiang Yang. Abnormal activity recognition based on hdp-hmm models. In *IJCAI*, pages 1715–1720, 2009.

[48] Marco F Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe D Hanebeck. On entropy approximation for gaussian mixture random vectors. In *Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008. IEEE International Conference on*, pages 181–188. IEEE, 2008.

[49] Tâm Huynh and Bernt Schiele. Analyzing features for activity recognition. In *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*, pages 159–163. ACM, 2005.

[50] APDM Inc. http://www.apdm.com/.

[51] H. Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000.

[52] Ahmad Jalal, Shaharyar Kamal, and Daijin Kim. A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments. *Sensors*, 14(7):11735–11759, 2014.

[53] F. Jelinek. *Statistical methods for speech recognition*. MIT press, 1997.

[54] Tim LM Kasteren, Gwenn Englebienne, and Ben JA Kröse. Human activity recognition from wireless sensor network data: Benchmark and software.

Activity recognition in pervasive intelligent environments, pages 165–186, 2011.

[55] Eunju Kim, Sumi Helal, and Diane Cook. Human activity recognition and pattern discovery. *IEEE Pervasive Computing*, 9(1), 2010.

[56] H. C. Kim and Z. Ghahramani. Bayesian Classifier Combination. In *International Conference on Artificial Intelligence and Statistics*, 2012.

[57] Mustafa Kose, Ozlem Durmaz Incel, and Cem Ersoy. Online human activity recognition on smart phones. In *Workshop on Mobile Sensing: From Smartphones and Wearables to Big Data*, volume 16, pages 11–15, 2012.

[58] Narayanan C Krishnan and Diane J Cook. Activity recognition on streaming sensor data. *Pervasive and mobile computing*, 10:138–154, 2014.

[59] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.

[60] Javier Ortiz Laguna, Angel García Olaya, and Daniel Borrajo. A dynamic sliding window approach for activity recognition. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 219–230. Springer, 2011.

[61] Ó. D. Lara and M. A. Labrador. A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys & Tutorials*, 15(3):1192–1209, 2013.

[62] O.D. Lara and M.A. Labrador. A survey on human activity recognition using wearable sensors. *Communications Surveys & Tutorials, IEEE*, 15(3):1192–1209, 2013.

[63] Jonathan Lester, Tanzeem Choudhury, and Gaetano Borriello. A practical approach to recognizing physical activities. In *International Conference on Pervasive Computing*, pages 1–16. Springer, 2006.

[64] Heike Leutheuser, Dominik Schuldhaus, and Bjoern M Eskofier. Hierarchical, multi-sensor based classification of daily life activities: comparison with state-of-the-art algorithms using a benchmark dataset. *PloS one*, 8(10):e75196, 2013.

[65] Yunji Liang, Xingshe Zhou, Zhiwen Yu, and Bin Guo. Energy-efficient motion related activity recognition on mobile devices for pervasive healthcare. *Mobile Networks and Applications*, 19(3):303–317, 2014.

[66] Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[67] Juliana Lockman, Robert S Fisher, and Donald M Olson. Detection of seizure-like movements using a wrist accelerometer. *Epilepsy & Behavior*, 20(4):638–641, 2011.

[68] D.J.C. MacKay. Ensemble learning for hidden markov models. Technical report, Technical report, Cavendish Laboratory, University of Cambridge, 1997.

[69] S.O.H. Madgwick, A.J.L. Harrison, and R. Vaidyanathan. Estimation of IMU and MARG orientation using a gradient descent algorithm. In *Rehabilitation Robotics (ICORR), 2011 IEEE International Conference on*, pages 1–7. IEEE, 2011.

[70] Miloš Marjanović, Miloš Kovačević, Branislav Bajat, and Vít Voženílek. Landslide susceptibility assessment using svm machine learning algorithm. *Engineering Geology*, 123(3):225–234, 2011.

[71] F.L. Markley, Y. Cheng, J.L. Crassidis, and Y. Oshman. Averaging quaternions. *Journal of Guidance, Control, and Dynamics*, 30(4):1193–1197, 2007.

[72] L. Martino, J. Read, and D. Luengo. Independent Doubly Adaptive Rejection Metropolis Sampling. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2014.

[73] H. Q. Minh, M. Cristani, A. Perina, and V. Murino. A regularized spectral algorithm for hidden markov models with applications in computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2384–2391. IEEE, 2012.

[74] Subhas Chandra Mukhopadhyay. Wearable sensors for human activity monitoring: A review. *IEEE sensors journal*, 15(3):1321–1330, 2015.

[75] K. P. Murphy. *Machine Learning*. MIT Press, 2012.

[76] Alfredo Nazábal and Antonio Artés. Active sensing in human activity recognition. In *International Work-Conference on Artificial Neural Networks*, pages 157–166. Springer, 2017.

[77] Alfredo Nazábal and Antonio Artés-Rodríguez. Discriminative spectral learning of hidden markov models for human activity recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 1966–1970. IEEE, 2015.

[78] Alfredo Nazábal, Pablo García-Moreno, Antonio Artés-Rodríguez, and Zoubin Ghahramani. Human activity recognition by combining a small number of classifiers. *IEEE journal of biomedical and health informatics*, 20(5):1342–1351, 2016.

[79] Wei Niu, Jiao Long, Dan Han, and Yuan-Fang Wang. Human activity detection and recognition for video surveillance. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 1, pages 719–722. IEEE, 2004.

[80] Andrew Ofstad, Emmett Nicholas, Rick Szcodronski, and Romit Roy Choudhury. Aampl: Accelerometer augmented mobile phone localization. In *Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments*, pages 13–18. ACM, 2008.

[81] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[82] Juha Parkka, Miikka Ermes, Panu Korpipaa, Jani Mantyjarvi, Johannes Peltola, and Ilkka Korhonen. Activity classification using realistic data from wearable sensors. *IEEE Transactions on information technology in biomedicine*, 10(1):119–128, 2006.

[83] Shyamal Patel, Konrad Lorincz, Richard Hughes, Nancy Huggins, John Growdon, David Standaert, Metin Akay, Jennifer Dy, Matt Welsh, and Paolo Bonato. Monitoring motor fluctuations in patients with parkinson's disease using wearable sensors. *IEEE Transactions on Information Technology in Biomedicine*, 13(6):864–873, 2009.

[84] Gian Paolo Perrucci, Frank HP Fitzek, and Jörg Widmer. Survey on energy consumption entities on the smartphone platform. In *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, pages 1–6. IEEE, 2011.

[85] Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45, 2006.

[86] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, January 2010.

[87] S.J. Preece, J.Y. Goulermas, L.P.J. Kenney, and D. Howard. A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. *Biomedical Engineering, IEEE Transactions on*, 56(3):871–879, 2009.

[88] L. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[89] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, pages 257–286, 1989.

[90] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L Littman. Activity recognition from accelerometer data. In *Aaai*, volume 5, pages 1541–1546, 2005.

[91] C. P. Robert, G. Celeux, and J. Diebolt. Bayesian estimation of hidden markov chains: A stochastic implementation. *Statistics & Probability Letters*, 16(1):77–83, 1993.

[92] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1):1–39, 2010.

[93] A.M. Sabatini. Estimating three-dimensional orientation of human body parts by inertial/magnetic sensing. *Sensors*, 11(2):1489–1525, 2011.

[94] K. Shoemake. Animating rotation with quaternion curves. *ACM SIGGRAPH computer graphics*, 19(3):245–254, 1985.

[95] S. M. Siddiqi, B. Boots, and G. J. Gordon. Reduced-rank hidden markov models. *arXiv*, 2009.

[96] E. Simpson, S. J. Roberts, I. Psorakis, and A. Smith. Dynamic Bayesian Combination of Multiple Imperfect Classifiers. In T V Guy, M Kárny, and D H Wolpert, editors, *Decision Making and Imperfection*, pages 1–35. Springer-Verlag, 2013.

[97] L. Song, B. Boots, S. M. Siddiqi, G. J. Gordon, and A. J. Smola. Hilbert space embeddings of hidden markov models. In *Proceedings of the 27th international conference on machine learning (ICML)*, pages 991–998, 2010.

[98] Xing Su, Hanghang Tong, and Ping Ji. Activity recognition with smartphone sensors. *Tsinghua Science and Technology*, 19(3):235–249, 2014.

[99] Adi L Tarca, Vincent J Carey, Xue-wen Chen, Roberto Romero, and Sorin Drăghici. Machine learning and its applications to biology. *PLoS Comput Biol*, 3(6):e116, 2007.

[100] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476), 2006.

[101] Douglas L Vail, Manuela M Veloso, and John D Lafferty. Conditional random fields for activity recognition. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, page 235. ACM, 2007.

[102] Annie Vallières and Charles M Morin. Actigraphy in the assessment of insomnia. 2003.

[103] TLM van Kasteren, Gwenn Englebienne, and BJA Kröse. Human activity recognition from wireless sensor network data: Benchmark and software. In *Activity Recognition in Pervasive Intelligent Environments*, pages 165–186. Springer, 2011.

[104] C.J. Van Rijsbergen. *Information retrieval*. Butterworths, 1979.

[105] P.H. Veltink, P. Slycke, J. Hemssems, R. Buschman, G. Bultstra, and H. Hermens. Three dimensional inertial sensing of foot movements for automatic tuning of a two-channel implantable drop-foot stimulator. *Medical engineering & physics*, 25(1):21–28, 2003.

[106] Aiguo Wang, Guilin Chen, Jing Yang, Shenghui Zhao, and Chih-Yung Chang. A comparative study on human activity recognition using inertial sensors in a smartphone. *IEEE Sensors Journal*, 16(11):4566–4578, 2016.

[107] Yi Wang, Jialiu Lin, Murali Annavaram, Quinn A Jacobson, Jason Hong, Bhaskar Krishnamachari, and Norman Sadeh. A framework of energy efficient mobile sensing for automatic user state recognition. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, pages 179–192. ACM, 2009.

[108] P. C. Woodland and D. Povey. Large scale discriminative training of hid-

den markov models for speech recognition. *Computer Speech & Language*, 16(1):25–47, 2002.

[109] Zhixian Yan, Vigneshwaran Subbaraju, Dipanjan Chakraborty, Archan Misra, and Karl Aberer. Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach. In *Wearable Computers (ISWC), 2012 16th International Symposium on*, pages 17–24. Ieee, 2012.

[110] Rong Yang and Baowei Wang. Pacp: A position-independent activity recognition method using smartphone sensors. *Information*, 7(4):72, 2016.

[111] P. Zappi, C. Lombriser, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, and G. Tröster. Activity recognition from on-body sensors: Accuracy-power trade-off by dynamic sensor selection. In *European Conference on Wireless Sensor Networks*, pages 17–33, 2008.

[112] Cha Zhang and Yunqian Ma. *Ensemble machine learning*, volume 1. Springer, 2012.

[113] Ting Zhang, Jiang Lu, Fei Hu, and Qi Hao. Bluetooth low energy for wearable sensor-based healthcare systems. In *Healthcare Innovation Conference (HIC), 2014 IEEE*, pages 251–254. IEEE, 2014.

[114] C. Zhu and W. Sheng. Recognizing human daily activity using a single inertial sensor. In *8th World Congress on Intelligent Control and Automation (WCICA)*, pages 282–287. IEEE, 2010.

[115] Zack Zhu, Ulf Blanke, Alberto Calatroni, and Gerhard Tröster. Human activity recognition using social media data. In *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*, page 21. ACM, 2013.

[116] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology.* Ravenio Books, 2016.