Acoustic Gait recognition Using Large Scale Feature Extraction and Support Vector Machines

Institute for Human-Machine Communication
Technische Universität München, Germany
Univ.-Prof. Dr.-Ing. habil. G. Rigoll

Master's Thesis

Author: Adriana Anguera Jordà
Advisor: Dipl.-Ing. Jürgen Geiger

Started on: 19.03.2012
Handed in on: 01.08.2012

Abstract

In this document we present the study of acoustic gait recognition using large scale feature extraction and support vector machines. This work includes the presentation of the database recorded in two phases, with a total of 305 people walking in three different ways that took part in the recordings, the classifier that has been used for the investigation, namely Suppot Vector Machines, and a section of experiments and results presented after all the experiments have been carried out. In addition, there is a conclusion with a summary of the work and some ideas that could be fulfiled by researchers interested in the theme.

A good people recognition has been achieved using this method, obtaining almost a 40% of correctly classified people walking in a normal way. However, the study has gone further trying also a gender and shoe type identification. Surprisingly, these results have proved to get a better classification, obtaining around a 70% of correctly classified people for the gender classification and around a 60% for the shoe type classification.

Contents

1

Introduction and Background

Ever since technology started to develop, human beings have always been interested in creating machines that can behave similar to humans. Nowadays, every company wants to be the pioneer in developing new technologies and there is a lot of competition. So here is when the fight for achieving not only the best results, but also the smallest device or the most attractive design appears. Researchers are in a constant competition for being the first ones to present the best results, which will shift the trend of technology.

Nowdays a lot of research is being done in the areas of image and speech recognition. The aim is to automate the process of human recognition either by using their images or sounds. Humans have the ability to recognise others from different sounds. For example one can distinghuish which family member is entering the house by means of sounds emitted such as the movement of the keys, their footsteps or the noise they make while walking. Males usually do not sound similar to females, neither do childern and so on. So, what if this information could be recorded and put into a machine so that it is able to distinguish a person. This is our challenge. Our research topic focuses on creating a system capable of recognising a person by the sound he or she makes when he or she is walking.

There have been several approaches to do person identification with image and speech recordings. Nevertheless, the audio field is still unexplored. For that reason, we do not have a base with results in which we can rely on or we can compare with. We are in front of a challenge, which we intend to achieve with success.

In our matter of study, we focus on building a model capable of identifying a person by its acoustic gait. We are interested in extracting information from audio files (recordings of people walking) and translating it into characteristics which can then be classified with the Support Vector Machines classifier (see Section 3.6).

1.1 Related work

As we have already mentioned, people identification with audio data is a field which still has to be investigated. However, much more research has been done in the fields of audio-visual, image and speech recognition. These fields have promoted a great interest and have resulted in big investigations and applications. In [WL12] a research on audio identification via fingerprint has been made. They use compressed-domain spectral entropy as the audio feature to implement a novel audio fingerprint algorithm. Moreover, in [DDS05] an experiment to test speaker identification systems using the CUAVE database can be seen. Some more studies of speech recognition and person recognition from image sequences can be found in [Lue97].

There has also been studies made on people identification using footstep detection. Some information about this topic can be found in [YSY04], [AI06], [AI08], [Bla06], [She04] and [SjR05].

1.2 Aim of the thesis

In this Master Thesis, our goal is to build a system capable of recognising people by the sound they make when they walk. We want to make a model of person identification and also extend the experiments to a gender and shoe type classification.

The motivation to do so, as well as accepting the challenge of being a field still not very studied, is to create a first baseline with the results of classifying people by means of their acoustic gait.

Next, we would like to give some ideas of the possible applications that people identification according to their acoustig gait could have.

The first application that corsses everyone's mind is to identify thieves. Imagine there has been a robbery and the video cameras have recorded both, image and audio, but for some reason the image has been damaged, so all the information we can get out of the recording is audio. In this occasion, we could try to identify the thief by his acoustic gait, comparing it to the samples we would have of the suspicious thieves. A second application could be used by security systems. In the same way that there exist retina scan, or fingerprint as passwords for alarms or security chambers, another way of creating this unique password could be by the acoustic gait. Moreover, it could be used to identify diseases. For example, the Parkinson Disease could be detected if we had a database of the normal type of walking and another one with samples of people affected by the disease in a premature stage. This could be really helpful in detecting the disease in time and act as soon as possible. As we are in the era of technology, smartphones and applications, a new phone application could be created to recognise people when they cross in front of the phone, for example, you could leave the smartphone in a room where the kids are not allowed to enter, and see whether they obey or not.

Much more applications could be listed in this thesis but it is not the aim. Nevertheless, as technology is developing in giant steps, there will appear more and more applications for acoustic gait recognition in the near future which probably we still cannot imagine.

2

Literature review

In this chapter a review of the current state of the research on the topic covered in this Master's Thesis is going to be made.

The main topic covered in this Master's Thesis is people identification according to the noise they make while they walk, using feature extraction and pattern classification. As many other human's characteristics, such as fingerprints, eye retina scan or voice, the gait, which is defined as a person's manner of walking, is a unique feature for every person. Therefore, such characteristic can be used to identify a person.

Previously to this, some other studies concerning people classification have been carried out. However, they have been more focused on the area of image or speech identificacion. During the last years, several approaches to being able to identify people by their voice have been implemented. First it was develped for military intelligence purposes and later on it was adopted by the police for solving criminal cases, such as murder, rape, extortion, burglary, and so on ([SCW12]). Because of that, a lot of investigation and progress on the topic was done.

The techniques used to do the classification have also changed over the years. During the late eighties, some of the technologies used include frequency estimation, hidden Markov models, Gaussian mixture models, pattern matching algorithms, artificial neural networks, matrix representation, vector quantization, multilayer perceptron and decision trees. However, a new tool appeared during the last decade in the field of machine learning and has proved to give better results for the same problems. This is the Support Vector Machines (SVM). The advantages of this effective classifier are that it can cope with samples with much higher dimensions, its solution is that with maximum margin and its convergence to the minimum of the associated cost function is guaranteed([RSUdM], [LME10], [THT00], [Phi02]).

Similarly, more investigation has also been done in the field of image recognition. The attempt is to identify people by the features extracted from the silhouette, taken from the shape and the colours of the people([DNTC09]). People identification

with video imaging is a very important task for analysing the content of a video in a surveillance system. This is also a challenging problem due to the fact that there are many variable sources in an image of a person, such as pose, scale, illumination, expression, motion blur, etc ([EZ04]).

Some studies on people identification using footstep detection have also been made during the past years. The investigations on this field that have been carried out have concentrated more on the type of shoes and the type of floor. The experiments done in this area have also used a smaller database than the one we use in this Master's Thesis. Some experiments on person identification using footsteps sound have been made by japenese researchers. Mainly, they have focused on the gait, the footwear and the floor [YSY04], [AI06] and [AI08]. Another research on footsteps detection has been made. In this occasion, it is described how indoor footsteps can be detected from other atmosphere sounds using sound wave sensor [She04].

Besides using video or image for people identification while walking, there are also some other methods. In [KK07] it is described how can a person be identified by means of an Acoustic Doppler Sensor (ADS) based technique for the characterization of gait. Furthermore, a new approach to footstep-based biometric identification by combining pattern classifiers with different feature sets is presented in [SjR05]. In this occasion, footsteps samples are obtained from a pressure-sensitive floor. Finally, in [Bla06], the problem of detecting footsteps using acoustic and seismic sensors is covered. The problem is approached from three different angles; modulating footstep signal energy, linear predictive modeling and a new method for blindly estimating the filters of a SIMO channel.

3

Methodology

In this chapter the database which we have been working on is presented. Furthermore, we will explain the methodology employed to perform the experiments, this is, how we got the information first of all, the database, then processed this information and finally obtained the results.

3.1 TUM-Kinekt Gait Database collection

In order to start with our research, first of all we had to collect a database, the TUM-Kinekt Gait Database. In our case, this was already done by Sebastian Bachmann, a student of the TUM who did his Diplomarbeit about "People Identification analysing their steps" [Bac12]. The TUM-Kinekt Gait Database was built in two sessions. The first one, which was recorded between 31.01.2012 and 02.02.2012, was composed of 176 people walking in three different ways. As it was the winter period, these people were dressed in thick coats and wore lots of clothes. The second session of recordings took place between 23.04.2012 and 25.04.2012. 129 more people were recorded while walking, 32 of which had also participated in the first session. During these dates, temperatures were higher so people dressed with fewer clothes, which make a difference in the audio recordings. So now in total our database is made up of 305 people. Each person from both of the sessions was asked to walk in three different ways (listed below) and more than once. So in total, we had six audio files of the normal type of walking, two of walking with a backpack and two of walking with coating shoes. Overall, we had more than 30.000 audio files in our database collection.

These different types of walking were:

- `Normal type of waking`
- `Walking with a backpack`

Figure 3.1: Normal type of walking.

Gender	Number of people	Percentage
male	186	60,98%
female	119	39,02%

Table 3.1: Gender percentage of male and female people who participated in the database.

- `Walking with coating shoes`

Now you can see that the picture in Figure 3.1 shows the normal type of walking, in Figure 3.2 we can distinguish a person walking with a backpack and finally in Figure 3.3 we can see a person walking with coating shoes.

Tables 3.1 and 3.2 show the percentage of the participants that were male or female and how many people wore the different types of shoes respectively. In addition, in Table 3.3 you can see the different nationalities that also took part in the database.

Shoe type	Number of people	Percentage
Sneakers	195	57,86%
High boots	61	18,10%
Low boots	38	11,28%
Loafers	28	8,31%
Others	15	4,46%

Table 3.2: Shoe type percentage of people who participated in the database.

Figure 3.2: Walking with a backpack.

Figure 3.3: Walking with coating shoes.

Nationality	Number of people	Percentage
Germany	220	73,37%
China	21	6,91%
Unknown	10	3,29%
Italy	4	1,32%
Greece	3	0,99%
Austria	3	0,99%
France	2	0,66%
Kazakhstan	1	0,33%
India	2	0,66%
Spain	3	0,99%
Bulgary	2	0,66%
Chile	1	0,33%
Egipt	2	0,66%
Iran	1	0,33%
Macedonia	1	0,33%
Cyprus	1	0,33%
Tunisia	1	0,33%
Mexico	1	0,33%
Latvia	1	0,33%
USA	1	0,33%
Slovakia	1	0,33%
South Korea	2	0,66%
Poland	3	0,99%
Brasil	3	0,99%
Serbia	1	0,33%
Vietnam	1	0,33%
Ukraine	2	0,66%
Morocco	1	0,33%
Rumania	2	0,66%
Thailand	1	0,33%
Lithuania	1	0,33%
Scotland	1	0,33%
Russia	1	0,33%
Afghanistan	1	0,33%
Luxembourg	1	0,33%
Croatia	1	0,33%
Australia	1	0,33%
Turkey	1	0,33%

Table 3.3: Nationality percentage of people who participated in the database.

3.2 Microsoft Kinect camera

The audio files of this study have been recorded with a Microsoft Kinect sensor. This camera was created by Alex Kipman [Pal11] and developed by Microsoft for the gameconsole Xbox 360 [Bis11] and later adapted for Windows PCs. We have chosen this camera because its software enables advanced gesture, facial and voice recognition. Furthermore, it is capable of tracking simultaneously up to six people with a feature extraction (3.4.2) of 20 joints per person [Cra12].

3.2.1 Kinect microphones

The Microsoft Kinect camera counts with four microphones in a line, three of them in the left side and one in the right side, located below the device. The aim of this microphone distribution is to capture the voices that are in front of the device and separate them from other sounds in the environment. Logically, if we put microphones at different places, the sound will not arrive at the same time, so taking into account the difference between the signals that the microphones get and the speed of the sound in the air, we can calculate where does the source of the sound come from. Moreover, we can also determine its approximate position.

Once the position of the sound is known, a complex algorithm merges the signals of the four microphones, getting as a result one signal that contains the sound, which comes from an imaginary cone that begins in the device and expands to the user. In addition to this, there is a filter that is in charge of deleting the signals outside of the human's voice frequency and amplifies it. Furthermore, the microphone also deletes the echoes produced because of the furniture and walls [Ano10] [McC01].

However, in our experiments we did not use such beamforming and echo cancellation techniques, but we converted the audio from four channels to a monochannel.

3.2.2 Sensor and Camera

The Kinect sensor is a horizontal bar, of approximately 23cm, connected to a small circular base with a motorized pivot and is designed to be positioned lengthwise above or below the video display. The device features an "RGB camera, depth sensor and multi-array microphone running proprietary software", which provide full-body 3D motion capture, facial recognition and voice recognition capabilities. The sensor has a motion and depth camera with 640x480 pixel resolution and @30 FPS. The depth sensor consists of an infrared laser projector combined with a monochrome CMOS sensor, which captures video data in 3D under any ambient light conditions [Cra12].

3.2.3 Field of view

The area required to use Kinect is roughly $6m^2$, although the sensor can maintain trackiing through an extended range of approximately 0.7-6m. The sensor has an angular field of view of 57°horizontally and 43°vertically, while the motorized pivot is capable of tilting the sensor up to 27°either up or down. The horizontal field of the Kinect sensor at the minimum viewing distance of 0.8m is therefore 87cm, and the vertical field is 63cm, resulting in a resolution of just over 1.3mm per pixel [Cra12], [Rob10].

3.2.4 Data streams

The microphone array features four microphone capsules and operates with each channel processing 16-bit audio at a sampling rate of 16kHz. The data streams are of 320x240 processing 16-bit in depth with @30 frames/sec and 640x480 processing 32-bit in colour with @30 frames/sec [Bis11].

3.3 Classifying methods

There are several options when it comes to choose a classifier for solving audio problems. One of them is using Artificial Neural Networks (ANN). They are a paradigm of learning and automatic processing inspired in how the animal's nervous system work. It has the ability to learn by means of a learning stage, which consists of providing the ANN with input data at the same time you indicate which is the desired output. After a good learning and training, and an adequate algorithm, the network will be able to classify a new entry [AI10].

Similarly, there is another type of classifier called Support Vector Machines (SVM). They consist of a group of supervised learning algorithms. It is a model able to predict the class of a new sample after having been trained. This classifier is detailed further on this document, in section 3.6.

As you can see, comparing these two types of classifiers (see Table 3.4), we conclude that for the problem we are dealing with, the most suitable classifier is the SVM. Not only SVM training always finds a unique global minimum and it has a very efficient training, but in practice, the SVMs are less prone to overfitting ([Ano12c], [Ano12d], [Wik12d]).

3.4 OpenSMILE

SMILE is an acronym which stands for *Speech & Music Interpretation by Large Space Extraction* [FE12]. OpenSMILE feature extraction is a toolkit that enables to extract large audio feature spaces in real time. This is possible due to the use

ANN vs SVM characteristics	
ANNs	SVMs
Hidden layers transform to any dimension	Kernel transform very superior dimension
Multiple local minima	Only one local minima
Hard training	Very efficient training
Very efficient classification	Very efficient classification
Design of hidden layers and nodes	Design of Kernel function cost parameter C
Very good behaviour in typical problems	Very good behaviour in typical problems
	Extremely robust for generalization, less need of heuristic for training

Table 3.4: Comparison of ANN and SVM

of the PortAudio [com12] library, which is a platform independent live audio input and live audio playback. It tries to combine features from Music Information Retrieval and Speech Processing. In other words, it is a modular and flexible feature extractor for signal proceessing and machine learning applications [FE10]. The main characteristics in which it focuses are the audio-signal features. Nevertheless, given suitable input components, they can also be used to analyse signals from other modalities, such as physiological signals, visual signals, and other physical sensors components thanks to their high degree of abstraction. It is written in C++, has a fast, efficient, and flexible architecture, and runs on various main-stream platforms such as Linux, Windows and MacOS.

Although openSMILE is designed for a real-time online processing, it can also be used off-line in batch mode for processing of large data-sets. This is not a very common feature found in related feature extraction software as most of related projects are designed for off-line extraction and require the whole input to be present. Moreover, openSMILE is able to extract features as new data arrives.

There are several data formats commonly used in the field of data mining and machine learning. OpenSMILE supports reading and writing in various data formats in order to facilitate interoperability. These formats include PCM WAVE for audio files, CSV (Comma Separated Value, spreadsheet format) and ARFF (Weka Data Mining) (see section 3.5) for text-based data files, HTK (Hidden-Markov Toolkit) parameter files, and a simple binary float matrix format for binary feature data.

Using the open-source software gnuplot [Ano12a], extracted features which are dumped to files can be visualised. A strength of openSMILE, due to its highly modular architecture is that almost all intermediate data which is generated during the feature extraction process (such as windowed audio data, spectra, etc.) can be accessed and saved to files for visualisation or further processing [FE10].

3.4.1 Use of openSMILE

Basically, openSMILE was designed to be used for research applications, demonstrators, and prototypes. However, the main target group of its users is researchers and system developers.

Currently, openSMILE is used by researchers and companies all around the world, which are working in the field of speech recognition (feature extraction front-end, keyword spotting, etc.), the area of affective computing (emotion recognition, affect sensitive virtual agents, etc.), and Music Information Retrieval (chord labelling, beat tracking, onset detection etc.) [FE10].

3.4.2 Feature extraction

Feature extraction is a dimensionality reduction of the input data when it is suspected to be notoriously redundant (much data, but not much information). The input data will be transformed into a reduced representation set of features (also named features vector). When the features extracted are carefully chosen, it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input.

Feature extraction consists in simplifying the amount of resources needed to describe a large set of data accurately [MR05].

3.4.3 Audio features

The audio baseline feature set consists of 1941 features, composed by 25 energy and spectral related low-level descriptors (LLD)x42 functionals, 6 voicing related LLDx32 functionals, 25 delta coefficients of the energy/spectral LLDx23 functionals, 6 delta coefficients of the voicing related LLDx19 functionals, and 10 voiced/univoiced durational features. The LLD details can be seen in Table 3.5. The set of LLD covers a standard range of commonly used features in audio signal analysis and emotion recognition. The functional set has been based on similar sets, but has been carefully reduced to avoid LLD/functional combinations that produce values which are constant, contain very little information, and/or high amount of noise.

All these features are collected in a configuration file, called openSMILE configuration file for AVEC 2011. AVEC stands for *Audio Visual Emotion Challenge*. This was a competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, visual and audiovisual emotion analysis [BSmP11]. In this file, we can find the functionalities extracted from the audio files. The features are related to different characteristics, such as voice, energy or spectral, among others.

Energy & spectral (25)
loudness (auditory model based), zero crossing rate, energy in band 250 - 650 Hz, 1kHz - 4kHz, 25%, 50%, 75% and 90% spectral roll-off points, spectral flux, entropy, variance, skewness, kurtosis, psychoacoustic sharpness, harmonicity, MFCC 1 - 10
Voicing related (6)
F_0 (sub-harmonic summation (SHS) followed by Viterbi smoothing), probability of voicing jitter, shimmer (local), jitter (delta: "jitter of jitter") logarithmic Harmonic-to-Noise Ratio (logHNR)

Table 3.5: Details of the 31 Low-Level Descriptors.

name of functional	**feature relation**
functionalsA	functionals for energy and spectral related LLD
functionalsAde	functionals for energy and spectral related LLD
functionalsF0v	functionals for pitch onsets/offsets (voiced)
functionalsF0p	functionals for pitch onsets (pauses)
functionalsNz	functionals for pitch and vq related lld in voiced regions
functionalsNzDe	functionals for pitch and vq related lld in voiced regions

Table 3.6: Functionals feature relation.

In Table 3.6 you can see the functionals from the configuration file which are related to different aspects in the audio recordings.

Not all the features are suitable for acoustic gait recognition. Some are more related to voicing features and others are better for audio-ambience features. As we are not interested in voice or speech identification, in order to carry out our experiments we have made a feature reduction by removing the voicing related features. In Table 3.6 we can see a list of the functionals followed by a brief description of what they are related to. We can clearly see that $functionals\ F0v$, Nz and $NzDe$ are related to voiced regions, so theorethically, these functionals will not influence in the result when classifying people. $Functionals\ F0p$ are related to pauses, so supposedly they will neither have a great effect for a correct classification. Last of all we have the $functionals\ A$ and Ade, which are both related to energy and spectral features. In theory, these are the features that have a greater impact in the acoustic gait identification. They contain the information about frequency, entropy, variance, harmonicity, loudness, among others. Mel-Frequency Cepstrum Coefficients (MFCC), are supposedly more commonly used in speech recognition and speaker

identification systems. However, MFCCs are increasingly finding uses in audio measures and genre classification. Further on, in section 4.2.2, we can see the results of applying this feature reduction.

3.5 Weka

Weka (Waikato Environment for Knowledge Analysis) is a software platform for the automatic learning and data mining written in Java and developed in The University of Waikato [oW12]. In 1993, the Weka University of New Zeland started the development of the original version of Weka in TCL/TK and C ([Wik12f], [Wik12a]). Four years later, the code was rewritten in Java including implementations of modeling algorithms. In 2005, Weka recieved from SIGKDD [Wik12e] (*Special Interest Group on Knowledge Discovery and Data Mining*) the award "*Data Mining and Knowledge Discovery Service*".

The Weka package contains a collection of visualisation tools and algorithms for data analysis and predictive modeling, as well as a user's graphic interface that enables to easily access to its functionalities. The original version was initially designed as a tool to analyse data proceeding from the agriculture field, but the recent version, based in Java, is used in much more different areas, particularly with investigation and educational purpose.

Detailed below are some strenghts which influenced in the decision of using Weka for our investigation:

- `Freely available under the general public licence of GNU`
- `Completely implemented in Java, it runs on any platform`
- `Large collection of techniques for data processing and modeling`
- `Easy to use for a beginner thanks to its user graphic interface`

On the other hand, we also have to mention an important weakness. At the present time, Weka does not cover the algorithms in sequence modeling.

As we already mentioned in section 3.4, Weka can support different data formats. In this investigation, we have used the ARFF format, which stands for *Attribute Relation File Format*. This is an ASCII text file that describes a list of instances sharing a set of attributes. These files were also developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato.

However, Weka holds several standard tasks of data mining, specially data processing, clustering, classification, regression, visualisation and selection. All Weka techniques are based on the assumption that the data are available in a flat file or in a relation, in which every data registered is described by a fixed number of attributes

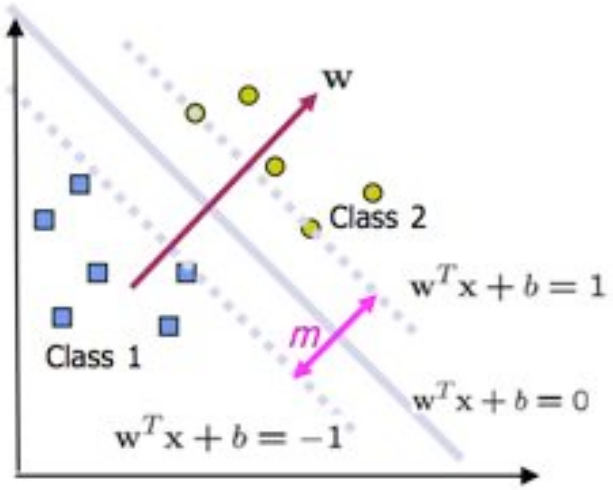

Figure 3.4: The decision fronteer must be at a maximum distance from the data as possible.

(usually numeric or nominal, although there are also other types). Weka also provides access to databases via SQL [Gro12] thanks to the connection JDBC [Wik12b] (*Java Database Connectivity*) and can process the returned result by a search done to this database.

The classifier used by Weka is the SMO (*Sequential Minimal Optimization*). It implements John Platt's sequential minimal optimization algorithm for training a support vector classifier.

3.6 Support Vector Machines (SVM)

This section has been inspired by the paper [Bet05].

Support Vector Machines were developed by Vladimir Vapnik [CLRC12] and his team in the AT&T laboratories. They were presented in 1992 and became famous when they gave much superior results than neural networks in handwriting recognition, using pixels as input. SVM is a very powerful tool which can be used for pattern recognition and classification problems. It pretends to predict from what is already known [Ano12b].

The basic idea of the SVM, in a binary classification task, consists in creating a hyperplane that separates our classes. As there are infinite planes that can make this separation, the SVM tries that the closer vectors to this separation (called support vectors) have a maximum distance to the hyperplane as possible (see Figure 3.4). It is expected that the larger the margin, the better generalisation of the classifier [ROD12].

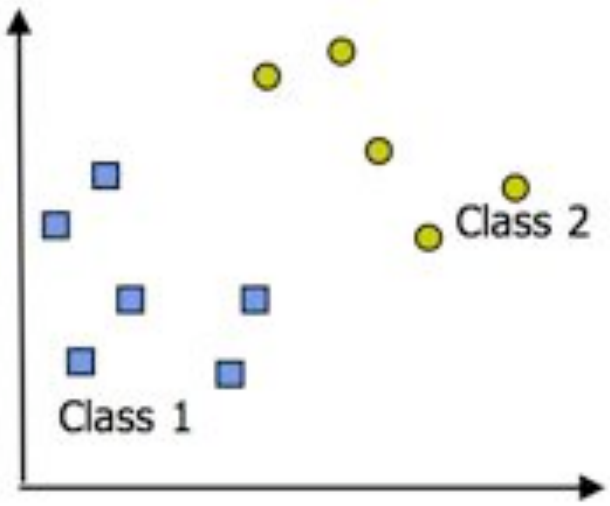

Figure 3.5: Linearly separable case.

Maximizing the margin m is a quadratic programming (QP) problem and can be solved by his dual problem introducing Lagrange multipliers. When the data cannot be linearly separated, a change of space is done by a function that transforms the data so that it can be linearly separated. Such function is called Kernel. There are three different types of kernels:

- Linear: $K(x_i, x_j) = (x_i)^T x_j$
- Polinomic: $K(x_i, x_j) = (\gamma(x_i^T x_j) + \tau)^d, \gamma > 0$
- RBF: $K(x_i, x_j) = e^{-(\gamma \|x_i - x_j\|^2)}, \gamma > 0$

We are going to do a review of the basic theory of SVM in classification problems.

3.6.1 Linearly separable case

Let's supose that we have been given a group S of points tagged for training as we can see in Figure 3.5.

Each training point $x_i \in \Re^n$ belongs to one of the two classes and it has been given a tag y_i for $i = 1...i$. In most of the cases, the search of an adequate hyperplane in an input space is too restrictive to be of practical use. A solution to this situation is to map the input space in a characteristic space of a higher dimension and look for the optimal hyperplane there. If $Z = \varphi(x)$ is the corresponding vector notation in the characteristic space with a map φ of $\Re^n$ to a characteristic space Z, we want to find the hyperplane:

$$wz + b = 0 \tag{3.1}$$

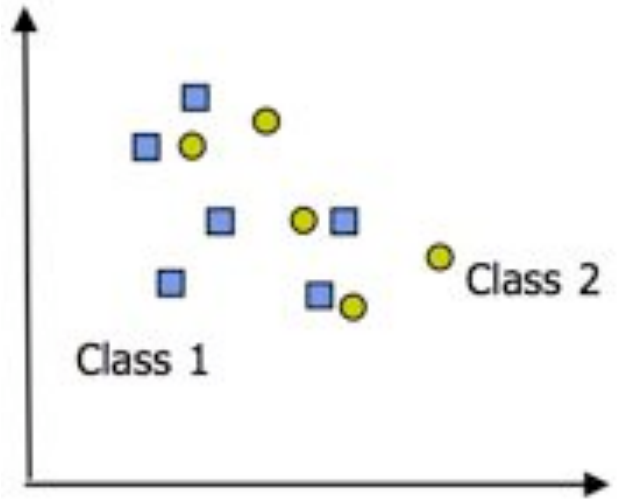

Figure 3.6: Non lineally separable case.

defined by the pair (w,b) so that we can separate the point x_i along with the function:

$$f(x_i) = sign(wz_i + b) = \left\{ \begin{array}{ll} 1 & y_i = 1 \\ -1 & y_i = -1 \end{array} \right\} \tag{3.2}$$

where $w \in Z$ and $b \in \Re$. More precisely, it is said that the group S is linearly separable if (w,b) exists so that the inequations:

$$\left\{ \begin{array}{ll} (wz_i + b) \geq 1 & y_i = 1 \\ wz_i + b) \leq -1 & y_i = -1 \end{array} \right\} i = 1, .., l \tag{3.3}$$

will be valid for all the elements of the group S. For the linearly separable case of S, we can find a unique optimal hyperplane for which the margin between both projections is maximized.

3.6.2 Non linearly separable case

In order to deal with data which is not linearly separable (see Figure 3.6), the previous analysis can be generalised introducing some non-negative variables $\xi_i \geq 0$ so that (3.3) is modified to:

$$y_i(wz_i + b) \geq 1 - \xi_i, i = 1, ..., l. \tag{3.4}$$

The $\xi_i \neq 0$ in (3.4) are those for which the point x_i does not satisfy (3.3). Then, the term $\sum_{i=1}^{l} \xi_i$ can be taken as an error measure in the classification.

The problem of the optimal hyperplane is then redefined as the solution to the problem:

$$min \left\{ \ \tfrac{1}{2} ww + C \sum_{i=1}^{l} \xi_i \ \right\} \tag{3.5}$$

subject to

$$y_i(wz_i + b) \geq 1 - \xi_i, i = 1, ..., l. \tag{3.6}$$

$$\xi_i \geq 0, i = 1, ..., l \tag{3.7}$$

where C is a constant. The parameter C can be defined as a regulation parameter, which controls how soft the margins in the hyperplane are. In our experiments, this parameter has been changed in order to see the different response to different values of it. More details about it can be found in references [Vap98] and [ROD12].

Looking for the optimal hyperplane in (3.7) is a QP problem that can be solved constructing a Lagrangean and transforming it into the dual:

$$maxW(\alpha) = \left\{ \ \sum_{i=1}^{l} \alpha_i - \tfrac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j z_i z_j \ \right\} \tag{3.8}$$

subject to:

$$\sum_{i=1}^{l} y_i \alpha_i = 0, 0 \leq \alpha_i \leq C, i = 1, ..., l \tag{3.9}$$

Where $\boldsymbol{\alpha} = (\alpha_1 ... \alpha_l)$ is a vector of positive Lagrange multipliers associated with the constants in 3.4.

The Khun-Tucker [Wik12c] theorema plays an important role in the SVM theory. According to this theorema, the solution $\overline{\alpha}_i$ of problem (3.8) satisfies:

$$\overline{\alpha}_i(y_i(\overline{w} z_i + \overline{b}) - 1 + \overline{\xi}_i) = 0, i = 1, ..., l \tag{3.10}$$

$$(C - \overline{\alpha}_i)\xi_i = 0, i = 1, ..., l \tag{3.11}$$

Out of this equality, we deduce that the only values $\overline{\alpha}_i \neq 0$ (3.11) are those that for the constants in (3.4) are satisfied with the equal signal. The point x_i corresponding to $\overline{\alpha}_i \geq 0$ is called *support vector*. But there are two types of support vectors in a non separable case. In the case of $0 \leq \overline{\alpha}_i \leq C$, the corresponding support vector x_i satisfies the equality $y_i(\overline{w} z_i + \overline{b}) = 1$ and $\xi_i = 0$.

And in the case of $\overline{\alpha}_i = C$, the corresponding ξ_i is different from zero and the corresponding support vector x_i does not satisfy (3.3). We refer to these support vectors as errors. The point x_i corresponding to $\overline{\alpha}_i = 0$ is correctly classified and it is clearly far away from the decision margin in Figure 3.7.

In order to build the optimal hyperplane $\overline{w} z + \overline{b}$, we use

$$\overline{w} = \sum_{i=1}^{l} \overline{\alpha}_i y_i z_i \tag{3.12}$$

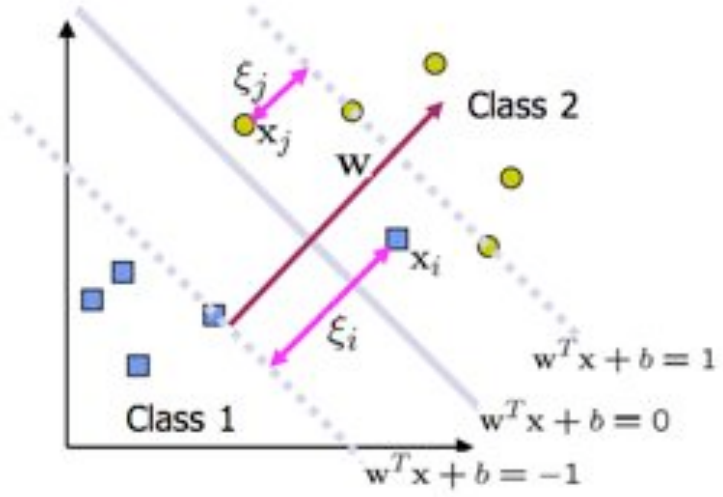

Figure 3.7: Appearance of the error parameter ξ_i in the classification error.

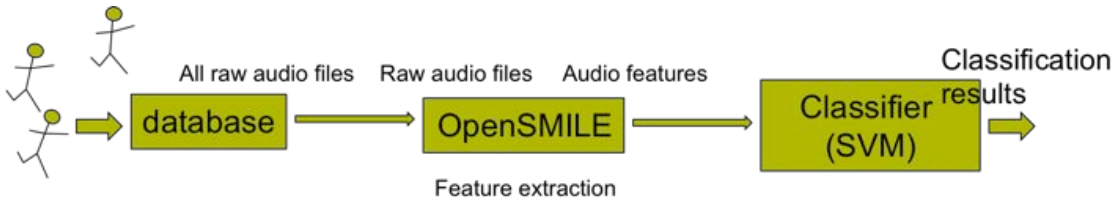

Figure 3.8: Diagram of the classification process with SVM classifier.

and the scalar b can be determined from the conditions of Kuhn-Tucker (3.11). The generalised decision function of (3.2) and (3.12) is such that

$$f(x) = sign(wz + b) = sign(\sum_{i=1}^{l} \alpha_i y_i z_i z + b) \tag{3.13}$$

3.7 Simulation

Once we know which classifier to use and under which program we are going to run the simulation, we are ready to start our first experiments. To make it clearer to the reader, in Figure 3.8 you can appreciate a diagram with all the steps followed since the recordings the recordings of the people are collected in the database, continuing with the audio format conversion and finally with the classification providing results.

We have two options in order to do the simulations; either we run a perl script specifying the classifier and its parameters, or we use the Weka interface explained in section 3.5.

Options specific to weka.classifiers.functions.SMO	
-M	fit logistic models to SVM outputs
-V	The number of folds for the internal cross-validation. (default -1, use training data)
-L	The tolerance parameter. (default 1.0E-3)
-P	The epsilon for round-off error. (default 1.0E-12)
-N	Whether to 0=normalize/1=standardize/2=neither. (default 0=normalize)
-C	The complexity constant C
-W	The random number seed. (default 1)
-K	The Kernel used.(default:weka.classifiers.functions.supportVector.PolyKernel)
Options specific to kernel .classifiers.functions.supportVector.PolyKernel	
-E	The exponent for the polynomial kernel. (default: 1.0)
-C	The size of the cache (a prime number), 0 for full cache and -1 to turn it off.

Table 3.7: Parameter options for the classification.

3.7.1 Perl script

In the event of using a perl script, a specific line has to be included in it. This command specifies the parameters of our SVM classifier and uses the Weka for making the classification.:

java -Xmx14000m -classpath ../etc/weka.jar weka.classifiers.functions.SMO -o -v -i -C $complexity -M -t \"$trainfile \" -T \"$testfile\" -L 0.001 -P 1.0E-12 -N $standardize -V -1 -W 1 -K \"weka. classifiers.functions.supportVector.PolyKernel -C 250007 -E $exponent\"

In Table 3.7 you can see the possible options of the parameters as well as their meaning.

3.7.2 Weka interface

Working with the Weka interface is simple due to its intuitive use. In order to change the functionalities of the classifier, we just open a drop-down menu and there will appear the different options that we can choose. These options are the same as explained above in Table 3.7.

4

Experiments and Results

In this chapter we are going to present the experiments we have been carrying out during the last five months as well as the results we have obtained. These experiments are based on a database which has already been introduced in chapter 3.1, and now we will describe more in detail which information was recorded. In order to perform the experiments, we created a training and a testing set out of the database. Generally, we used the Sequential Minimal Optimization (SMO) for the classification (see section 3.5). In any other case, it will be specified.

4.1 Description of the TUM-Kinekt Gait Database

As we mentioned before in section 3.1, the database was collected in two different sessions. The difference between them was the number of recordings we had in each of them. The information that was gathered contained specifications for the gender, age, height, type of shoe, country of origin and number and date of the recording. These characteristics will be used to either identify who is walking, or distinguish between a man or a woman, or which type of shoe they are wearing.

According to the type of walking we had a different number of audio files; for our first recording session, the recordings consisted of six audio files for the normal type of walking and two audio files for both, walking with a backpack and with coating shoes. Then our database was increased, this is, we had up to 305 people. For the first 32 people who appeared in both recordings sessions, we have twelve audio files for the normal type of walking, and four for walking with a backpack and with coating shoes. For the rest of the people it is just the same as the first session. These files will then be useful to separate the database in a training set and a testing set to make the classification.

Training and Testing files	
Training files	4 audio files from normal type of walking
Testing files	2 audio files from walking with backpack 2 audio files from normal type of walking 2 audio files from walking with coating shoes

Table 4.1: Description of audio files in training and testing sets

Options specific to weka.classifiers.functions.SMO	
-M	true
-V	-1
-L	1.0E-3
-P	1.0E-12
-N	1=standardize
-C	1
-W	1
-K	weka.classifiers.functions.supportVector.PolyKernel
Options specific to kernel .classifiers.functions.supportVector.PolyKernel	
-E	1
-C	25007

Table 4.2: Parameter options for the SMO classification.

4.2 Experiments

This section shows how have we proceeded with our experiments, starting with the simplest configurations up to the most complex ones in order to obtain more interesting results.

First of all, we are going to explain how we created two different sets out of the database so that we could train with one and test with the other. As we have mentioned above, we had six audio files of the normal type of walking, so we used four of them for our training set. Then, the rest of the audio files, this is two for each type of walking (normal, backpack and coating shoes), were used for the testing set. You can look at Table 4.1 which will make a clearer idea. Although in the second session we increased the database, we still used the same training and testing sets to carry out our experiments, the difference was the number of people.

For most of the experiments, the parameters chosen for the SMO classification are specified in Table 4.2. If in any case they are changed, it will be specified.

Next, we are going to present each experiment we performed, detailing in each case the training and testing sets used, as well as the results in a graphic or table form in order to make them more understandable.

	Correctly classified instances
10 classes	41,67%
50 classes	35,33%
110 classes	28,48%
176 classes	26,67%

Table 4.3: Comparison results increasing database.

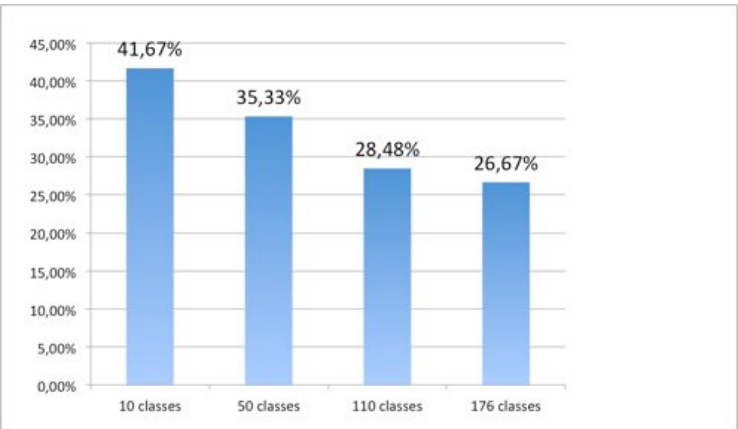

Figure 4.1: Comparison of the classification according to different number of classes.

4.2.1 Increasing database

The first experiment consisted in starting the classification of people with a small amount of our database and slowly increase it, to see the difference in the results when we tried to classify a different number of people. We started by classifying 10 people, followed by 50, 110 and finally 176. We used the recordings from the first session. In this case, we did not alter the openSMILE configuration file for AVEC 2011 (see subsection 3.4.3) which contains the features related to different characteristics.

We trained with four audio files of the normal type of walking and we tested with the remaining audio files from the other types of walking all together. This is, with two audio files from the normal type of walking, two from the backpack and two from the coating shoes (six audio files in total). The classifier chosen was the SMO and its parameters were the ones specified in Table 4.2.

In Table 4.3 we can see the results of the classification as we increased the number of classes. The same results are presented in a graph form in Figure 4.1. We can appreciate that the number of correctly classified instances decreases as we increase the number of classes. This is a coherent result due to the fact that as more people in the training set you have, there is more people to compare with when it comes to make the test.

Training and Testing files	
Training files	4 audio files from normal type of walking
Testing files with backpack	2 audio files from walking with backpack
Testing files noral walking	2 audio files from normal type of walking
Testing files with coating shoes	2 audio files from walking with coating shoes

Table 4.4: Description of audio files in training and testing sets

	Correctly classified instances
walking with backpack	30,11%
normal walking	47,71%
walking with coating shoes	3,45%
mean of the three types of walking	27,09%

Table 4.5: Comparison results reducing functionals.

4.2.2 Reducing functionals

As our interest is to be able to identify as much people as possible, from now on we will stick to experiments with 176 classes trying to improve the results. Our first attempt is reducing the number of functionals. This is, removing the ones that have less influence in the audio identification. So we kept only with the functionals for energy and spectral related LLD (functionalsA and functionalsAde from the avec2011.conf file, see Table 3.6 in Subsection 3.4.3).

In this occasion, we also trained with four audio files of the normal type of walking, and we tested separately three testing sets each of them containing two audio files of the normal type of walking, walking with a backpack and with coating shoes. (See Table 4.4).

The results are presented in Table 4.5.

As we can see from the Graph 4.2 the best classification is for the people walking in a normal way. We can think that it is a reasonable result due to the fact that we are training with audios from the normal type of walking, so they are better recognised. We will see that this situation repeats almost along all the experiments.

We can have a general view in Table 4.6 comparing the results with all the functionals and with the functionals reduction. We can appreciate a slight increase in the results. To understand the importance of the difference between the two results, we can refer to the statistical significance [**?**]. We can claim a result to be statistically significant if it is unlikely to have occurred by chance. To determine this significance, we have used a perl script which calculates the significance between two values. This value is calculated in the following way; we define the difference between the two values and the standard deviation as $\mu_{diff} = p_1 - p_2$ and $\sigma_{diff} = \sqrt{\frac{1}{N}p_1^*(1-p_1)p_2^*(1-p_2)}$ respectively, being p_1 and p_2 the values we want to test

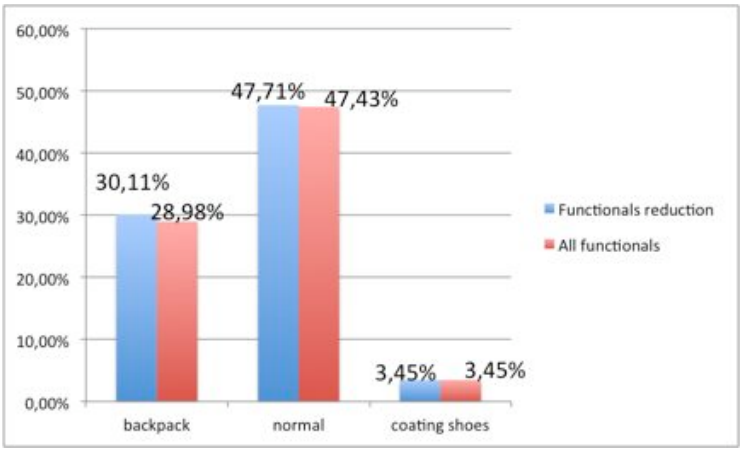

Figure 4.2: Comparison of people classification using all the functionals and the functionals reduction.

	All functionals	**Reduced functionals**
walking with backpack	28,98%	30,11%
normal walking	47,43%	47,71%
walking with coating shoes	3,45%	3,45%
mean of the three types of walking	26,67%	27,09%

Table 4.6: Comparison results between all functionals and reduced functionals.

whether one is better than the other. Then we compute the normalized z-score as $z = \frac{p_1 - p_2}{\sigma_{diff}}$. To test the process, we use the one-tailed significance values which give us the amount of evidence required to accept that an event is unlikely to have arisen by chance.

So if we look at our results for the mean of the three types of walking, $p_1 = 0,2667$ and $p_2 = 0,2709$. We evaluated them in the perl script prepared to do this calculation and the results show out that there is only one number of patterns differing, so we can conclude that there is no statistical significance between these two values. Nevertheless, from now on we are going to work with the functionals reduction, this is, removing the voiced related features, as we obtained almost the same results using less information and the experiments ran quicker.

4.2.3 Increasing training set

In order to understand the influence of the number of files used for the training set, we made an experiment that consisted in training with only one audio file from the normal type of walking, and increasing the training set one by one until we had the four audio files. The testing sets were formed by two audio files of the normal type

	backpack	normal walking	coating shoes
1 audio file training	15,34%	31,43%	3,45%
2 audio files training	21,88%	31,43%	3,45%
3 audio files training	28,13%	39,43%	4,02%
4 audio files training	30,11%	47,71%	3,45%

Table 4.7: Comparison results increasing training set.

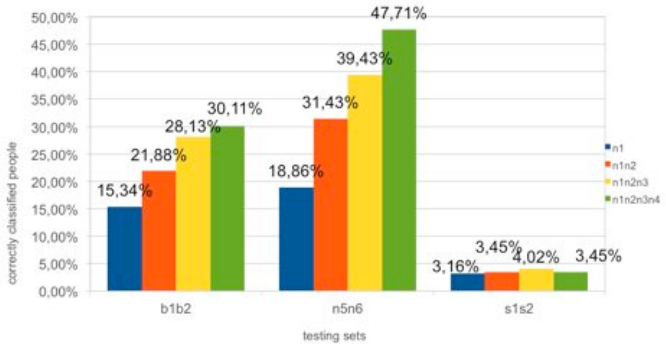

Figure 4.3: Comparison of people classification using different number of audio files in the training set.

of walking, two audio files of walking with a backpack and two audio files of walking with coating shoes. In Table 4.7 we can see the result.

In Graph 4.3 you can see the difference in training with a different number of audio files. These results prove us that it is better to classify with most of audio files as possible. Once again, the best results are obtained for the normal type of walking.

4.2.4 Gender classification

Another of our motivation was to try to make a gender classification. We wanted to see whether it was possible to make an acceptable gender recognition or not.

In this occasion, we divided the database into two. One half was for the training set and the other for the testing set, in such a way that no same person could be at the same time in both sets. In Table 4.8 you can see how was the database divided and the audio files corresponding to the training and testing set.

Looking at Table 4.9 we can say that we have reasonable results, being the correctly classified instances above a 50% which we can consider good results, although they could be improved. This means, for example, that for the people walking in a normal way, we classify correctly whether it is a man or a woman in the 58,62% of the cases.

Training and Testing files	
88 training files	4 audio files from normal type of walking
88 testing files	2 audio files from walking with backpack 2 audio files from normal type of walking 2 audio files from walking with coating shoes

Table 4.8: Description of audio files in training and testing sets for the gender classification.

	Correctly classified instances
walking with backpack	62,50%
normal walking	60,80%
walking with coating shoes	52,84%
mean of the three types of walking	58,71%

Table 4.9: Gender classification.

We can also have a look at Graph 4.4 to have a better vision of the gender classification.

In this case, we also did the experiment of training with a different number of audio files. The testing set was created in the same way as before. The results are presented in Table 4.10 and Graph 4.5.

Surprisingly, in this case we cannot claim that as we increase the number of samples in the training set the results are also increased.

4.2.5 Shoe type classification

As well as the gender classification, we also wanted to make a shoe type classification. The database was divided in the same way as in the gender classification (see Table 4.8). In the database collected, among other information, we had the detail of which type of shoes were the people wearing. In Table 4.11 we can see that, for example, for the normal type of walking we classify correctly with a 60,80% which type of shoes our participants are wearing.

In Table 4.12 we can also see the results of the shoe type classification when

	backpack	**normal walking**	**coating shoes**
1 audio file for training	57,95%	59,66%	60,92%
2 audio files for training	61,93%	59,77%	58,05%
3 audio files for training	59,66%	66,67%	47,73%
4 audio files for training	62,50%	60,80%	52,84%

Table 4.10: Comparison gender results increasing training set.

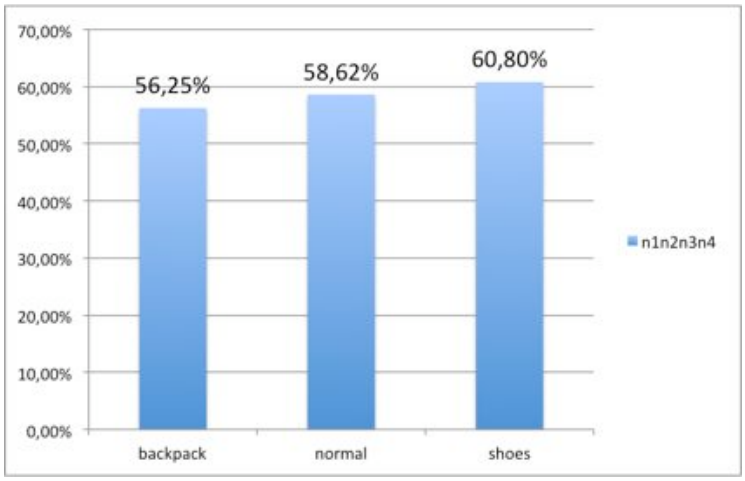

Figure 4.4: Gender classification of 176 people.

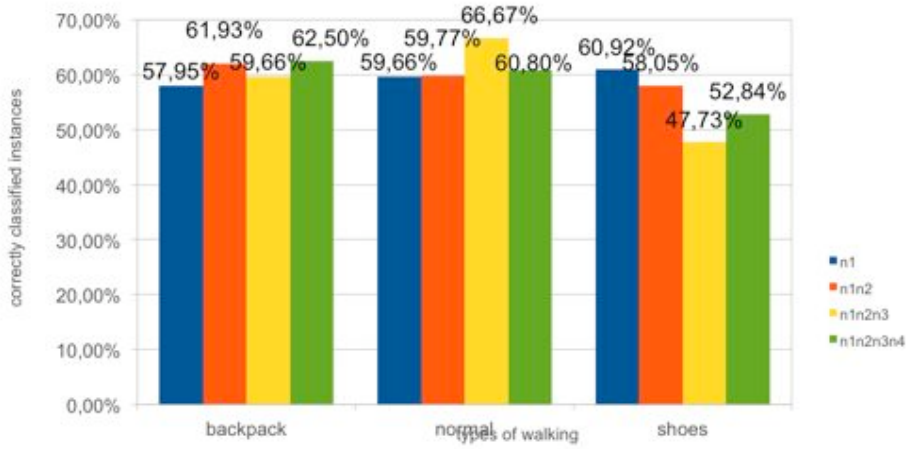

Figure 4.5: Comparison gender classification using different number of audio files in training set.

	Correctly classified instances
walking with backpack	38,64%
normal walking	31,82%
walking with coating shoes	27,59%
mean of the three types of walking	32,68%

Table 4.11: Shoe type classification.

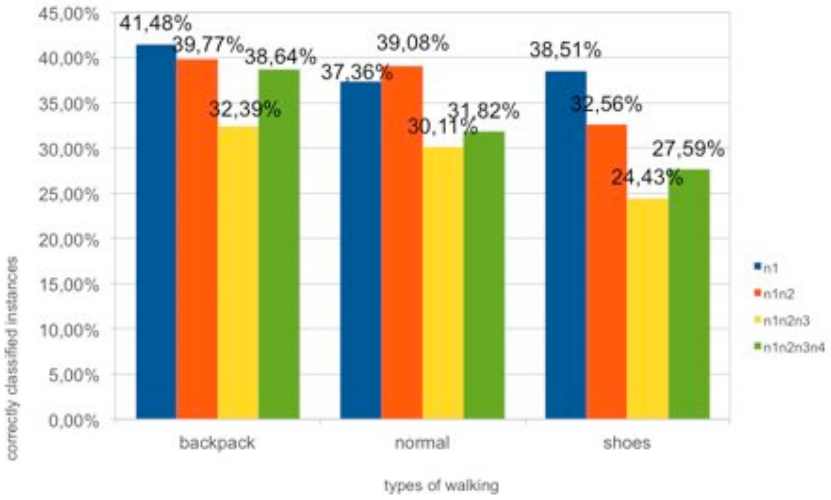

Figure 4.6: Comparison shoe type classification using different number of audio files in training set.

	backpack	**normal walking**	**coating shoes**
1 audio file for training	41,48%	37,36%	38,51%
2 audio files for training	39,77%	39,08%	32,56%
3 audio files for training	32,39%	30,11%	24,43%
4 audio files for training	38,64%	31,82%	27,59%

Table 4.12: Comparison shoe type results increasing training set.

we increased the number of audio files in the training set. In Graph 4.6, we can appreciate the comparison of the results.

As the case for the gender classification, in the shoe type classification the results do not improve when we increase the number of audio files in the training set.

4.2.6 Combining old and new database

In this point, we are going to use the audio files from the second session of recordings combined with the ones in the first session. However, we are only going to use the first 32 people. This is because these people participated in both of the sessions and we want to compare how good can it classify the same people but recorded in different times. As we had justified before, we still work with the reduced features. We trained with the same four audio files of the normal type of walking as in the previous experiments, but this time we tested with the audio files of every type of walking from the second session of recordings. The results, which we can see in

	Correctly classified instances
walking with backpack	3,25%
normal walking	1,56%
walking with coating shoes	4,69%
mean of the three types of walking	3,17%

Table 4.13: 32 people classification from new and old recording session.

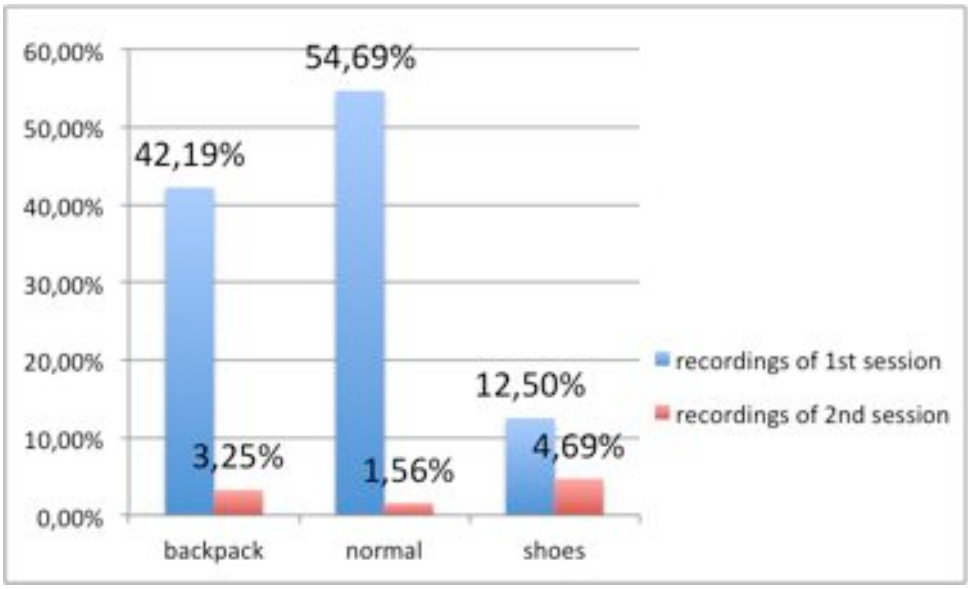

Figure 4.7: Comparison of results testing with recordings from the first and second session.

Table 4.13, show us that we do not get a very successful classification.

In order to compare the results, we did the same experiments, with 32 people, but with all of the recordings from the first session. We can clearly see the difference in Table 4.14 and in a graph form in Figure 4.7. The results are much better when we use only the audio files from the first session of recordings.

4.2.6.1 Gender classification

Furthermore, continuing with the 32 classes, we also experimented a gender classification. Still, with the training set from the first session of recordings and the testing set from the second one. In Table 4.15 we can see the results.

Moreover, we would like to compare the results for the gender classification with 32 people, when we tested with people from the first and second session. In Graph 4.8 we can see the difference.

	testing with first session of recordings	testing with second session of recodings
backpack	42,19%	3,25%
normal walking	54,69%	1,56%
coating shoes	12,50%	4,69%

Table 4.14: Comparison 32 people classification according to session recorded.

	Correctly classified instances
walking with backpack	59,38%
normal walking	71,89%
walking with coating shoes	46,89%
mean of the three types of walking	59,39%

Table 4.15: 32 gender classification from new and old recording session.

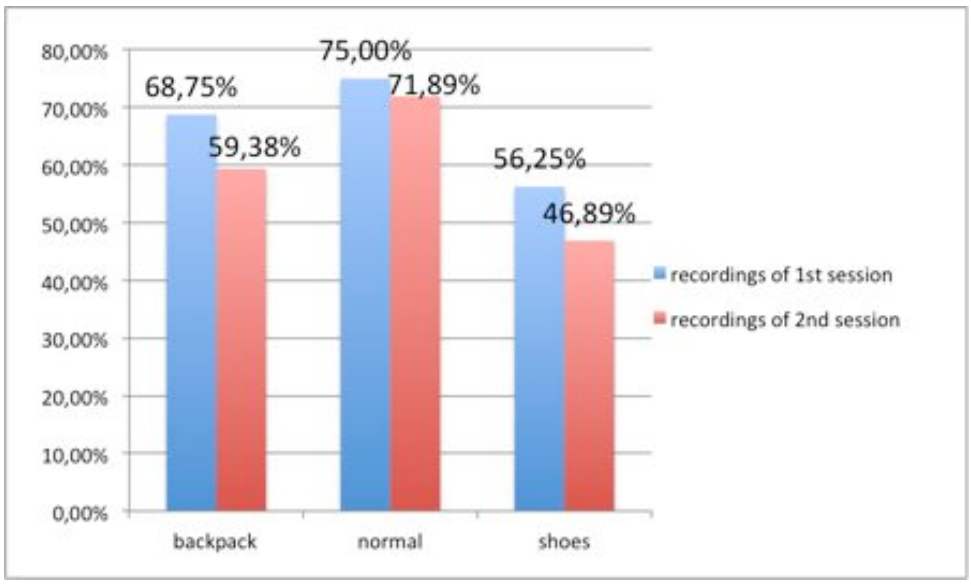

Figure 4.8: Comparison of results testing with recordings from the first and second session.

	backpack	normal walking	coating shoes
1 audio file for training	9,51%	15,25%	3,11%
2 audio files for training	18,52%	26,07%	2,95%
3 audio files for training	21,97%	35,25%	4,26%
4 audio files for training	25,74%	39,84%	3,44%

Table 4.16: Comparison results increasing training set with the whole database.

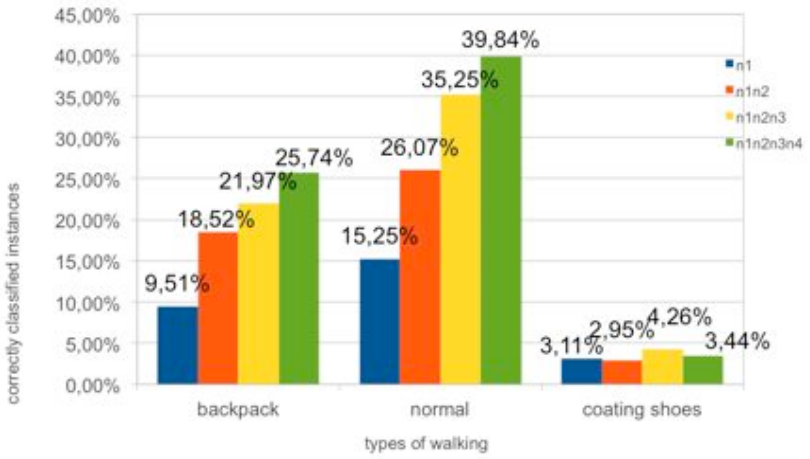

Figure 4.9: Comparison of 305 people identification using different number of audio files in training set.

4.2.7 Whole database

As we made progress in our work, we reached to a point where we increased our database. From now on, we used the samples from the second session of recordings in which we had the participation of 305 people. We also continued with the feature reduction, this is working only with the functionals related to energy and spectral, as explained in subsection 4.2.2. So, our first experiments with the whole database, as in subsection 4.2.3, consisted in training with one audio file and increase it until we reached the four audio files from the normal type of walking. The testing set consisted in two audio files from the normal type of walking, two from the backpack and two from the coating shoes. In Table 4.16 we can see the results.

Judging by the results, we can claim that as we increased the training set, the results of the classification also improved, being the normal type of walking the best classified. In Figure 4.9 we can see the difference between training with a different number of audio files.

Training and Testing files	
153 training files	4 audio files from normal type of walking
152 testing files	2 audio files from walking with backpack 2 audio files from normal type of walking 2 audio files from walking with coating shoes

Table 4.17: Description of audio files in training and testing sets for the gender classification.

Gender	Number of people	Percentage
male	186	60,98%
female	119	39,02%

Table 4.18: Gender percentage of male and female people who participated in the database.

4.2.7.1 Gender classification

Continuing with the procedure, we now tested how well can our system make a gender classification when we have increased the database. In Table 4.17 you can see how we divided the database.

In Table 4.18 you can see the percentage of male and female people who participated in the database.The results of the classification are shown in Table 4.19.

If we make a comparison between the gender classification from the whole database and the half of the database (see subsection 4.2.4), we can say that we have better results when the database is increased. In Figure 4.10 you can appreciate this difference.

4.2.7.2 Shoe classification

We made the same experiment but with the shoe type classification using the whole database. The database was divided in the same way as in the gender classification case (see Table 4.17). You can see the results in Table 4.20.

Comparing the results of the shoe type classification between the recordings from the first session and the second session, we can see in Figure 4.11 that they are better

	Correctly classified instances
walking with backpack	59,21%
normal walking	62,83%
walking with coating shoes	58,22%
mean of the three types of walking	60,09%

Table 4.19: Gender classification with the whole database.

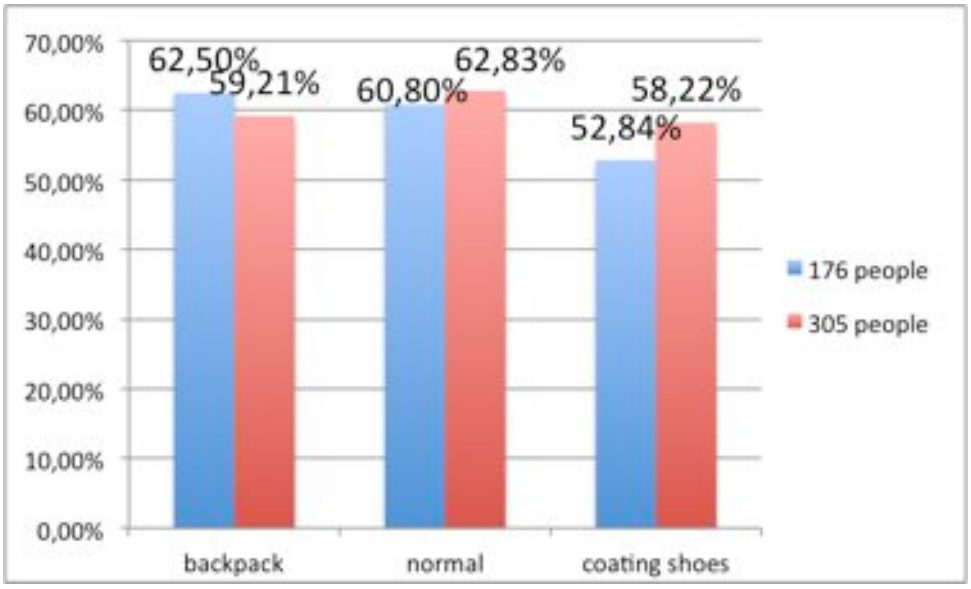

Figure 4.10: Comparison of gender classification with 176 people and 305 people.

	Correctly classified instances
walking with backpack	54,61%
normal walking	53,62%
walking with coating shoes	45,40%
mean of the three types of walking	51,21%

Table 4.20: Shoe type classification with the whole database.

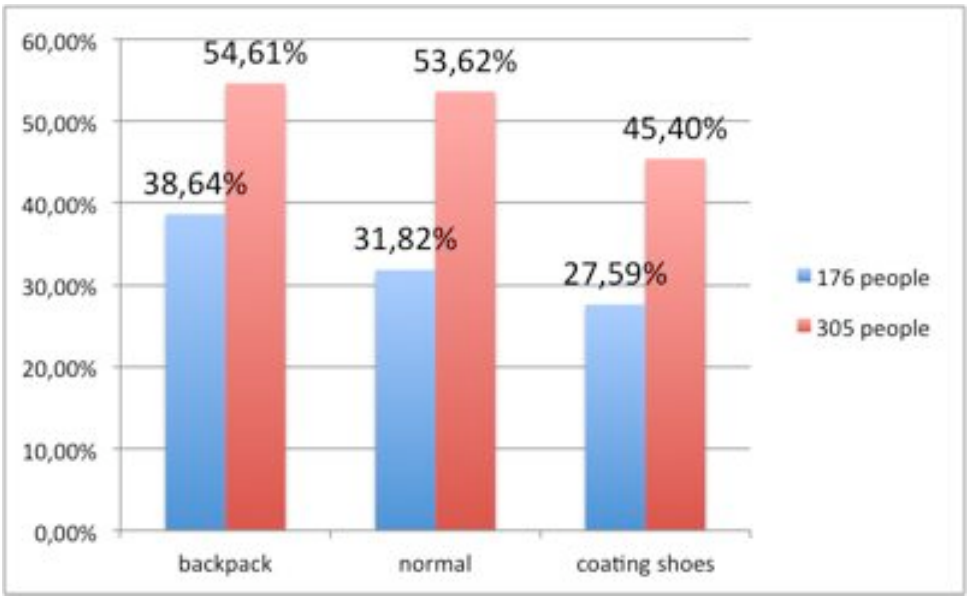

Figure 4.11: Comparison of shoe type classification with 176 people and 305 people.

Shoe type	Number of people	Percentage
Sneakers	195	57,86%
High boots	61	18,10%
Low boots	38	11,28%
Loafers	28	8,31%
Others	15	4,46%

Table 4.21: Shoe type percentage of people who participated in the database.

when we have a larger database.

4.2.8 Only a shoe type experiments

In order to understand the influence of the shoe type in the classification, we made some separate experiments getting people who were wearing only one type of shoe. We wanted to know how well could we classify people wearing only sneakers, low boots, loafers, high boots or others. In Table 4.21 you can see how many people were wearing these types of shoes. The "Others" group included the sandals, ballerinas and rubber boots which very few people wore them.

Again, we trained with four audio files from the normal type of walking and we tested with two audio files from the normal type of walking, two from walking with a backpack, and two from walking with coating shoes. In Graph 4.12 you can see the result of this classification.

If we compare these results to the ones we obtained when we did the people

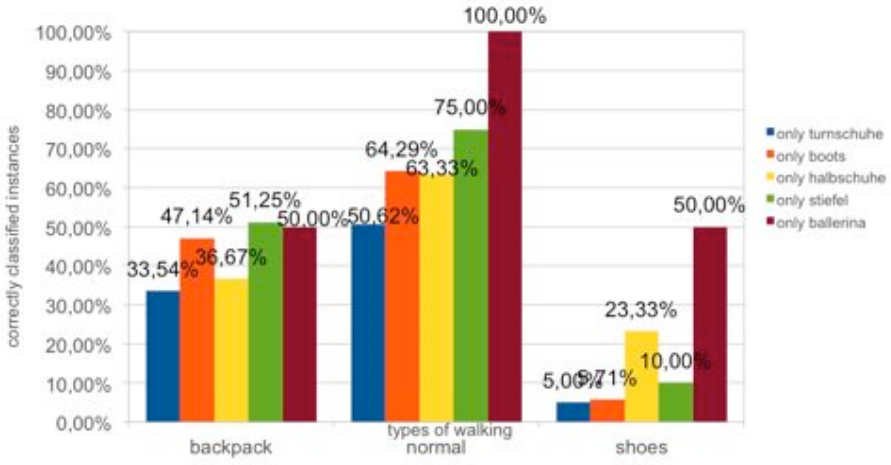

Figure 4.12: Classification of people who wore the same shoe type.

classification, we can say that the shoe type has a great influence as we obtained better results for the classification of people who were only wearing sneakers (who are the majority) than when we tested people walking in a normal way using people with all types of shoes (see Table 4.5).

4.2.9 Including a development set

The last of our experiments consisted in creating a new setup. In this setup, we had a training set, an intermediate development set and a testing set. The development set was used to tune the algorithm. In this occasion, the training set was composed of six audio files from the normal type of walking and the testing set included also the six audio files of the normal type of walking, two audio files of walking with a backpack and two audio files of walking with coating shoes. With this setup we will try a gender and shoe type classification, so the experiments are person independant, this means that no person in the training set appears in the testing set and viceversa. The training set consisted of 105 people, the development set of 50 people and finally the testing set contained 150 people.

A new concept has been introduced in this experiment, this is the weighted and unweighted average. In the weighted average, the data points do not contribute equally to the final average. Each value has been assigned a weight and these weights determine the relative importance of each quantity on the average. It is calculated in the following way:

We have different values: x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9 and x_{10}
And their weights are: w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9 and w_{10}
To calculate its weighted average you have to:

1. Multiply each value by its weight. $x_i * w_i$.

2. Add up the products of the values times the weight. $\sum_{i=1}^{10} x_i * w_i$.
3. Add the weight themselves to get a total weight. $\sum_{i=1}^{10} w_i$.
4. Divide the total value by the total weight: $\frac{\sum_{i=1}^{10} x_i * w_i}{\sum_{i=1}^{10} w_i}$

The unweighted average is simply the sum of all the values, divided by the number of values, taking into account the number of classes.

Furthermore, in these experiments we changed some of the SMO's parameters, such as the flter type (-N) and the complexity (-C) in order to optimize them. The complexity parameter, which SMO support vector machine uses to buld the hyperplane between any two target classes as explained in Section 3.6.2, controls how soft the margins of the hyperplane are. In practice how many instances are used as 'support vectors' to draw the linear separation boundary in the transformated euclidean feature space.

Below you can see an example of the command line used in the perl script in order to run the experiments.

my $complexity=0.005;
my $standardize = 0;
my $exponent = 1;
my $cmd = "java -Xmx14000m -classpath ../etc/weka.jar weka.classifiers.functions SMO -o -v -i -C $complexity -M -t ¨$trainfile¨T ¨$testfile¨L 0.001 -P 1.0E-12 -N $standardize -V -1 -W 1 -K ẅeka.classifiers.functions.supportVector.PolyKernel -C 250007 -E $exponent¨";

Our experiments in this occasion consisted in looking for an optimal value for the complexity parameter of the SVM as well as trying different combinations of the training set and the development set. In the following Tables 4.22, 4.23, 4.24 and 4.25 you can see the results combining different values of complexity and whether we trained with or without the development set, as well as testing with or without the development set. Similarly, in Graphs 4.13, 4.14, 4.15 and 4.16 you can also see the results for the weighted average recall and different values of complexity in a graph form.

In Tables 4.26 and 4.27 we can see the best results once we obtained the optimal value for the complexity and with the weighted and unweighted average recall. In Graphs 4.17 and 4.18 we the best results we have obtained up to now with the gender and shoe type classification, with the optimal value of the complexity and the weighted average recall. In both cases, the best results were when we used the development set for training.

After all these experiments, we can conclude that having used a development set has definitely improved our results, as well as finding an optimal value of the complexity parameter.

	C=0,01		**C=0,001**		**C=0,005**	
	WAR	UAR	WAR	UAR	WAR	UAR
Test dev B	61,00	58,85	65,00	62,40	63,00	60,65
Test dev N	66,00	64,60	68,00	65,40	67,67	66,20
Test dev S	52,00	54,20	58,00	55,90	55,00	56,05

Table 4.22: Gender classification results in percentage using training set, and testing with development set, and weighted and unweighted average recall.

	C=0,01		**C=0,001**		**C=0,005**		**C=0,0001**	
	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR
Test B	67,33	63,50	68,33	62,95	68,00	63,05	68,33	62,75
Test N	68,22	64,40	68,67	63,05	68,78	64,50	69,22	63,40
Test S	41,00	50,75	50,67	52,80	42,00	50,20	53,67	54,05

Table 4.23: Gender classification results in percentage using training set with development set, and testing set, and weighted and unweighted average recall.

	C=0,01		**C=0,001**		**C=0,005**		**C=0,00001**	
	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR
Test dev B	43,00	34,02	42,00	33,24	41,00	33,10	50,00	36,38
Test dev N	45,33	41,98	42,33	39,96	44,33	41,50	49,67	36,54
Test dev S	38,00	30,60	37,00	29,52	39,00	31,06	47,00	34,06

Table 4.24: Shoe type classification results in percentage using training set, and testing with development set, and weighted and unweighted average recall.

	C=0,01		C=0,001		C=0,005		C=0,00001	
	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR
Test B	54,67	33,82	55,33	33,10	56,33	34,82	65,33	36,74
Testing N	54,67	33,92	54,00	33,84	54,89	35,10	61,33	31,20
Test S	43,33	40,24	43,33	39,38	44,00	38,04	60,67	35,98

Table 4.25: Shoe type classification results in percentage using training set with development set, and testing set, and weighted and unweighted average recall.

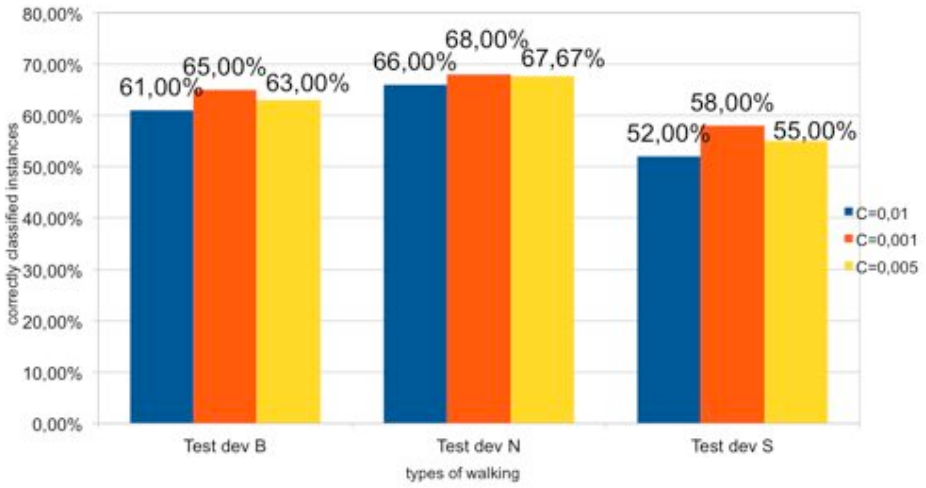

Figure 4.13: Gender classification with different complexity values, normal training and testing with development set.

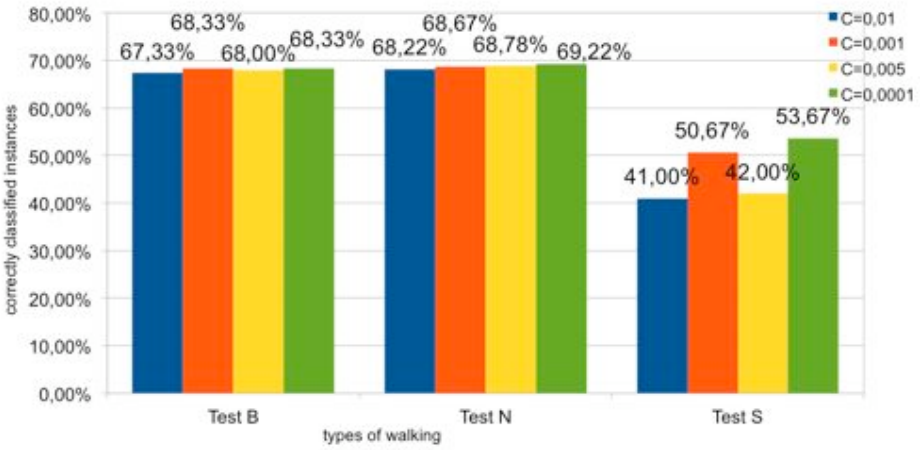

Figure 4.14: Gender classification with different complexity values, training with development set and normal testing.

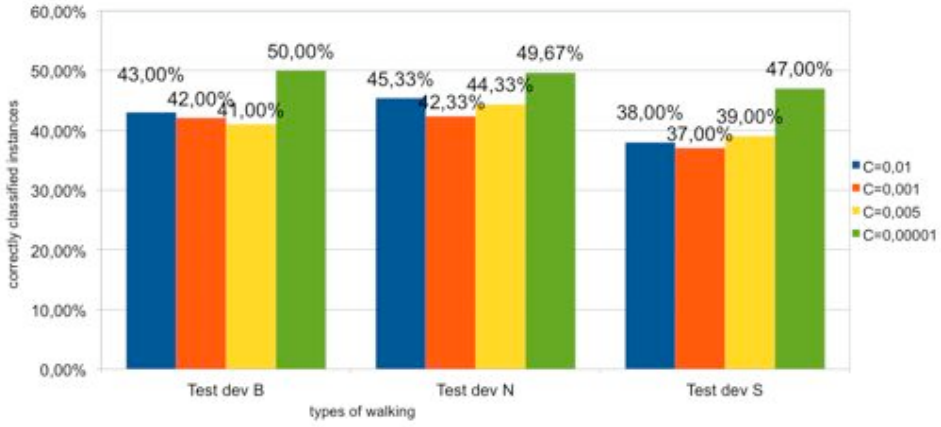

Figure 4.15: Shoe type classification with different complexity values, normal training and testing with development set.

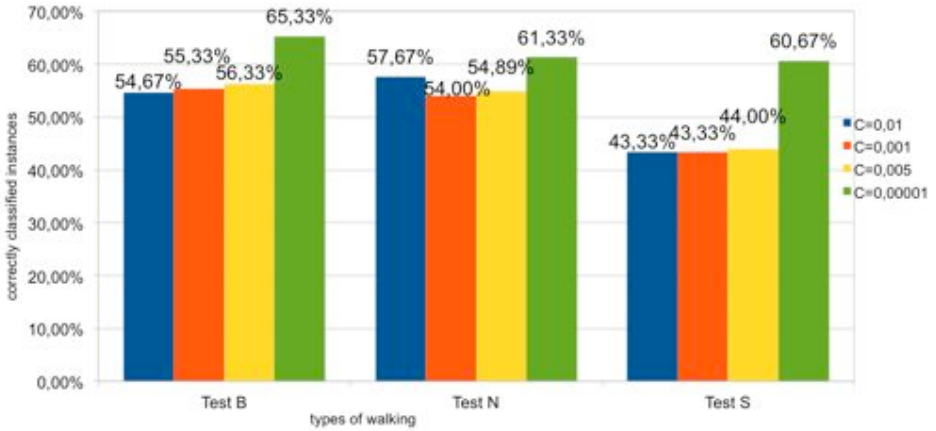

Figure 4.16: Shoe type classification with different complexity values, training with development set and normal testing.

	C=0,001		**C=0,0001**	
	WAR	UAR	WAR	UAR
Gender test B	67,67	62,20	68,33	62,75
Gender test N	69,00	63,10	69,22	63,40
Gender test S	59,67	57,80	53,67	54,05
Gender dev B	65,00	62,40		
Gender dev N	68,00	65,40		
Gender dev S	58,00	55,90		

Table 4.26: Gender classification results in percentage with optimal complexity values and development set, and weighted and unweighted average recall.

	C=0,00001		C=0,00001	
	WAR	UAR	WAR	UAR
Shoe type tes B	65,00	34,80	65,33	36,74
Shoe type test N	60,78	32,34	61,33	31,20
Shoe type test S	59,67	36,10	60,67	35,98
Shoe type dev B	50,00	36,38		
Shoe type dev N	49,67	36,54		
Shoe type dev S	47,00	34,06		

Table 4.27: Shoe type classification results in percentage with optimal complexity values and development set, and weighted and unweighted average recall.

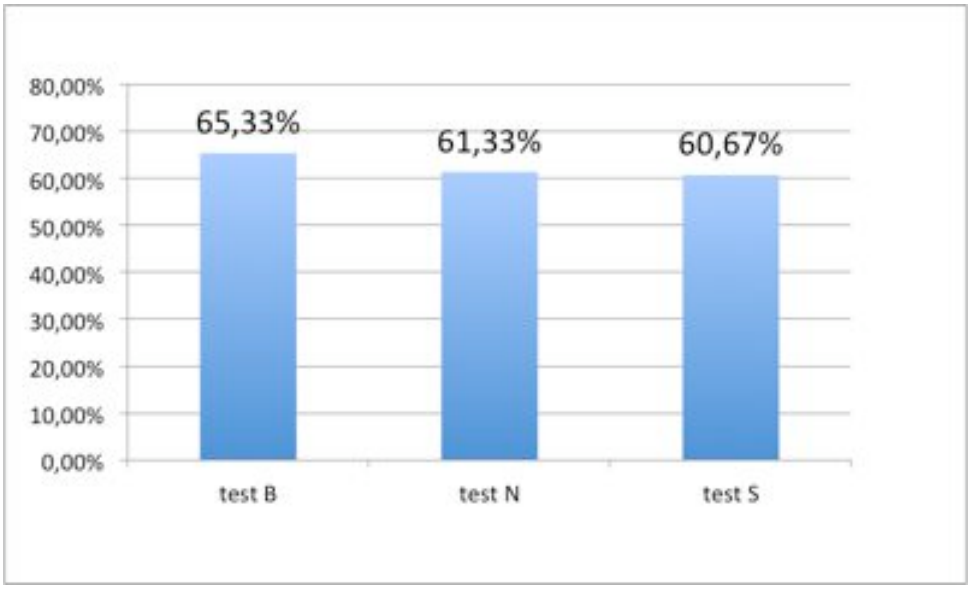

Figure 4.17: Best weighted average results for the shoe type classification, using the development set for training and normal testing, with the complexity value of C=0,00001.

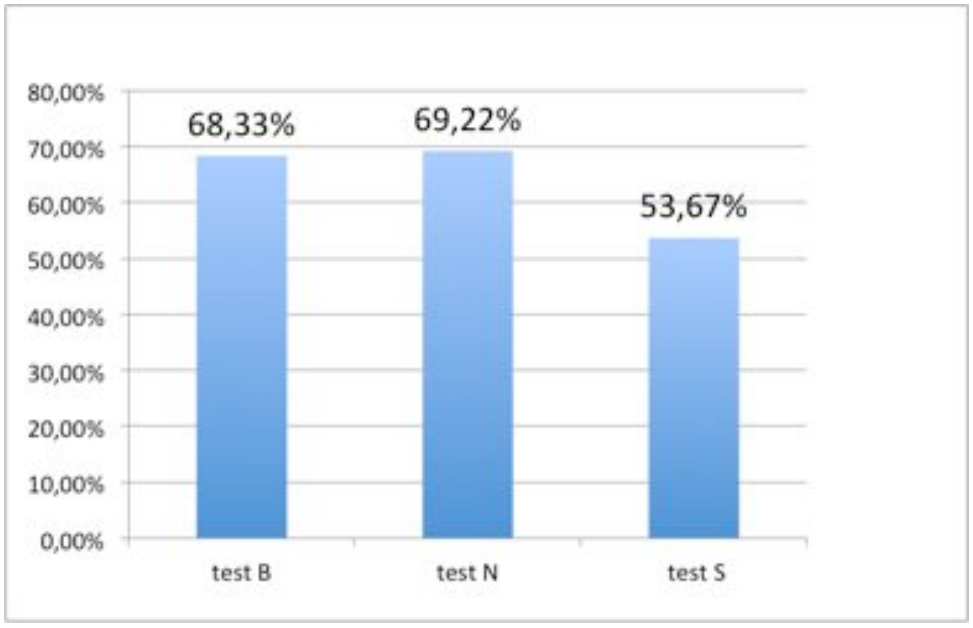

Figure 4.18: Best weighted average results for the gender classification, using the development set for training and normal testing, with the complexity value of C=0,0001.

5

Conclusion

In this chapter we are going to make a reminder about the aim of this thesis, as well as a review of all the information and results we have collected. Furthermore, there is also a section which pretends to give some ideas for future researchers interested in this theme.

5.1 Conclusion

In the beginning of this Master's Thesis we wondered whether it was possible to classify people according to the noise they made when they walked or not. After our research and experiments, we are now in the position to claim that it is possible to make a people identification by means of the acoustic gait. In order to achieve our goal, we first did some literature review on the field of acoustic identification. Unfortunately, this field is still quite unexplored so we have not been able to compare the results we obtained with any previous collection of information. Nevertheless, our results proved to be reliable and can be used as a base for future research.

However, we have not only focused on people recognition, but we have also tried a gender and shoe type classification. This is, we tried to identify whether it was a man or a woman walking from our audio files and which type of shoe they were wearing (sneakers, high boots, low boots, loafers or others). In Tables 5.1, 5.2 and 5.3 you can see the most relevant experiments we have carried out and their results.

Out of these experiments we can see that we had a higher percentage of correctly classified instances when we did the gender and shoe type classification than when we classified people. So, according to our results, perhaps instead of making a people classification we could focus more on trying to identify smaller characteristics such as height, weight, age, shoe type or gender.

305 People classification	
	Correctly classified instances
walking with backpack	25,74%
normal walking	39,84%
walking with coating shoes	3,44%

Table 5.1: People classification using the whole database (305 people), training with 4 audio files of the normal type of walking and testing with two audio files of the normal type of walking, two of the backpack and two of the coating shoes.

Gender classification with 305 people	
	Correctly classified instances
walking with backpack	68,33%
normal walking	69,22%
walking with coating shoes	53,67%

Table 5.2: Gender classification using the whole database (305 people), training with development set with six audio files of the normal type of walking and testing with six audio files of the normal type of walking, two of the backpack and two of the coating shoes.

Shoe type classification with 305 people	
	Correctly classified instances
walking with backpack	65,33%
normal walking	61,33%
walking with coating shoes	60,67%

Table 5.3: Shoe type classification using the whole database (305 people), training with development set with six audio files of the normal type of walking and testing with six audio files of the normal type of walking, two of the backpack and two of the coating shoes.

5.2 Summary

Overall, we have tried to give an answer to our motivation, which was trying to identify people by their acoustic gait. We have achieved our goal and we have even extended our experiments to a gender and shoe type classification successfully.

5.3 Future work

This section is intended for future researchers who may be interested in taking the following ideas further. We have created a starting point from which future experiments can be carried out.

In the same way as we have made a gender and shoe type classification, it would be also interesting to try if it could be possible an age and height classification. Further on, a new experiment could be introduced by adding noise to the testing sets and see if a coherent classification can still be done. Similarly, as we already did in section 4.2.6, we could have recordings from the same people walking but in different atmospheres and try then the identification. Moreover, another very interesting experiment would be to combine image and acoustic files to see if in this way we could obtain better results. In addition, all our experiments have been using the SVM classifier, however there are several other pattern classifiers that might give a good result. A good idea would be to make the same experiments but with other classifiers and compare the results. Finally, we can still increase the database by adding new recordings and trying to improve the results by optimizing parameters.

Bibliography

[AI06] Hiroshi Yasukawa Akitoshi Itai. *Footstep Recognition with Psyco acoustics parameter*. Aichi Prefectural University Nagakute Aichi Japan, 2006.

[AI08] Hiroshi Yasukawa Akitoshi Itai. *Footstep Classification Using Simple Speech Recognition Technique*. Aichi Prefectural University Nagakute 480-1198 Japan, 2008.

[AI10] Neuro AI. *Artificial Neural Networks*. http://www.learnartificialneuralnetworks.com/, 2010. accessed on 20/06/12.

[Ano10] Anonymous. *Kinect microphones*. http://www.computableminds.com/, post/Kinect/multiarray/microphone/how-works/xbox-360, 2010. accessed on 20/06/12.

[Ano12a] Anonymous. *Gnuplot*. http://www.gnuplot.info/, 2012. accessed on 19/06/12.

[Ano12b] Anonymous. *Maquinas con vectores de soporte*. http://www.google.com/urlsa=trct=jq=esrc=ssource=webcd=2ved=0, CFkQFjABurl=http3A2F2Fdis.unal.edu.co2Ffgonza2Fcourses2F20032, Fpmge2Fpresent2FExpoSVM.pptei=HoPhT9aUF8Pesgas-vhwusg=AFQjCNGTCV1qS1qAlZImh-EhP7rDxaNgRQsig2=sBLbm0e 2012. accessed on 19/06/12.

[Ano12c] Anonymous. *Overfitting*. http://www.dtreg.com/svm.htm, 2012. accessed on 27/07/12.

[Ano12d] Anonymous. *SVM vs Neural Networks*. http://www.svms.org/anns.html, 2012. accessed on 19/06/12.

[Bac12] Sebastian Bachmann. *Diplomarbeit: Personenidentifikation anhand der Gangart mittels Tiefendaten*. 2012.

[Bet05] Gustavo A. Betancourt. *Las maquinas de soporte vectorial (SVMs)*. Grupo de Instrumentacion y Control, Facultad de Ingenieria Electrica Universidad Tecnologica de Pereira, 2005.

[Bis11] Sriharsh Biswal. *Xbox 360 Kinect: technology description*. http://www.techibuzz.com/xbox-360-kinect-review/, 2011. accessed on 18/06/12.

[Bla06] Ross E. Bland. *Acoustic and Seismic Signal Processing for Footstep Detection*. B.S Massachusetts Institute of Technology, 2006.

[BSmP11] Florian Eyben Gary McKeown Roddy Cowie Björn Schuller, Michel Valstar and maja Pantic. *AVEC 2011 - The First International Audio/Visual Emotion Challenge*. Technische Universität München, Institute for Human-Machine communication, Munich, Germany, Imperial college London, Queen's University, School of Psychology, Belfast, BT7 1NN, UK, Twente University, EEMCS, Twente, The Netherlands, 2011.

[CLRC12] Royal Holloway University of London Computer Learning Research Centre. *Vladimir Vapnik*. http://www.clrc.rhul.ac.uk/people/vlad/, 2012. accessed on 19/06/12.

[com12] PortAudio community. *Port Audio*. http://www.portaudio.com, 2012. accessed on 19/06/12.

[Cra12] Stephanie Crawford. *Microsoft Kinect Camera*. http://electronics.howstuffworks.com/microsoft-kinect2.htm, 2012. accessed on 20/06/12.

[DDS05] Patrick Lucey David Dean and Sridha Sridharan. *Audio-Visual speaker identification using the CUAVE database*. Queensland University of Technology, Australia, 2005.

[DNTC09] L. Khoudour L. Douadi D-N. Truong Cong, C. Achard. *Video sequences association for people reidentification across multiple nonoverlapping cameras*. French National Institute for Transport and Safety Research (INRETS) Villeneuve d'Ascq France UPMC Univ Paris Institute of Intelligent Systems and Robotics Sur Seine France, 2009.

[EZ04] Mark Everingham and Andrew Zisserman. *Automated Person Identification in Video*. Visual Geometry Group Department of Engineering Science University of Oxoford, 2004.

[FE10] Björn Schuller Florian Eyben, Martin Wöllmer. *OpenSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor*. 25-29-10-2010.

[FE12] Björn Schuller Florian Eyben, Martin Wöllmer. *OpenSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor*. http://opensmile.sourceforge.net/, 2012. accessed on 19/06/12.

[Gro12] SQL Maestro Group. *SQL*. http://www.sqlmaestro.com/products/, mssql/maestro/?gclid=COrc54ayxrECFTMhtAodFzEAhQ, 2012. accessed on 19/06/12.

[KK07] Bhiksha Raj Kaustubh Kalgaonkar. *Acoustic Doppler Sonar for Gait Recognition*. Mitsubishi Electric Research laboratories, 2007.

[LME10] Mumtaj Begam Lindasalwa Muda and I. Elamvazuthi. *Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques*. Journal of computing, volume 2, 2010.

[Lue97] Juergen Luettin. *Visual Speech and Speaker Recognition*. Department of Computer Science University of Sheffield, 1997.

[McC01] Iain McCowan. *Microphone Arrays: A Tutorial*. April 2001.

[MR05] Oded Maimon and Lior Rokach. *Feature Extraction*. The Data Mining and Knowledge Discovery Handbook, 2005.

[oW12] The University of Waikato. *The University of Waikato*. http://www.waikato.ac.nz/, 2012. accessed on 19/06/12.

[Pal11] Elvira Palomo. *Kinect*. http://www.vadejuegos.com/noticias/2011/10/20/alex-kipman-creador-de-kinect-mi-trabajo-es-inventar-el-futuro-095659.html, 2011. accessed on 18/06/12.

[Phi02] P.J. Phillips. *Paper appearance in: Pattern Recognition 2002 Conference on*. Conference publications Nat. Inst. of Stand. and Technol. Gaithersburg MD USA, 2002.

[Rob10] I Heart Robotics. *Field of view*. http://www.iheartrobotics.com/2010/12/limitations-of-kinect.html, 2010.

[ROD12] David G. Stork Richard O. Duda, Peter E. Hart. *Pattern Classification*. second ed. edition, 2012.

[RSUdM] D. Martín-Iglesias A. Gallardo-Antolín C. Peláez-Moreno R. Solera-Ureña, J. Padrell-Sendra and F. Díaz de María. *SVMs for Automatic Speech Recognition: A Survey.* Signal Theory and Communications Department, EPS Universidad Carlos III de Madrid SPAIN.

[SCW12] Lonnie Smrkovski Steve Cain and Mindy Wilson. *Voiceprint identification.* http://expertpages.com/news/voiceprintidentification.htm, 2012.

[She04] Bin She. *Framework of footstep detection in indoor environment.* HCI Lab, Samsung Advanced Institute of Technology, Beijing Samsung Telecom RandD Center China, 2004.

[SjR05] Jaakko Suutala and juha Röning. *Combining classifiers with different footstep feature sets and multiple samples for person identification.* Intelligent Systems Group, Infotech University of Oulu, Finland, 2005.

[THT00] Lee Hotraphinyo Tejaswini Hebalkar and Richard Tseng. *Voice Recognition and Identification System.* Digital Communications and Signal Processing Systems Design, 2000.

[Vap98] V.Ñ. Vapnik. *Statistical learning theory.* New York, 1998.

[Wik12a] Wikipedia. *C, lenguaje de programacion.* http://es.wikipedia.org/wiki/C-28lenguaje-de-programaciC3B3n29, 2012. accessed on 19/06/12.

[Wik12b] Wikipedia. *Java Database Connectivity.* http://es.wikipedia.org/wiki/JDBC, 2012. accessed on 19/06/12.

[Wik12c] Wikipedia. *Karush-Kuhn-Tucker conditions.* http://en.wikipedia.org/wiki/KarushE28093KuhnE28093Tucker-conditions, 2012. accessed on 19/06/12.

[Wik12d] Wikipedia. *Maquinas de Vectores de Soporte.* http://es.wikipedia.org/wiki/MC3A1quinas-de-vectores-de-soporteComparativa-SVM-vs-ANN, 2012. accessed on 19/06/12.

[Wik12e] Wikipedia. *SIGKDD.* http://en.wikipedia.org/wiki/SIGKDD, 2012. accessed on 19/06/12.

[Wik12f] Wikipedia. *TCL.* http://es.wikipedia.org/wiki/Tcl, 2012. accessed on 19/06/12.

[WL12] Xiangyang Xue Wei Li, Yaduo Liu. *Robust Audio Identification for MP3 Popular Music.* School of Computer Science and Technology, Fudan University, Shanghai, China., 2012.

[YSY04] Takashi Takasuka Yasuhiro Shoji and Hiroshi Yasukawa. *Personal Identification Using Footstep Detection.* Aichi Prefectural University Nagakute Aichi Japan, 2004.