M 2014

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

CONTRIBUTIONS TO THE AUTOMATIC RECOGNITION OF PORTUGUESE SIGN LANGUAGE

CARLOS JORGE ALVES DA HORA MARTINS

DISSERTAÇÃO DE MESTRADO APRESENTADA
À FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO EM
ENGENHARIA DA INFORMAÇÃO

Resumo

Este trabalho insere-se num projeto de investigação com vista ao desenvolvimento de uma plataforma móvel de reconhecimento automático de línguas gestuais e foi realizado sobre uma base de dados da Língua Gestual Portuguesa (LGP) criada recentemente. As contribuições do mesmo para o projeto passaram pelo estudo diferentes tipos de características que podem ser usadas para classificar gestos estáticos e dinâmicos, e pela avaliação do impacto que a inclusão de informação de profundidade pode ter no reconhecimento de línguas gestuais.

No que concerne aos gestos estáticos, verificou-se que as características regionais, baseadas na silhueta dos objetos, são as que melhor descrevem a mão, em detrimento de características baseadas em medidas estatísticas ou de histogramas de orientação de gradiente. Além da base de dados da LGP, foi também utilizado o conjunto de dados de Triesch que corresponde a uma base de dados de referência para o estudo de gestos estáticos. A comparação dos resultados obtidos evidenciou que a naturalidade de execução dos gestos da base de dados LGP influenciou negativamente a classificação obtida.

Para os gestos dinâmicos, com componente de movimento, foram avaliados dois tipos de características: baseadas na posição absoluta das mãos e no seu movimento relativo. As características baseadas no movimento relativo foram as que retornaram melhores resultados, mostrando-se mais resilientes às velocidades de execução dos gestos e posição das mãos relativamente ao corpo.

No presente estudo a informação de profundidade existente não revelou melhorias significativas na classificação de gestos estáticos ou dinâmicos, embora se reforce a sua importância para a definição de novas representações das mãos, baseadas na estrutura das mesmas e não apenas na sua aparência.

Abstract

The present work is integrated in a research project which aims the development of a mobile platform of sign language automatic recognition and was realized over a recently created Portuguese Sign Language (LGP) database. The contributions for the project were the study of features that can be used in the classification of static and dynamic gestures and the impact evaluation of the depth information inclusion in the recognition of sign languages.

In what concerns the static gestures, it was observed that the regional features, based on the object's silhouette, are the ones which best describe the hands, surpassing the performance of features based on statistical measures and gradient orientation histograms. Besides the LGP database, the Triesch dataset, which corresponds to a benchmark database for the study of static gestures, was also addressed. The obtained results comparison revealed that the LGP gestures execution naturalness affected negatively the obtained classification.

For the dynamic gestures, with movement component, two types of features were evaluated: the first based on the hands absolute position and the second on the relative movement. The relative movement based features returned the best results, proving to be more resilient to the gesture execution speed and hand position with respect to the body.

No meaningful improvements in the classification of static and dynamic gestures were revealed by the inclusion of depth information. However, this type of information has a high potential to be used in new forms of hand representation, based not only on the hands' appearance, but also on their structure.

Agradecimentos

Este trabalho não teria sido possível sem o apoio...

... do Professor Jaime Cardoso e da Doutora Ana Rebelo que me orientaram ao longo de todo o trabalho, sendo decisivos nos momentos chave, e que estiveram disponíveis sempre que foi necessário;

... do Doutor Ricardo Sousa e do Pedro Ferreira que integram o projeto e que contribuíram para a definição inicial dos objetivos;

... das Professoras Ana Sofia Paiva e Ana Rio que despenderam algum do seu tempo para nos ensinar Língua Gestual Portuguesa;

... da Escola Artística de Soares dos Reis e do Agrupamento de Escolas Eugénio de Andrade: Escola EB 2/3 de Paranhos cujos alunos colaboraram na construção da base de dados;

... do VCMI que me acolheu e em particular do Pedro Monteiro que me auxiliou no tratamento da informação de profundidade;

... do INESC enquanto instituição de acolhimento.

A todos o meu sincero agradecimento,

Carlos Martins

Contents

List of Figures

List of Tables

List of Acronyms

DF	Depth Features
DTW	Dynamic Time Warping
GOH	Gradient Orientation Histograms
HM	Hu Moments
HMM	Hidden Markov Models
LGP	Língua Gestual Portuguesa (Portuguese Sign Language)
RF	Regional Features
SL	Sign Language
SLR	Sign Language Recognition
SVM	Support Vector Machines
VCMI	Visual Computing and Machine Intelligence

Chapter 1

Introduction

Communication is one of the most fundamental characteristics of the human evolution. Its implications go far beyond the communication with each other. It defines how our brain is structured and so the way how we relate to the world around us. Most of us have the ability of communicate with the voice and audition, using the sound as medium. For those who are not able to use these senses the relation with their environment is a bit different. Sign Languages emerged inside the deaf communities as the preferential way of communication. They are spread around the world with a wide variety and complexity and can be as rich as spoken languages.

The technological evolution to which the world has assisted since the end of past century, created several forms of everyday life simplification. We have cellphones and computers for communication and work, medical devices to treat us when we are diseased or internet for almost everything, to name just a few examples. It is quite natural to imagine that we can use these resources to facilitate our communication with who does not know our language or cannot use the same communication medium. Sign Language computational recognition has a fundamental role in the communication with deaf people in the upcoming future.

Sign Language Recognition (SLR) systems have several applications. They can be used in translation systems to convert signs in sound and vice versa. This is important not only to ease the communication between deaf and hearing people but also increase the amount of contents to which the deaf can access. The creation of a visual dictionary is another interesting possibility. There are already dictionaries which show us the gestures correspondent to a given spoken word but it is not possible yet to know the meaning of a given gesture. Academically, SLR systems are a particularly interesting case study of the gesture recognition field which has applications in the new man-machine interaction systems.

This work is a contribution to the study of SLR systems and is integrated in a research project which aims the development of a mobile platform of SLR. A recently created database of Portuguese Sign Language (LGP) was used. This database was created at INESC with the

cooperation of some native signers, and is composed of several isolated signs and sentences. At the moment there are no studies or applications in LGP automatic recognition and so, this work provides the first tools and will enhance new paths of investigation.

The study is divided in two main types of gestures: the static gestures, which only have hand shape component, and the dynamic gestures which have also movement and position with respect to the body. Several features that can be used to describe each type are presented associated to the state-of-the-art models that can be used to classify them. Given the existence of depth information in the LGP database, the use of this type of information is also evaluated.

This work is divided in the following way: Chapter 2 is described briefly what is a Sign Language and is presented the LGP in more detail. In Chapter 3 is summarized the State-of-the-Art in SLR. Chapter 4 presents the LGP database and the preprocessing applied to the data. In Chapter 5 the static gestures classification performed are explained and in Chapter 6 the same is done to the dynamic gestures. In the end Chapter 7 presents the final conclusions of this work.

Chapter 2

Sign Languages Overview

2.1 The Sign Languages Importance

The communication development is one of the most important aspects that enabled the human evolution as we know. It defines our structure of thought and the way how we develop ourselves inside our communities and societies. Languages are in a process of constant evolution and result directly from the interaction between their users. By just looking to the different spoken languages across the world it is possible to see these evidences and the distinction between them is more sharp as more isolated the communities are. While very important to the development of spoken languages, these aspects of the communication evolution contributed to the isolation and marginalization of deaf and mute people.

All children are born with the innate ability to learn a language. Especially during the first years it is fundamental, to the cognitive development, that the children are surrounded of a communication environment. In the case of the deaf children however this condition may be hard to guarantee. A deaf can learn spoken languages, training lipreading. This is not possible however while they are growing and need to learn fundamental concepts as the notion of father or mother. The SL is then the best option to teach and help these children in their development because it uses a medium of information transmission that is accessible to them.

The importance of this form of communication is enormous to the deaf communities and is mandatory that a society that prizes the integration of everyone and especially of the disadvantaged, that provides the means to teach and disseminate the Sign Languages to whoever needs it.

2.2 The Sign Languages History

There is not much information on how deaf people behave inside their communities before sixteenth century. It is quite easy however to imagine that they probably grouped themselves and

developed their own ways of communicate with each other. Even the ones that born deaf inside hearing families, without contact with other hearing impaired people, might have created some form of making their relatives understand them. The ability of communicate is natural to the human condition, independently of the medium in which it is carried. Even the hearing people execute unconscious gestures when explaining an idea to the others, sometimes on the telephone.

One of the first teachers known of hearing impaired people was Frey Ponce de Leon to who was trusted the education of Spanish nobles on the middle of sixteenth century. The deaf education at that time was a luxury because a deaf without a title was not even considered as a person according to the law in force. To his successor, João Pablo Bonet, is due the first SL known book [1].

According to Stokoe [2], until the eighteen century the teaching of deaf people was performed by tutors that taught few pupils with unclear or secret methods. The language taught was the local spoken language, with the same grammar and lexicon, and the words where spelled letter by letter using the correspondent sign. L'Épée demonstrated however that it was possible to habilitate someone to decode a letter or a sign without really understand it, showing the disadvantages of the methodologies used until then. He was one of the first to understand the naturalness of sign communication and created a teaching method based on signs from deaf communities. L'Épée founded a school to teach sign languages to deaf people in 1750 and his work was one of the most important contributions to the evolution of sign languages around the world as we know now.

Some years later, Gallaudet was sent from Connecticut, USA, to Europe to acquire knowledge about the teaching methods of sign languages. He was received by Sicar, one of the L'Épée successors, which sent Clerc with him to aid in the creation of school for hearing impaired people in USA. This is the reason why the American Sign Language (ASL) is much more similar to the French Sign Language (LSF) than with the British Sign Language (BSL) and is also a proof that there is not necessary correspondence between some local sign and spoken language. The same happen with the Portuguese Sign Language (LGP) and the Brazilian Sign Language (LIBRAS).

The discussion around the Sign Language existence and definition as a language was far from being over. A few years after the creation of the SL school in USA, it was opened a lively discussion opposing the combined method of Edward Gallaudet to Alexander Bell, a supporter of the oral method. This discussion was carried in 1880 to the Second International Congress on Education of the Deaf, the famous congress of Milan, where several resolutions were approved, from which the following stand up [3]:

Resolution 1: *"The Convention, considering the incontestable superiority of articulation over signs in restoring the deaf-mute to society and giving him a fuller knowledge of language, declares that the oral method should be preferred to that of signs in education and the instruction of deaf-mutes."*

Resolution 2: *"The Convention, considering that the simultaneous use of articulation and signs has the disadvantage of injuring articulation and lip-reading and the precision of ideas, declares that the pure oral method should be preferred."*

While in Europe every country accepted and implemented the Congress resolutions, USA and Britain opted to maintain their combined method. The technological progress observed after the Second World War brought new devices able to increase the auditory capacity of the deaf. However, the results were not very satisfactory and the deaf continued showing a retarded development when compared to the hearing people, contributing to their social exclusion.

Only in 1962, with the study of Stokoe [2], the first SL linguist, the ideas behind the Milan resolutions which defended that the SL was poor, rudimentary and without structure were abolished. Stokoe demonstrated that the SL is equivalent to the verbal languages, with own grammar and phonetic level. As result of this remarkable work the SL was adopted as the first language of deaf people all over the world.

In Portugal the LGP definition starts only in the decade of 1980 when it was collected the first signs of the LGP lexicon and defined a first grammar. In 1997 the Portuguese Sign Language gain the legal status of official language of the deaf community. It is now, hand in hand with the Portuguese Language and Mirandês, an official language of Portugal.

2.3 The Portuguese Sign Language

Since the Portuguese sign language was recognized as a language there were developed several works [4-6] with the purpose of homogenize all known "dialects" in a single language. Given the communication isolation of the deaf communities, it is quite normal the existence of regional variations, sometimes so pronounced that almost constitute a different language. Wittmann [7] did a comparative study of several sign languages across the world and classified them according to their origin and similitude. He states that LGP was created from Swedish Sign Language (STS) that, by its turn, is related to the BSL.

The Portuguese Sign Language has its own grammar, lexicon and syntax. The information is organized in sentences that in turn are composed of isolated signs. An isolated sign can be divided in cheremes, the equivalent to the phonemes of spoken languages. These cheremes are the basic structural unit of sign language gestures and are constituted of 5 elements: hand configuration, place of articulation, movement, orientation and facial or body expressions [6].

- **Hand Configuration**

In the LGP the hands do not have the same importance. There is a dominant hand, which carry the most relevant information, and a supporting hand that is used to complement some gestures. The choice of the right or left hand as dominant hand is free to the signer.

Hands can adopt countless configurations by just moving the fingers. There are several gestures which require specific hand configurations as the alphabet or the ordinal numbers.

Many words Portuguese words, like the names of people or cities, do not have a correspondent gesture in the LGP. It is usual in this case to sign them character by character using the correspondent signs.

- **Orientation**

The hand orientation is directly related to the hand configuration and is defined by the palm placement. In some cases the hand orientation inversion alone indicates the gesture opposite.

- **Place of Articulation**

Since the sign language information is carried through the visual medium, the place where the gestures are realized is extremely important. There are two main places where the gestures can be realized: the virtual rectangle in front of the signer and the plane of the signer body where some spots can be used as references to some signs.

- **Movement**

Some gestures are static, as the cardinal numbers which do not use movement to transmit information, while others are dynamic, as verbal subjects that have the movement as the main component. The gestures speed and duration may have also an interpretation.

- **Facial and Body Expression**

The facial and body expression has also an important role. This kind of information can be used to complement other gestures and from them may depend the distinction of a statement from a question.

The organization of sentences follows a Subject-Object-Verb (SOV) structure in opposition to what happen in the spoken languages. The next example shows how the sentence - "The cat eats fish" - would be pronounced in the LGP. The sentence is in the gloss format that is normally used to represent signs with text.

EL: The cat eats fish.

LGP: **CAT + FISH + EAT//**

Nevertheless, there are some situations where the structure Object-Subject-Verb (OSV) can be used.

In the same sentence the negation can be transmitted just by adding the gesture /NOT/ after the verb neutral form.

EL: The cat does not eat fish.

LGP: **CAT + FISH + EAT + NOT//**

The time notion is transmitted using a similar approach. The following two examples show how to express the idea of past or future respectively.

EL: The cat ate fish.

LGP: **CAT + FISH + EAT + PAST//**

EL: The cat will eat fish tomorrow.

LGP: **CAT + FISH + EAT + TOMORROW//**

The inclusion of facial and body expression can be used as a complement of other gestures. The time notion for example can be reinforced by performing the same gesture in a plane further away from the signer, to the front to represent future or to the back to the past.

Chapter 3

State-of-the-Art

SLR is an interesting case study of gesture recognition and there are already many studies and research teams working on it. This section aims to present some of the most relevant contributions in this field that serve as basis to the current work.

The first step to be considered is data acquisition method. There are two main ways of gather information from sign language gestures: using wearable hardware equipment or in a vision based way [8]. The wearable hardware equipment corresponds to the use of data gloves or similar equipment that store information of the hand and fingers position and their relative movement. This is the most accurate way of recording gesture information and some studies [9, 10] have reported good results in the SLR with this kind of data. However, this approach is not real world scenario. A system that aims to ease the interaction with deaf people should be vision based.

Vision based data is harder to handle in terms of features extraction because of the difficulties of correctly segment the hands and face in the images. Most of the analyzed studies, used data recorded in controlled environments and some of them with visual markers to aid the hand and finger tracking. It is the example of Holden and Owens [11] that used gloves with colored finger joints to ease the features extraction. The majority of the studies, e. g. [12-14], used images recorded in studios with dark background and signers wearing clothes with non-skin color and long sleeves. There are few examples of systems created to be used in uncontrolled environments without background restrictions [15]. More recently, with the emergence of depth cameras like Kinect from Microsoft, some authors presented systems that used this type of information [16]. Figure 3.1 shows different data acquisition methods used in SLR.

The state-of-the-art techniques exposed are mostly based on systems that used vision based data without optical markers. This field can be divided in three main areas as presented in Figure 3.2: segmentation and tracking, features extraction and recognition. Section 3.1 describes briefly the segmentation stage, section 3.2 presents the most common features used to represent gestures according to the literature and in section 3.3 is shown a review of the recognition

methodologies. In the end, in section 3.4, is presented a selection of available databases that can be used in SLR.

Figure 3.1 - Data acquisition methods: data gloves; gloves with colored fingers; gloves with distinct color; studios with contrasting background and long sleeves; unconstrained environment and depth images.

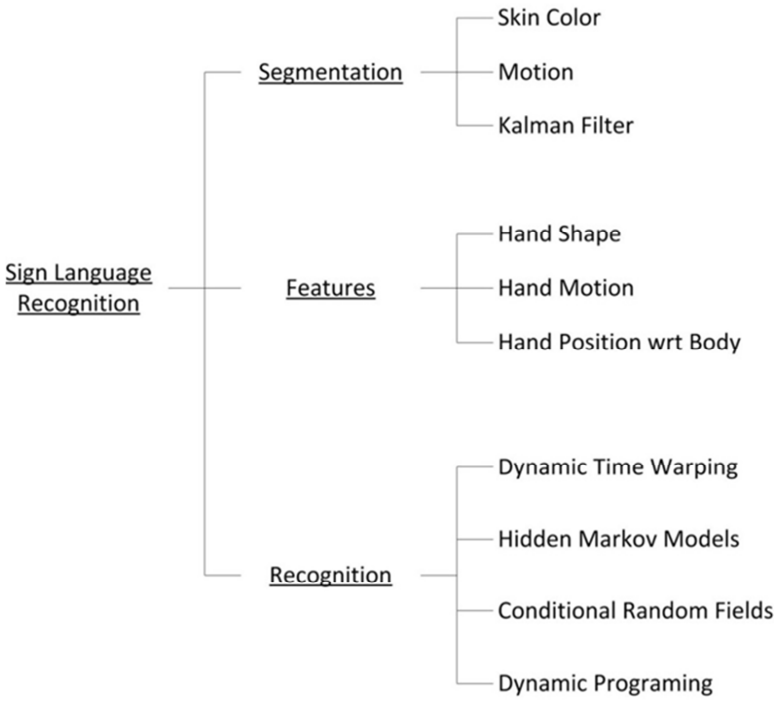

Figure 3.2 - Some vision based hand gesture recognition techniques.

3.1 Segmentation

In order to create a system able to recognize SL gestures it is necessary to get information from images. An image is a set of pixels with values representing colors. The direct appreciation of this type of data by a computer is not very informative. To use images it is necessary the application of some kind of transformation. In the particular case of gesture recognition the first step of this transformation is the face and hands segmentation. The segmentation can be divided in two main areas: detection and tracking. Detection comprises all methodologies applied to individual frames in order to identify the hands, face and other body parts considered relevant. The tracking in other hand, corresponds to the use of previous or posterior information of the position of the objects in study and predict the current position.

In gestures that are performed without gloves, a good characteristic that can be used is the color of the skin. Most of the studies in SLR used skin color models to distinguish skin pixels from the rest of the image. Jones and Rehg [17] presented a histogram based model to identify skin pixels in the RGB space. The model got good performances in the distinction of images with or without people for Web filtering purposes. Posterior studies suggested the use of color spaces with the chromatic component separated from the luminance. It is the case of Khan, et al. [18] that did a comparative study between several clustering methods in different color spaces and determined that the 2D color spaces based on the hue and saturation returned the best results. Cooper and Bowden [13] presented a color model that is refined by the color of the face. The faces are detected using Viola-Jones algorithm and from them a skin color Gaussian model of the signer skin is obtained.

In real world scenarios these models may detect background objects as skin or even detect other people. There are not, yet, satisfactory answers in the literature to surpass this difficulty in approaches based in the skin color. A solution is proposed by Von Agris, et al. [15] that suggest a background model, corresponding to the median of a set of initial images, which can be used to filter the background in the sequence.

Motion and shape information are also addressed in some works but without satisfactory results in terms of accuracy and computation speed. Awad, et al. [19] presented a system to segment the skin that combines skin color, motion and position models. The authors defined motion in a two-step process: first they computed the difference between the correspondent pixels of two subsequent images and then they transformed these distances in probabilities, assigning higher probabilities to pixels that were previously defined as skin. A regular Kalman filter was used to estimate the position of the hands from one frame to the other. After that, the distance between the image and the predicted template was computed and normalized in order to be in the form of a probability. The final decision was done by combining the results of the three models. Another segmentation difficulty is deal with occlusion. Many gestures in all sign languages are performed in front of the face or with the hands touching or hiding each other and forming a blob in the computer vision point of view. Some authors [15, 20] choose to deal with these blobs as single objects without trying to divide it. Holden, et al. [12] proposed the use of

an active contour model combined with temporal variance information to distinguish overlapped object. Figure 3.3 is an example of this method results.

Figure 3.3 - Holden, et al. [12] proposal to deal with occlusion.

Segmentation still is a critical area of the SLR. The majority of the studies analyzed have strong restrictions in terms of background and clothing which ease the objects detection, improving the performance of the respective systems, but are far to be a real world scenario.

3.2 Manual Features

Messages in sign languages have two main components: the manual and non-manual. Manual component includes all information that can be carried by the hands. It can be divided in: hand shape, motion and position with respect to the body. The non-manual component by its turn includes the facial expressions and the movement of the head and body. Below are described some of the state-of-the-art features that can be extracted from the hands.

The field of gesture recognition starts with the identification of static hand shapes. Freeman and Roth [21] propose a system based on the histogram of orientations that was independent of luminance but it was not independent of hand orientation and identifies different postures as the same. Chang, et al. [22] used Zernike Moments (ZMs) and Pseudo-Zernike Moments (PZMs) in the identification of 6 hand gestures. They worked with binary images and got some miss classifications due to segmentation inefficiencies. Kelly, et al. [23] used size functions based on hand contour and Hu moments as features to represent the hand shapes. The authors claim accuracies above 95 % in the identification of the alphabet letters from the ISL dataset with 24 signers.

Most of the SL gestures include motion together with the hand shape. Frequently, different hand motion with the same hand shape represents a totally different sign and vice-versa. Starner, et al. [20] suggested the use of regional characteristics as features. After detecting the hands blobs the authors extracted from each the *x* and *y* position, their derivatives, the blobs area, the angle of the eigenvector of least inertia and its length and the eccentricity. This corresponds to feature vector of sixteen elements that was further used to training. These values were not normalized turning this method dependent of the signer and of the distance to the camera. To deal with occlusion the authors propose the computation of the features of the joint blob and the use of it to both hands.

A similar approach was used by Von Agris, et al. [15] but with 11 features of each hand and non-manual features from the facial expressions. To turn their model independent, the authors normalized the absolute features according to the head position and shoulders distance estimation. Occlusion was also treated differently. The features of each hand during the occlusion periods were linearly interpolated from the conditions before and after occlusion. This method however is still dependent of a correct normalization that may be difficult due to signer morphological independence, e. g. in terms of hand size. It also did not address the importance that the relative position of the hands with each other and with the head. Holden, et al. [12] propose the use of features based only on the relative position of the hands and face as the left scheme of Figure 3.4 shows. The feature vector was: *cos* θ_1, *sin* θ_1, *cos* θ_2, *sin* θ_2, *cos* θ_3, *sin* θ_3, *DRt*, *DLt and SRt / SLt*. *DRt* and *DLt* correspond to the roundness of the right and left hand respectively and S corresponds to the area of the hands. Yang, et al. [14] used a similar method but with a different set of features.

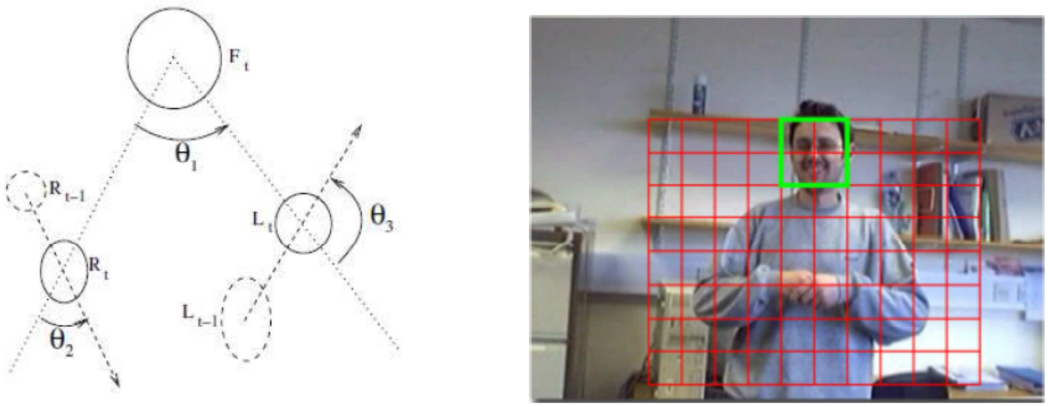

Figure 3.4 - (left) Angles extracted by Holden, et al. [12] to compute the feature vector. (right) Cooper and Bowden [13] grid division.

A different approach was introduced by Cooper and Bowden [13]. The authors proposed a method to classify each viseme according to its category [24]: placement, movement and

arrangement. Their method consists on binarizing the image according to the skin and then quantizing the result with a grid defined by the face location. Each square of the grid has a quarter of the area of the face that is located in the upper center of the grid as shown in the right image of Figure 3.4. Each square is turned white if more than 50% of it is skin and in black otherwise. From the resultant image they computed several measures like the image moments.

With the emergence of depth cameras some authors propose systems based on 3D information. It is the case of Kurakin, et al. [16] that used a feature vector with speed, rotation and shape information. The authors modeled the shape using two components. The first consisted on dividing the hand bounding box with a grid and building a feature vector with the information of hand percentage in each square. The second corresponds to the division of the same bounding box in a polar grid centered in the mass center of the segmented hand and registration of the distance from this center to the hand boundary in each polar division.

The diversity of features used in SLR studies is a reflex of the embryonic state in which this area is. Almost all studies use a different set of features, sometimes corresponding to variations of the existent but usually being entirely different. The problem with this diversity is that these studies do not compare the results of their features with others previously proposed and since almost all of them use different gestures, it is hard to take conclusions about the features quality and applicability. This lack of normalization in terms of procedures make this area very dispersed and hampers the use of previous works to develop new ones.

3.3 Recognition

The SL recognition can be divided in two main areas: the recognition of isolated signs and the recognition of continuous signs. An isolated sign corresponds to the execution of a single gesture that represents a single word or idea. Martínez, et al. [25] divided the isolated signs in three parts: 1) movement from rest position to the place where the sign starts; 2) the sign itself; 3) movement of returning to the rest position. In continuous sign language recognition the objective is the detection of sign language sentences. These sentences correspond to a sequence of gestures that alone represent words. The gestures execution is realized in a continuous manner without returning to the rest position during the gestures transition. The transition is done naturally from the end position of one gesture to the start position of the next and it is called sign epenthesis [26]. This is a source of confusion to the recognition task and it is the reason why many initial studies only work with isolated signs. Additionally, the study of isolated signs provides interesting clues of what type of features should be used in SLR.

One of the simplest methods to recognize gestures is the motion comparison, ignoring the hands shape and position with respect to the body. Athitsos, et al. [27] used the Dynamic Time Warping (DTW) metric to compare the movements of the test set of gestures to the training and then classifying according to the smallest value obtained. Chai, et al. [28] proposed a similar

approach with 3D movement data from Kinect with an equivalent comparison measure. Instead of using the standard DTW the authors computed the Euclidean distance between trajectories.

The most common method to recognize signs are the Hidden Markov Models [12, 15, 20]. One of the first works that applied this method in a computer vision environment was proposed by Starner, et al. [20]. After tuning the model, the authors determine that the best results correspond to an HMM with four states. A similar approach was proposed by Von Agris, et al. [15] but with parallel HMMs classifying different aspects of the same gesture. The left scheme of Figure 3.5 shows an example of how this method works.

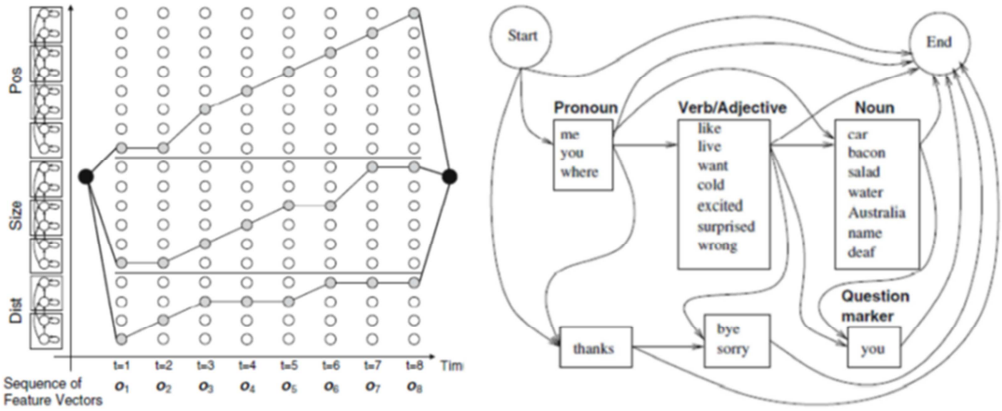

Figure 3.5 - (left) Example of parallel HMM from Von Agris, et al. [15]. (right) Grammar structure used by Holden, et al. [12].

Continuous sign language recognition brings additional difficulties. A continuous gesture corresponds to sentence composed of two or more words. These words are normally organized according the grammar structure of the correspondent sign language. Holden, et al. [12] included grammar constraints to ease the sentence level recognition as right image of Figure 3.5 shows. This type of imposition prevents the model of identifying some words in the wrong position of the sentence. Sign languages have characteristics that do not exist in spoken languages. One of them is the existence of non-intentional signs, as the previously mentioned epenthesis, that might confuse the recognition system or even the interlocutor. An automatic system must be able of distinguish each individual word in a sentence. Recently several studies proposed methodologies that are able to handle with this problem. Yang, et al. [14] proposed a model based on Conditional Random Fields able to distinguish between vocabulary signs from non-vocabulary signs. A different approach is proposed by Yang, et al. [14] who handled the problem using Dynamic Programing.

3.4 Databases

The sign language recognition with computer vision data requires the acquisition of a significant amount of different gesture videos. To improve the classification accuracy, each gesture should be performed several times and, in systems that intent to be signer independent, the acquisition process should be repeated by each subject. Additionally, videos must be annotated in a frame-by-frame basis to allow posterior classification. This is an onerous and expensive process that requires the cooperation of native signers or interpreters.

Many of the presented studies on this field work with their own databases that are not easily or totally accessible to consult. This hampers the comparison between systems that work with different datasets and so the conclusions to take about their performances. To surpass these difficulties some benchmark datasets were created with the respective annotations. They can be organized in two groups depending on the recognition task to which they can be used for: the isolated sign recognition and the continuous sign recognition. In the following subsections are presented some of the more relevant databases in SLR.

3.4.1 Purdue RVL-SLLL ASL Corpus

The Purdue RVL-SLLL ASL Corpus (2002) is an available database based of the American Sign Language (ASL) presented in [25]. It has 2 parts, one composed of isolated signs and the other of continuous. The first part contains a set 39 gestures that the authors called primitive motion, corresponding to basic hand movements in ASL, and 62 gestures representing the English alphabet and the numbers from one to twenty. Each gesture was done only once by each signer in the most accurate way possible. The second part includes 10 groups of two sentences (paragraphs) performed sequentially and recorded in the same video. The two sentences, corresponding to different utterances, were not divided which can present some problems to the recognition task.

The experiment was performed by 14 native ASL signers in two different light environments: with diffused light to reduce the shadows and with directed light to increase contrast. The studio of the experiment had a uniform and contrasting background but it is not referred any clothing constraint.

3.4.2 RWTH-BOSTON Corpora

The RWTH-BOSTON Corpora is a set of databases summarized in [29]. These databases were created to serve as benchmark datasets of ASL to SLR studies. The acquisition conditions were the same to all databases: dark studio background and the signers clothing was constrained with long sleeves and non-skin color. Each gesture was recorded by 4 cameras: two in the front, to stereo vision purposes, one on the side and the other in the front, with close zoom on the face. The last one stored the images in RGB and the others in grayscale.

RWTH-BOSTON-50 Corpus (2005) was the first database created to be used in isolated sign recognition. It is smaller than the correspondent of Purdue, with only 3 signers and a vocabulary

of 50 words. In contrast to what was done in the Purdue database, each gesture was repeated more than once in a total of 483 utterances.

The RWTH-BOSTON-104 database (2007) was one of the first benchmark databases created to continuous sign recognition and is still one of the more relevant and used. It has more than 15k annotated frames corresponding to 201 sentences performed by 3 native signers. More recently [30], on this database was annotated information about hand and head position for tracking purposes.

The last and larger database is the RWTH-BOSTON-400 database (2008). It has a total of 843 sentences and vocabulary of 104 words, performed by 4 native signers. This corresponds to almost 70k images that, for now, do not have ground-truth information. One particularity of this dataset is that it does not have only the annotation of the glosses but also of the English translation.

3.4.3 SIGNUM Corpus

The presented databases are not yet suitable to signer-independent continuous sign language recognition. The Purdue database has a small number of sentences and the RWTH-BOSTON because it has only 4 signers. A more complete database is presented by von Agris and Kraiss [31], the SGNUM database of the German Sign Language (DGS). It is the larger database recorded until now for recognition and tracking purposes with approximately 55 hours of videos, performed by 25 native signers, which correspond to almost 31k videos. It can be divided in two parts: one with only isolated signs and the other with sentences.

The isolated sign part has a vocabulary of 450 words selected with the requirement of be the most common in several books and videos used in DGS learning. The continuous part is composed of 780 sentences that vary in number of words from 2 to 11. In terms of video recording, it was performed in a very controlled environment with dark blue background and dark clothing with long sleeves.

3.4.4 Other Databases

There are already other SL databases available. Dreuw, et al. [30] present and compare some of them in terms of study application and annotated number of frames. Table 3.1 summarizes some important database characteristics.

Table 3.1 - Some characteristics of the most important databases.

Nome	SL	Recognition Type	Number of Signers	Annotated Frames	Vocabulary Size	Sentences	Hand Tracking
Purdue RVL-SLLL [25]	ASL	Isolated/ Continuous	14	-	-	10	no
Boston-50 [29, 30]	ASL	Isolated	3	1450	50	-	yes
Boston-104 [29, 30]	ASL	Continuous	3	15764	103	201	yes
Boston-400 [29]	ASL	Continuous	4	68555	406	843	no
SIGNUM [30, 31]	DGS	Isolated/ Continuous	25	-	450	780	yes
Phoenix [31, 32]	DGS	Continuous	7	293077	911	1980	yes
ATIS ISL [32, 33]	ISL	Continuous	24	~6000	400	680	yes

Fields with value '-' represent information that does not exist or is not clear.

Chapter 4

Data and its Treatment

4.1 The LGP Database

The LGP database was recently created at INESC and is the first database oriented to the SLR based on the Portuguese sign language. This database is composed of 182 isolated signs, including the alphabet and the ordinal numbers as well as pronouns, verbs or common expressions, some realized with one hand and others with both. These signs include not only the informative part but also the movement from rest position and the return to it, since it is hard to determine the exact frame where the gesture start and ends. It also has 40 sentences with a variable number of isolated gestures. The list of all isolated signs and sentences used can be consulted in Appendix A.

The gestures were performed by 13 native deaf, of male and female gender, in a free and natural expression environment, without clothing restrictions but with a uniform background. Some of the subjects realized their gestures standing up while others were seated in a chair. In Figure 4.1 are shown two examples of gestures performance conditions.

This dataset is one of the few with depth information associated to the RGB images obtained using the Microsoft Kinect camera. The images were captured with rate of 20 frames per second and with a resolution of 640x480 pixels. The depth capturing technology is still in an initial stage and so, there exist some noise associated to surfaces normal to the capturing plane and non-rigid materials as hair. Figure 4.2 is an example of a pair of color and depth images.

It does not have yet annotated frames for classification or marks for tracking but the existence of depth information may open new paths of investigation which make the use of this database very promising.

Figure 4.1 - Capturing conditions of the LGP database.

Figure 4.2 - Color and depth pair of images.

4.2 Data Treatment

One of the most important tasks of SLR systems is the segmentation of the images to isolate the hand or the face from the rest of the image. This is not however the focus of this study. The segmentation performed was done manually using the software *Interactive Segmentation Tool* [34] developed by Kevin McGuinness from the Dublin City University. As the name indicates, this software allows the segmentation of an object from its background interactively by simply indicating with lines the areas corresponding to the foreground and background. It possesses several segmentation algorithms as the *Seeded Region Growing Segmenter* [35], that had the best performances in the segmentation of the body from the background, and the *Interactive Graph Cuts Segmenter* [35] that was used to segment the hands. Figure 4.3 shows two examples of the utilization of this software.

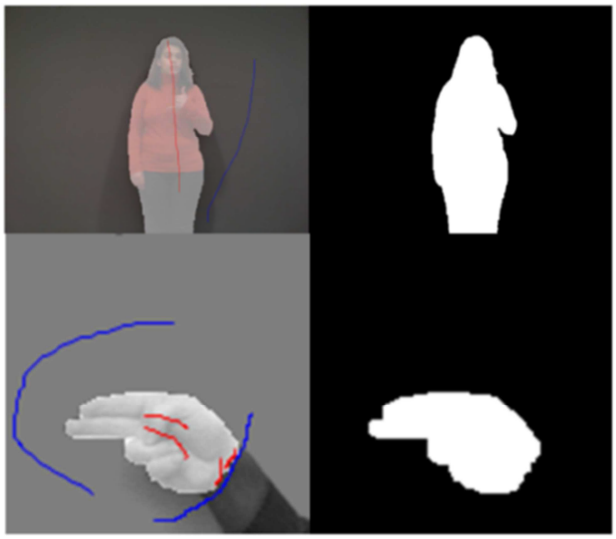

Figure 4.3 - Examples of the utilization of the software Interactive Segmentation Tool.

In some cases however, it is hard to segment the images even with this tool. When the cloth color is similar to the subject skin, when the hand is in front of the face or when it is moving and appears associated to blurring effects, it is difficult to determine the hand boundaries. To surpass this difficult over the color images, the edges of the correspondent depth images was superimposed. The existence of boundaries from other sources improved the software performance and increased the segmentation speed.

This approach raised a new problem. The depth and color images had different enlargements and the subjects were not in the same positions. Figure 4.4 is an example of type of discrepancy existent.

Figure 4.4 - Example of the discrepancy between the color and depth images.

The solution found was to transform the depth images using an affine transform. This operation remaps the points (x, y) to the location to the new location (x', y') as shown in Equation (4.1). It is a combination of linear and translation operations which keep the parallelism between lines and the ratios between them. This involves some loss of information but, since the difference between images is small, it is negligible.

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{4.1}$$

To compute the coefficients of this transformation it is necessary to select manually some matching points of both images. Since the used transformation has 6 parameters at least 3 points of both images are needed. *The Image Processing Toolbox* of *MatLab* has already an implementation of this transformation technique that is shown in Figure 4.5.

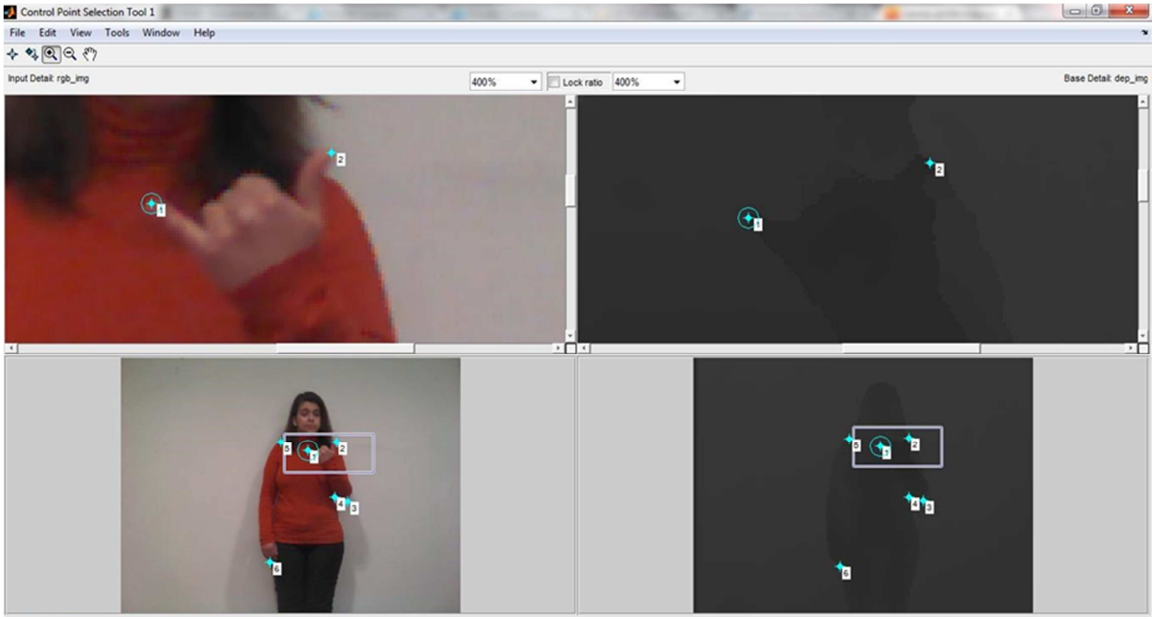

Figure 4.5 - MatLab tool to apply the affine transform.

The edge image was then obtained using a *Sobel* edge detector and overlapped in the color image. This configuration was chosen because the segmentation software has better performances over the color images than over the depth images.

Chapter 5

Static Gestures

The current chapter presents a study of several features that can be used to classify hand images according to the shape. Two distinct datasets were selected with this purpose: the Triesch dataset and the LGP dataset. The former is a benchmark dataset created by Triesch and von der Malsburg [36] to be adopted in hand shape classification systems while the second corresponds to a selection of images from the LGP Corpus.

Several features were computed from the hand images which are the following: Regional Features, Hu Moments and Gradient Orientation Histograms. These features were used to classify the images individually and combined with each other. In the case of the LGP dataset the depth information was also added in the classification process.

Two distinct classifiers were evaluated in order to understand their behavior with each type of features: the Naive Bayes and Support Vector Machines. Each classifier and the validation procedures are briefly described below.

This chapter is divided into the following way: section 5.1 describes both datasets in more detail and the manual segmentation procedure adopted. In section 5.2 are exposed the different features used and section 5.3 presents the classifiers used and the validation methodology adopted for each dataset. In the end, section 5.4 presents the results obtained and their discussion.

5.1 Datasets

The Triesch dataset was created to evaluate the performance of hand shape recognition systems. It is composed by a set of 10 different hand-shapes, which represent the letters *a, b, c, d, g, h, i, l, v* and *y* of the American Sign Language. These signs were performed by 24 different subjects against 3 distinct backgrounds: white, dark, and complex. The backgrounds variety was

introduced to test the performance of recognition systems in different conditions, with the complex background corresponding to the most challenging scenario. Figure 5.1 shows the sign *d* performed with different backgrounds.

Figure 5.1 - Different backgrounds used by Triesch.

This study did not involve the application of automatic segmentation and so it was realized manually as previously described. However, in order to enable the comparison between the results obtained and the ones presented by Kelly, et al. [23] the images with complex background were excluded.

To evaluate the robustness of the features, the images were segmented with two distinct levels of detail. The first, that from now on will be called segmentation level 1, corresponds to the most accurate contour possible of the hand with the wrist separating the hand from the arm, similarly to what Kelly, et al. [23] did in their study. The second level, segmentation level 2, corresponded to a coarser segmentation with less attention to the contour detail and inclusion of the arm. Figure 5.2 shows the same image with segmentation levels 1 and 2.

Figure 5.2 - Different levels of segmentation applied. Segmentation level 1 on the left and Segmentation level 2 on the right.

Ten gestures from five subjects were selected from the LGP dataset, corresponding to the cardinal numbers from zero to nine. These gestures were chosen by their relationship an absence of movement. In order to improve the classification conditions five images of each subject gesture were included, corresponding to five consecutive frames.

The naturalness of the gestures realization has a high cost in terms of classification performance. While in the Triesch dataset the gestures were realized with some precision by the signers, the LGP gestures were performed in a natural manner which means that to the same gesture may correspond different hand shapes. Figure 5.3 is an example of this variation to the gesture eight. From it can be seen not only different hand orientations and openings, but also different positions with respect to the body. This was the reason why the images of this dataset were only subjected to the first level of segmentation previously mentioned.

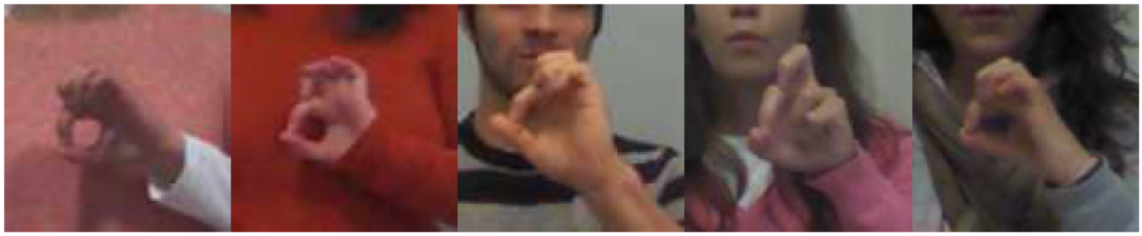

Figure 5.3 - Gesture eight of the LGP performed by five subjects.

5.2 Hand features

Several studies with the purpose of classify gestures based on the hand shape were developed. The features considered more relevant to the SLR study are described in the current section. They can be divided into the following way: Regional Features, Hu Moments and Gradient Orientation Histograms. The following subsections describe each group in detail.

5.2.1 Regional Features

Regional Features (RF) are the most widely used features in SLR systems. Starner, et al. [20] and Von Agris, et al. [15] are just two examples of studies that used this kind of features. They can be computed directly from the segmented hand contour and correspond direct measures that can be taken from it. In order to enable the comparison of hands with different sizes or distances to the camera, only relative features were selected. The Regional Features (RF) used are the following:

- **Solidity:** Area of the segmented hand divided by the correspondent convex area;
- **Extent:** Area of the segmented hand divided by the bounding box area;

- **Orientation 1 and 2:** Sine and Cosine, respectively, of the angle between the major axis and the horizontal direction of the bounding box. This angle varies between -90º and 90º;
- **Inertia ratio:** Ratio between the minor and major axis;
- **Eccentricity:** Corresponds to the inertia ratio of an ellipse with the same second order moments of binary object;
- **Perimeter proportion:** Hand perimeter divided by the area;
- **Bias:** Euclidean distance between the segmented object centroid and the bounding box center, divided by the object Area.

The Bias feature was created with the purpose of measure the deviation of the object center of mass relatively to the bounding box center. Some of the Hu Moments carry a similar type of information. It is important to denote at this point that this type of features does not provide any information about the interior of the segmented object.

5.2.2 Hu Moments

Image moments are statistical measures of an image which can be used as descriptors of the hand shape. Equation (5.1) represents the general expression of an image moment m_{pq}.

$$m_{pq} = \sum_{x=0}^{x=M-1} \sum_{y=0}^{y=N-1} (x)^p (y)^q f(x,y) \tag{5.1}$$

The moment m_{pq} is the *(p+q)-th* order moment of an *M x N* image.

Central moments μ_{pq} of an image are invariant to translation and can be computed as shown in Equation (5.2).

$$\mu_{pq} = \sum_{x=0}^{x=M-1} \sum_{y=0}^{y=N-1} (x-\bar{x})^p (y-\bar{y})^q f(x,y), \tag{5.2}$$

where $\bar{x}$ and $\bar{y}$ are the coordinates of the image centroid:

$$\bar{x} = \frac{m_{10}}{m_{00}} \qquad \bar{y} = \frac{m_{01}}{m_{00}} \tag{5.3}$$

The normalization of the central moments η_{pq} results on their transformation into scale invariant:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \tag{5.4}$$

$$\gamma = \left[\frac{(p+q)}{2}\right] + 1 \tag{5.5}$$

Hu Moments (HM) were introduced by Hu [37] and correspond to a set of combinations of normalized central moments which are independent of the object position, size and orientation. The seven Hu Moments are represented by the following expressions.

$M_1 = (\eta_{20} + \eta_{02})$,

$M_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$,

$M_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$,

$M_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$,

$M_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{30} + \eta_{12})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{21} + \eta_{03})^2 - (\eta_{21} + \eta_{03})^2]$,

$M_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$,

$M_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$.

5.2.3 Gradient Orientation Histograms

Gradient Orientation Histograms are widely applied nowadays in image analysis to match key points of different images and detect if both correspond to a representation of the same object. Freeman and Roth [13] proposed a hand recognition system based on the histogram of the image gradient orientations, computed over the entire hand bounding box. Despite of the weak results obtained, it was decided to include these features in this study because they are among of the few that provide information about the interior of the segmented hand and not only about the contour.

The image gradient ∇f is computed over grayscale images. It represents the magnitude and orientation of pixels intensity variation in each direction of the image and has the following expression (Equation (5.6)):

$$\nabla f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right]^t = \left[I_x, I_y\right]^t \tag{5.6}$$

From it can be computed the magnitude *G* and orientation θ of the image intensity variation as shown in the Equations (5.7).

$$G = \sqrt{{I_x}^2 + {I_y}^2} \qquad \theta = \tan^{-1}\frac{I_y}{I_x} \tag{5.7}$$

The Gradient Orientation Histograms (GOH) were computed from orientation matrix using 8 bins from 0 to $^{7\pi}/_{4}$ radians with a bin size of $^{\pi}/_{4}$ radians and normalized in order to sum 1. The gradient was computed using only the hand region – see Figure 5.4 for an example.

Figure 5.4 - (left) Masked grayscale image. (right) Gradient orientation image representation.

5.2.4 Inclusion of depth and features combination

The existence of depth information in the LGP dataset raised the possibility of including it as a feature. Depth cameras are a relatively recent technology and there are not yet many studies with this type of information. Kurakin, et al. [16] created an hand shape recognition system based on depth images. Nevertheless, they did not use the depth as a direct feature but as mean to segment the hand from the rest of the body.

In this work the selected Depth Features (DF) correspond to the percentile 0.1, 0.5 and 0.9. The 0.5 percentile is the depth median while the others aim the representation of the limit values of depth. The 0.1 margin was used to exclude some outliers in the form of noise.

The depth extraction was preceded of a normalization procedure. The value of the closest pixel to the camera was subtracted to the whole depth image. This procedure allowed the comparison of different images, independently of the distance to the camera.

The features were used alone and combined with each other. The adopted combinations were the following:

- RF + HM;
- RF + GOH;
- RF + HM + GOH;
- RF + DF;
- RF + HM + DF.

Other combinations of features were also tested but without relevant results.

5.3 Classification

To test the quality of the features, two distinct classifiers were encompassed in terms of behavior: Gaussian Naive Bayes and Support Vector Machines (SVMs). Naive Bayes is a probability based model as the Hidden Markov Models (HMM) which will be used in the next chapter with moving gestures. This common characteristic makes its evaluation interesting in the static gestures context. SVM is a linear classifier with very good performances and was used in several studies of hand shape recognition. Sections 5.3.1 and 5.3.2 present a brief description of these classifiers.

The validation was done in different ways to each dataset because of the available amount of data. The adopted classification systems are described in Section 5.3.3.

5.3.1 Gaussian Naïve Bayes

Gaussian Naive Bayes [38] is a probabilistic classifier based on the Bayes Theorem. It assumes that the features are independent from each other which is a strong assumption and the reason why it is called naive.

In the training stage, the prior probabilities of each class are estimated from the training set. Then the distribution that best fit the data according to each feature is determined to each class. In the case of Gaussian distribution, its expression is the one presented in Equation (5.8) and it requires the estimation the mean and variance from the training data.

$$p(x|C_j) = \frac{1}{(2\pi)^{D/2}|\sum_j|^{1/2}} exp\left\{-\frac{1}{2}(x - \mu_j)^T \sum_j^{-1}(x - \mu_j)\right\}, \quad (5.8)$$

D is the number of features, μ_j is the average vector of class *j* and Σ_j represents its covariance matrix.

It is then possible to compute the posterior probability by the Bayes Theorem of Equation (5.9).

$$p(C_j|x) = \frac{p(x|C_j)p(C_j)}{\sum_{i=1}^{D} p(x|C_i)p(C_i)}, \tag{5.9}$$

The classification of test set consists in determine the class which maximizes the posterior probability given the features values – see Equation (5.10).

$$Class = \underset{j}{arg\max}\, p(C_j|x), \tag{5.10}$$

Naïve Bayes is one of the simplest classifiers available and, despite of the assumption of features independence, it is fast in both training and classification stages, can handle real and discrete data and it is not sensitive to the effect of irrelevant features.

5.3.2 Support Vector Machines

The Support Vector Machine [38] classifier (SVM) was created based on the work of Vladimir Vapnik and in its initial formulation it is a linear classifier. To understand the concept behind SVM lets define the dataset *D*, represented by Equation (5.11), which has *n* points with dimension *p* and two classes.

$$D = \{(x_i, y_i);\ x_i \in \mathbb{R}^p, y_i \in \{-1,1\}\}, \forall\ i = 1, 2, \dots, n \tag{5.11}$$

SVM consists in find the hyperplane which best splits the data according to the correspondent class. This hyperplane can be defined by the normal vector *w* and the offset to the origin $\frac{b}{\|w\|}$. Equation (5.12) represents the general expression of such hyperplane.

$$w \cdot x + b = 0 \tag{5.12}$$

It is now possible to determine the class of a given point just by looking to its position in relation to this hyperplane. However, the number of hyperplanes that respect this condition is infinite and some of them may classify new points wrongly. The solution is to find the hyperplane with the higher margin or, in other words, the hyperplane which has the higher distance to the closest points of both classes, the support vectors. The margin is limited by the two hyperplanes, represented in Equation (5.13), which are parallel and cross the support vectors. Figure 5.5 is a schematic representation of this method.

$$\begin{cases} w \cdot x + b = 1 \\ w \cdot x + b = -1 \end{cases} \tag{5.13}$$

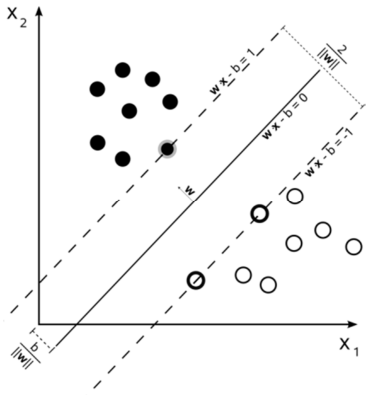

Figure 5.5 - SVM hyperplane and margin representation.

The vector *w* and parameter *b* can be determined by solving the quadratic problem of Equation (5.14).

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|^2 \\ &subjec\ to:\ \ y_i(w \cdot x + b) - 1 \geq 0, \qquad \forall i = 1, \dots, n \end{aligned} \tag{5.14}$$

Nevertheless, usually the data is not perfectly linearly separable. In many cases the classes are mixed and so, to find the hyperplane which separates them better is hard. To ease this difficulty it can be attributed some weights $\xi_i \in (0,1]$, allowing the existence of some points inside of the margin or misclassified. This procedure softens the margin rigidity and is the reason why it is called SVM with Soft margins. The problem can be given by Expression (5.15).

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n} \xi_i\right) \\ &subjec\ to:\ \ y_i(w \cdot x + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0, \qquad \forall i = 1, \dots, n \end{aligned} \tag{5.15}$$

The parameter *C* regulates the margin relaxation and the parameter optimization complexity.

As defined until now, SVM is able to classify with good results data that is linearly separable, even if there is some noise and is necessary to use soft margins. In many problems, despite this, the separation is non-linear. SVM works with linear separations and so, in order to be useful it is necessary to project the data to a higher dimensional space where the data is linearly separable. This operation is called the kernel trick. There are several types of kernels like the Polynomial, RBF or Sigmoidal and its choice should be based in some prior knowledge about the data. Nonetheless, when the data distribution or behavior is unknown, to test different Kernels and select the one that returns the best results is a normal procedure.

5.3.3 Validation

The validation is very important to evaluate the classification performance. Since the Triesch dataset is larger than the LGP, different validation procedures were used.

Let us first look to the Triesch dataset validation. Following the work of Kelly, et al. [23] our procedure of validation adopted was the *leave-k-subjects-out* technique. As the name indicates it consists in separate the data of *k* subjects to test and train with the remaining. Given that the

dataset has 24 subjects, the following proportions of training/test were used: 21/3, 12/12, 8/16 and 3/21. The two last are the same that were presented in [23].

Different approaches were used with the Naive Bayes and SVM models. Since Naïve Bayes does not require parameter optimization, the model was created directly over the training set and used to classify the test set. SVM required the optimization of parameter *C* and the choice of the most adequate kernel. A *10-fold cross validation* procedure was adopted to reduce the training set overfitting. This procedure consists in mix the training set at random and split it in 10 equal parts or bins. Then, 9 bins were selected to build the model and the remaining is used to test. This step was repeated until that all bins have been used to test. The accuracy of the training stage corresponded to the average of the 10 accuracies obtained. The SVM parameters optimization was done maximizing the average accuracy described. A schematic representation of this method is shown in Figure 5.6.

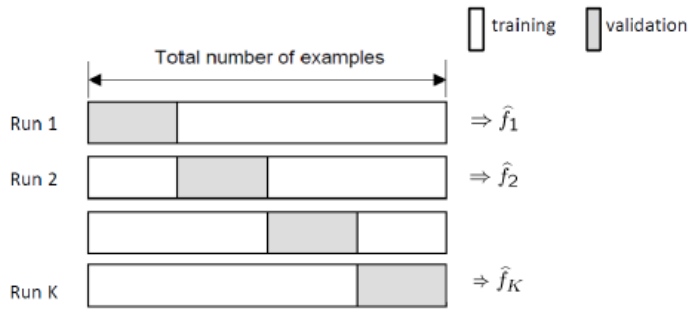

Figure 5.6 - *k*-fold cross validation.

The LGP validation was slightly different since the number of subjects was substantially smaller. Two procedures were adopted with this dataset. The first consisted in disregard the test and use all the data to train with *10-fold cross validation*, appreciating the model over the training error. The second procedure is widely used in SLR and is known as *leave-one-subject-out*. It follows the same principles of the *leave-k-subjects-out* procedure but all subjects are iteratively used to test the model.

5.4 Results and Discussion

In this section the classification results are presented. The results were obtained using the implementations of the *RapidMiner* [39] which is a Data Mining oriented software. This software has an intuitive interface and several data treatment and classification options. In the operations where some randomness was needed, it was used always the same random seed to avoid comparison errors of the results.

This section is divided in two. In Section 5.4.1 the results obtained to the Triesch dataset are presented and compared to the results of [23]. Section 5.4.2 is dedicated to the appreciation of the LGP dataset. In the end, Section 5.4.3 presents some conclusions about the data.

5.4.1 Triesch Dataset Results

Table 5.1 shows the classification accuracies of the Naïve Bayes classifier to the Triesch dataset with segmentation 1.

Table 5.1 - Naïve Bayes accuracies to Triesch dataset with segmentation level 1.

Features	21/3	12/12	8/16	3/21
RF	98,33%	91,63%	92,48%	88,31%
HM	80,00%	82,43%	82,13%	79,95%
GOH	83,33%	80,75%	81,50%	79,00%
RF + HM	96,67%	92,47%	93,42%	88,54%
RF + GOH	98,33%	93,72%	94,36%	90,69%
RF + HM + GOH	100,00%	93,72%	94,04%	90,69%

Looking to the results of the isolated features it is easy to see that the RF returned the best results. As expected, the scenario with more subjects in the training set has the best model performances. It is still remarkable that the worst scenario has accuracies of 88% to the RF, given the amount of data used to train. The features combination also returned interesting results. All of the selected combinations returned accuracies similar to the RF alone.

Table 5.2 below has the SVM results to the same data. Different Kernels were tested but was the Linear Kernel that returned the best performances.

Table 5.2 - SVM accuracies to Triesch dataset with segmentation level 1.

Features	21/3	12/12	8/16	3/21
RF	100,00%	92,47%	94,04%	88,78%
HM	96,67%	89,54%	88,71%	83,77%
GOH	86,67%	83,68%	81,19%	77,33%
RF + HM	93,72%	93,72%	94,67%	89,02%
RF + GOH	100,00%	94,56%	94,98%	88,78%
RF + HM + GOH	100,00%	94,56%	94,67%	89,02%

As expected, the SVM results are slightly better than the ones obtained with Naive Bayes.

To understand the source of misclassifications, the SVM classification confusion matrix was computed with Train/Test proportion of 8/16 to the combination of all features. This

combination was chosen because it is the most complete in terms of feature amount and has a fair proportion of train and test data. Table 5.3 presents the confusion matrix obtained.

Table 5.3 - Confusion Matrix of the Triesch dataset.

True \ Predicted	A	B	C	D	G	H	I	L	V	Y
A	1									
B		0,97								
C			0,91					0,06		0,03
D			0,06	0,97						
G					0,88	0,16				
H					0,13	0,84				
I				0,03			1			
L								0,94		
V		0,03	0,03						1	
Y										0,97

From the analysis of this confusion matrix it can be seen that the gestures *G* and *H* are frequently confused with each other. However, it is not an unexpected result since these two gestures are quite similar as shown in Figure 5.7 from the work of Kelly, et al. [23].

Figure 5.7 - Triesch dataset gestures.

The use of a rough segmentation returned worse results as expected. Table 3.4 and Table 3.5 show these results to the Naive Bayes and SVM classifiers respectively.

Table 5.4 - Naïve Bayes accuracies to Triesch dataset with segmentation level 2.

Features	21/3	12/12	8/16	3/21
RF	90,42%	81,59%	82,76%	73,75%
HM	56,67%	54,81%	54,55%	48,93%
GOH	81,67%	73,64%	72,41%	69,69%
RF + HM	91,67%	83,68%	84,33%	74,22%
RF + GOH	95,00%	83,68%	84,33%	78,76%
RF + HM + GOH	95,00%	84,52%	84,64%	75,18%

Table 5.5 - SVM accuracies to Triesch dataset with segmentation level 2.

Features	21/3	12/12	8/16	3/21
RF	93,33%	87,28%	85,58%	78,04%
HM	90,00%	86,61%	79,31%	68,02%
GOH	80,00%	71,97%	73,98%	68,47%
RF + HM	95,00%	88,70%	86,52%	77,33%
RF + GOH	95,00%	90,38%	88,40%	79,24%
RF + HM + GOH	95,00%	93,31%	90,91%	78,76%

The results maintain the tendencies observer with segmentation level 1 but, as expected the accuracies obtained are slightly lower.

To get an idea about the quality of the results, the comparison of the best results to the ones presented in [23] is shown in Table 5.6. The segmentation conditions used in this work are closer to what was defined as segmentation level 1. Table 5.6 also shows the results of other works over the same data.

Table 5.6 - Comparison of the results from different studies.

Study	Proportion	
	8/16	3/21
Triesch, et al. [36]	-	95,20%
Just, et al. [40]	89,90%	-
Kelly, et al. [23]	91,80%	85,10%
Best obtained	94,98%	90,69%

The analysis of this table allows to conclude that the adopted features returned good results when compared with studies with the same objective. The results presented by Triesch and von der Malsburg [36] are substantially better than the rest. However, the elastic graph matching

technic used by these authors has an higher computational complexity and takes much more time, spending some seconds in the analysis of each image, as stated by Kelly, et al. [23].

5.4.2 LGP Dataset Results

The LGP dataset was tested with segmentation level 1 but, since it has depth information, it was possible to add features and form new combinations. Table 5.7 has the training accuracies obtained by the classifiers Naive Bayes and SVM with the whole data used for training.

Table 5.7 - Classification accuracies of the LGP dataset.

Features	Naive Bayes	SVM
RF	88,40%	91,20%
HM	52,80%	76,00%
GOH	20,80%	10,80%
DF	18,40%	35,60%
RF + HM	89,20%	91,60%
RF + GOH	87,60%	91,60%
RF + DF	86,40%	96,00%
RF + HM + GOH	90,00%	91,60%
RF + HM + DF	87,60%	96,40%

The classifiers present a behavior similar to the Triesch dataset. The results show that the RF are still the ones that provide the best performances and in the case of GOH and DF the low accuracies obtained suggest that these features should not be considered alone. The features combination did not improve the results significantly with the Naïve Bayes classifier. The SVM classification however, returned good accuracies with RF + DF features, suggesting that this combination provides the best description of these gestures. To determine the source of misclassifications, the confusion matrix of the features RF + DF with SVM classification was computed – results can be seen in Table 5.8.

Table 5.8 - Confusion Matrix of the LGP dataset.

True \ Predicted	0	1	2	3	4	5	6	7	8	9
0	0,96								0,04	
1		0,88						0,04		
2			0,96							
3				1						
4					1					
5			0,04			0,92				
6							1			
7		0,12						0,92		
8	0,04					0,08			0,96	
9								0,04		1

The most expressive confusion shown in this table is between the gesture 1 and 7. As it happened in the Triesch dataset study, this misclassification is due to the similarity between different gestures. As shown in Figure 5.8, in some cases it is difficult to classify even for humans unfamiliarized with these signs.

Figure 5.8 - Gestures 1 (left) and 7 (right) of the LGP dataset performed by the two subjects.

The second part of this dataset analysis was the evaluation of the models performances when trained with four subjects and tested with the remaining. Using the previous analysis insights, it was used the RF alone and combined with the DF in this evaluation. The results for both datasets are shown in the Table 5.9 and Table 5.10 respectively.

Table 5.9 - Accuracies of the leave one subject out classification with the RF.

Test Subject	Naive Bayes	SVM
1	56,00%	66,00%
2	90,00%	84,00%
3	80,00%	54,00%
4	66,00%	70,00%
5	62,00%	52,00%
Average	**70,80%**	**65,20%**

Table 5.10 - Accuracies of the leave one subject out classification with the RF + DF.

Test Subject	Naive Bayes	SVM
1	42,00%	56,00%
2	88,00%	74,00%
3	78,00%	56,00%
4	74,00%	64,00%
5	60,00%	58,00%
Average	**68,40%**	**61,60%**

Both classifiers got worse results in this procedure. The results show high discrepancies between subjects. These differences are, for sure, not only related to the naturalness with which the LGP gestures were performed in comparison to the Triesch dataset, but also with some segmentation difficulties. In opposition to the Triesch dataset, the wrist of these images was not always visible: in some cases due to the hand orientation and in the others because the sleeves hide almost the half of the hand. The addition of depth features in this case did not provide improvements. Instead, the classifier's performance fell down. The reason for this performance reduction might be related to the training overfitting. The existence of more features improved the training results but the model became overfitted to the training set.

5.4.3 Results Conclusions

From the Triesch results it can be concluded that the adopted features can describe the hand shape with good performances when the gesture execution is rigid. Gestures performed with naturalness are more difficult to classify by this type of features. This shows that the hand shape classification should incorporate some structural features of the hands that allow inferring the correct hand shape even when the hand is rotated. The use of depth information may allow to compute this structural features but it is necessary the evolution of the capturing technology to increase the detail and reduce the observed noise.

Chapter 6

Dynamic Gestures

Gestures with movement component are very different from the static gestures analyzed until now. The information is composed by movement, which includes the direction and speed, position with respect to the body and hand shape which may vary during the gesture execution. The features to be used with these gestures should be able to carry this kind of information.

Another difficulty, that might be hard to handle, is the temporal evolution of the gestures. The fact that the features values change during the gesture execution motivates the use of a different type of classifiers.

This chapter is divided in the following way: Section 6.1 presents the dataset, Section 6.2 describes the two types of movement features selected, Section 6.2.2 has a description of the classifiers applied and Section 6.4 presents the obtained results and discussion.

6.1 Gestures selected

The choice of dynamic gestures to this study followed principles similar to the ones that were implemented with static gestures. The personal pronouns: I, YOU (singular), HE, WE, YOU (plural), THEY, and the possessive pronouns: MY, YOUR, OUR, YOURS were selected. These gestures can appear in the same contexts and, for that reason, they should be distinguishable.

These gestures are hard to describe with a single image. Below, the description of these gestures done by Mesquita and Silva [6] is presented.

- **I (first person of the singular)**

  Hand closed with the index finger stretched and directed to the chest. Palm directed to the chest and fist moving in the signer direction.

- **YOU (second person of the singular)**

Hand closed with the index finger stretched and directed to the receptor. Palm directed to the receptor and fist moving in the receptor direction.

- **HE (third person of the singular)**

Hand closed with the index finger stretched directed to the side of the dominant hand. Palm directed to the receptor and fist moving away from the signer in the direction of the receptor dominant hand. The arm stays bent.

- **WE (first person of the plural)**

Hand closed with the index finger stretched and directed to the signer and palm directed to the ground. The gesture start with the index finger touching the chest in the opposite side of the dominant hand and then the fist moves away, in the direction of the receptor and of the dominant hand side, with a rotary movement until get back to the initial position.

- **YOU (second person of the plural)**

Hand closed with the index finger stretched and directed to the signer and palm directed to the ground. The gesture starts with the index finger touching the chest in the side of the dominant hand and then the fist moves away in the direction of the receptor and for the opposite side of the dominant hand with a rotary movement. During the execution the finger changes direction and ends pointing to the receptor.

- **THEY (third person of the plural)**

Hand opened with the palm directed to the ground. The gesture is performed on the side of the dominant hand shoulder with arm bent and corresponds to the circular movement of the hand in a small range.

- **MY**

Hand closed with the palm directed to the opposite side of the dominant hand. The gesture starts with the fist in the front and away of the signer and then is moved until touch the chest. This gesture is exactly equal to the gesture WE but during the execution the hand closes completely. The gesture should be accompanied by the facial expression of pronounce the sound *fff*.

- **YOUR**

Hand closed with the palm directed to the receptor. The gesture starts with the fist closer to the chest and then is moved in the receptor direction. This gesture is exactly equal to the gesture YOU but during the execution the hand closes completely and the palm ends

directed to the receptor. The gesture should be accompanied by the facial expression of pronounce the sound *fff*.

- **OUR**

This gesture is exactly equal to the gesture WE but during the execution the hand closes completely. The gesture should be accompanied by the facial expression resultant of pronounce the sound *fff* in the end of the gesture.

- **YOURS**

This gesture is exactly equal to the gesture YOU (Plural) but during the execution the hand closes completely and the palm ends directed to the receptor. The gesture should be accompanied by the facial expression resultant of pronounce the sound *fff* in the end of the gesture.

The gestures recorded and used in this study show some variations to what is described. Some subjects prolonged their gestures more than the others and some repeated the dynamic part. This happened due to the naturalness of the sign execution.

It should also be noticed the similitude between some of the gestures which is the source of some of the misclassifications observed. As a preventive reminder it was decided to group here the gestures used according to their apparent resemblance.

- Group 1 - I + MY
- Group 2 - YOU (singular) + YOUR;
- Group 3 - HE + THEY
- Group 4 - WE + OUR
- Group 5 - YOU (plural) + YOURS.

6.2 Features Extracted

In this section the two sets of elected dynamic features are described. The first set corresponds to the Cartesian coordinates of the hand and it was evaluated both in absolute value and in frame variation. The second set is a variation of the features used by Holden, et al. [12]. All feature sets were combined with the best hand features set used in the previous chapter that was considered to be RF + DF. Despite of used in the previous section to represent static gestures, these hand features were computed in each frame of the dynamic gestures in order to express the hand shape variation.

6.2.1 Cartesian Coordinates

Due to the differences in the subjects posture and location, these features were computed using the face center as the origin of the coordinate system. Figure 6.1 shows the 2D representation of the features.

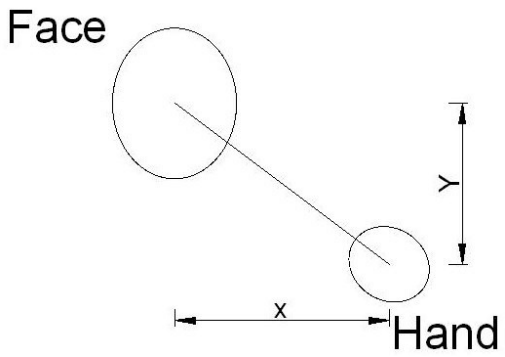

Figure 6.1 - Cartesian coordinates features.

The distance in depth *z* between the hand and the face was also included. To evaluate if the variation of the hand position is more important than the absolute position, the relative positions were considered as features. The feature vector was thus: $x, y, z, \Delta x, \Delta y, \Delta z$.

The study of DTW was composed by some variations of this set:

- S1 - $x \; and \; y$;
- S2 - $x, y \; and \; z$;
- S3 - $\Delta x, \Delta y \; and \; \Delta z$;
- S4 – *All features including RF and DF.*

In the case of HMM different combinations of these features, including the hand shape features, were assessed:

- Mov1 (x, y, z);
- Mov2 $(\Delta x, \Delta y, \Delta z)$;
- Mov1 + Mov2;
- RF + DF + Mov1;
- RF + DF + Mov2;
- RF + DF + Mov1 + Mov2;

6.2.2 Relative Features

The Relative features aim to be independent of the image scale since they are based only on the angles between hand positions. Figure 6.2 shows a schematic representation of this set of features.

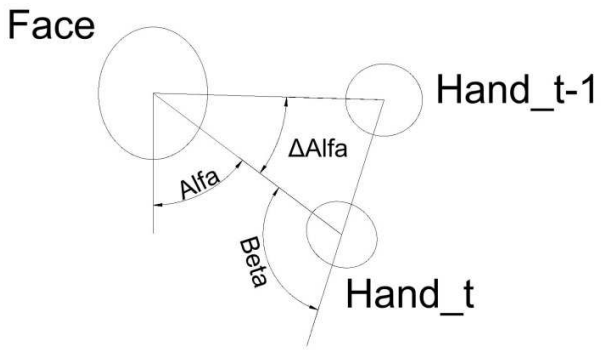

Figure 6.2 - Angle based features representation.

The angle *Alfa* was computed using the x and y distances of the previous set. The angle *Beta* together with the variation of *Alfa* indicates the direction of the movement of the hand. As depth features were used the same features of the previous sets. The final feature vector was: $Alfa, Beta, \Delta Alfa, z, \Delta z$.

The combinations of features used were the following:

- Mov $(Alfa, Beta, \Delta Alfa, z, \Delta z)$;
- RF + DF + Mov;

6.3 Classifiers

6.3.1 Dynamic Time Warping

Despite DTM cannot be considered a classifier, it can be used as a measure of shape similarity between two series of observations. Its conception allows comparing series with different sizes or with lag between each other. The resultant measure cannot be called a metric by the reasons that are further discussed but it can be used to determine among a set of pairs of observations, which ones are more similar and more discrepant. These properties motivate its application in the field of speech recognition [41] and more recently in the SLR by Athitsos, et al. [27].

The first step of this method is the computations of the cost matrix A. Given the data series ***x*** with size *p* and ***y*** with size *q*, both with dimension *N*, this matrix stores the Euclidean distance between all pair of points as shown in Equation (6.1). The use of Euclidean distance is not mandatory and other distance measures can be used depending of the problem in hands.

$$A = \begin{bmatrix} A_{x_1,y_1} & \cdots & A_{x_1,y_q} \\ \vdots & \ddots & \vdots \\ A_{x_p,y_1} & \cdots & A_{x_p,y_q} \end{bmatrix}$$
$$A_{x_i,y_j} = d(x_i, y_j) = \sqrt{\sum_{k=1}^{N} (x_{i,k} - y_{j,k})^2} \tag{6.1}$$

Defined the cost matrix is now easy to define what DTW is. This method consists in finding the path with lower cost from the beginning of the series until the end. It can be achieved by computing the matrix ***B*** defined by the Equation (6.2).

$$B_{11} = A_{x_1,y_1}$$
$$B_{ij} = A_{x_i,y_i} + min(B_{i-1,j-1}, B_{i,j-1}, B_{i-1,j}), if\ i\ \wedge j\ \neq 1 \tag{6.2}$$

The value corresponding to the position (p, q) of matrix ***B*** is the total accumulated cost and is over this value that is computed the distance between the series ***x*** and ***y*** as represented in Equation (6.3).

$$DTW(x, y) = \sqrt{B_{pq}} \tag{6.3}$$

The reason why DTW cannot be called a metric was pointed by Müller [42] which say that DTW does not respect the fourth metric condition also known as triangular inequality. This means that $DTW(a, b) + DTW(b, c)$ is not necessarily higher or equal than $DTW(a, c)$.

This method is scale dependent and so does not make sense to compare series expressed in different units. The same happens with the dimension units. In the particular case of this study, the use of original units affects strongly the final results since the ***x*** and ***y*** dimensions (parallel to capturing plane) have the pixel as unit, while the depth *z* (normal to the execution plane) has the unit of depth images that is completely different from the pixel. This forced the normalization of the data that is further described.

6.3.2 Hidden Markov Models

Classifiers as the standard SVM or Neural Networks are widely applied and with good performances in the classification of isolated data where each value does not depend on the previous values. However this assumption is not always true. In many problems as weather forecasting, the prediction does not depend only of the current observations but also on what was observed previously. Hidden Markov Models have gained in the recent years an important role in the solution of this type of problems. Below is presented the general description of this classifier and the customizations applied to this study.

General Description

HMM assume that the data being modeled follows a Markov process but the states are unknown and the observations are associated to these states. The probability of a sequence is always associated to the sequence of hidden states. Figure 6.3 shows a schematic representation of an HMM of a sequence of observations *x* with size *T*.

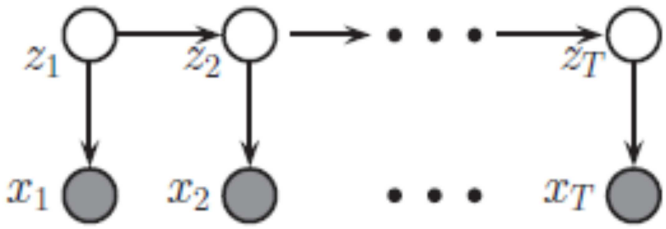

Figure 6.3 - HMM representation.

In this representation *z* is the discrete sequence of hidden states. The states are connected in a left-to-right HMM but others topologies can be used. Given this representation, the probability of the sequence of observations *x* with sequence of states *z* is show in Equation (6.4).

$$p(x_1, \dots, x_T, z_1, \dots, z_T) = p(z_1) \prod_{t=2}^{T} p(z_t|z_{t-1}) \prod_{t=1}^{T} p(x_t|z_t) \tag{6.4}$$

An HMM is defined by three parameter sets: the initial state probability vector π, the state transition matrix *A* and the emission probabilities η of the observation model.

The state transition matrix *A* follows the same principles of the Markov models which means that $A_{ij} = p(x_t = j|z_t = i)$ and $\sum_j A_{ij} = 1$. Figure 6.4 shows an example of a three state left-to-right HMM with the transition matrix of Equation (6.5).

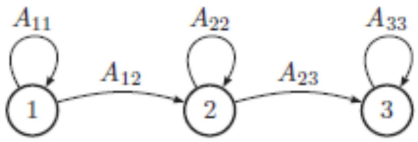

Figure 6.4 - Three state left-to-right HMM.

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 \\ 0 & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix} \tag{6.5}$$

The observations may be discrete or continuous. This affects the selected observation model. In the discrete case this model corresponds to the joint distribution matrix *B* of observations and states shown in Equation (6.6). With continuous observations the procedure is to find the distribution that best fits the data given the state. It is usual the use of Gaussian distributions as presented in Equation (6.7).

$$p(x_T = l | z_T = k) = B(\mathrm{k}, \mathrm{l}) \tag{6.6}$$

$$p(x_T | z_T = k) = \mathcal{N}(x_t | \mu_k, \Sigma_k) \tag{6.7}$$

The training of HMM can be supervised or unsupervised in terms of state sequence. In the supervised case the sequence of states is known and the problem corresponds to estimate the parameters directly over the sequences of states and observations. When the sequence of states is unknown, the unsupervised scenario, the Baum-Welch algorithm which is based on the EM algorithm (Expectation-Maximization) is usually applied. This method starts with the best guess of the initial parameters θ and state sequence *z* and iteratively updates the parameters and determines the sequence of states that maximizes $p(x|\theta)$. This iterative process converges to a local optimum which may be affected by the initial guess.

Each class of the training data corresponds to a different HMM model. The classification of the test set is realized determining which model maximizes $p(x|\theta)$. Since the test state sequence is unknown the computation of $p(x|\theta)$ is not straight-forward. This probability is often computed using the Forward algorithm. In problems where it is important to determine the sequence of states, it can be used the Forward-Backward algorithm or the Viterbi algorithm which the return the sequence that maximizes $p(x|\theta)$ can be preferable. For a complete and accurate description of this algorithm is recommended the consultation of [38, 43].

Model Customization

Beyond the parameter estimation, HMM has several other characteristics which can be adjusted. The transition matrix can be conditioned to impose the certain chain connection structure to the states. In this work three types of transition matrices were testes and are shown in Equation (6.8) for a four state problem.

$$\pi 1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad A1 = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ 0 & a_{22} & a_{23} & 0 \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\pi 2 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad A2 = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{6.8}$$

$$\pi 3 = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{bmatrix}, A3 = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

The first type corresponds to a four state left-to-right HMM where transitions to the same state or to the next are allowed. The initial probabilities of each state were conditioned in order to start always in the first state. Otherwise the model could end with a smaller number of states.

Type *A2* is widely applied in speech recognition systems and is called Bakis model. It is similar to the previous but allows the transition to two states ahead. Figure 6.5 show the schematic representation of this model.

The last type is the most complete allowing the transitions between all states.

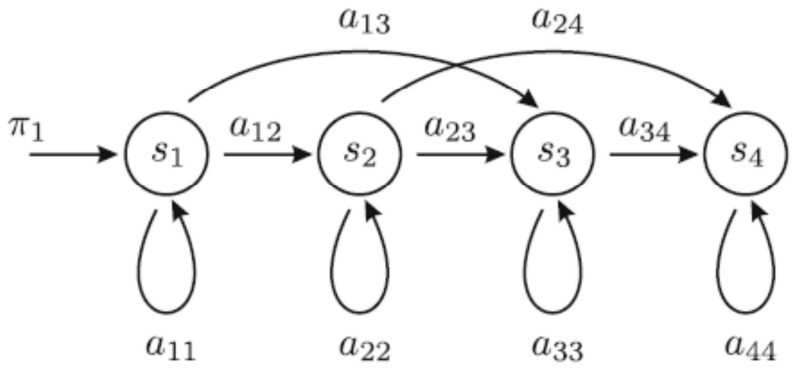

Figure 6.5 - Bakis model [15].

The observation model represents how the observed data is distributed according to each state. In this work the Gaussian Mixture of Models (GMM) was chosen to represent the data. The basic idea is that the data might be best fitted by several distributions instead of only one, assuming that inside the same class can coexist more than one population. The model was trained with a GMM with two components.

6.3.3 Validation

The DTW distances between all pairs of gestures sequences was firstly computed. The sequences of each subject were then isolated and the closest *k* signs to each were determined. Two different values of *k* were tested: 1 and 5. *K* equal to 1 corresponds to the closest gesture while to *k* equal to 5, the class mode of the five closest gestures was computed. This classification approach corresponds to the *leave-one-subject-out* method mentioned previously.

The HMM accuracy was evaluated in a leave-one-subject-out fashion which followed the same principles presented in Section 5.3.3. The study was performed not only over 10 classes but also over groups based on the gestures apparent similarities, presented at the end of Section 6.1. Given that the HMM training starts with random values of the transition matrix and random initial state probabilities, some variability in the results of different runs was observed. To decrease the uncertainty about the results, each model was run 5 times and was computed average and variance obtained.

6.4 Results and Discussion

6.4.1 DTW Results

As mentioned in Section 6.3.1, DTW is scale dependent. Given that the data had features in different scales, the normalization was necessary. The normalization was performed over all gestures of each subject, with each feature normalized between 0 and 1.

The class of gestures with smaller DTW of the training set to each subject classification is shown in Table 6.1. For simplification purposes the gestures were numbered from 1 to 10 by the order in which they are presented in Section 6.1.

Table 6.1 - Class of the gesture with smaller DTW.

Gesture	Subject																			
	Subset 1					Subset 2					Subset 3					Subset 4				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	1	1	7	7	7	8	1	7	7	7	2	1	1	1	1	1	1	1	1	1
2	1	7	8	2	8	10	7	7	7	8	2	1	2	3	2	2	1	2	2	2
3	3	3	3	6	3	3	3	3	3	6	2	3	1	3	1	3	3	3	3	3
4	1	4	4	8	5	2	2	5	7	8	5	9	4	3	4	5	4	4	2	4
5	4	4	10	9	5	4	8	7	7	1	4	9	10	7	4	1	2	5	1	5
6	6	6	6	6	3	6	6	6	6	6	1	3	6	3	6	6	6	6	3	6
7	1	1	7	1	7	8	1	7	1	7	2	1	7	1	7	1	1	7	7	7
8	2	1	7	2	2	8	7	7	2	8	7	1	2	3	1	1	8	8	8	8
9	1	5	7	10	7	10	5	1	7	7	5	9	1	7	1	1	4	7	7	1
10	3	1	7	10	5	4	2	7	10	8	4	9	9	10	1	8	8	1	8	1
Average Accuracy	0,34					0,3					0,34					0,6				

Results revealed low accuracies to the movement features alone and that the inclusion of depth was a source of misclassification instead of improving the results. The use of the position variation did not returned significant differences. *S4* returned the best results which emphasize the importance of the hand shape features even to this dataset. In Table 6.3 is shown the confusion matrix of *S1* features classification.

Table 6.2 - Subset S1 confusion matrix.

		Predicted									
		1	2	3	4	5	6	7	8	9	10
True	1	0,4						0,6			
	2	0,2	0,2					0,2	0,4		
	3			0,8			0,2				
	4	0,2			0,4	0,2			0,2		
	5				0,4	0,2				0,2	0,2
	6			0,2			0,8				
	7	0,6						0,4			
	8	0,2	0,6					0,2	0,0		
	9	0,2				0,2		0,4		0,0	0,2
	10	0,2		0,2		0,2		0,2			0,2

This confusion matrix shows that the most significant source of misclassifications is closely related to the gestures similarities pointed in Section 6.1. See for example the percentage of

times that the gesture *I* (1) is confused with *MY* (7) and vice-versa. In fact these gestures are only distinguishable by the hand shape.

An additional approach was realized to determine not only the gesture with smaller DTW but the mode of the five with smaller DTW. In the multimodal cases was selected the smaller one. Table 6.3 presents the results obtained.

Table 6.3 – Class mode of the five gestures with smaller DTW.

Gesture	Subject																			
	Subset 1					Subset 2					Subset 3					Subset 4				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1	1	1	7	7	8	8	1	7	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	8	2	8	2	1	7	7	1	2	1	2	1	7	2	1	2	2	2
3	3	6	3	3	6	3	6	3	3	3	1	3	1	3	1	3	3	2	3	3
4	1	5	5	8	8	2	2	5	2	8	1	9	4	1	4	5	4	4	1	4
5	4	4	10	9	2	4	8	7	1	1	4	9	10	7	4	1	2	5	1	5
6	6	6	6	6	6	6	6	6	6	6	1	3	3	6	3	6	3	6	3	6
7	1	1	7	1	7	2	7	7	1	7	1	1	7	1	1	1	1	7	7	7
8	2	1	7	2	2	8	7	7	1	1	1	1	2	1	7	1	1	8	1	8
9	8	5	7	10	8	8	7	7	8	7	7	4	1	7	1	1	1	7	7	1
10	3	1	7	10	5	3	8	7	10	1	8	9	9	10	1	8	8	1	7	8
Average Accuracy	0,28					0,36					0,28					0,52				

The results are worse for *S1, S3* and *S4* but better to *S2*. In the perfect scenario the four gestures with smaller DTW should correspond to the same class of the test gesture. Given that the mode of the five first gestures returned worse results than using the closest gesture alone. It can be concluded that the elected gestures are very different from each other, even belonging to the same class. This might be related to the fact of some subjects perform their gestures stand up while the others were seated which affects significantly the position of the hand with respect to the body.

6.4.2 HMM Results

HMM is dependent of some variables as the number of states and the transition matrix type. In order to observe the influence of these variables, the model was trained and tested over the whole data. Table 6.4 shows the accuracy results for all Cartesian features, the three transition matrices types described in Equation (6.8) and three different amounts of states: 3, 4 and 5.

Several conclusions can be taken from these results. Looking to the relation between the number of states and the complexity of the transition matrix it can be said that they are inversely related. In almost all of the classifications using *A1*, the simplest transition matrix, a higher number of states result in lower accuracies while *A3* revealed the opposite tendency. Due to this observation, 4 states were implemented in further classifications since it is not overfitted to any of the used transition matrices.

Besides this results show that the models with most complete transition matrix returned the better results. This is not unexpected since this transition matrix allows jumps between all states, leading to a better fit. Nevertheless, as it is shown in the leave-one-subject-out classification below, this corresponds to an overfitting to the training data and is not always true if the test data was not part of the training set.

The feature comparison did not show, to this classification, any tendency or disparity which deserves consideration.

In Table 6.5 the leave-one-subject-out classification results are shown with hand shape features. The obtained accuracies decreased significantly. This is an evidence of the implicit differences which may exist between gestures performed by different subjects. However, this set of features is not preponderant in the study of dynamic gestures. Table 6.6 presents the average accuracies obtained with the rest of the Cartesian features.

Table 6.4 - Training set classification accuracies with Cartesian features.

Features	# of States	A1		A2		A3	
		Average	Variance	Average	Variance	Average	Variance
RF + DF	3	99,2%	0,000	95,6%	0,000	98,0%	0,000
	4	93,6%	0,003	96,4%	0,002	98,8%	0,000
	5	95,2%	0,002	97,2%	0,002	99,2%	0,000
Mov1	3	94,0%	0,001	90,4%	0,002	97,6%	0,000
	4	93,2%	0,003	92,4%	0,008	100,0%	0,000
	5	88,0%	0,011	90,0%	0,001	100,0%	0,000
Mov2	3	74,4%	0,003	73,2%	0,004	88,0%	0,002
	4	79,2%	0,006	83,2%	0,007	94,8%	0,002
	5	74,8%	0,007	79,6%	0,004	99,2%	0,000
Mov1 + Mov2	3	95,6%	0,002	96,8%	0,001	99,6%	0,000
	4	99,2%	0,000	94,8%	0,004	100,0%	0,000
	5	92,0%	0,006	95,2%	0,003	100,0%	0,000
RF + DF + Mov1	3	94,8%	0,006	98,0%	0,002	100,0%	0,000
	4	94,8%	0,003	96,0%	0,003	100,0%	0,000
	5	88,8%	0,005	95,2%	0,002	100,0%	0,000
RF + DF + Mov2	3	99,6%	0,000	99,2%	0,000	99,6%	0,000
	4	99,2%	0,000	100,0%	0,000	100,0%	0,000
	5	97,6%	0,003	100,0%	0,000	100,0%	0,000
RF + DF + Mov1 + Mov2	3	99,6%	0,000	100,0%	0,000	100,0%	0,000
	4	98,4%	0,001	99,2%	0,000	100,0%	0,000
	5	95,2%	0,005	98,0%	0,001	100,0%	0,000

Table 6.5 - Leave-one-subject-out results with Hand shape features.

Subject	A1		A2		A3	
	Average	Variance	Average	Variance	Average	Variance
1	66,0%	0,028	68,0%	0,007	56,0%	0,018
2	40,0%	0,010	46,0%	0,003	46,0%	0,003
3	68,0%	0,007	70,0%	0,010	72,0%	0,002
4	58,0%	0,007	58,0%	0,007	58,0%	0,007
5	72,0%	0,012	70,0%	0,005	60,0%	0,005
Average	60,8%	0,013	62,4%	0,006	58,4%	0,007

Table 6.6 - Leave-one-subject-out classification results with Cartesian features.

Features	A1		A2		A3	
	Average	Variance	Average	Variance	Average	Variance
Mov1	26,4%	0,024	28,0%	0,015	26,0%	0,009
Mov2	33,2%	0,018	27,2%	0,015	29,6%	0,011
Mov1 + Mov2	27,2%	0,017	25,2%	0,006	25,6%	0,005
RF + DF + Mov1	40,0%	0,018	42,4%	0,019	41,6%	0,009
RF + DF + Mov2	53,2%	0,016	56,0%	0,015	45,6%	0,013
RF + DF + Mov1 + Mov2	35,6%	0,022	32,8%	0,011	29,2%	0,007

Table 6.6 shows very low accuracies, mainly in the dynamic features. Even combined with the hand shape features, the returned results are worse than the ones which were obtained with these features alone. The transition matrix *A3* is no longer the one with the best results.

The model was tested with the gestures grouped by their similarity using the groups presented in Section 6.1. Table 6.7 shows the results obtained with the same data but with the grouped classes.

Table 6.7 - Leave-one-subject-out classification results with Cartesian features and gestures grouped by their similarity.

Features	A1		A2		A3	
	Average	Variance	Average	Variance	Average	Variance
RF + DF	62,4%	0,011	65,6%	0,010	60,8%	0,008
Mov1	50,0%	0,014	50,8%	0,012	53,6%	0,012
Mov2	53,6%	0,020	49,2%	0,018	52,8%	0,004
Mov1 + Mov2	54,4%	0,012	54,0%	0,022	56,8%	0,012
RF + DF + Mov1	71,6%	0,010	66,8%	0,011	62,8%	0,009
RF + DF + Mov2	71,6%	0,006	67,2%	0,013	68,4%	0,006
RF + DF + Mov1 + Mov2	64,0%	0,012	61,6%	0,022	60,8%	0,010

The obtained results are better than previously, but the improvement is not very significant since the used class grouping implied that the new classes should be much more distinct than the original ones. These results demonstrate that the this set of features does not represent very well the dynamic gestures used and lead to the use of the second set of features which are independent of scale and less affected by the execution speed and subject posture. In Table 6.8 the results of the training set accuracies obtained with the Relative features and 4 states are outlined.

Table 6.8 - Training set classification accuracies with Relative features.

Features	A1		A2		A3	
	Average	Variance	Average	Variance	Average	Variance
Mov	73,6%	0,001	73,2%	0,005	86,4%	0,004
RF + DF + Mov	100,0%	0,000	99,6%	0,000	100,0%	0,000

The single *Mov* features did not return very good results when compared to the Cartesian features of Table 6.4. The success of these features emerges when the test data is unknown to the training. The leave-one-subject-out classification results for the original and grouped labels can be seen in Table 6.9.

Table 6.9 - Leave-one-subject-out classification results with Relative features.

Labels	Features	A1		A2		A3	
		Average	Variance	Average	Variance	Average	Variance
Original Labels	Mov	37,2%	0,010	38,0%	0,011	38,0%	0,009
	RF + DF + Mov	71,2%	0,013	68,8%	0,016	70,8%	0,005
Grouped Labels	Mov	56,8%	0,018	61,6%	0,010	57,2%	0,009
	RF + DF + Mov	76,0%	0,009	78,4%	0,006	83,2%	0,004

As it was observed in the previous analysis, the single *Mov* features did not improve the results but this table shows that their combination with the hand shape features returns good accuracies when compared to the Table 6.6 and Table 6.7. To show the types of misclassifications, Table 6.10 presents the average confusion matrix of the features *RF+DF+Mov* with the original labels and transition matrix of type *A3*.

Table 6.10 - Confusion matrix of relative features.

		Predicted									
		I	You	He	We	You(P)	They	My	Your	Our	Yours
True	I	0,84	0,08						0,08		
	You	0,08	0,64							0,28	
	He			0,84			0,16				
	We				0,88	0,08				0,04	
	You(P)				0,4	0,4					0,2
	They						1				
	My	0,16						0,76		0,08	
	Your	0,24				0,04			0,4	0,08	0,24
	Our				0,28	0,04		0,16		0,48	0,04
	Yours					0,16					0,84

Table shows that most of the obtained misclassifications are related to the possessive pronouns. The gestures *WE* and *THEY* are the most distinct of the rest.

6.4.3 Results Conclusions

The used single movement features returned low accuracies in all classifications performed with test data outside of the training set. This might be related to the naturalness of the gestures execution and different postures of the subjects. However, it should be noticed that the gestures used are quite similar to each other and some of them represent in fact variations of the others. These results are not unexpected since, as was pointed it Section 6.1 some of the gestures have exactly the same hand movement varying only in terms of hand shape. The combination with hand shape features improved the classification significantly to both features sets. Relative features proved to be better to represent these gestures and to be used in classification systems.

Chapter 7

Conclusions

Sign Language Recognition systems are in an embryonic stage of development despite of all the work realized until now. This field can be decomposed in three main areas: Segmentation, Feature Extraction and Recognition, each one with enough complexity to form a field by itself.

The first part of this work was focused on the study of features that can be extracted from static gestures, which do not have movement component. Two datasets were used: the Triesch dataset and the cardinal numbers from the LGP dataset. For the first, which is to a benchmark dataset, good performances were obtained even with challenging segmentation conditions. The second dataset returned significantly worse results even with the inclusion of depth information. These performance differences are related to the execution naturalness of the gestures. The gestures of the first dataset were realized with rigidity while in the second dataset the gestures were performed freely. This indicates that the standard appearance based features, widely implemented in many studies, are not enough to represent signs in real world applications. The study of features which represent the hand structure may improve the hand representation and is left for further studies.

The study of gestures with movement was the target of the second part. Two distinct operations were performed. The first consisted in evaluate the gestures similarity within the same class but performed by different subjects. The similarity was measured using the Dynamic Time Warping metric. A high level of misclassification was observed, mostly between gestures which are naturally similar. Again, the inclusion of depth information did not improve the results. The second operation was the classification of the gestures sequences with Hidden Markov Models. This classifier has been broadly used to classify sequential data and got good results when compared with the DTW similarities assessed. The HMM classification showed that the use of relative position features to describe movement returns better performances than the use of global position features.

The depth information did not add relevant improvements in this work but this conclusion should not be generalized. As stated the gestures studied were performed freely and some of the

misclassifications identified are not easy to detect, even for humans unfamiliarized with the language or out of context. The depth information may be used to infer the 3D hand shape and compute the structure features mentioned.

References

[1] J.-P. Bonet, *Reduction de las letras, y arte para enseñar a ablar los mudos*: Abarca de Angulo, 1882.

[2] W. C. Stokoe, "Sign language structure: An outline of the visual communication systems of the American deaf," *Journal of deaf studies and deaf education,* vol. 10, pp. 3-37, 2005.

[3] M. 1880. (16-07-2014). *Milan 1880 Congress*. Available: http://milan1880.com/milan1880congress/eightresolutions.html

[4] Portugal and A. V. Ferreira, *Gestuário: língua gestual portuguesa*, 1997.

[5] M. A. Amaral, A. Coutinho, M. R. D. Martins, and R. Johnson, *Para uma gramática da língua gestual portuguesa*, 1994.

[6] I. Mesquita and S. Silva, "Guia prático de Língua Gestual Portuguesa: Ouvir o silêncio," *Nova Educação. Braga,* 2007.

[7] H. Wittmann, "Classification linguistique des langues signées non vocalement," *Revue québécoise de linguistique théorique et appliquée,* vol. 10, pp. 215-288, 1991.

[8] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review,* pp. 1-54, 2012.

[9] R.-H. Liang and M. Ouhyoung, "A real-time continuous gesture recognition system for sign language," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, 1998, pp. 558-567.

[10] G. Fang, W. Gao, X. Chen, C. Wang, and J. Ma, "Signer-independent continuous sign language recognition based on SRN/HMM," in *Gesture and sign language in human-computer interaction*, ed: Springer, 2002, pp. 76-85.

[11] E.-J. Holden and R. Owens, "Visual sign language recognition," in *Multi-Image Analysis*, ed: Springer, 2001, pp. 270-287.

[12] E.-J. Holden, G. Lee, and R. Owens, "Automatic recognition of colloquial Australian sign language," in *Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on*, 2005, pp. 183-188.

[13] H. Cooper and R. Bowden, "Large lexicon detection of sign language," in *Human-Computer Interaction*, ed: Springer, 2007, pp. 88-97.

[14] H.-D. Yang, S. Sclaroff, and S.-W. Lee, "Sign language spotting with a threshold model based on conditional random fields," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 31, pp. 1264-1277, 2009.

[15] U. Von Agris, J. Zieren, U. Canzler, B. Bauer, and K.-F. Kraiss, "Recent developments in visual sign language recognition," *Universal Access in the Information Society,* vol. 6, pp. 323-362, 2008.

[16] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 2012, pp. 1975-1979.

[17] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *International Journal of Computer Vision,* vol. 46, pp. 81-96, 2002.

[18] R. Khan, A. Hanbury, J. Stöttinger, and A. Bais, "Color based skin classification," *Pattern Recognition Letters,* vol. 33, pp. 157-163, 2012.

[19] G. Awad, J. Han, and A. Sutherland, "A unified system for segmentation and tracking of face and hands in sign language recognition," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 2006, pp. 239-242.

[20] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer based video," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 20, pp. 1371-1375, 1998.

[21] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," in *International Workshop on Automatic Face and Gesture Recognition*, 1995, pp. 296-301.

[22] C.-C. Chang, J.-J. Chen, W.-K. Tai, and C.-C. Han, "New approach for static gesture recognition," *Journal of information science and engineering,* vol. 22, pp. 1047-1057, 2006.

[23] D. Kelly, J. McDonald, and C. Markham, "A person independent system for recognition of hand postures used in sign language," *Pattern Recognition Letters,* vol. 31, pp. 1359-1368, 2010.

[24] D. Brien, U. o. D. D. S. R. Unit, and B. D. Association, *Dictionary of British Sign Language/English*: Faber & Faber, Limited, 1992.

[25] A. M. Martínez, R. B. Wilbur, R. Shay, and A. C. Kak, "Purdue RVL-SLLL ASL database for automatic recognition of American Sign Language," in *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, 2002, p. 167.

[26] R. Yang, S. Sarkar, and B. Loeding, "Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 32, pp. 462-477, 2010.

[27] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, A. Thangali, *et al.*, "Large lexicon project: American Sign Language video corpus and sign language indexing/retrieval algorithms," in *Proc. Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, 2010.

[28] X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang, X. Chen, *et al.*, "Sign Language Recognition and Translation with Kinect," in *IEEE Conf. on AFGR*, 2013.

[29] P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and H. Ney, "Benchmark Databases for Video-Based Automatic Sign Language Recognition," in *LREC*, 2008.

[30] P. Dreuw, J. Forster, and H. Ney, "Tracking benchmark databases for video-based sign language recognition," in *Trends and topics in computer vision*, ed: Springer, 2012, pp. 286-297.

[31] U. von Agris and K.-F. Kraiss, "Towards a video corpus for signer-independent continuous sign language recognition," *Gesture in Human-Computer Interaction and Simulation, Lisbon, Portugal, May,* 2007.

[32] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. H. Piater, *et al.*, "RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus," in *LREC*, 2012, pp. 3785-3789.

[33] J. Bungerot, D. Stein, P. Dreuw, H. Ney, S. Morrissey, A. Way, *et al.*, "The ATIS sign language corpus," 2008.

[34] K. McGuinness. (16-07-2014). *Interactive Segmentation Tool*. Available: http://kspace.cdvp.dcu.ie/public/interactive-segmentation/

[35] K. McGuinness and N. E. O'connor, "A comparative evaluation of interactive segmentation algorithms," *Pattern Recognition,* vol. 43, pp. 434-444, 2010.

[36] J. Triesch and C. von der Malsburg, "Classification of hand postures against complex backgrounds using elastic graph matching," *Image and Vision Computing,* vol. 20, pp. 937-943, 2002.

[37] M.-K. Hu, "Visual pattern recognition by moment invariants," *Information Theory, IRE Transactions on,* vol. 8, pp. 179-187, 1962.

[38] C. M. Bishop, *Pattern recognition and machine learning* vol. 1: springer New York, 2006.

[39] R. a. contributors. (16-07-2014). *RapidMiner*. Available: http://rapidminer.com/

[40] A. Just, Y. Rodriguez, and S. Marcel, "Hand posture classification and recognition using the modified census transform," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, 2006, pp. 351-356.

[41] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition* vol. 14: PTR Prentice Hall Englewood Cliffs, 1993.

[42] M. Müller, *Information retrieval for music and motion* vol. 2: Springer, 2007.

[43] K. P. Murphy, *Machine learning: a probabilistic perspective*: MIT press, 2012.

Appendix A

Table A.1 - LGP Database isolated signs.

Category	Sign
Alphabet	A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q R, S, T, U, V, W, X, Y, Z
Cardinal number	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Pronouns	Eu, Tu, Ele, Nós, Vós, Eles, Meu, Teu, Nosso, Vosso, Como, Onde, Porquê, Qual, Quando, Quem
Verbs	Andar, Aprender, Beber, Cair, Comer, Comprar, Condizir, Correr, Ensinar, Escrever, Estudar, Falar, Gostar, Ir, Jogar, Ler, Ouvir, Partir, Perder, Trazer, Vender, Ver, Vir
Adverbs	Bom, Bonito, Feio, Grande, Mau, Novo, Pequeno, Sujo, Velho
Basic Expressions	Adeus, Ajudar, Com licença, Desculpe, Não, Obrigado, Olá, Por favor, Sim
Feelings / Emotions	Aborrecido, Amor, Cansado, Doente, Feliz, Triste, Zangado
Colors	Amarelo, Azul, Branco, Castanho, Cinzento, Laranja, Preto, Rosa, Roxo, Verde, Vermelho
Family	Avó-Avô, Bebé, Casado, Divorciado, Filho, Irmão-Irmã, Mãe, Pai, Avó-Avô, Bebé, Casado, Divorciado, Filho, Irmão-Irmã, Mãe, Pai, Rapaz-Rapariga, Solteiro
Professions	Advogado, Arquiteto, Bombeiro, Cientista, Enfermeiro, Engenheiro, Médico, Músico, Policia, Professor
Places	Casa, Casa de Banho, Cidade, Cozinha, Escola, Hospital, Hotel, Igreja, Loja, País, Praia, Quarto, Restaurante
Food	Água, Café, Carne, Copo, Maça, Peixe, Prato, Queijo, Talheres
Animals	Cão, Cavalo, Gato, Inseto, Ovelha, Pássaro, Porco, Vaca
Time	Amanha, Ano, Dia, Hoje, Mês, Noite, Ontem, Tarde
Weather	Calor, Chuva, Frio, Neve, Nevoeiro, Sol

Table A.2 - LGP Database sentences.

Number	Sentences
1	Tudo bem?
2	A tua família como está?
3	Como te chamas?
4	Qual é o nome da tua mãe?
5	Que idade tens?
6	Onde vives?
7	Qual é o teu número de telefone?
8	Onde trabalhas?
9	Qual é a tua Profissão?
10	És surdo ou ouvinte?
11	Tens irmãos?
12	Tens animais de estimação?
13	Queres ir ao cinema?
14	Onde fica a estação de metro mais próxima?
15	Qual é o melhor hotel da cidade?
16	Eu sou médico.
17	Eu tenho 30 anos.
18	O meu irmão quer ser jogador de futebol.
19	Eu tenho dois irmãos e uma irmã.
20	O meu irmão está à procura de trabalho.
21	Os meus avós vivem em nossa casa.
22	Eu vivo em Portugal mas nasci em Itália.
23	Eu vou de autocarro para a escola.
24	Hoje de manha fui de carro para o trabalho.
25	A minha irmã joga basquetebol.
26	Ontem fui ver um jogo de futebol.
27	Eu costumo ler antes de adormecer.
28	O meu pai comprou um livro novo.
29	A minha irmã adora fazer compras.
30	O meu primo vai casar no próximo ano.
31	No próximo fim-de-semana vou ao cinema com os meus amigos.
32	O meu amigo tem uma casa de praia.
33	Eu conheci uma rapariga muito bonita.
34	Ontem senti-me doente e fui ao hospital.
35	Eu gosto de tomar o pequeno-almoço no café.
36	Hoje de manha comi leite com cereais.
37	Eu prefiro comer carne do que peixe.
38	Amanha vou jantar a um restaurante indiano com a minha família.
39	Amanhã vai chover.
40	Hoje está frio.