Universidade do Minho

Escola de Engenharia

Carlos Miguel Freitas Sotelo

PORTUGUESE SIGN LANGUAGE RECOGNITION FROM DEPTH SENSING HUMAN GESTURE AND MOTION CAPTURE

October 2014

Universidade do Minho

Escola de Engenharia

Carlos Miguel Freitas Sotelo

PORTUGUESE SIGN LANGUAGE RECOGNITION FROM DEPTH SENSING HUMAN GESTURE AND MOTION CAPTURE

Dissertação de Mestrado
Mestrado em Engenharia Informática
Trabalho realizado sob a orientação de:
Prof. Carlos Silva (Universidade do Minho)
Prof. Miguel Sales Dias (Microsoft Language Development Center, Instituto Universitário de Lisboa, ISCTE-IUL)

October 2014

"Would you kindly..." – Frank Fontaine

Acknowledgments

Now that this journey is about to end, I need to thank to all those who helped me, directly or indirectly, through the making of this thesis.

First of all, my humble acknowledgements to my supervisors Miguel Sales Dias and Carlos Silva. Miguel Dias, thank you for making this experience possible in Microsoft and for believing in me. Carlos Silva, I am grateful for giving me good advice and for all review of this research.

Secondly, my greatest acknowledgement to João Freitas and Hélder Abreu, for all the "Friday Talks", advises and companionship. João, thank you for all your support and patience during these months. Your guidance and teachings were crucial for me to bring this thesis to fruition. Hélder, thanks for being a great pillar and friend, for feeding my healthy discussions with distinct ideals and perspectives during our stay at MLDC.

I could not forget to thank Rui Almeida. Your help, work and support at the beginning of this research were crucial, helping it to become true.

Also, best regards to all MLDC personnel, particularly the MLDC Porto, who were always available and always step forward intending to support and teach me in many ways.

Thanks to my roommates – Hélder Abreu (again), Nuno Morais and Ana Silva – for all the cooperation within our stay in Porto. For brightening my nights and dinners.

Last but not least, a big "thank you" to my mother and father, to my sister and to my little brother for always being there, despite the trouble. Filipa Lima, thank you for supporting me emotionally throughout this period, for being there when needed, for bearing all my humors, and thank you for all the possible help revising this thesis.

Resumo

Tal como as línguas faladas, as línguas gestuais evoluíram ao longo do tempo, contendo gramáticas e vocabulários próprios, sendo assim oficialmente consideradas línguas. A principal diferença entre as línguas faladas e as línguas gestuais é o meio de comunicação, sendo dessa forma as línguas gestuais línguas visuais. Sendo que a principal língua falada entre a comunidade surda é a língua gestual, construir uma ferramenta que funcione como uma ligação que facilite a comunicação entre a comunidade surda e o resto das pessoas é o principal objetivo e motivação desta dissertação.

O nosso sistema tem como característica principal não ser intrusivo, descartando o uso de sistemas de "*Data Gloves*" ou sistemas dependentes de múltiplas câmaras ou outros aparelhos. Isto é conseguido usando um único aparelho, o *Kinect One* da *Microsoft*, que consegue captar informações de cor e profundidade.

No desenvolvimento deste trabalho, quarto experiências foram realizadas: reconhecimento simples da configuração da mão; reconhecimento da configuração da mão em sinais; reconhecimento de sinais usando somente informação dos trajetos das mãos; reconhecimento de sinais com o trajeto e as configurações das mãos. A primeira e terceira experiências foram realizadas de forma a conferir o método de extração de características, enquanto a segunda e quarta experiências foram conduzidas de forma a adaptar os primeiros sistemas ao problema real do reconhecimento de sinais em LGP.

A primeira e segunda experiências obtiveram taxas de acerto de 87.3% e 64.2% respetivamente enquanto as experiências respetivos ao reconhecimento de sinais obtiveram taxas de 91.6% para a experiência contendo só o trajeto da mão, e 81.3% com o trajeto e a configuração das mãos.

Abstract

Just like spoken languages, Sign Languages (SL) have evolved over time, featuring their own grammar and vocabulary, and thus, they are considered real languages. The major difference between SL and other languages is that the first one is signed and the second one is spoken, meaning that SL is a visual language. SL is the most common type of language among deaf people as no sense of hearing is required to understand it.

The main motivation of this dissertation is to build a bridge to ease the communication between those who are deaf (and hard-of-hearing) and those not familiarized with SL. We propose a system whose main feature is the absence of intrusion, discarding the usage of glove like devices or a setup with multiple cameras. We achieved this using the Kinect One sensor from Microsoft. Using a single device, we can acquire both depth and colour information, yet this system makes usage only on the depth information.

Four experimental situations have been performed: simple posture recognition, movement postures recognition, sign recognition using only hand path information, and sign recognition using hand path and hand configuration information. The first and third experimental classes were conducted, in order to confirm the feature extraction method’s eligibility while the second and fourth experiments were conducted to address our hypothesis. Accuracy rates reached 87.4% and 64.2% for the first and second experiments, respectively. In the experiments concerning signs, accuracy rates of 91.6% for hand path data only, and 81.3% for hand path and hand configuration data were achieved.

CONTENTS

List of figures

List of Tables

Acronyms and Abbreviations

AGR	– Automatic Gesture Recognition
ASL	– American Sign Language
APSLR	– Automatic Portuguese Sign Language Recognition
CRF	– Conditional Random Fields
CSL	– Chinese Sign Language
GUI	– Graphical User Interface
HCI	– Human-Computer Interaction
HCRF	– Hidden Conditional Random Fields
HMM	– Hidden Markov Models
LDCRF	– Latent Dynamic Conditional Random Fields
ME	– Movement Epenthesis
MLDC	– Microsoft Language Development Center
NMF	– Non-manual features
PSL	– Portuguese Sign Language
RGB	– Red, Green, Blue
RGB-D	– Red, Green, Blue - Depth
SL	– Sign Language
SLR	– Sign Language Recognition
SRN	– Simple Recurrent Network
SVO	– Subject-Verb-Object
SOV	– Subject-Object-Verb
2D	– Two Dimensional
3D	– Three Dimensional

1 Introduction

1.1 Sign Language

Just like any spoken and written language, Sign Languages (SL) have evolved over time, featuring their own grammar and vocabulary, and thus, they are considered real languages. The major difference between SLs and other languages is that the first ones are signed while the second ones are spoken, which means that SLs are **visual** languages. SLs are the most common type of languages among deaf people since no sense of hearing is required to understand it.

Up until the late 1960s, SL were not considered real languages, often being assumed as sets of gestures that could be loosely connected to convey meaning to simple relations. **Dr. William C. Stokoe**, with the help of some of his deaf students from the University of Gallaudet, published in 1960 the monograph Sign Language Structure (a version can be found in (Stokoe, 2005)) where the author proposed that signs could be analysed as the composition of three different elements without meaning: shape of the hand, motion of the hand, and position occupied by the hand. This assumption permitted him to consider SL as a natural language. Although at the beginning his affirmations were seen with some repulsion due to the novelty of his ideas, this study had a very important role in the publication of the first American Sign Language (ASL) dictionary based on linguistic principles. In this first dictionary, Stokoe organized the signs depending on its shapes (position of the hand, shape, motion, etc.) and not depending on its English translation. This publication was the beginning of SL linguistics research.

The Portuguese government only recognized the Portuguese Sign Language (PSL) as an official Portuguese language, along with Portuguese and Mirandese, as in 1997.

1.1.1 Stokoe’s Model

In spoken language, the phonology refers to the study of physical sounds present in human speech (known as phonemes). Similarly, the phonology of SL can be defined. Instead of sounds, the “phonemes” are considered as the different signs present in a row of hand signs. They are analysed taking into account the following parameters (Stokoe, 2005):

1. **Hand Configuration**[1] - hand shape configuration when doing the sign;
2. **Orientation of the hand –** orientation where the palm of the hand is pointing to;
3. **Position** - where the sign is done according to the rest of the body (mouth, forehead, chest, shoulder);
4. **Motion -** movement of the hand when doing the sign (swaying, circularly).
5. **Contact point**: dominant part of the hand that is touching the body (palm, fingertip, back of the fingers).
6. **Plane -** where the sign is done, depending on the distance with respect to the body (first plane being the one with contact to the body and fourth plane the most remote one).
7. **Non-manual features** (NMF) - refers to the information provided by the body (facial expression, lip movements, or movements of the shoulders). I.e. when the body leans to the front, it expresses future tense. When it is leaned back, expresses past tense. Also, non-manual signs such has face expression, show grammatical information such as question markers, negation or affirmation, localization, conditional clauses, and relative clauses.

1.1.2 Movement-Hold model

While Stokoe’s work was the first to model and detail the SL, other models followed.

In 1989 Lidell and Johansen (Liddell, Johnson, 1989) developed the movement-hold model which was summarized by Valli and Lucas (Valli, Lucas, 1992):

[1] Throughout this work, Hand Configuration will be multiple times referred as **posture** or **hand posture;**

“The basic claim about the structure of signs in the Movement-Hold Model is that signs consist of hold segments and movement segments that are produced sequentially. Information about the handshape, location, orientation, and non-manual signals is represented in bundles of articulatory features...Holds are defined as periods of time during which all aspects of the articulation bundle are in a steady state; movements are defined as periods of time during which some aspect of the articulation is in transition. More than one parameter can change at once. A sign may only have a change of handshape or location, but may have change of both handshape and location, and these changes take place during the movement segment.”

This model contrasts to the work of Stokoe where different components of the sign are described in different channels. While Stokoe’s model can be seen as a parallel model, in which the properties take their values, the movement-hold model is a sequence of many properties changing between Holds and Movements.

Both models have similar approaches and conclusions and despite not being obvious how best to include these higher level linguistic constructs of the language, it is obviously essential for Sign Language recognition - SLR to become reality. Within SLR both the movement-hold, sequential information from Liddell and Johnson and the parallel forms of Stokoe, are acceptable annotations.

Inter-signer differences are very large; every signer has their own style, in the same way that everyone has their own accent or handwriting. Signers can be either left handed or right handed. For a left handed signer, most of the signs will be mirrored.

1.1.3 Portuguese Sign Language

According to (Bela Baltazar, 2010), the PSL a sign is composed by 5 features, where the first 3 compose the base of any sign: **hands configuration**, **place of articulation**, **hands orientation,** while the other 2: **facial expression** and **body movement**, with equal importance, can distinguish signs with similar execution. In (Bela Baltazar, 2010) there are identified 14 facial expressions and 57 hand configurations for PSL.

There are other properties similarly to what happens in the models described previously:

Gender – the occurrence of the gender modifier only happens in the specific case of animated beings. Usually it is done with the usage of the signs *“man”* or “*woman”*. However, the masculine is usually denoted by the absence of the modifier, while the feminine is predominantly marked by prefixing i.e. *“queen”* is the conjunction of the signs *“woman”* and “*king” in that order.* Other cases exist in which the feminine has a different sign than the masculine, as in “*father”* and “*mother”*.

Number – there are multiple ways of denoting the plural. The repetition of the sign (as in *“coisa”/”coisas”),* can be achieved by doing the sign with both hands, if originally is performed by only one (as in *“pessoa”/”pessoas”),* the usage of a numeral to specify small quantities (as in “*quatro filhos*” that is *“filho”* and *“quatro”*) or the usage of a determinative, to non-countable amounts (as in “*muitos homens”*, sign composed by “*homem”* and “*muito”).*

Order of the elements in the phrase – as in other pairs of SL and its matching spoken/written language, PSL has a structure distinct from the Portuguese Language (PL). The predominant structure of a phrase in PL is the subject–verb–object (SVO), while in PSL the predominant structure is subject-object-verb (SOV). Some examples can be:

Table 1 – Differences between a phrase in PL and PSL. While the PL predominantly follows a SVO structure, PSL uses SOV.

Language	Sentences	
PL	*“O aluno deu uma flor à professor.”*	*“Eu vou para casa.”*
PSL	*“Aluno flor professora dar.”*	*“(Eu) casa ir.”*

The meaning of the left sentence in PL is *“The student gave the teacher a flower”*, while the right one is *“I go home” while in PSL is “Student flower professor give”* and *“(I) home go”,* respectively.

From these examples it is possible to see that PSL does not use prepositions such as “*o”,* “*uma” and “à”* (“the”, “one” and “to”) and that in some cases, for instance, if the subject is implicit in the context, it is not always necessary to perform the sign of said subject (in the right sentence (*“Eu*”).

Also observable in the examples is the property that all verbs are signed in the infinitive form (In the examples it is only observable in Portuguese and not in the English translation, since in English, the conjugation of the verbs “go” and “give”, in the first person for these particular verbs, match the infinitive form. The same does not happen in Portuguese). To show other temporal conjugation of the verbs, the time adverbs are added. In the absence of these adverbs the body moves, leaning forward to represent future or backwards to represent past.

Type of sentence – to perform a question, the signer resorts to facial expression that can be combined with the use of interrogative pronouns, which appear at the end of the sentence. For the exclamatory sentence, other facial expressions are used as well as the posture of the torso and head can change.

Negative Form – the negation of a sentence is accomplished with the usage of body expression such as the movement of the head, the execution of the gesture “no” or through the facial expression combined with the movement of the head.

1.2 Motivation

The major motivation of this thesis is to contribute to build a bridge and ease the communication between the deaf and hearing impaired people and people not familiarized with SL, especially in the case of communication between speakers of PSL and of Portuguese, being this quest a few years old with only recent and significant breakthroughs (Almeida 2011).

1.3 Problem Description

In Automatic Gesture Recognition (AGR), one of the most difficult challenges is to classify the meaning of the sensed and acquired gesture raw data, for example, in the context of a gesture-based Human-Computer Interaction (HCI) system for an application, and the same happens with SL. The sequences of raw static or moving data, that comprise a gesture or a sign, must be understood by the application.

As explained before, a sign in SL is composed by smaller parts that, despite being generally acknowledged in any SL model, are not entirely addressed yet in a single academic work with full proficiency and good results, namely considering Stokoe 7 fundamental sign parameters: **hand configuration, orientation of the hand, position, motion, contact point, plane** and **non-manual features.**

Most of the works in the literature, usually with very interesting results (Almeida, 2011; Capilla, 2012; Chai et al., 2013; Vogler, Metaxas, 1999), focus solely on the gesture[1] part of the sign that is performed by the hands (usually called the hand path), independently of being isolated sign or continuous sign recognition. Some contributions simply address the problem of recognizing the hand configuration, once again in isolated postures (Almeida, 2011; Kollorz et al., 2008) or in the purpose of finger spelling (Uebersax, Gall, 2011). Fewer works address both hand posture and hand path problems at once (Souza, Pizzolato, 2013), and even fewer do so in a non-intrusive and simple way, like with the Microsoft Kinect sensor (Souza, Pizzolato, 2013).

The main problem that this thesis addresses, is how to successfully perform Automatic Portuguese Sign Language Recognition, for signs that observe only the manual features of the Stokoe model, that is, hand configuration, orientation, position, motion, contact point and plane of the hand.

[1] A gesture is: “A motion of the limbs or body to express thought or to emphasize speech.” (Dictionary, 2014)

1.4 Thesis Hypothesis

To properly address the identified problem, we state the following thesis hypothesis:

- **H1** - In the first hypothesis, we state that it is possible to perform Automatic Portuguese Sign Language Recognition (APSLR) of many signs, by analysing the manual features of the Stokoe model, namely, the hand configuration, orientation, position, motion, contact point and plane of the hand, where the last 6 parameters are strictly connected and can be implicitly observed by analysing the hand path (or hand spatial trajectory), which is the movement performed by both hands. We further state that hand configuration and hand path can be automatically recognized by using a suitable machine learning-based classification technique, trained and tested with data collected by a low-cost non-invasive RGB-D sensor.
- **H2** – Our second hypothesis is that an approach that classifies both hand configuration and hand path, can outperform a system based only on hand path classification using, like for hypothesis H1, a suitable machine learning-based classification technique, trained and tested with data collected by a low-cost non-invasive RGB-D sensor.

1.5 Objectives

To demonstrate the hypothesis described above**,** the following thesis objectives were enunciated:

O1 – Automatic Hand Posture Recognition System - Specify, develop and test an **Automatic Hand Posture Recognition** system for PSL hand configurations, that uses fully 3D data structures to define, describe, record and classify in real-time the hand posture of both hands in a PSL sign, whose data is captured by a RGB-D sensor (Kinect One). We have used a k-fold cross-validation technique to train and test a Support Vector Machine - SVM classifier. This objective is related with hypothesis **H1**.

O2 – Automatic Hand Path Recognition System - specify, develop and test an **Automatic Hand Path Recognition** system, that uses joints positions from both hands trajectories, to define, describe, record and classify in real-time the paths of both hands in a PSL sign, whose data is captured by a RGB-D sensor (Kinect One). We have used a k-fold cross-validation technique to train and test a Support Vector Machine - SVM classifier. This objective is related with hypothesis **H1**.

O3 – Automatic Sign Recognition System – specify, develop and test a system that, using the sub-systems developed to fulfil objectives O1 and O2, defines, describes, records and classifies some signs in PSL using imaging, depth and joints hand data collected by a RBG-D sensor. We have also used a k-fold cross-validation technique to train and test a Support Vector Machine - SVM classifier. This objective is related with hypothesis **H1**. By comparing the results of O3 and O2, it also addresses the second hypothesis (**H2**).

O4 – Software Application - PhySaLiS - Portuguese Sign Language Recognition System– develop and test a software application that, using sub-systems developed for O1, O2 and O3, lets the user record and view data on hand configurations and PSL signs in a useful way.

The description of Portuguese Sign Language has its specific meanings and symbols, which differs from other sign languages. In this sense, it is important to verify if the work and results reported in the literature regarding other Sign Languages, are also valid and possible to achieve in the case of PSL. Our intention is to show that our research is on pair and event extends the literature results, for simple and limited problems of Automatic Sign Language Recognition, with a specific application to PSL.

Apart from its scientific goals, this work has the social aim to increase the social inclusion of more than 100.000 hearing impaired people that live in Portugal (Bela Baltazar, 2010). Yet, we don´t claim in this thesis to propose a final and unique solution to the immense problem of Automatic Portuguese Sign Language Recognition, but rather to contribute with an original approach, to such research challenge.

1.6 Document Structure

After presenting some fundamental concepts about SLs for a good understanding of the context of the research presented in this thesis, the remaining chapters of this document are structured as follows:

Chapter 2: This chapter describes some of the critical related works, taken from the state-of-the-art literature, regarding the two steps that usually compose SLR systems, namely methods for data collecting and analysis and classification methods.

Chapter 3: In this chapter, the details of the proposed system architecture and its application Graphical User Interface (GUI) are presented, its design and implementation are discussed and the description of the components for the developed system, detailed. We also present the data collection methods and properties, as well as details about the collected corpora and the recording specifications.

Chapter 4: This chapter presents the results of the methods used for the Posture, Hand Path recognition and Sign recognition processes. The comparison analysis with other literature works is also addressed in this chapter.

Chapter 5: Conclusions and considerations about our thesis hypothesis coverage, the fulfillment of our thesis objectives, and the recommendations for future work, are presented in this last chapter.

2 State of Art on Sign Language Recognition

Automatic SLR is divided into two major problems, namely extracting/detecting features, and recognizing them. This section is divided into 2 subsections. The first one, “Existing Data Collection Methods” addresses the feature extraction problem, this is, the way the data is captured and what data is captured. The data represents the information that the system has, in the previous states, before identifying the sign (for instance).

The second subsection, “Analysis and Classification Methods”, corresponds to the second major problem named before, the “recognizing”. This represents the problem of giving meaning to the data collected in the first phase. Deciding which sign/gesture represents, or even conveying meaning to entire “sentences” is the result of the classification.

2.1 Sign Language Recognition

SL is not merely a mirror of spoken language, it has a sentence structure and grammar that can be quite different to the language it is derived from (Kadhim Shubber, 2013). Used worldwide by a multitude of individuals, from those from the deaf communities, to their teachers, friends, and families. These communities often have their citizens segregated from the rest of society due to the difficulties in communicating with the rest (Almeida, 2011).

Typing and/or writing in Portuguese, or any other written language, is not straight forward for deaf and hard-of-hearing people. For those who have been deaf their whole lives, having learned as their first language a SL, learning spoken languages is comparable to learning a new language.

Currently, it is not possible for deaf and hard-of-hearing people to communicate with others in their native language using computers. Essentially, they have to communicate in a foreign language whenever they need to communicate with someone unfamiliar with SL, by typing or writing for instance.

Although, explored for many years, it is still a challenging problem in practice. A more cohesive and robust approach was recently developed by Microsoft (Chai et al., 2013) for SL. Despite this, in particular the case of

Portuguese Sign Language, an efficient system to perform automatic recognition of PSL still remains unknown.

The Kinect, while initially being developed for a gaming purpose, rapidly saw its original purpose adapted to various usages because of its low-cost as depth sensor, with very distinct activities, one of those, and most important to the matter, was the Sign and Gestures recognition.

The main idea is to use the Kinect to capture the gestures by retrieving information from the depth sensor, while machine learning and pattern recognition programming helps to interpret the meaning of those gestures.

By using the Kinect depth sensor to retrieve information from the scene, a lot of problems caused by, for instance, bad lighting in the scenario disappear once depth information and colour information are analysed. The older version of the Kinect device had most of its problems confined to the low available resolution, turning the recognition process difficult (Khoshelham, Elberink, 2012). This process is expected to suffer a substantial change with the anticipated Kinect to be used in our work, the Kinect One sensor.

By using the segmentation from one posture to another, and also combining the trajectory of the sign (Chai et al., 2013), it is possible to use machine learning technology and pattern recognition technology to make the final decision of what is the meaning of the gesture.

2.2 Existing Data Collection Methods

Data Collection is the first step for a SLR system, being for that one of the big areas in the SLR studies done for some time.

Some early SLR systems used “data gloves” and accelerometers to acquire specific hand features. The measurements (position, orientation, velocity, others) were directly measured using a sensor such as the DataGlove (Kadous, 1996; Vogler, Metaxas, 1997). Usually the data captured by the sensor was sufficiently discriminatory enough that feature extraction was almost inexistent and the measurements were directly used as features.

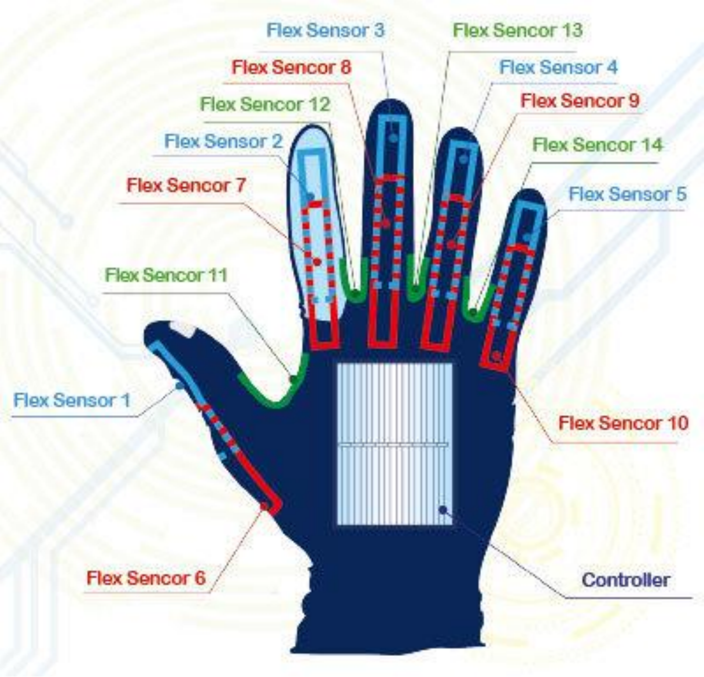


Figure 1 – Data glove example. Usually, this glove devices feature precise data about the hands parts positioning in 3D space and also including accelerometers, giving other information like velocity, acceleration, etc

These glove systems had several advantages when compared to simple video methods(Kadous, 1996):

- The processing power and bandwidth, required for real time video processing, were extremely high in contrast to the data extracted from the glove systems, which were concise and accurate compared to the information from video cameras.
- Some specific data such as hand orientation, forward/backward motion and finger position and information (due to fingers overlapping/ occlusion) are very difficult to extract from one simple video camera.
- Glove systems can be used regardless of the environment, whether complex backgrounds or signer attire.

While glove systems gave the advantage of accurate positions, they had an obvious downside, they constricted the mobility of the signer, altering the signs performed. Some efforts were made to modify the glove-like device in order to make a less constricting device, but the evolution in video devices (both in costs and performance) made the use of vision more popular to address the problem of SLR. Along with the previous facts, the community began to acknowledge that the hand tracking stage of the system does not attempt to produce a fine-grain description of hand shape, and therefore the use of such detailed information could be less relevant for humans to interpret SL (Fang et al., 2004).

The usage of vision input to address SLR problems started with a single camera. For these systems to solve the hands segmentation issue, algorithms such as “skin detection algorithms” or other methods to segment the hand were needed. Many works followed this or similar approaches, ranging from (Freeman, Roth, 1995; Parish et al., 1990) to (Pashaloudi, Margaritis, 2002; Wilson, Bobick, 2000) and (Wang, Quattoni, 2006; Yang et al., 2010). Another approach to solve this problem was the usage of coloured gloves to ease the segmentation issue. The 2D image usage as data input to solve the problem was also used in combination of multiple 2D cameras.

Sequence of images are captured from a combination of cameras. Some examples are systems that use one or more cameras such as: monocular (Zieren, Kraiss, 2004), stereo (Hong et al., 2007), orthogonal (Starner, Pentland, 1995) or other non-invasive sensors such as small accelerometers. In 1999 Segen and Kumar calibrated a light source (along with a camera) to compute depth through the shadow projections of the hands (Segen, Kumar, 1999). Other works (Brashear et al., 2003; McGuire, 2004; Starner, Pentland, 1995) used a front view camera in conjunction with a head mounted camera facing down on the subject’s hands to aid recognition, the last one also used accelerometers to aid the process.

In addition to the previous methods, depth can also be inferred using side and vertical mounted video cameras (Athitsos et al., 2010) or other combination of positions, such as cameras in the 3 axis, as with in (Vogler, Metaxas, o. J.).

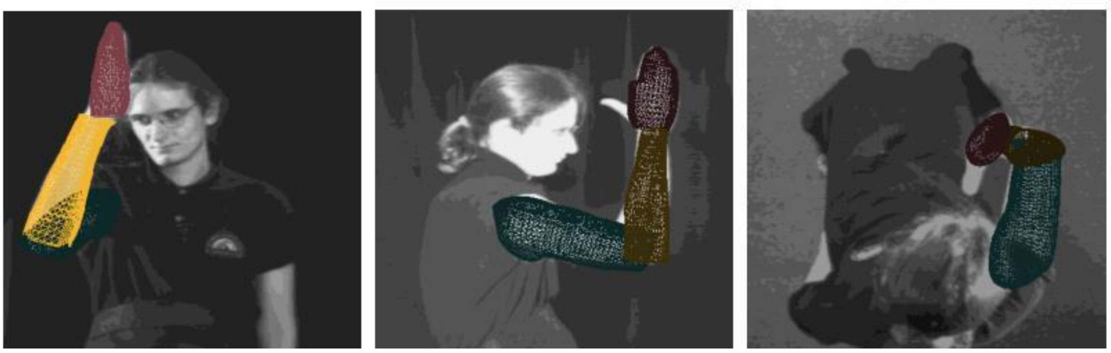


Figure 2 – 3 axis camera system (Vogler, Metaxas, o. J.)

These systems do not give much flexibility with “where to use” the system and are often accompanied with other restrictions, because most of them are created for controlled environments, and in the case of multiple video cameras, require specific calibrations and settings for the cameras positions, which in turn also require more space

Another data collection system that was used for SLR purposes was the Time Of Flight – TOF – camera (Kollorz et al., 2008). Despite this special camera being able to get depth information alone, it wasn’t extensively used due to its costs.

2.2.1 Microsoft Kinect

In 2008, Microsoft released the Kinect for Windows (v1) device for public use, along with an open library that allowed multiple uses for the device. The Kinect sensor featured a RGB video camera, an Infra-Red sensor and a multi-array microphone, which contains four microphones for capturing sound. Because of these four microphones, it is possible to record audio as well as find the location of the sound source and the direction of the audio wave. The RGB video camera had a resolution of 1280 x 960 pixels with a FOV (Field of View) of 43° vertical by 57° horizontal, with a depth image resolution of 640 x 480 pixels. In optimal conditions, this sensor managed to obtain 30 FPS of both colour and depth data.

Because of these specifications, the Kinect sensor was adopted by multiple researchers to address the SLR, usually in a multimodal approach. It also had the ability to follow up to 2 persons with a complete skeleton composed by 20 joints. This way, in a single device, researchers can have both RGB and depth data and even some joint information, given by the Kinect’s library.

Examples of Kinect usage for SLR systems are (Almeida, 2011; Capilla, 2012b; Chai et al., 2013; Zafrulla et al., 2011). It was mostly used because of the cheap way of acquiring depth information, simplifying the process to obtain the hands' positions, as well as other body parts

For this work, the new Kinect One sensor will be used. Released to the public this last September. It features a RGB camera that outputs 1920x1080 pixels of colour data, and an infrared camera that produces a 514x424 pixels depth image. With this information, the Kinect is able to estimate the body position and even 25 joints. For each joint the sensor gives it's positioning in a 3D space, in Cartesian coordinates (X, Y and Z). Despite having lower dimensions on the depth image, the new sensor achieved an improved accuracy on the depth values, as can be seen on Figure 3.

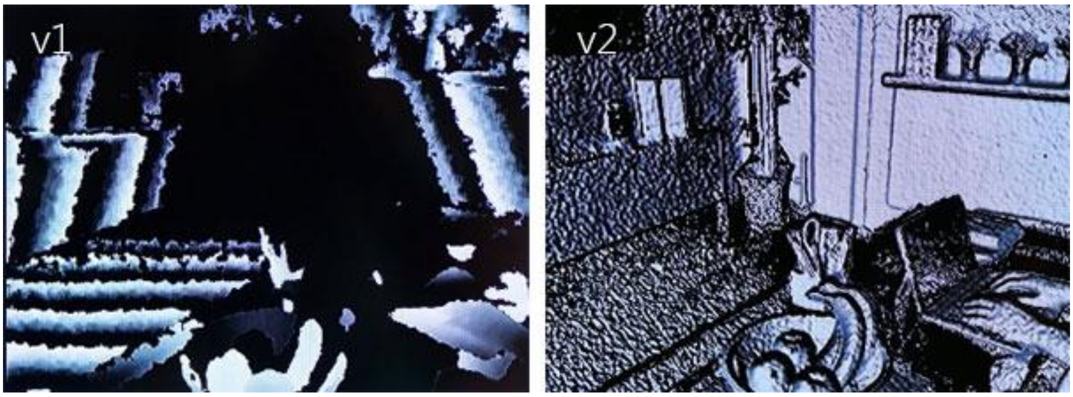


Figure 3 – Kinect versions depth comparison. The left image corresponds to the first version of the Kinect sensor, released in 2008, while in the right it is the same scene with the new Kinect One sensor. (Microsoft, 2014a)

2.3 Analysis and Classification Methods

Due to the nature of the problem of SLR, it soon became usual to establish the comparison between the speech recognition and SLR. Since both systems had much in common: both aim to recognize some language conveyed through a medium (one audible, the other visual); both processes vary with time (are time-varying) showing statistical variations, thus making the use of Hidden Markov Models (HMM) plausible for modelling both processes.

Both systems have to consider not only the context but also the coarticulation effects. However, there are also important differences. Speech signals are well-suited for analysis in the frequency domain, whereas SL signals, due to their spatial nature, do not show such a suitability (Vogler, Metaxas, 1997). Another problem that distinguish both systems is the coarticulation. While in the speech (audible) problem, the coarticulation is denoted by silence between words or one or more words affecting the pronunciation, and therefore the sound of following ones. This does not exactly add new sounds to the problem, but change the existing ones. While in the sign recognition problem, the coarticulation between signs is often visible, as when one gesture/sign, ends in a determined pose and the next meaningful sign starts in a complete different pose. In order to position the hands for the second sign, the signer must make a new movement, one which conveys no meaning to what the signer wants to express. This way, the problem of coarticulation in SLR is quite different from the Speech recognition because it adds new movements to the phrases. This problem is denominated "movement epenthesis" (ME).

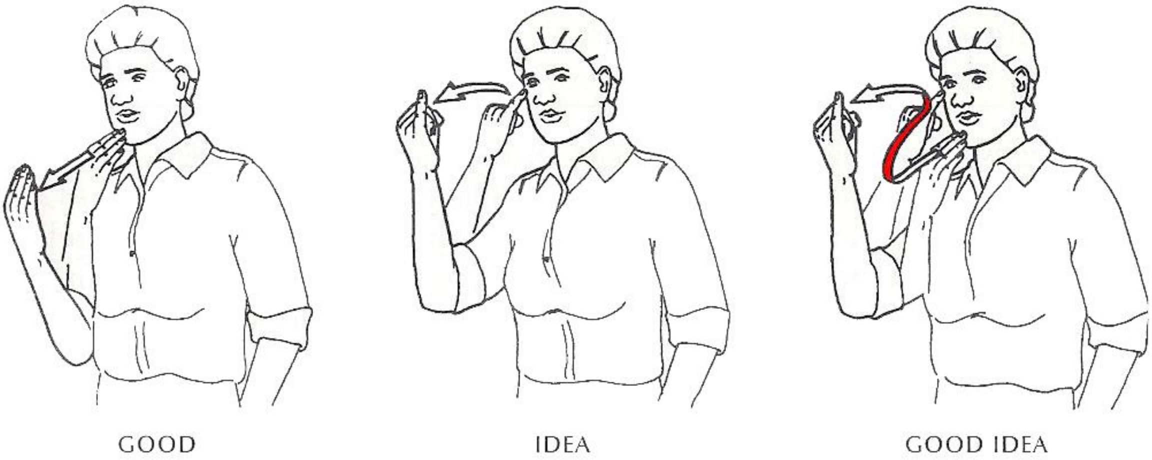


Figure 4 – The red path in the left image represents the movement epenthesis

With such similarities, most of the early approaches applied the use of HMM from the speech recognition research to the SLR problem. Many examples of HMM can be found in multiple projects, many with distinct forms of data collection.

In 1995, Starner and Pentland (Starner, Pentland, 1995) did not focus on the typical finger signing usually focused until then and instead, focused on gestures, which represent whole words, since real SL can only usually proceed at the normal pace of spoken conversations due to these kind of gestures. Through the use of HMM, this work achieved a low error rate on both training set and an independent test without invoking complex models of the hands (without modelling the fingers). They also conclude that with a larger training set and context modelling lower error rates are expected.

In (Vogler, Metaxas, 1998) improved upon their previous approach (Vogler, Metaxas, 1997) overcoming some limitations of the HMM method had by itself, by using Context-Dependent Modelling. They also used three-dimensional data for features, over the typical two-dimensional feature system. They also concluded that for continuous sign recognition, larger training sets were required.

Pashaloudi and Margaritis (Pashaloudi, Margaritis, 2002) achieved 85.7% recognition rate for continuous recognition of Greek Sign Language sentences. They used a 26 Greek words vocabulary, amongst them, nouns, pronouns, adjectives and verbs. Again, this work concluded that their training was insufficient and gave low recognition rates for the continuous method.

Despite the good results of HMM for isolated recognition, the HMM method by itself is not able to produce good results in continuous recognition due to the ME problem. Another problem of the HMM methods is the scalability. As the word count in the vocabulary increases, both combinations number and learning data for each sign is needed (in order to differentiate similar signs).

(Fang, Gao, 2002) aid the typical HMM system with an improved Simple Recurrent Network (SRN) to segment the continuous Chinese Sign Language (CSL). Up until this work, a signer-independent SLR for continuous recognition was inexistent. This work demonstrated the use of HMM aided with other methods, in this case SRN, could be implemented to solve some SLR problems.

A novel approach was presented in (Bowden et al., 2004) using Markov chains combined with Independent Component Analysis (ICA). In the first stage of classification, a high level description of hand shape and motion was extracted and then “fed” to the combination of methods previously mentioned. This procedure tried to work one of the bigger problems of the HMM methods, the huge amount of training data needed for good results. Due to the generalisation of features, and therefore the simplification in training, chains could be trained with new signs “on the fly” with immediate classification. The true important achievement with this work was the ability to produce high classification results on ‘one shot’ training and to demonstrate real time training on one individual with successful classification performed on a different individual performing the same signs.

Other machine learning techniques were used that were not based in HMM models. An example of such different techniques is the work of Capilla in (Capilla, 2012), which using Kinect in the data collection part and Nearest Neighbour - Dynamic Time Warping algorithm, achieved an accuracy of 95 percent for a vocabulary of 14 homemade signs. Also using Kinect, for its obvious advantage of getting depth information, is the work of (Almeida, 2011), which implemented 3D Path Analysis for isolated SR problem, achieving perfect recognition rates for the 10 word dictionary used.

More recent approaches, with importance to the ME problem in the SLR was the work of (Yang et al., 2010) and (Chai et al., 2013).

Yang et al. (Yang et al., 2010) developed an approach based in dynamic programming-based matching, as it does not place demands on the training data as much as probabilistic models such as HMM do. With this method they also allowed the incorporation of grammar models. They compared the performance of their method with Conditional Random Fields (CRF) and Latent Dynamic-CRF- (LDCRF) based approaches. The results showed more than 40 percent improvement over CRF and LDCRF approaches in terms of frame labelling rate. They also got a 70 percent improvement in sign recognition rate over the unenhanced DP matching algorithm that did not accommodate the ME effect.

By using 3D Motion Trajectory Matching with 3D data from the Kinect, (Chai et al., 2013) achieved recognition rates up to 96 percent in a 239 word dictionary containing CSL words. In their database, each word was recorded 5 times.

2.4 Summary

In this section, we first introduced some of the more relevant and used data collection methods used in SLR from *DataGloves* to simple Video Cameras, ending with the chosen method to be used in this project, the Kinect.

After explaining the data collection, the most commonly used machine learning techniques in SLR, with special attention and focus to the HMM usage, since it is one of the most used methods in AGR.

3 PhySaLiS - Portuguese Sign Language Recognition System

In this chapter, we present the developed system, referred to as **PhySaLiS** (Portuguese Sign Language Recognition System). The system requirements derived from the objectives are enunciated in the first section, followed by the description of the system architecture and of the application GUI, in the second section. Section 3.3 describes the Corpus used for both automatic postures recognition and automatic signs recognition systems. This section also specifies the data collection, such as number of signers, repetition of words, etc. Section 3.4 describes the pre-processing done to the raw data before feature extraction is performed. Section 3.5 explains the methods used for feature extraction of hands configurations and hand movement data and, finally, section 3.6 details the classifier creation methods and its specifications.

3.1 System Requirements

To achieve the objectives purposed for this thesis, multiple requirements were derived from each of one. Both system requirements (SR) and GUI requirements (GR) are presented in Table 2.

Table 2 – System Requirements definition. SR and GR codes correspond to System Requirements and GUI Requirements accordingly. For each requirement the requirement id, description, objective and status is shown.

Requirement	Description	Objective	Status
SR1	Develop a technique to extract and normalize hands information from the depth image to be used for the automatic hand posture recognition system	O1,O3	Completed
SR2	Develop/Apply a technique to extract and normalize hands information from the joints information to be used for the automatic sign recognition system	O2,O3	Completed
SR3	Collect hand configurations performed by multiple users, using depth data, allowing the	O1, O3	Completed

	creation of a structured dataset for training and testing		
SR4	Collect signs performed by multiple users, using depth data, allowing the creation of a structured dataset for training and testing	O1,O2	Completed
SR5	Develop a classification technique appropriate for the automatic sign recognition task	O2, O3	Completed
SR6	Develop a classification technique appropriate for the automatic hand posture recognition task	O1, O3	Completed
SR7	Use a previously collected dataset to train and test the hand posture classification system	O1, O3	Completed
SR8	Use a previously collected dataset to train and test the sign classification system	O2, O3	Completed
GR1	Allow sign recording for multiple distinct signers	O4	Completed
GR2	Create tool to calibrate arm size for the signer	O4	Completed
GR3	Start and stop sign recording, manually or automatically	O4	Completed
GR4	Start and stop posture recording.	O4	Completed
GR5	Load previously saved recordings, either for postures or signs systems	O4	Completed
GR6	Present the classification for both hand configurations for each frame in the recorded signs	O4	Completed
GR7	Allow tuning of parameters for the classifier creation, such as kernel to use and kernel parameters, as well as type data to be used for the classifiers	O4	Not Completed

3.2 System Architecture and GUI

As previously mentioned, this thesis is focused on tackling a specific problem of recognising some signs that observe only the manual features of the Stokoe model, and to enable that, a two layered system architecture is proposed.

In the diagram of Figure 5 – Automatic Sign Recognition System Architecture, the first layer represents the **Posture system** (depicted in teal colours), while the second layer represents the **Sign system** (depicted in reddish tones) which depends on the first layer.

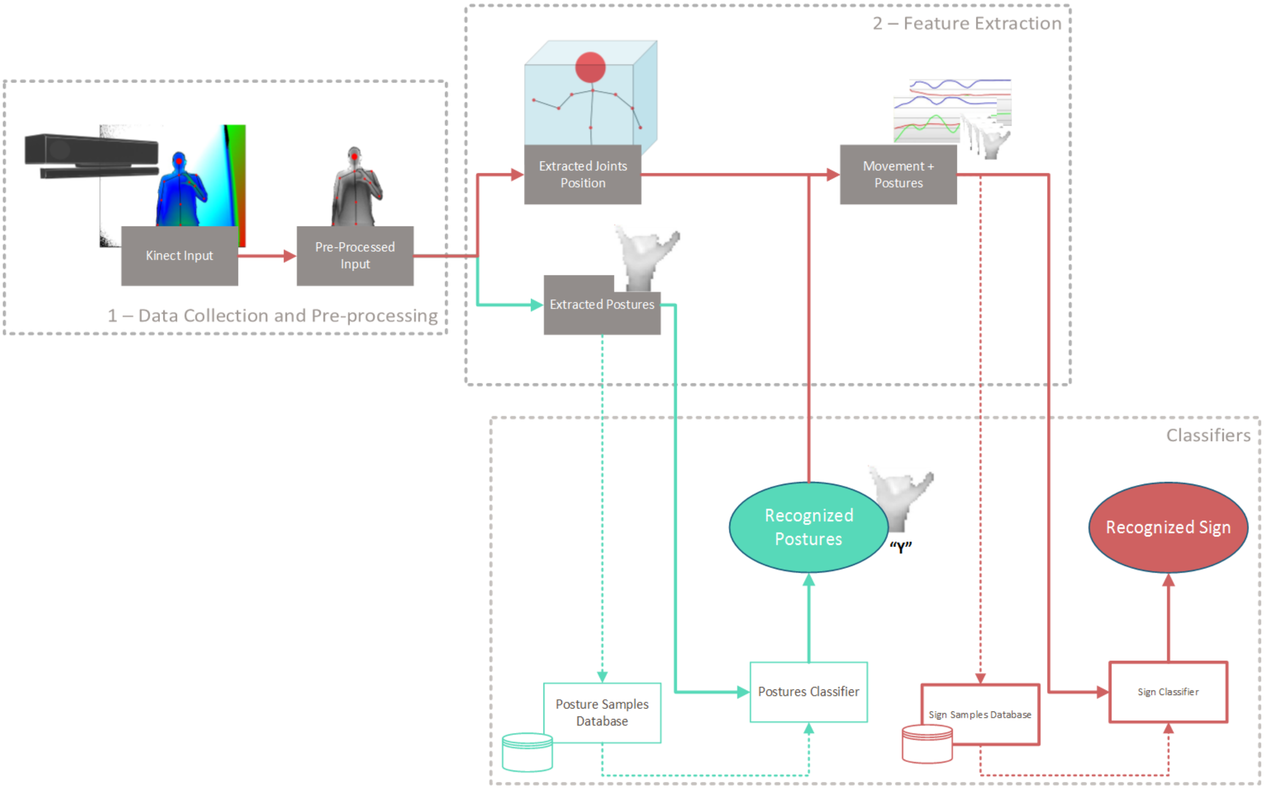


Figure 5 – Automatic Sign Recognition System Architecture. The first module comprises the data collection and pre processing methods. The second one handles Feature Extraction, while the last one includes the training and testing o the classifiers.

These layers are divided in three modules, each of those represented by a dashed container. The independent Module 1 performs signal acquisition and pre-processing of the Kinect One input streams, namely, depth frames and "tracked body" data. The second module handles feature extraction. Finally, the third module produces and stores training data samples for the classifier and handles the hand posture or sign recognition process.

The Pre-processing module takes care of the depth and joint data collection process and some pre-processing before passing the data to the Feature Extraction module. It provides functions to estimate *global speed* of movements as well as determining when a movement starts or ends, while also allowing the calibration of the system to a particular signer, tuning parameters like the hand size, the arm size, or the application of Erode and Dilate filters to the depth stream (for noisier environments).

The second module, “**Feature Extraction”,** performs feature extraction and processing. This feature extraction will provide information for real-time automatic posture or sign recognition and provides the training samples set, a critical data source for the gesture recognition (GR) process, which will be stored in a database.

Like in (Almeida, 2011) **postures** refer to static hand configurations and **signs** to complete hand gesture representations. Both of these representations require a data structure and a set of methods to manipulate and provide the tools for the feature extraction process.

The module also comprises a set of posture acquiring functions, namely, to get and convert posture dimensions, to get posture images from the respective joints (using the joint 3D position and the Depth Data frame), and also more general methods to perform adjustments, such as resizing, several types of translations on different axes as well as scaling and ratio transformation, and enabling *dexel*[1] data normalization.

Finally, the **“Classifiers**” module handles the management of the collected data and drives the system classifiers. It allows the training of Support Vector Machines classifiers and the viewing and deleting of postures and signs from the training sets.

The module also implements the creation of bitmap images from *postures* and data charts from the hands movements for the signs**,** to aid in later analysis for automatic GR performance, reliability and accuracy.

The recognition task, which is the usage of methods from the modules 1 and 2 and the classifiers created in this same module, output a sign, or a hand configuration label.

[1] Dexel is a concept introduced in (Almeida, 2011) and refers to “Depth picture element” or “Dexel”

3.2.1 GUI - PhySaLiS Application

PhySaLiS is the software application developed through this thesis. It got its name for being the fruit of hard work and for enclosing the acronym PSLS – PSL System, that is one of the objectives of this work.

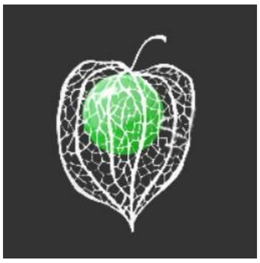
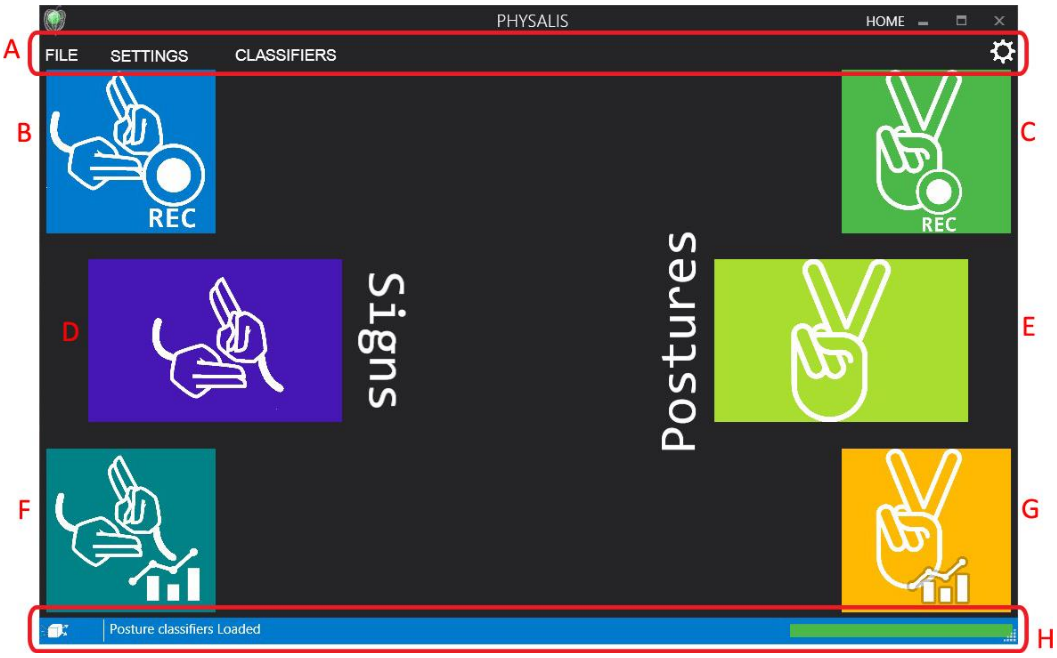


Figure 6 – PhySaLiS logo

The berry in the logo is green instead of the most commonly known images of the *physalis* fruit, which usually is red or orange. This is due to the application being able to grow and mature to produce a better system able to aid the sign language community, particularly, the PSL community.

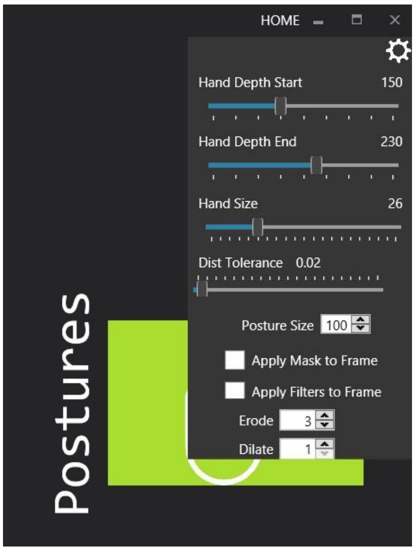
The application was developed in the *Metro* style, and is composed by simple lines and effects. The "Home" screen looks as follows:



Posture classifiers Loaded

Figure 7 – PhySaLiS home screen. On the left we have buttons for the signs functionalities, and on the right, buttons for hand postures.

In the Home screen it is possible to go to any of the major functionalities of the system. These can be divided in 2 major groups: Signs (left side) and Postures (right side) – labels in red. For each of those groups it is possible to do **Data Recording** (B and C), **Data Analysis** (F and G) and **Recognition** (D for **Postures** and E for **Signs**).



Hand Depth Start 150

Hand Depth End 230

Hand Size 26

Dist Tolerance 0.02

Posture Size 100

□ Apply Mask to Frame

□ Apply Filters to Frame

Erode 3

Dilate 1

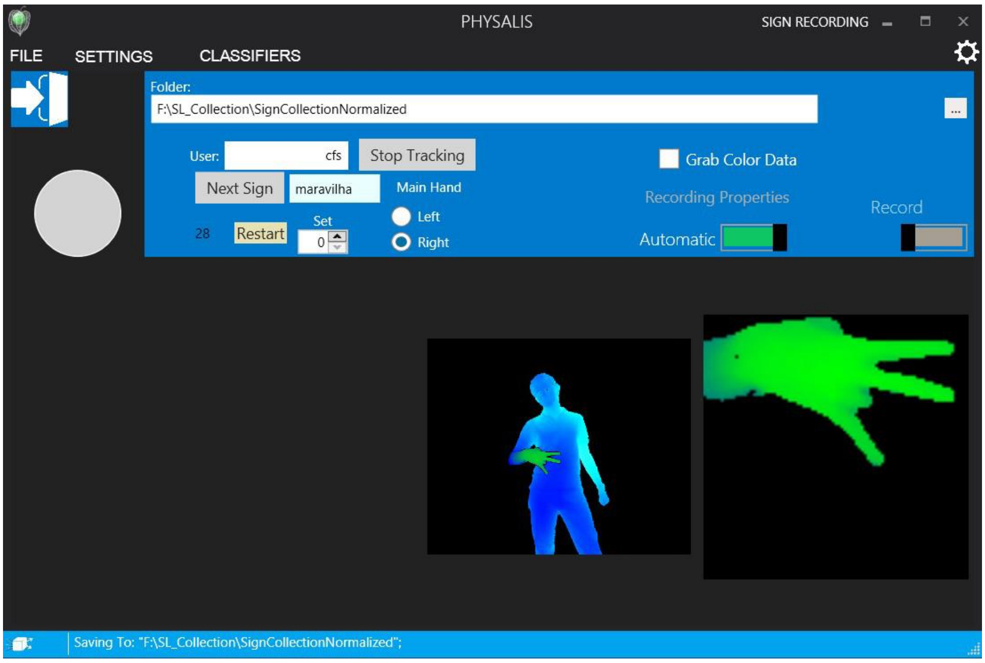
Figure 8 –Data collection options. These options configure the normalizations performed on both postures and signs systems. The hand depth start, end hand size define the bounding box that extracts the hand depth image. The distance Tolerance detects if the signer is moving or not.

Menu bar A in the main screen, represents the header bar which contains the features to load data of both for signs and postures recognition subsystems, change settings for the application and to load and create classifiers. Loaded classifiers will be used in screen areas D and E. The loaded data can be used either to create classifiers or to be subject of analysis (features F and G). It also contains the dropdown container with options concerning the Data Collection feature. This options include: control of the volume size that extracts the hand depth image; the movement tolerance for the sign system; the size of the depth image of the hand (“*posture size”*), the size of the erode and dilate filters to be applied for background extraction and the option to apply or not the background extraction. Throughout this menu is also possible to access the tool to calibrate the signer arm size.

In the bottom of the application (label H), there is a status bar that shows information on the operations performed by the system and other status messages (Figure 7).

3.2.1.1 Data Recording Screens

The data recordings for both signs and postures are done accessing B and D buttons, respectively (Figure 7).



Folder: F:\SL_Collection\SignCollectionNormalized

User: cfs

maravilha

28

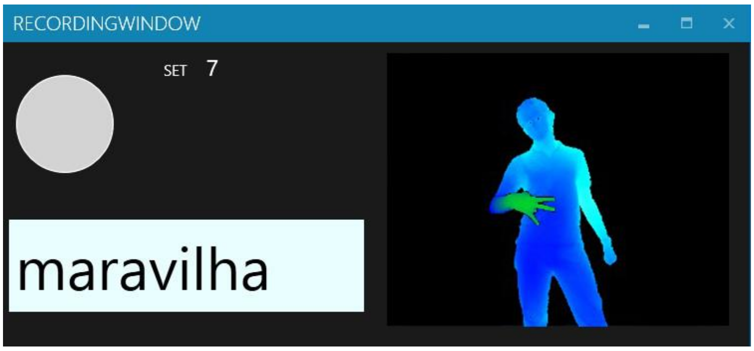
Set 0

Saving To: "F:\SL_Collection\SignCollectionNormalized";

Figure 9 – PhySaLiS sign recording screen. In this screen the user can create a new folder for a new signer, which will act as a repository of the recordings of new instances of signs. The user can define which hand behaves as the main hand, the sign to be recorded and if the sign is to be recorded automatically or manually via a start/stop button.

In B - sign recording screen (Figure 9) - it is possible to give an alias to the signer and to create a folder with that name, where to save the sign (each signer data goes into distinct folders). It is also possible to record the signs either automatically or manually. In automatic recording, the “start recording” and “stop recording” signals are given by estimation of the movement of both hand joints. In this method, it is required that the signer is in a standing pose (Figure 16) before the activation of the start signal. This way all signs recordings start from the same place and pose of the signer. For the manual mode, a simple recording button is available, so the user can start and stop the recording at any time.

The user needs also to associate a word to the sign being recorded, so that the recording has the right label for the classifiers. A “*helper*” window (Figure 10), helps in the recording process. It contains simple information, suhc as the sign to perform and its instance number, a visual feedback indicating weather the system is recording, ready to record or idle and a visualization of the depth data stream.

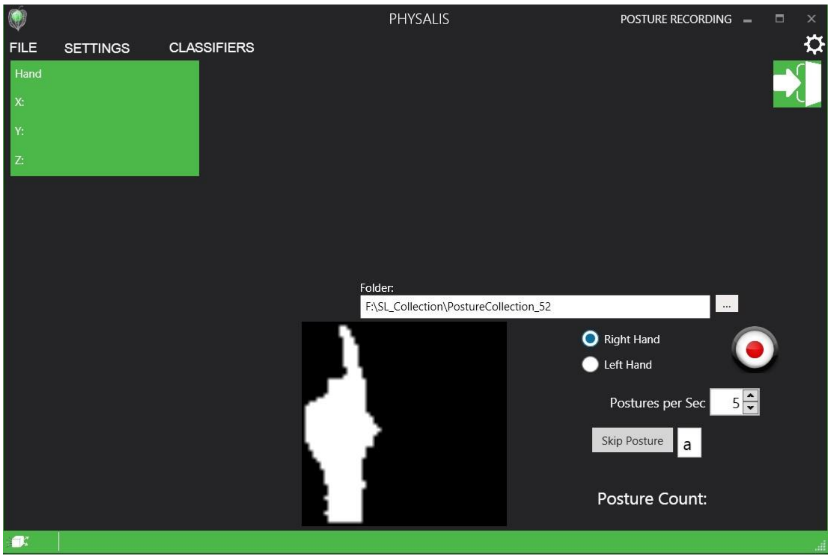


SET 7

maravilha

Figure 10 – Sign recording helper window. The purpose of this window is to help the signer to know what sign to perform and when to perform it.

For D – posture recording screen (Figure 11) – much like in the previously depicted screen, it is possible to select the folder in which the postures are going to be stored, as well as to select which posture to record and how many captures per second are required. This way it is possible to save multiple instances of the



Folder: F:\SL_Collection\PostureCollection_52

● Right Hand
○ Left Hand

Postures per Sec 5

a

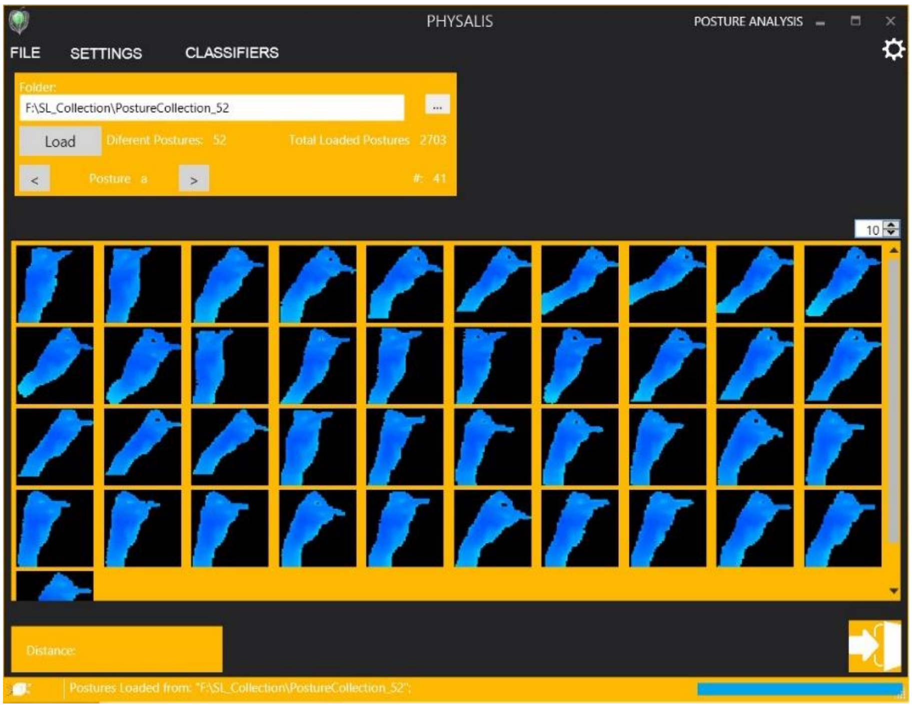
Posture Count:

Figure 11 – PhySaLiS posture recording screen. In this screen the user can select which hand to track for the recording, select which posture to record and how many images per second need to be captured.

posture in an efficient way (with small variations or not). By default, the right hand is recorded, but it is possible to choose which hand to record, either left or right.

3.2.1.2 Data Analysis Screens

The F and G screen areas (Figure 7) give the user access to both Postures and Signs analysis features, respectively. These screens serve the purpose of inspecting the corresponding recorded data sets for both signs and postures, as the names suggest. In both, is possible to select the folder where to fetch the datasets, to inspect each instance of each class of the dataset, and to navigate through classes and, for the sign analysis, through users.



Folder: F:\SL_Collection\PostureCollection_52

Diferent Postures: 52

Total Loaded Postures 2703

Posture a

#: 41

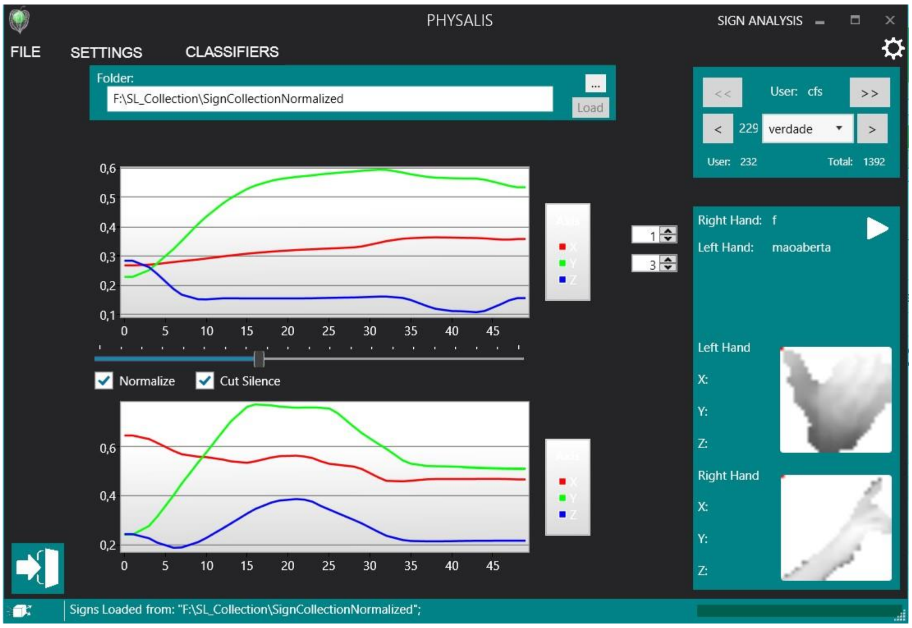
10

Distance:

Postures Loaded from: "F:\SL_Collection\PostureCollection_52";

Figure 12 – PhySaLiS posture analysis screen. In this screen it is possible to visualize the recorded postures as well as to eliminate selected posture instances.

In the Postures Analysis screen (Figure 12), each instance corresponds to a depth image of the hand and each class is composed by multiple images. In the case of Signs, each of the instances is composed by two movement plots and two normalized depth images per frame, which correspond to both hands. It is possible to navigate throughout each of the frames composing the movement.



Folder: F:\SL_Collection\SignCollectionNormalized

User: cfs

229 verdade

User: 232 Total: 1392

Right Hand: f

Left Hand: maoaberta

☑ Normalize ☑ Cut Silence

Signs Loaded from: "F:\SL_Collection\SignCollectionNormalized";

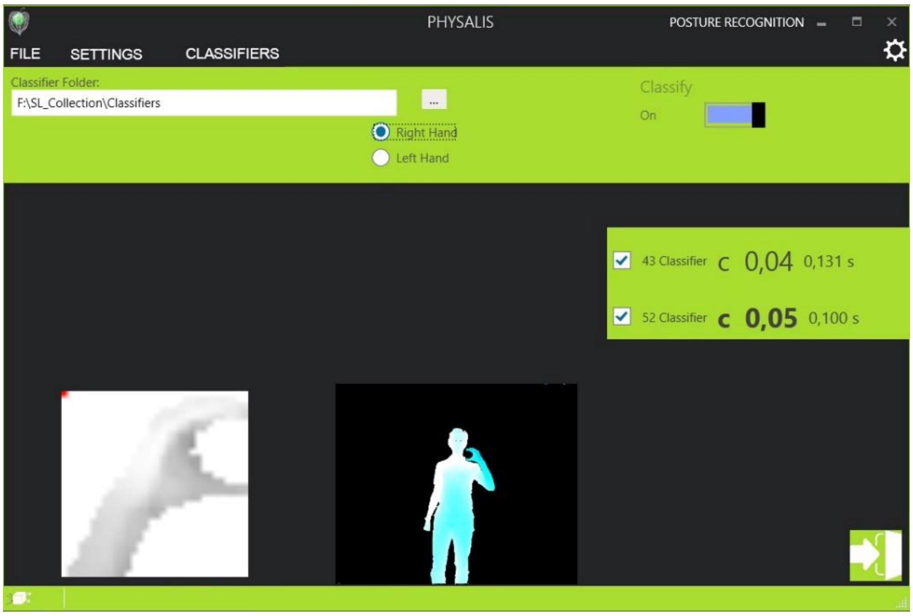
Figure 13 – PhySaLiS sign analysis screen. This screen gives visual representation of the acquired and processed sign data. It is also possible to view any sign, for any recorded user.

In the movement analysis screen (Figure 13), it is possible to navigate between all the instances of the signs previously recorded and loaded in the system. It is also possible to observe the collected movements of both hands. In each frame, and for each of the two movements, de system depicts the label of each hand, classified with our 43 Postures classifier. The user can view the movement data for both hands, with or without applying the normalization or cut silence methods.

3.2.1.3 *Recognition Screens*

The recognition screens, accessible through areas D and E of the home screen (Figure 7), lets the user to practically test the classifiers developed with the system, in real time.

For the Posture Recognition Screen (Figure 14), the user can choose which hand to track and recognize and both posture classifiers are used (namely, the 43 class classifier and the 52 class classifier). The result shown in the classification is the hand configuration in the spelled alphabet in PSL, that most occurred in the last 10 recorded frames.



Classifier Folder: F:\SL_Collection\Classifiers

● Right Hand
○ Left Hand

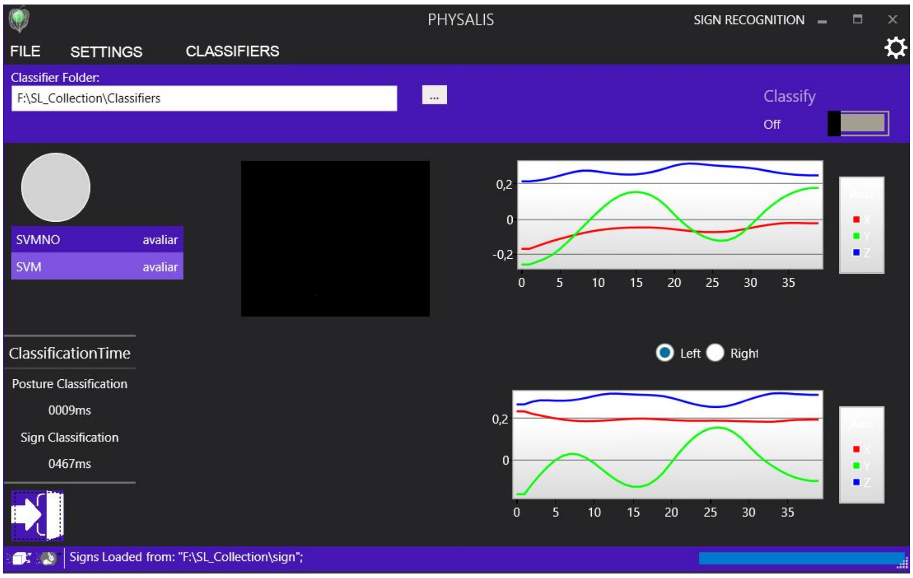
Classify
On

☑ 43 Classifier C 0,04 0,131 s

☑ 52 Classifier C 0,05 0,100 s

Figure 14 – PhySaLiS posture recognition screen. In this screen is possible to perform hand postures in real-time with either left or right hand and to experiment the available posture classifiers (43classifier and 52 classifier).

Similarly, in the Sign Recognition Screen (Figure 15), the user can choose the main hand of the signer and two SVM classifiers are used to recognize the sign performed by the signer. It is also showed the time each of the classifiers took to recognize the sign.



Classifier Folder: F:\SL_Collection\Classifiers

Classify Off

SVMNO avaliar

SVM avaliar

ClassificationTime

Posture Classification 0009ms

Sign Classification 0467ms

Signs Loaded from: "F:\SL_Collection\sign";

Figure 15 – PhySaLiS sign recognition screen. In this screen is possible to perform signs with either left or right hand as the main hand and to experiment the two SVM sign classifiers developed with the system. At the right it is possible to observe the movement chart of the performed sign.

3.3 Data Collection

3.3.1 Setup

To collect the gesture data for this thesis, we have used a setup with the Kinect One sensor, with the following requirements:

1. No direct sunlight in the room in which the recordings are to take place. This is needed so the Kinect depth information may work with the less noise possible (Andersen et al., 2012);
2. The sensor is at about 1.3 meters from the ground and placed on a stable and horizontal surface;
3. The signer/user is between 1.5 and 3.5 meters away from the sensor, facing it with a standing and frontal pose (Figure 16);
4. No object is between the signer and the sensor.
5. Other sources of infrared light should not be present in the room since they may produce artefacts in the depth image;

Other constraints such as specific artificial illumination are not an issue since the only information used is from the depth stream, which works on infrared light.

The system requires calibration regarding the signer size, task performed by a simple tool developed for that purpose, which, by capturing the signer in a standing and frontal pose, could estimated the arm length. After having the arm span information, the collection can proceed by simply making the selected sign. All signs start to be recorded automatically from the standing pose (Figure 16), where the start and stop signals are computed by estimations of the amount of movement.

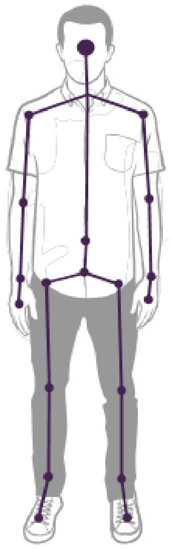


Figure 16 – Example of a standing pose (Microsoft, 2014b) *to initiate data collection. The back and legs stay straight, both arms fall along the torso*

3.3.2 Corpora

An important aspect for any study involving human participants is to obtain the approval of a regulated ethics committee. In this thesis, the carried experimental activities, included detailed descriptions of the data collections and associated studies/experiments, which were submitted to the ethics committee of Instituto Universitário de Lisboa ISCTE-IUL, regulated by the dispatch nº7095/2011.[JF1] and approved. In the case of the experiments described in this thesis, all data collections participants gave informed written consent and were properly informed of the purpose of the experiment, its main characteristics and that they could quit at any time.

Signs

The Corpus, or vocabulary, used in our system concerning **signs,** was chosen to better address and better illustrate the problem addressed in this thesis. It was not chosen by the meaning of the word in Portuguese, but by the sign properties in PSL that represented the word. The criteria used in the selection of 29 signs in PSL, were:

- Concerning individual signs :

   A. The “auxiliary” hand behaves similarly to the main hand (like a mirror);

B. Only the main hand moves (sign performed with only one hand);

C. The “auxiliary” hand acts as a support for the main hand.

- Across Signs:
   1. Signs with the same or similar movement, but different configurations for the hand(s);
   2. Signs with the same posture but different movements of the hands;
   3. Signs with the same hand configuration and the same movement but different locations in relation to the human body, i.e. According to (Bela Baltazar, 2010) the sign “*mesa*” is described as “Both hands in configuration ‘*zeta’*, **positioned bellow the chest,** start together at the centre and move apart to the sides” as for the sign “*balcão*”, its definition is precisely “gesture identical to ‘mesa’ **but done higher** (above the chest)”.

The criteria A, B and C (Moita et al., 2011), are observed in every sign, being those properties a way to model sings. Properties 1, 2 and 3, describe some type of signs that observe the problem that this thesis addresses.

The signs used are depicted in table 3.

Table 3 – List of signs selected to compose the train and test corpus. Each group of colour stacked together represent signs in which the path of the hands are the same, or are very similar. i.e. the signs “balcão” and “mesa” have the exact same movement but in distinct positions: the first one is done above the chest, while the second one, under the chest.

Apagar (*to erase*) Escrever (*to write*) Graxa (*shoe polish*)
Balança (*scales*) Avaliar (*to evaluate*)
Discutir (*to discuss*) Guerra (*war*)
Gritar (*to scream*) Cantar (*to sing*)

Eclipse (*eclipse*) Morrer (*to die*)
Fio (*wire*) Tubo (fino) (*thin pipe*) Tubo (médio) (*medium pipe*)
Balcão (*counter*) Mesa (*table*)
Testemunha (*witness*) Verdade (*truth*)

Apoio (*support*)
Cadeira (*chair*)
Quente (*hot*)
Maravilha (*wonderfull*)
Ajudar (*to help*)
Receber (*to welcome*)
Comunicar (*to communicate*)
Trabalhar (*to work*)
Não (*no*) Televisao (*television*)

Nadar (*to swim*)

The signers were previously briefed about the data collection method and no signer had any background on PSL. Each signer had to learn each of the performed signs prior to the data collection.

The group of signers were composed by 5 men and 1 women, with ages between 23 and 31 with an average of 26 and heights between 1,52m and 1.95m. For each signer, eight repetitions of each of the 29 signs were collected.

Postures

For the hand configuration recognition sub-system, two different data collections and consequently two classifiers were developed. For the first recordings, 52 different hand postures were recorded from 2 signers. Each of those postures were recorded around 50 times with some small variations, such as small hand orientation or fingers angles. Those variations were minimal in order not to change the posture. Basically, the produced hand postures were in the form of finger spelling (alphabet spelling in PSL), with the configuration clearly facing the sensor. This first experiment was conducted only to test the method used to extract the features and to have a comparison with previous works in PSL recognition (Almeida 2011).

The second data collection, for the purpose of hand configuration recognition sub-system, was also exploited for PSL sign recognition and was composed by 43 postures. For these recordings, 2 signers were used, and about 350 instances for each posture were recorded. This time, the postures could change its orientation to better accommodate what happens in real sign production. One example of this property is the word "*abdicar" – abdicate* (Figure 17)*.*

While the description for the word is *“Dominant Hand in configuration ‘b’ passes along the non-dominant hand in configuration ‘1’”* (Bela Baltazar, 2010), the hand configuration observed can be described as ‘q’.

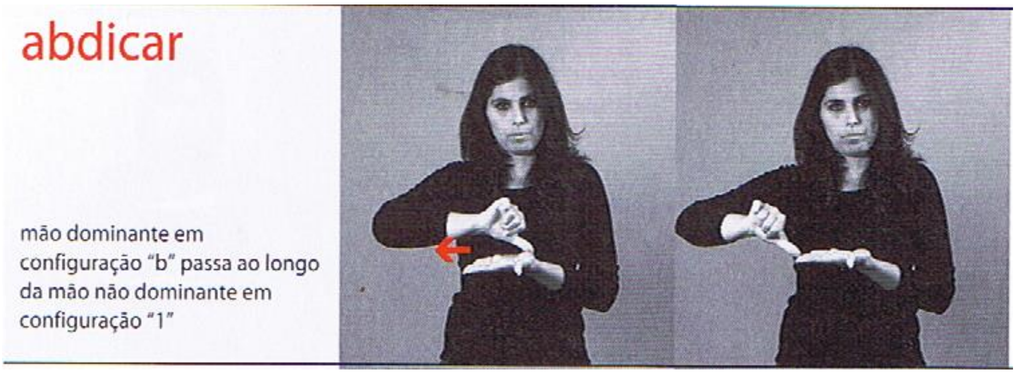

abdicar
mão dominante em configuração “b” passa ao longo da mão não dominante em configuração “1”

Figure 17 – Description of the sign "abdicar" (Bela Baltazar, 2010). The description of the sign “abdicar” in PSL is: “Dominant Hand in configuration ‘b’ passes along the non-dominant hand in configuration ‘1’”

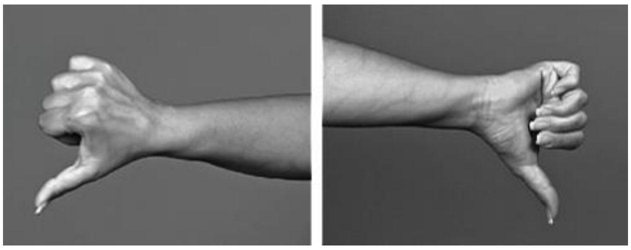


Figure 18 - Hand configuration “q” (Bela Baltazar, 2010). *On the left is the view of the signer, and on the left is the view of the “receiver”*

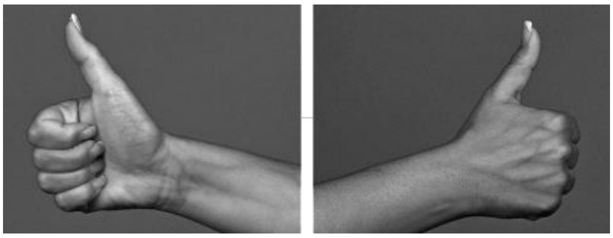


Figure 19 - Hand configuration "b" (Bela Baltazar, 2010) *. On the left is the view of the signer, and on the left is the view of the “receiver”*

For both signs and hand configurations data collections, the pre-processing conditions described in the next section, apply.

3.4 Pre-processing

In our set-up he data is acquired via a Kinect One Sensor. From this sensor, only the output depth stream and the derived body stream, are used in our computer vision system.

Generally, one of the first critical steps in computer vision systems is the background removal. In our system, we require such background removal operation, to properly segment user, or signer.

3.4.1 Background Removal

Consider $D_1, D_2, \ldots, D_{30}$ as a sequence of depth images with $512\, x\, 424$ pixel values (16 bits), for the time instants $1, 2 \ldots 30,$ respectively and, DM the target background removal Depth Mask, to be created. As a reminder, the depth of each pixel is coded in its 16 bit value, and is computed by the sensor, as a depth value in the sensor reference frame.

- In the first instant we have ($n = 1$):

$$DM = D_1$$

- For the instants *n* such as $1 < n < 30$:
  - Let D_n be the depth map for the instant *n*.
  - Consider the depth values $p(x, y)$, where $x \in \{0..512\}$ and $y \in \{0..512\}$ from D_n and DM represented by dn_{xy} and md_{xy} respectively.
  - $If\ dn_{xy} < dm_{xy}\ then\ dm_{xy} = dn_{xy}$

At the end of the 30 instants, DM will be the depth image containing the minimum (or closest) depth values observed during those frames. We assume that the sensor only “sees” the scenario, which means that the user needs to be away from the sensor field of vision in the first 30 frames.

- For the instants n such as $n > 30$:
  - Let $\lambda = 50$ be a sensor noise tolerance factor that represents 5 cm.
  - Considering D_n as being the depth map for the instant $n > 30$.
  - $If\ d_{xy} \geq (dm_{xy} - \lambda)\ then\ d_{xy} = 0$

This is the same process as applying a mask, where values from D_n that are higher than the values on the Depth Mask image - λ will be eliminated.

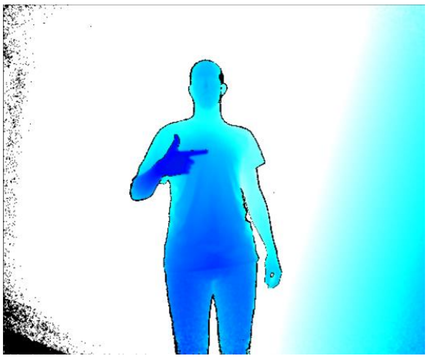


Figure 20 – Kinect sensor original depth input. The colour in the depth images is simply a form of representation: the depth images contain one value per pixel, corresponding to the depth in millimetres of that point in space, in the Kinect sensor reference frame.

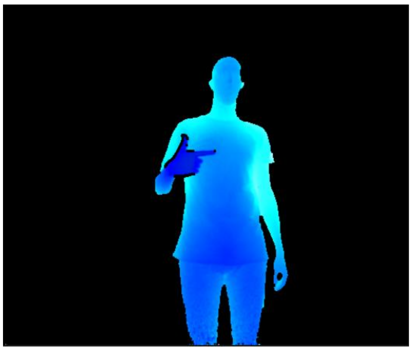


Figure 21 – Kinect sensor depth input after application of the background subtraction process. With background subtraction - the body silhouette becomes the only visible object, as it is the only object that moves in the Kinect sensor field of view.

After the background is subtracted, an erode filter is applied in order to remove some noise introduced in the depth image by sunlight, since the infrared sunlight reflected by other objects or refracted in the room windows, might “damage” the Kinect depth image.

3.4.2 Hands Segmentation

The next step before the feature extraction process is to extract the hands region, from the depth image.

For this process, both the depth image and the hands position are needed, where the hands position are given by the Kinect sensor SDK, in the Body structure, and the depth image is the result of the previous step, Background Removal.

By using a fixed size and a fixed depth, it is possible to extract the hand depth image from the global image doing simple math and making use of the coordinates mapping from the Kinect SDK, that allows a conversion from the real space (Kinect camera reference frame) to the depth space (depth image reference frame), mapping camera coordinates into pixels in the depth image.

3.5 Feature Extraction

The features to train and test the SVM classifier, are computed from the 3D coordinates (joint position) of the hands and from the depth images of both hands.

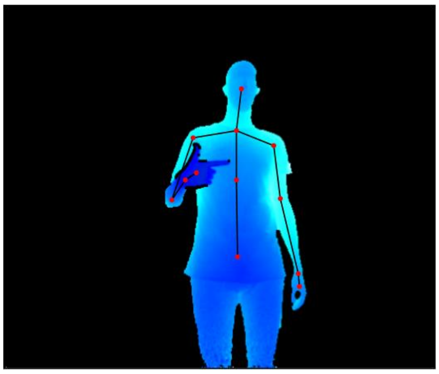


Figure 22 – Both depth and body inputs from Kinect in one frame. The red dots are the joints given by the Kinect SDK. Each joint is represented by a point in 3D space.

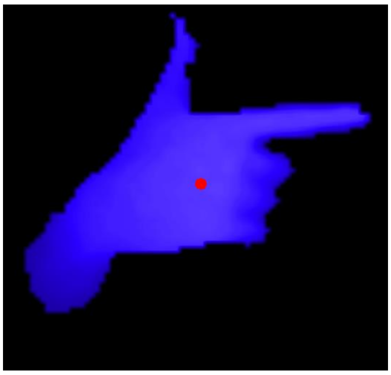


Figure 23 – Extracted hand depth image width the hand center joint depicted as a red dot

In our framework, a sign is discretized in several consecutive frames, where, in each frame, we have both hands positions and both hands depth images. To compute the features of the sign, we need to perform several normalization procedures, before passing such feature data onto the classifier, whether for training or recognizing (testing) purposes.

These normalization procedures occur in two forms:

- Normalization of both hands positions, for all frames of the production of the sign, which results in **hand path normalization**.
- Normalization of both hands depth images, for all frames of the production of the sign, which results in **hand depth images normalization.** This represents the hands configurations for each frame and will be used in a first classifier, resulting in a **posture label** per frame and for each hand,

3.5.1 Hand Path Normalization

The Hand Path is also the moving part, or the gesture part, of the sign structure. The position values given by the Kinect sensor for each of the hands joints, have as a centre of the reference frame, the body **Spine Centre** joint depicted in (Figure 24).

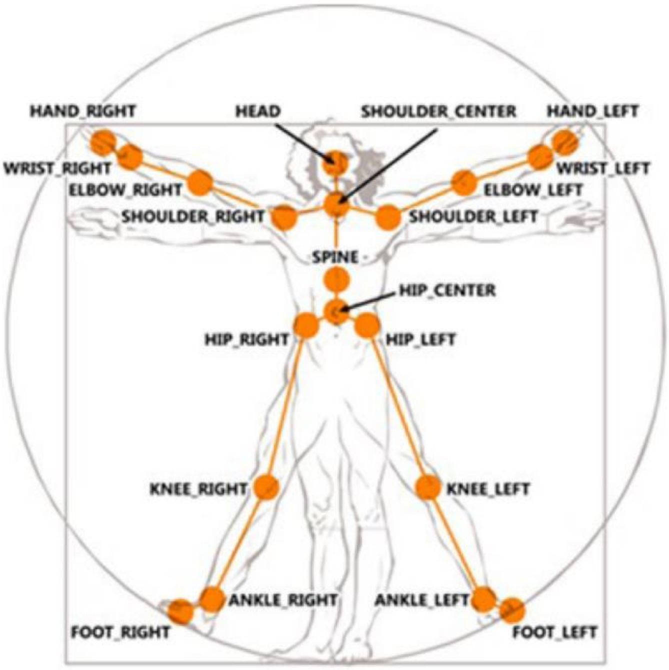


Figure 24 – Kinect body joints. The Spine Centre joint used in the normalization method is the joint above the “HIP_CENTER” and below the “SHOULDER_CENTER”

Captured at 30 frames per second, the raw data resulting from recording both hands during a sign, is represented in Figure 25.

In this raw data, it is possible to see that the only the X and Y coordinates of each hand centre joint, are defined in the spine centre joint reference frame. This brings an obvious problem, that is, if from recording to recording, the signer is at varying distances from the sensor, the hand centre joint Z value will vary greatly. In order to represent the (X, Y, Z) coordinates of the hands centre joints in the same reference frame, and to be able to recognize signs further away or closer to the sensor, the Z coordinate for each hand is normalized according to the Body Absolute Position, which is the same as the **Spine Centre** joint.

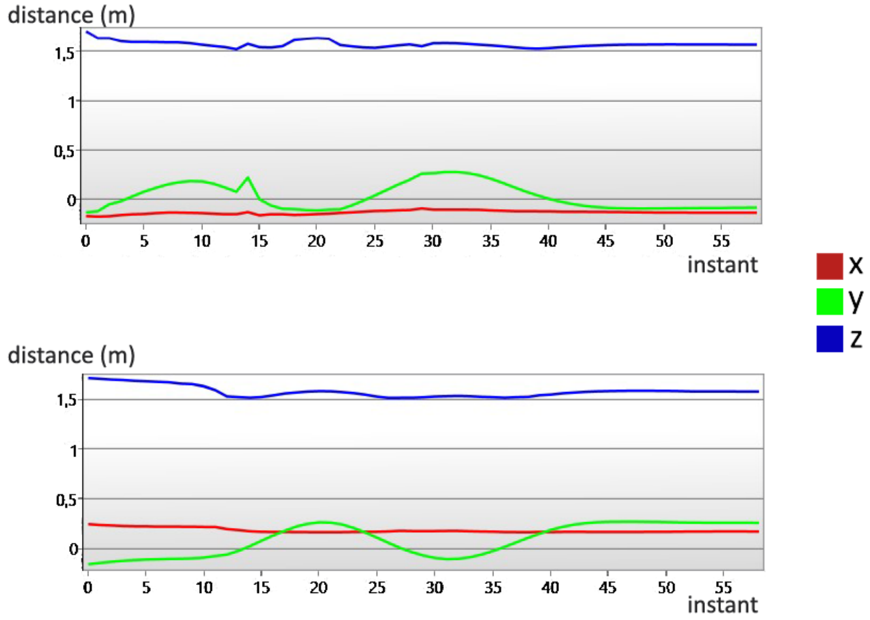


Figure 25 – Raw hand centre joint (X, Y, Z) coordinate data, of an instance of the sign "avaliar”. The top chart corresponds to the left hand path while the bottom one is the right hand path chart.

Let $H_{(x,y,z)}$ be any of the hands centre joints points given by the sensor and let $SpineC_{(x,y,z)}$ be the Spine Centre Point. The first normalization transformation for both hands is:

$$H = \begin{cases} H_x = H_x \\ H_y = H_y \\ H_z = SpineC_z - H_z \end{cases}$$

This way, the problem of the signer distance to the sensor is eased

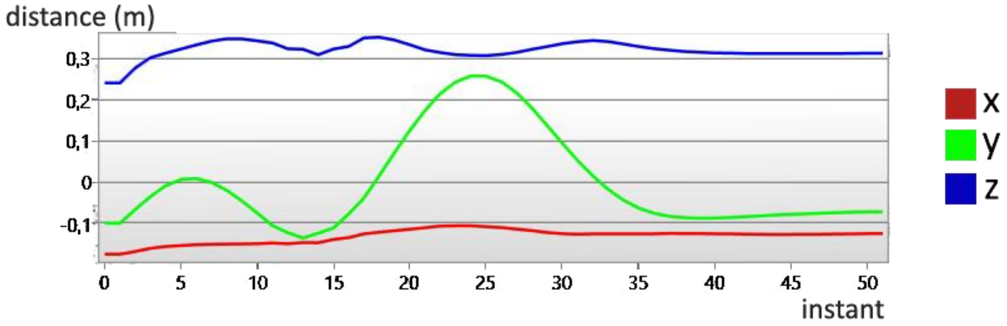


Figure 26 – Hand centre joint (X, Y , Z) coordinate data of the same instance of the sign "avaliar” shown in Figure 25 after the first step of normalization. Only the left hand chart is shown.

After the classifier was trained and tested with this data normalization approach, we´ve noted that signers with distinct heights, hence distinct arm span, showed distinct results. As the previous normalization step did nothing to solve this problem, another method was needed. This issue was addressed by warping the hands position space to a predefined value according to each signer **arm size.**

The signer arm size is estimated with a method that takes the joints from the hand to the shoulder and calculates the distance between such joints.

This is the same as creating a virtual box around the signer, which varies with the signer arm span, and for that, we need first to define the boundaries of said box, in each frame:

$$Min_x = SpineC_x - AS\ , Max_x = SpineC_x + AS$$
$$Min_y = SpineC_y - AS\ , Max_y = SpineC_y + AS$$
$$Min_z = AS\ , \qquad Max_z = -AS$$

Where AS is the Arm Size value

Having the arm size, the new coordinates for any hand for each instance become:

$$H_\alpha = \frac{OldH_\alpha - Min_\alpha}{Max_\alpha - Min_\alpha}\ , where\ \alpha\ \in \{x, y, z\}$$

$where\ H_\alpha\ is\ the\ normalized\ joint\ coordinate\ \alpha\ and\ OldH_\alpha\ is\ the\ old\ one$

After this normalization, signers with distinct heights are less of a problem, since now, the hands positions along any movement with any signer are normalized to the same normalized space.

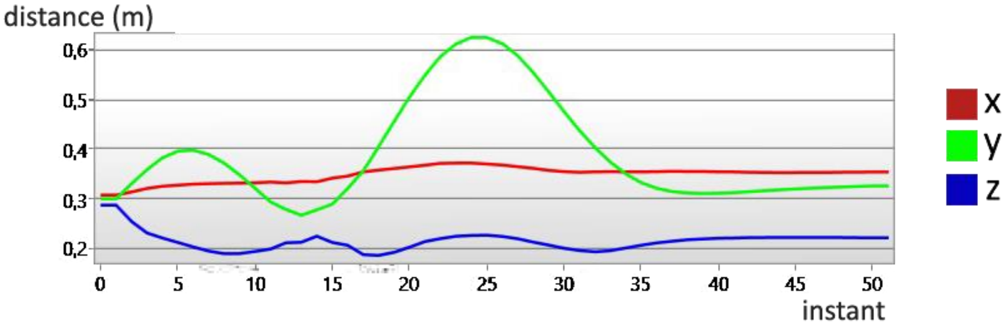


Figure 27 – Hand centre joint (X, Y, Z) coordinate data, of the same instance of the sign "avaliar" shown in Figure 25 after the second step of normalization. Only the left hand chart is shown.

The recording of the movement starts and stop automatically, with the start signal given by the start of movement and the end signal, by an estimation of movement. This technique is not too precise due to noise on the joint data, introduced by the sensor. The next step is then to remove the frames at the end of the recorded movement with irrelevant information, considered to be “silence”. To compensate for this noise, some frames of the end part of the movement need to be discarded. Starting from the end of the movement, 6 coordinates (3 coordinates from each hand) are observed in a sliding window of 3 frames. If all the coordinates in this 3 frame window have variations lower than a fixed threshold, then the center frame is eliminated from the movement. It is possible to see, comparing Figure 27 and Figure 28, that after this method is applied, all the frames in Figure 27 from the final to near the frame 40, were eliminated.

The last step to create the feature vector, is to normalize the hand movements’ size, that is, make all movements to have the same number of frames. Hands movements are described by an array of normalized coordinates (X, Y and Z) for each hand centre joint, with as many positions as frames that discretized the sign, giving a total of six arrays.

A recorded hand movement, after passing through the previous processes, has the following representation:

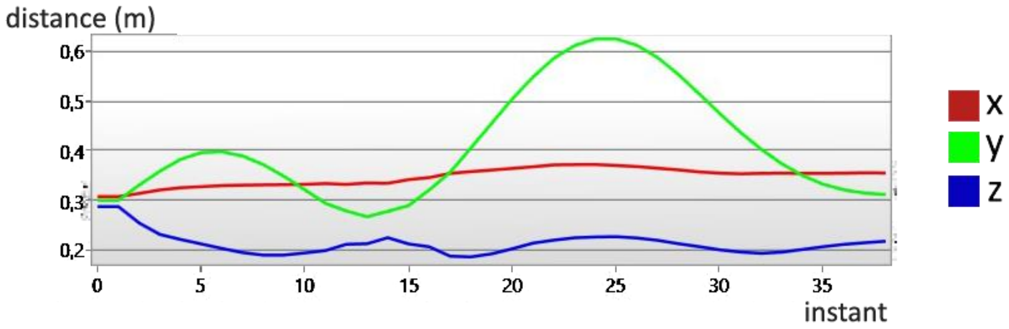


Figure 28 – Hand centre joint (X, Y, Z) coordinate data of the same instance of the sign "avaliar" shown in Figure 25 after the third step of normalization. This third step removed information at the end of the sign, in which the hands are halted, hence considered "silence". Only the left hand chart is shown.

In this case of (Figure 29) the recorded sign has about a little less than 40 frames, while in others cases might we might observe more or less, so we need to normalize all movements to the same number of frames, in order to work within the classifier.

When normalizing the movement, depending on its duration, one of two situations will occur. If the original frame size is bigger than the target size, or normalized size, the average values of the removed positions are used. When the frame size of the original is smaller than the target's, the inserted joint positions will have a value linearly interpolated with the previous and next positions. The inserted or removed joint positions are defined by the relation between the original and normalized sizes.

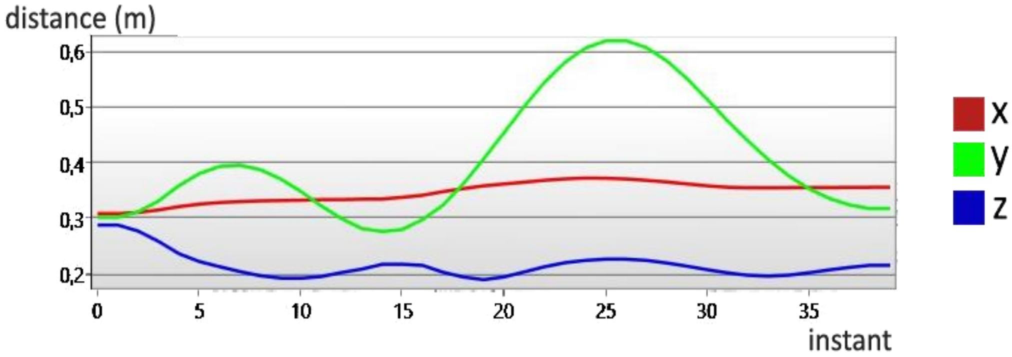


Figure 29 – Hand centre joint (X, Y, Z) coordinate data of the same instance of the sign "avaliar" shown in Figure 25 after the final step of normalization. This fourth step normalizes all signs to the same frame length. Only the left hand chart is shown.

In the end of the process, the movement diagram looks like the one in Figure 29.

3.5.2 Hand Depth Image Normalization

The original Hand depth image data stream, introduced in the subsection Hands Segmentation and illustrated by Figure 23, is not enough to solve some simple problems, such as:

a) Left and right hands are not the same – in PSL both hands can take any configuration required for the sign and, as in a written language, there are signers who are right handed while others are left handed;
b) Signer distance to sensor - The hand depth image is taken from the full depth image of the Kinect output, in which objects closer to the sensor have values closer to 0, and objects farther have increasingly larger values;
c) Signer hand size – once the volume that is used to extract the hand image is a fixed sized volume, a smaller hand occupies a smaller image proportion than a bigger hand, resulting in images with larger areas without information for smaller hands.

To address problem (a), when the hand depth image corresponds to the right hand, we just need to flip or mirror the image by its vertical axis, to obtain

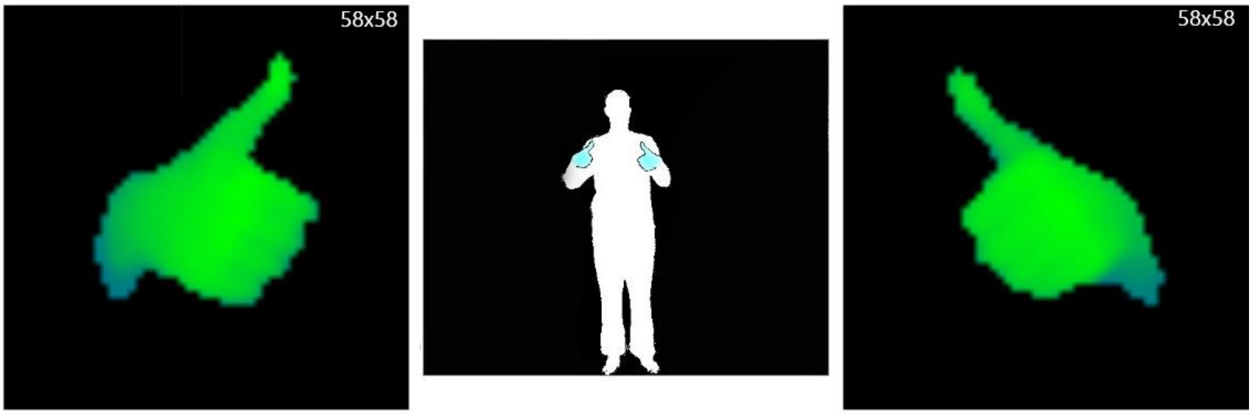


Figure 30 – Depth data input for the both hands. At the left side is the left hand and at the right side the right hand. The middle image is the original depth input after applying the background extraction. At the top right corners we depict the original size of the hand image. The colour representation is merely visual since the input for depth values varies from 0 (the sensor) to the max range the sensor can infer depth (accurately this range goes up to 4.5 m, hence 4500).

the left hand image. As a result, left and right hand images become equivalent and therefore comparable.

After the mirroring, both hands become comparable as an image (Figure 31).

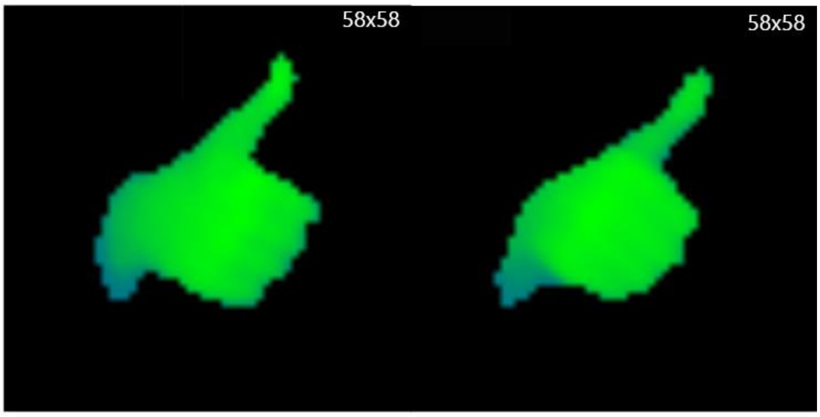


Figure 31- After mirroring one of the hands, the images become very similar. At the top right corners it is possible to see the original size of the hand image.

To address problem (b), we came up with a solution to eliminate the variable distances of the hand to the sensor, assuming that the closest value of the image, hence the one with the lower value different from 0, corresponds to the minimum value, that is, 1. This method is simply a shift on all the values of the image pixel values (depth values). This shift is equal to the minimum pixel value of the original image minus 1. This operation is not much observable from an image point of view, as in previous operations because of the representation used (converting 16bits grey image to 32bit RGBA image), yet it normalizes the distances of the hands to the sensor.

To solve (C), a scaling operation was done to the hand image so the hand would occupy the largest area possible. As the hand images used in the posture configuration recognition are 32x32 pixel and the hand image size recorded is usually considerably larger (varies with the signer hands distance to the sensor), a resizing is done in the image size. Two different methods were tested, one that would conserve the original aspect ratio of the hand portion in the image (called “Stretch Ratio”), and another method that does not conserve the aspect ratio of the hand (simply called “Stretch”). Figure 32 shows the results from the first method – “Stretch Ratio”.

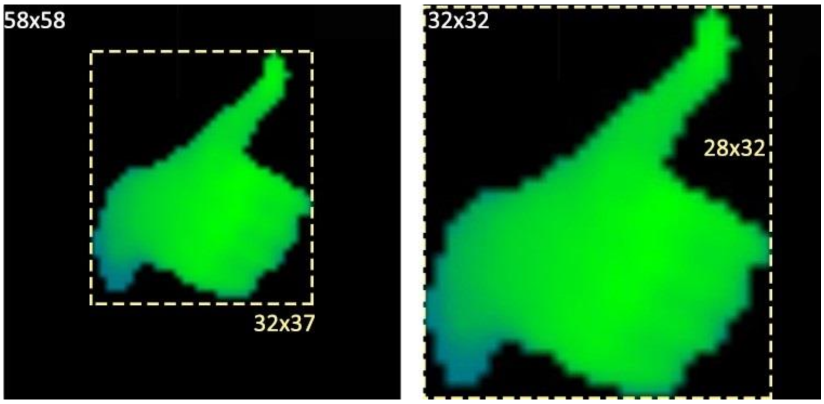


Figure 32 – At the left side is the hand depth image before applying any stretch method while in the right side is the result of the “Stretch Ratio” method. In both images, in white is the total size of the image while in yellow is the size of the square that contains the hand part in the image. The size is in the form AxB where A is the width and B the height.

The stretch method conserves the aspect ratio of the hand portion. The pixel values are calculated by linear interpolation according to the width values, this is, by horizontal lines.

The other tested method, not preserving the aspect ratio of the hand portion of the image, distorts the original image, but in some cases in an insignificant amount. Despite seeming to be a less promising approach to the problem, tests were conducted to verify its validity. Figure 33 shows the result of the “Stretch” method. Similarly to what happened in the previous method, the pixel values are calculated by linear interpolation according by horizontal lines.

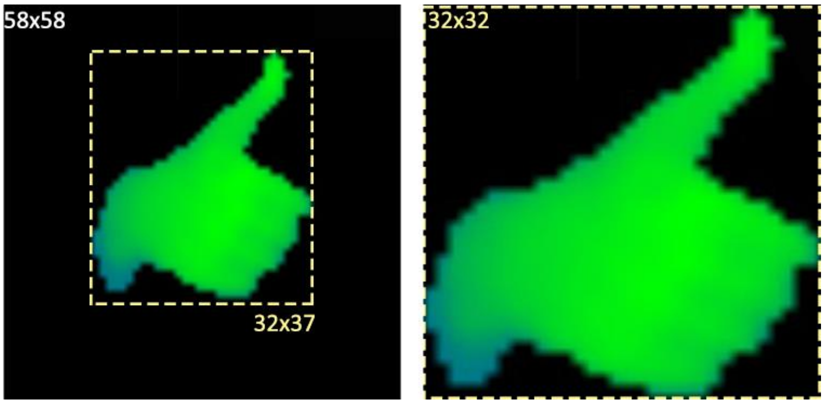


Figure 33 - At the left side is the hand depth image before applying any stretch method while in the right side is the result of the “Stretch” method. In both images, in white is the total size of the image while in yellow is the size of the square that contains the hand part in the image. The size is in the form AxB where A is the width and B the height.

3.6 Classifiers

This section specifies the methods required to train and test the **hand posture classifier** as well as the **sign classifier**, using SVM as the core machine learning technique for such classifiers and the data collected (see previous sections), as the training and testing datasets.

To be able to classify recorded postures and signs, such classifiers need first to be trained, and for that we need to identify a training dataset.

SVM can only solve binary problems, however, several approaches have been suggested to perform multi-class classification using SVM. In this thesis, we have adopted a one-against-one strategy for multi-class classification, dividing the multi-class problem into a set of binary problems. This set of binary problems compare all classes between each other. Redundant options can be discarded, such as comparing one class with itself (i.e. A vs A) and one of the two pairs of the same comparison (i.e. in the case of comparing A vs B and B vs A, B vs A can be ignored). Removing this redundant comparisons, a typical decision problem can be decomposed in the following subset of binary problems:

$$S = (n \times (n - 1))/2$$

Where S is the number of necessary SVMs and n is the number of classes. To decide for a class, a voting scheme is used. The class which receives more votes wins the decision process.

3.6.1 Postures Classifier

For the postures classifying system, two classifiers were created.

- The first hand posture classifier (52 classes) was created to verify the capture normalizations method.
- The second one (43 classes) was created to be merged with the sign classifier, in the same logical pipeline (see Figure 5).

The first classifier was trained with a dataset composed by 52 different postures, where this postures had minor variations. The second one had 43 different postures to address the problem of hand configurations varying position and palm orientation in signs, as described before. This way, the first and second classifiers are composed by 1326 and 861 machines respectively.

Any recorded posture from any class (hand configuration) used in both classifiers was transformed in a feature vector by transforming the hand depth image (normalized by the methods described acima), which is a two dimensional image with pixel values varying from 0 to 65536, 16bit, in a one dimensional vector of real values.

This classifier gives a result ranging from [0...52] and [0...43] for the first and second cases respectively, that correspond to the recognized hand posture.

To train and test the classifier, a K-fold cross-validation method was applied. In K-fold cross-validation, the original dataset is randomly partitioned into K equal size sample datasets. A K = 10 value was used and even though the sample datasets were randomly generated, we assured that each sample had various instances from each hand configuration.

In each fold of the algorithm, of the 10 sample datasets, 1 datasets was used as the test dataset for validating the model and the remaining 9 datasets were used as training data. This process was repeated 9 more times (performing a total of 10 folds), were each of the 10 datasets was used exactly once as test data. The recognition results from the folds were then averaged to produce a single estimation. The advantage of the cross validation method is that all observations are used for both training and validation, and each observation is used for testing exactly once.

Table 4 – Kernels tested for the SVM (implementations from Accord.net Framework (Souza, o. J.))

Linear	Gaussian	Quadratic
Inverse Multiquadratic	Histogram Intersection	Polynomial(2)
Polynomial(3)	Laplacian	Power

A set of 9 different kernels were tested for the SVM (see Table 4). We´ve created and trained the respective classifiers and retained the one that obtained the best accuracy results, to be used by the system. The tolerance value used on the sequential minimal optimization was set to 0.01. To be able to identify misrecognition patterns, over fitting, dominant classes and to assess the classifier’s accuracy for each sign of the vocabulary, a confusion matrix was created, and we´ve computed the training and testing accuracies for each of the eight folds.

3.6.2 Signs Classifier

To compare approaches and to address thesis hypothesis **H1**, two distinct sign classifiers were used. Both were created with the same SVM techniques explained for the postures classifiers cases. Both sign classifiers, have 29 classes and are composed by 406 machines.

- The first sign classifier relies solely on the hands path (not trained with posture features), hence, in handles only the movement component of the sign, or the gesture part.
- The second sign classifier uses both hand path and hands postures.

For each sign on the dataset, the feature vector is an array with a fixed number of positions, each one corresponding to a frame. This fixed dimension of the feature vector is chosen as the normalized frame duration of a sign (normalized method explained acima).

Each position of this vector contains, for the first case of sign classifier, 6 double values, corresponding the first three to the right hand centre joint (X, Y, Z) coordinates in that order and the following three positions, to the same coordinates of the left hand.

For the second signs classifier, each frame of the sign is described with 8 feature values, where the first three, correspond to the right hand centre joint (X, Y, Z) coordinates, the fourth value is set to the label of the hand configuration of the right hand (recognized with the posture classifier with 43 classes described before), and the four remaining values follow the same scheme as the first four values, but for the left hand.

The posture labels, feature values in the fourth and eight positions of the feature vector, are normalized to have a similar range as the other feature values. The hand centre joint (X, Y, Z) coordinates, range from [0...1] for all 3 coordinates. The normalization done to the original posture label value, which originally varied from [0...42], divided the value by 43. This way, the posture label value given to the feature vector for this component, has the range [0...1].

To test both classifiers, a K-fold cross-validation method was applied, similarly to what was done with the postures classifiers. A K = 8 value was used to ensure a correct division of the dataset, since it includes 48 repetitions of each sign (8 repetitions for each signer, times 6 signers). Despite the subsamples being randomly generated, we have assured that each subsample had one instance of a sign, from each of the signers. Again, the same kernels tested in the posture classifiers, were used to create both sign SVM classifiers.

Although it could be possible to achieve better results with a thorough exploration of the tested kernels parameters and even with other kernels or distinct machine learning techniques, if we consider the available time for realization of this thesis and if we also take into account, that its main aim was not to specifically explore the machine learning field, but rather to adopt existing and established techniques by the community, we can argue that was enough for this work, to utilize each kernel with the default values provided by the available SVM framework implementation.

3.7 Summary

In this chapter we have introduced the methodology used in the development of the proposed SLR system - PhySaLiS. We have started by presenting the system requirements derived from the objectives and presenting the system architecture and GUI. PhySaLiS has a minimalistic GUI that allows to easily record new signs or postures, manage the recorded data and use the classifiers for real time recognition.

The data collection subsection presented the data collection setup, corpora and detailed the data captured by our system. After the raw data was captured, pre-processing methods were applied, namely, background subtraction to isolate the signer from the background, and hands segmentation. The Kinect SDK was used to track the hands and extract the hands depth images. In the process of feature extraction, the hands images were carefully normalized, solving 3 main problems: left and right hands images were required to be comparable, the signer hand size could vary and the signer distance to the sensor could vary too. For the hand path features, the normalizations applied solved the following problems: signers distance to the sensor could vary arbitrarily; signers' heights are always different; the duration of the generated signs, vary from person to person.

After the normalization of the extracted features, the feature vectors were created and passed onto the different classification systems, for training and testing. Finally, we made a description about the classifiers, how they were created and their purpose for both hand posture recognition and sign recognition systems. All the classifiers used SVMs with different kernels and we´ve adopted a one-against-one strategy for multi-class classification. For the testing of the system, a K-fold cross-validation technique was employed.

4 Results and Evaluation

In this chapter we present the obtained results from the several approaches for the sign classification system and for the posture recognition sub system. The first section introduces the posture recognition system evaluation, using two distinct datasets, as well as the results from different normalization approaches, for the same system. The second section presents and compares the results obtained with two sign recognition systems: (1) sign system that uses only hand path information and (2) sign system that uses hand path and hands configuration information.

4.1 Hand Posture Recognition

After collecting the data and creating the classifiers for the hand posture system, the cross validation method was used to test the classifiers accuracy. For the 9 tested kernels, the 2 best kernels were the Gaussian Kernel and the Histogram Intersection Kernel, but only the results of the second one will be shown in order to simplify the visualization. For the 52 postures classifier multiple experiments were conducted in order to compare our approach with literature works. We´ve created classifiers with features computed from depth information, and others with features computed with binary (bitmap) information (shape of the hand). Variations in the hand depth image normalization process were also tested, were in the first experiments, classifiers where tested without addressing the problem of the signer hand size normalization (section 3.5.2), while in other experiments, the full normalization process was included. Finally, variations in the feature vector size were also tested, with feature vector sizes of 64 (images with 8x8 pixel), 256 (16x16 pixel), 1024 (32x32 pixel) and 4096 (64x64 pixel). No significant differences were obtained for the last three sizes. Since the image size of 32x32 pixel obtained usually higher accuracy, the selected feature vector size was 1024 (32x32 pixel).

Table 5 – Postures classifiers accuracy results for the data with 52 postures and with the 3 distinct normalization methods. The kernel used was the Histogram Intersection. Underlined we depict results of the same experiment, but with different feature sized vectors (32x32 and 64x64). The training and testing values are the averaged accuracy values from all K folds. Only the testing values provide reliable accuracy results. Training values are presented just for completeness.

		Normalization Method					
Features		No Stretch		Stretch Ratio		Stretch	
Type	**Size**	Training	Testing	Training	Testing	Training	Testing
Binary	**8x8**	0,554	0,475	0,584	0,498	0,666	0,583
	16x16	0,730	0,630	0,713	0,606	0,791	0,700
	32x32	0,758	0,654	-	-	0,815	0,720
	64x64	0,763	0,654	-	-	0,818	0,725
Depth	**8x8**	0,800	0,650	0,852	0,704	0,907	0,800
	16x16	0,910	0,770	0,913	0,772	0,961	0,861
	32x32	0,930	0,800	-	-	0,968	0,873
	64x64	0,942	0,809	-	-	0,970	0,871

By analysing the results (Table 5), firstly, it is possible to see that it is best to use the depth information than the binary information in our problem ("No Stretch" condition). This was expected since depth features contain more information than the binary (bitmap) ones. The feature size selected to be used in our case, was 32x32. The additional processing workload needed by the feature vector of size 64x64, does not increment the results in a statistically relevant amount. Size 32x32 addresses better our requirement of using hand the posture system within the sign recognition system pipeline, namely, by having to classify both hands in each of frame. In top of that, and because of the data collection method and sensor, the original raw input hand data, has sizes varying from 130x130 pixel to 40x40 pixel, in cases where the signer is farther away from the sensor but still in the acceptable range (1 to 3 meters). So, in some scenarios, if the size of 64x64 had been used, the original Kinect data for the hand would have been smaller and therefore the feature vector would need to be scaled up.

Considering the features of size 8x8 and 16x16, with the "Stretch" method applied, the performance of the classifiers had and an absolute average increase of 9.1% to the “Stretch Ratio” classifiers. Comparing the classifiers with the “Stretch” method against those without any stretch method (“No Stretch”), the first one had an absolute increase of 8.6% regarding features of sizes 8x8, 16x16, 32x32 and 64x64.

Analysing the data in Table-5, we can see that the best classifier uses the stretch method for the normalization of the hand depth image, for 8x8 and 16x16 feature vector sizes. With that conclusion at hand, we decided to adopt the stretch method for the normalization of the hand depth image and perform experiments with 32x32 feature size, for the reasons explained before, achieving a best result of **87.29%** of accuracy with 52 PSL hand postures. In appendix 0 we depict the cross matrix table for the 52 classifier. It is possible to see that we had 340 false positives out of 2703 recognitions, 30 postures had a recognition rate above the average 87.29%, while the standard deviation was 9.12%.

Comparing with similar works in the literature, (Almeida, 2011) made use of Kinect v1 using only depth information. This work achieved a 100% recognition rate on the 26 letters from the PSL alphabet (against our example with 52 postures) using a Skeletal-based Template Matching adaptation. His data set contained only one signer, and the testing was done with the same signer present in the data set, but with new recordings, thus showing over fitting. Also with the Kinect v1 and relying only on depth information (Souza, Pizzolato, 2013) achieved a recognition rate of 95.0% for 46 postures of the Brazilian Sign Language, also known as LIBRAS. Souza’s system was multi-user and used SVM with a Gaussian Kernel, using an estimated parameter δ. Using also depth information, but instead of the Kinect sensor, a TOF camera, (Kollorz et al., 2008) achieved a recognition rate of 95.12% for 12 hand configurations. Also making use of a TOF camera, (Uebersax, Gall, 2011), achieved an average recognition rate of 76.1% for the 26 letters of the ASL alphabet.

After concluding what was the best method to use, in the 52 postures case, we have created and tested a second SVM classifier with a dataset containing 43 postures, used also in the sign classification pipeline.

Table 6 – Testing results for the posture classifier for the dataset containing 43 postures. With the best method for normalization, depth data and Histogram Intersection kernel chosen, only the feature vector changes were experimented, mainly due to the computational and time costs of creating new classifiers. The training and testing values are averaged accuracy values from all folds. Only the testing values provide reliable accuracy results. Training values are presented just for completeness.

Features		Stretch	
Type	**Size**	Training	Testing
Depth	**8x8**	0,619	0,536
	16x16	0,723	0,616
	32x32	0.80	0.642

Since this dataset is composed by less postures (43) than the previous one, but that observe a lot more variation concerning the hand orientation, its computed accuracy after testing, dropped considerably (64.2% vs 87.3) if we compare it with the previous classifier. A similar approach and comparison, using only depth information from the hands, was done in (Souza, Pizzolato, 2013). To address the same problem of classifying hand postures in sign production, and with a set of 46 hand postures, it achieved a testing accuracy of 47.90%. Both performances are rather low when compared to the previous classifier because of the nature of the problem addressed in this second case (3.3.2). Despite the results, we believe that this is the only suitable classifier to address the problem of recognizing varying hand configurations along the motion of both hands, while a sign is being produced.

4.2 Sign Recognition

For the sign recognition system, that was the main challenge of this thesis, we have were created 2 SVM classifiers, as mentioned. In both cases, we have tested 9 different kernels, as explained in section 3.6.2, but only the 3 best results will be analysed.

For the first classifier, only concerned with the moving (gesture) part of the sign, that is, the hand path, the results are depicted in Table 7.

Table 7 – Testing accuracy results for the signs classifier using only the hand path as features. Various normalization sizes were tested, as well as 9 kernels. Only the best 3 are presented. The training and testing accuracy values are averaged accuracy values from all folds. Only the testing values provide reliable accuracy results. Training values are presented just for completeness.

		Kernel					
Features		Gaussian		Quadratic		Laplacian	
Type	**Size**	Training	Validation	Training	Validation	Training	Validation
Movement	**10**	0,813	0,779	0,866	0,836	0,945	0,88
	20	0,834	0,8	0,932	0,878	0,98	0,901
	30	0,841	0,802	0,962	0,9	0,99	0,907
	40	0,843	0,805	**0,976**	**0,916**	**0,996**	**0,916**
	50	0,844	0,805	0,982	0,916	0,997	0,915

Different feature normalization sizes (in frames) for the movements were tested. The movements of the dataset were normalized to that frame size with

Table 8 – Confusion matrix created in the classifier testing phase using the cross validation method. An average recognition rate of 91.6% was achieved for the selected 29 sign vocabulary.

	maravilha	apagar	escrever	graxa	balanca	avaliar	discutir	guerra	eclipse	morrer	fio	tubo(fino)	tubo(medio)	testemunha	verdade	mesa	balcao	gritar	cantar	apoio	cadeira	quente1	televisao1	ajudar	receber3	comunicar	trabalhar	nao	nadar	Hit	Total	HIT %
maravilha	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	48	1,0000
apagar	0	46	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	48	0,9583
escrever	0	1	41	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	41	48	0,8542
graxa	0	3	9	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	48	0,7500
balanca	0	0	0	0	42	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	42	48	0,8750
avaliar	0	0	0	0	11	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	35	48	0,7292
discutir	0	0	0	0	0	0	43	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	43	48	0,8958
guerra	0	0	0	0	0	0	6	41	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	41	48	0,8542
eclipse	0	0	0	0	0	0	0	1	44	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	44	48	0,9167
morrer	0	0	0	0	0	0	0	0	2	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	48	0,9583
fio	0	0	0	0	0	0	0	0	0	0	47	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	47	48	0,9792
tubo(fino)	0	0	0	0	0	0	0	0	0	1	1	37	6	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	37	48	0,7708
tubo(medio)	0	0	0	0	0	0	0	0	0	0	0	6	40	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	40	48	0,8333
testemunha	0	0	0	0	0	0	0	0	0	0	0	0	0	37	10	0	0	0	0	0	0	0	0	0	0	0	0	0	1	37	48	0,7708
verdade	0	0	0	0	0	0	0	0	0	0	0	0	0	12	35	0	0	0	0	0	0	0	0	0	0	0	0	0	1	35	48	0,7292
mesa	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	47	0	0	0	0	0	0	0	0	0	0	0	0	0	47	48	0,9792
balcao	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	45	0	0	0	0	0	0	0	0	0	0	0	0	45	48	0,9375
gritar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	0	0	0	0	0	0	0	0	0	0	0	48	48	1,0000
cantar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	0	0	0	0	0	0	0	0	0	0	48	48	1,0000
apoio	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	47	0	0	0	1	0	0	0	0	0	47	48	0,9792
cadeira	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	0	0	1	0	0	0	0	0	46	48	0,9583
quente1	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	0	0	0	0	0	0	0	45	48	0,9375
televisao1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	47	0	0	0	0	0	1	47	48	0,9792
ajudar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	0	0	0	0	0	48	48	1,0000
receber3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	0	0	0	0	48	48	1,0000
comunicar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	0	0	0	48	48	1,0000
trabalhar	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	0	1	46	48	0,9583
nao	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	0	48	48	1,0000
nadar	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	46	48	0,9583
																														1275	1392	0,9159

the method described in 3.5.1. By analysing the data in Table 7, it is possible to see that the best classifier uses features resized to 40 frames with the *Laplacian* kernel (testing accuracy of **91.6**%).

In the confusion matrix table for this classifier (Table 8). It is possible to see that were 117 false positives out of 1392 recognitions, 10 postures had a recognition rate below the average **91.6**% while the standard deviation was **8.93**%. In Table 3, we present the corpora, were we also highlight signs with similar or with the same hand movement. As it was expected, the signs that had lower recognition rate were precisely those that have the same hand movement. These are the pairs “escrever” and “graxa”, “balança” and “avaliar”, “discutir” and “guerra” and lastely “tubo(fino)” and “tubo(médio)”. Also, signs with similar hand movement as “eclipse” and “morrer” and “testemunha” and “verdade”, that have small differences only in the positioning of the sign, observed a lower accuracy rate. The classifier was able to distinguish signs with a significant positioning difference but with the same movement, as the case of the pair “mesa” and “balcão”: the sole difference in this pair, is that the first one is performed bellow the chest, while the second one above the chest. Yet, notwithstanding the fact that “balcão” has similarity with “mesa”, it also shares similar positioning with the sign “tubo”.

Comparing again with (Almeida, 2011), this author achieved a 100% recognition rate on just 10 signs from the PSL alphabet (against our example with 29 signs) using an algorithm of 3D Path Analysis. His data set contained only one signer, and the testing was done with the same signer present in the data set, thus observing the problem of over fitting. In Almeida’s work, of the 10 signs, only one pair shared similar hand paths. Similar approach to this thesis problem took (Souza, Pizzolato, 2013), testing first the system using only the hand trajectory information. They achieved a recognition rate of 55.24% for 13 signs of LIBRAS. Souza’s system is multi-user and he used HCRF to address this problem.

After validated the method of recognizing the movement part of the sign, by analysing the results and comparing with other works, the final step towards the demonstration of the hypothesis **H1,** can be done.

The second sign classifiers created have to merge both movement information (hand path information) with the hand configuration. The classifier explained in 3.6.2 yield the following results:

Table 9 – Testing results for the signs classifier using the hand path and the hand posture labels as features in each frame. Various normalization sizes were tested, as well as 9 kernels but only the best 3 are presented. The training and validation values are averaged accuracy values from the 8 folds. Only the testing values provide reliable accuracy results. Training values are presented just for completeness.

		Kernel					
Features		Gaussian		Quadratic		Histogram Intersection	
Type	**Size**	Training	Validation	Training	Validation	Training	Validation
Movement + Hand Labels	**10**	0,866	0,65	0,987	0,654	0,862	0,774
	20	0,918	0,731	0,99	0,732	0,885	0,796
	30	0,936	0,751	0,99	0,751	0,897	0,813
	40	0,942	0,759	0,99	0,746	0,896	0,809
	50	0,944	0,764	0,99	0,752	0,903	0,808

This approach, uses both movement and hand configurations information and we´ve obtained a best accuracy result of **81.3**%. Comparing this method with the previous one that used only hand movements, there is an absolute decrease in accuracy of **10.3**%. A possible justification for this decreasing in the sign recognition accuracy is due to the hand labels, that, despite normalized to fit the feature vector, introduce instability in the dataset. In the confusion matrix for this second sign classifier (Table 10), is possible to see that no sign increased recognition. Besides introducing more error in the signs that were not supposed to benefit from this approach (signs that had no similar nor equal hand movement), the prior hand label classification were not helpful in distinguishing the pair of signs (Table 3), which should actually benefit from this a-priori hand posture classification.

*Table 10 – Confusion matrix created in the sign classifier testing phase using the cross validation method. This classifier used both hand movement and hand configuration information. An average recognition rate of **81.3**% was achieved for the selected 29 sign vocabulary.*

	maravilha	apagar	escrever	graxa	balanca	avaliar	discutir	guerra	eclipse	morrer	fio	tubo(fino)	tubo(medio)	testemunha	verdade	mesa	balcao	gritar	cantar	apoio	cadeira	quente1	televisao1	ajudar	receber3	comunicar	trabalhar	nao	nadar	Hit	Total	HIT %
maravilha	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	48	1,0000
apagar	0	28	10	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	28	48	0,5833
escrever	0	5	25	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	25	48	0,5208
graxa	0	10	10	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	0	25	48	0,5208
balanca	0	1	0	0	36	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	36	48	0,7500
avaliar	0	0	0	0	13	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	23	48	0,4792
discutir	0	0	0	0	0	0	35	5	0	1	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	35	48	0,7292
guerra	0	0	0	0	0	0	5	39	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	39	48	0,8125
eclipse	0	0	0	0	0	0	0	0	42	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42	48	0,8750
morrer	0	0	0	0	0	0	0	5	4	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	39	48	0,8125
fio	0	0	0	0	0	0	1	0	0	0	40	1	0	0	0	1	5	0	0	0	0	0	0	0	0	0	0	0	0	40	48	0,8333
tubo(fino)	0	0	0	0	0	0	0	0	0	1	3	27	9	0	0	3	5	0	0	0	0	0	0	0	0	0	0	0	0	27	48	0,5625
ubo(medio	0	0	0	0	0	0	0	0	0	0	1	13	29	0	0	1	4	0	0	0	0	0	0	0	0	0	0	0	0	29	48	0,6042
testemunha	0	0	0	1	0	0	0	0	0	0	0	0	0	36	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	48	0,7500
verdade	0	0	0	0	0	3	0	0	0	0	0	0	0	13	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32	48	0,6667
mesa	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	46	0	0	0	0	0	0	0	0	0	0	0	0	0	46	48	0,9583
balcao	0	0	0	0	0	0	0	0	0	0	2	2	5	0	0	0	39	0	0	0	0	0	0	0	0	0	0	0	0	39	48	0,8125
gritar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	0	0	0	0	0	0	0	0	0	0	0	48	48	1,0000
cantar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	0	0	0	0	0	0	0	0	0	0	48	48	1,0000
apoio	0	2	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	3	0	0	0	0	0	40	48	0,8333
cadeira	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	1	0	0	0	0	1	0	0	46	48	0,9583
quente1	0	0	0	0	3	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	37	48	0,7708
televisao1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	47	0	0	0	0	0	0	47	48	0,9792
ajudar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	1	0	43	0	0	0	0	0	43	48	0,8958
receber3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	47	0	0	0	0	47	48	0,9792
comunicar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	47	0	0	0	47	48	0,9792
trabalhar	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	0	0	46	48	0,9583
nao	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	0	48	48	1,0000
nadar	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	46	48	0,9583
																														1132	1392	0,8132

The most similar work to our´s is by (Souza, Pizzolato, 2013), that for the sign recognition with hand information, also included the hand and face orientations information. Souza's achieved an **84.41**% accuracy (compared to our **81.3**% for 29 PSL signs), using a SVM with a Quadratic kernel for classifying the hand configuration and Hidden Conditional Random Fields (HCRF) to merge all the information. In their work, there was a substantial increase of the accuracy when comparing with their other approach without the hand posture information (55.24%)

5 Conclusions and Future Work

5.1 Conclusions

In this thesis, we have started by describing main components of the Portuguese Sign Language and by highlighting the main problems this work proposed to tackle.

After analysing the state of the art, we have presented in detail several experiments conducted to address the problem of performing Automatic Portuguese Sign Language Recognition, for signs that observe only the manual features of the Stokoe model, that is, hand configuration, orientation, position, motion, contact point and plane of the hand, where the last 6 parameters are strictly connected and can be implicitly observed by analysing the movement performed by both hands.

The experiments used our proposal for an Automatic Sign Recognition System Architecture. In this architecture, depth imaging data from the Kinect One sensor is consumed, and the system generates automatic PSL sign recognition results, by adopting a machine learning classification technique based in SVM and a one-against-one strategy for multi-class classification. To generate Sign recognition results, a K-fold cross-validation technique was employed.

The first hypothesis (**H1**) suggested that it would be possible to perform Automatic Portuguese Sign Language Recognition (APSL) of many signs, by analysing the manual features of the Stokoe model, namely, the hand configuration and hand path. We further hypothesised that hand configuration and hand path could be automatically recognized by using a suitable machine learning-based classification technique, trained and tested with data collected by a low-cost non-invasive RGB-D sensor.

Regarding the initial sub-problem of hand posture recognition, we have carried two major experiments, respectively, with a dataset of 52 PSL hand postures produced by 2 signers and 43 hand postures, generated by 2 signers too, where, in this last case postures could change its orientation to better accommodate what happens in real sign production. We´ve obtained, best average hand posture accuracy results of, respectively, **91.6**% and **81.3**%, in 10 folds, which compares well with the literature.

Subsequently, we performed two other major experiments with 6 signers (5 men and 1 women), where for each signer, we´ve collected eight repetitions of 29 PSL signs. These 29 signs were chosen to observe the following characteristics:

1. Signs with the same or similar movement, but different configurations for the hand(s);
2. Signs with the same posture but different movements of the hands;
3. Signs with the same hand configuration and the same movement but different locations in relation to the human body

With features observing hand paths only, we have achieved an average sign recognition accuracy in 8 folds, of **91.6**% for 29 PSL signs, which surpassed a similar experience of (Souza, Pizzolato, 2013), that achieved a recognition rate of **55.24**% for 13 signs of LIBRAS (Brazilian Sign Language).

When we merged merge both movement information (hand path information) with hand configuration, in the feature vector, our average sign recognition accuracy in 8 folds, decreased to **81.3**% for 29 PSL signs. In a similar experiment with Brazilian Sign Language by the same authors (Souza, Pizzolato, 2013), where for sign recognition with hand information, they also included the hand and face orientations information, Souza et al., achieved a **84.41**% accuracy for 13 signs of LIBRAS, which compares well with our **81.3**% for 29 PSL signs.

In general our automatic PSL sign recognition system, was able to correctly distinguish the paired signs of the classes 2 and 3 above, but showed less convincing results in distinguishing pairs of the type 1, possibly due to instabilities in the feature vector introduced by the hand posture recognition, that still need to be investigated. Despite this fact, our thesis extended the state-of-art in automatic recognition of PSL by using 52 hand and 29 signs of PSL, improving the previous 26 hand postures and 10 signs addressed by (Almeida, 2011). For these reasons, we believe that Hypothesis **H1** was verified in this thesis

The second hypothesis (**H2**) suggested that an approach that classifies both hand configuration and hand path, can outperform a system based only on hand path classification. In our experiments, that hypothesis was not verified, since the sign recognition system that analysed only the hand path of the sign, was the most reliable with an average accuracy of **91.6**%, compared to **81.5**% for the recognition system trained with features that merged hand configuration and hand path. While (Gineke, Reinders, 2009; Souza, Pizzolato, 2013) have stressed that hand information was crucial (but not exclusive, since Souza and Pizzolato add also hands and facial orientation), and that it could increase the sign recognition accuracy, this was not verified in our system. Therefore, with our studies, we were not able to demonstrate Hypothesis **H2**.

5.2 Future Work

As a future research in the topic of automatic recognition of sign language, and, namely, of PSL, we believe that the hand posture classification, could be improved. Other promising approaches could address: coupling other information with the hand depth image, such as hand orientation; adopt an appearance (geometrical an topological) based model of the hand for each frame or set of frames; try to assert the hand configuration by analysing a set of frames of the movement or even the whole movement, instead of classifying a hand posture in each frame.

Other line of research, could address the problem of facial expressions in the production of signs, in sync with the hands postures and hands movements. Interesting results in the literature, could be taken into account to address this problem, like in (von Agris et al., 2008), or with the ViKi (Visual Kinect) system developed by Hélder Abreu (Abreu, 2014), tackling the problem of automatic visual speech recognition, based on an articulatory approach to detect facial expressions in the region of the lips. Analysing other body parts and movements would also be relevant for a Sign Language system, aiming at identifying all classes of signs. Concerning the machine learning technique, other approaches could be tested, either by a thorough investigation and manipulation in the kernels used in the SVMs or for new kernels, or even by using other machine learning techniques such as HMM or Deep Neural Networks. Another interesting challenge to address, is the problem of continuous sign language recognition

.

.

Bibliography

Abreu, Hélder (2014): „Visual Speech Recognition for European Portuguese". Universidade do Minho - Escola de Engenharia.

Von Agris, Ulrich; Knorr, Moritz; Kraiss, Karl-Friedrich (2008): „The significance of facial features for automatic sign language recognition". In: *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*. Ieee, pp. 1–6, DOI: 10.1109/AFGR.2008.4813472. — ISBN: 978-1-4244-2153-4

Almeida, Rui (2011): „Portuguese Sign Language Recognition via Computer Vision and Depth Sensor". ISCTE-IUL Lisbon University Institute.

Andersen, MR; Jensen, T; Lisouski, P (2012): „Kinect depth sensor evaluation for computer vision applications". In:, p. 37, DOI: Technical Report ECE-TR-6.

Athitsos, Vassilis; Neidle, Carol; Sclaroff, Stan (2010): „Large lexicon project: American Sign Language video corpus and sign language indexing/retrieval algorithms". In: *Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT)*., pp. 11–14.

Bela Baltazar, Ana; Porto Editora (ed.) (2010): *Dicionário de Língua Gestual Portuguesa*. o.V. — ISBN: 978-972-0-05282-7

Bowden, Richard; Windridge, David; Kadir, T (2004): „A linguistic feature vector for the visual interpretation of sign language". In: *Computer Vision-ECCV*., pp. 1–12, DOI: 10.1007/978-3-540-24670-1_30.

Brashear, Helene; Starner, Thad; Lukowicz, Paul; et al. (2003): „Using multiple sensors for mobile sign language recognition". In: *Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings. (2003)*., pp. 45–52, DOI: 10.1109/ISWC.2003.1241392.

Capilla, DM (2012): „Sign Language Translator using Microsoft Kinect XBOX 360". University of Tennessee (Knoxville - USA).

Chai, Xiujuan; Li, Guang; Lin, Yushun; et al. (2013): „Sign Language Recognition and Translation with Kinect". In: *Vipl.Ict.Ac.Cn (2013)*.

Dictionary, The Free (2014): „Gesture definition". Retrieved am 27.10.2014 from http://www.thefreedictionary.com/gesture.

Fang, Gaolin; Gao, Wen (2002): „A SRN/HMM system for signer-independent continuous sign language recognition“. In: *Proceedings - 5th IEEE International Conference on Automatic Face Gesture Recognition, FGR 2002 (2002)*., pp. 312–317, DOI: 10.1109/AFGR.2002.1004172. — ISBN: 0769516025

Fang, Gaolin; Gao, Wen; Zhao, Debin (2004): „Large vocabulary sign language recognition based on fuzzy decision trees“. In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans (2004)*. 34 (3), pp. 305–314, DOI: 10.1109/TSMCA.2004.824852.

Freeman, WT; Roth, Michal (1995): „Orientation histograms for hand gesture recognition“. In: *International Workshop on Automatic Face and Gesture Recognition (1995)*. 12 , pp. 296–301.

Gineke, A; Reinders, MJT (2009): „Influence of handshape information on automatic sign language recognition“. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2009)*. 5934 , pp. 301–312, DOI: 10.1007/978-3-642-12553-9_27.

Hong, Seok-ju; Setiawan, Nurul Arif Setiawan; Lee, Chil-woo (2007): „Real-Time Vision Based Gesture Recognition for Human-Robot Interaction“. In: *Knowledge-Based Intelligent Information and Engineering Systems: KES 2007 - WIRN 2007, Pt I*. 4692 , pp. 493–500.

Kadhim Shubber (2013): „Microsoft Kinect used to live-translate sign language into text“. *July 18th*. Retrieved am from http://www.wired.co.uk/news/archive/2013-07/18/sign-language-translation-kinect.

Kadous, MW (1996): „Machine recognition of Auslan signs using PowerGloves: Towards large-lexicon recognition of sign language“. In: *Workshop Integration of Gestures in Language and Speech (1996)*., pp. 165–174.

Khoshelham, Kourosh; Elberink, Sander Oude (2012): „Accuracy and resolution of Kinect depth data for indoor mapping applications.“. In: *Sensors (Basel, Switzerland)*. 12 (2), pp. 1437–54, DOI: 10.3390/s120201437.

Kollorz, Eva; Penne, Jochen; Hornegger, Joachim; et al. (2008): „Gesture recognition with a Time-Of-Flight camera“. In: *International Journal of Intelligent Systems Technologies and Applications*. 5 (3/4), p. 334, DOI: 10.1504/IJISTA.2008.021296.

Liddell, SK; Johnson, RE (1989): „American sign language: The phonological base“. In: *Sign Language Studies (1989)*. 64 , pp. 195–278, DOI: 10.1353/sls.1989.0027.

McGuire, RM (2004): „Towards a one-way american sign language translator“. In: *Proceedings - Sixth IEEE International Conference on Automatic Face and Gesture Recognition (2004)*., pp. 620–625, DOI: 10.1109/AFGR.2004.1301602.

Microsoft (2014a): „Kinect One Specifications“. Retrieved am 27.10.2014 from http://www.microsoft.com/en-us/kinectforwindows/meetkinect/features.aspx.

Microsoft (2014b): „Microsoft Developer Network - Skeletal Tracking“. Retrieved am 21.10.2014 from http://msdn.microsoft.com/en-us/library/hh973074.aspx.

Moita, Mara; Carmo, Patrícia; Carmo, Helena; et al. (2011): „Preliminary studies for a Portuguese signing AVATAR modelization“. In: *Cadernos de Saúde (2011)*. 4 , pp. 25–35.

Parish, D H; Sperling, G; Landy, M S (1990): „Intelligent temporal subsampling of American Sign Language using event boundaries.“. In: *Journal of experimental psychology. Human perception and performance*. 16 (2), pp. 282–94.

Pashaloudi, Vassilia; Margaritis, Konstantinos (2002): „Hidden markov models for greek sign language recognition“. In: *Proceedings of 2nd WSEAS International Conference on Speech, Signal and Image Processing (ICOSSIP) (2002)*.

Segen, Jakub; Kumar, Senthil (1999): „Shadow gestures: 3D hand pose estimation using a single camera“. In: *Proceedings. 1999 IEEE Computer*

Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149) (1999). 1 , DOI: 10.1109/CVPR.1999.786981.

Souza, César Roberto De; Pizzolato, Ednaldo Brigante (2013): „Sign Language Recognition with Support Vector Machines and Hidden Conditional Random Fields Going from Fingerspelling to Natural Articulated Words“. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2013)*. 7988 LNAI , pp. 84–98, DOI: 10.1007/978-3-642-39712-7_7.

Starner, Thad; Pentland, Alex (1995): „Visual Recognition of American Sign Language Using Hidden Markov Models.“. In: *International Workshop on Automatic Face and Gesture Recognition (1995)*., pp. 189–194.

Stokoe, William C (2005): „Sign language structure: an outline of the visual communication systems of the American deaf. 1960.“. In: *Journal of deaf studies and deaf education*. 10 (1), pp. 3–37, DOI: 10.1093/deafed/eni001.

Uebersax, D; Gall, J (2011): „Real-time sign language letter and word recognition from depth data“. In: *Proceedings of the IEEE International Conference on Computer Vision (2011)*., pp. 383–390, DOI: 10.1109/ICCVW.2011.6130267.

Valli, Clayton; Lucas, Ceil (1992): *The linguistics of American sign language: An Introduction*. *Sign language studies*. o.V. — ISBN: 1563681137

Vogler, Christian; Metaxas, Dimitris (1997): „Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods“. In: *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation (1997)*. 1 , pp. 156–161, DOI: 10.1109/ICSMC.1997.625741.

Vogler, Christian; Metaxas, Dimitris (o. J.): „ASL recognition based on a coupling between HMMs and 3D motion analysis“. In: *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271) (1998)*., pp. 363–369, DOI: 10.1109/ICCV.1998.710744.

Vogler, Christian; Metaxas, Dimitris (1999): „Parallel hidden markov models for american sign language recognition“. In: *1999. The Proceedings of the*

Seventh IEEE International Conference on Computer Vision. 1 , pp. 116 – 122, DOI: 10.1109/ICCV.1999.791206.

Wang, SB; Quattoni, A (2006): „Hidden conditional random fields for gesture recognition“. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2 , pp. 1521 – 1527, DOI: 10.1109/CVPR.2006.132.

Wilson, AD; Bobick, AF (2000): „Realtime online adaptive gesture recognition“. In: *Pattern Recognition, 2000. Proceedings. 15th International Conference on (Volume:1)*. vol.1 (505), pp. 270 – 275, DOI: 10.1109/ICPR.2000.905317.

Yang, Ruiduo; Sarkar, Sudeep; Loeding, Barbara (2010): „Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming.“. In: *IEEE transactions on pattern analysis and machine intelligence*. 32 (3), pp. 462–77, DOI: 10.1109/TPAMI.2009.26.

Zafrulla, Zahoor; Brashear, Helene; Hamilton, Harley; et al. (2011): „American sign language recognition with the kinect“. In: *Proceedings of the 13th international conference on multimodal interfaces*., pp. 279–286, DOI: 10.1145/2070481.2070532. — ISBN: 978-1-4503-0641-6

Zieren, J; Kraiss, KF (2004): „Non-intrusive sign language recognition for human-computer interaction“. In: *Proceedings of the 9th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems, September 7-9, Atlanta, Georgia*.

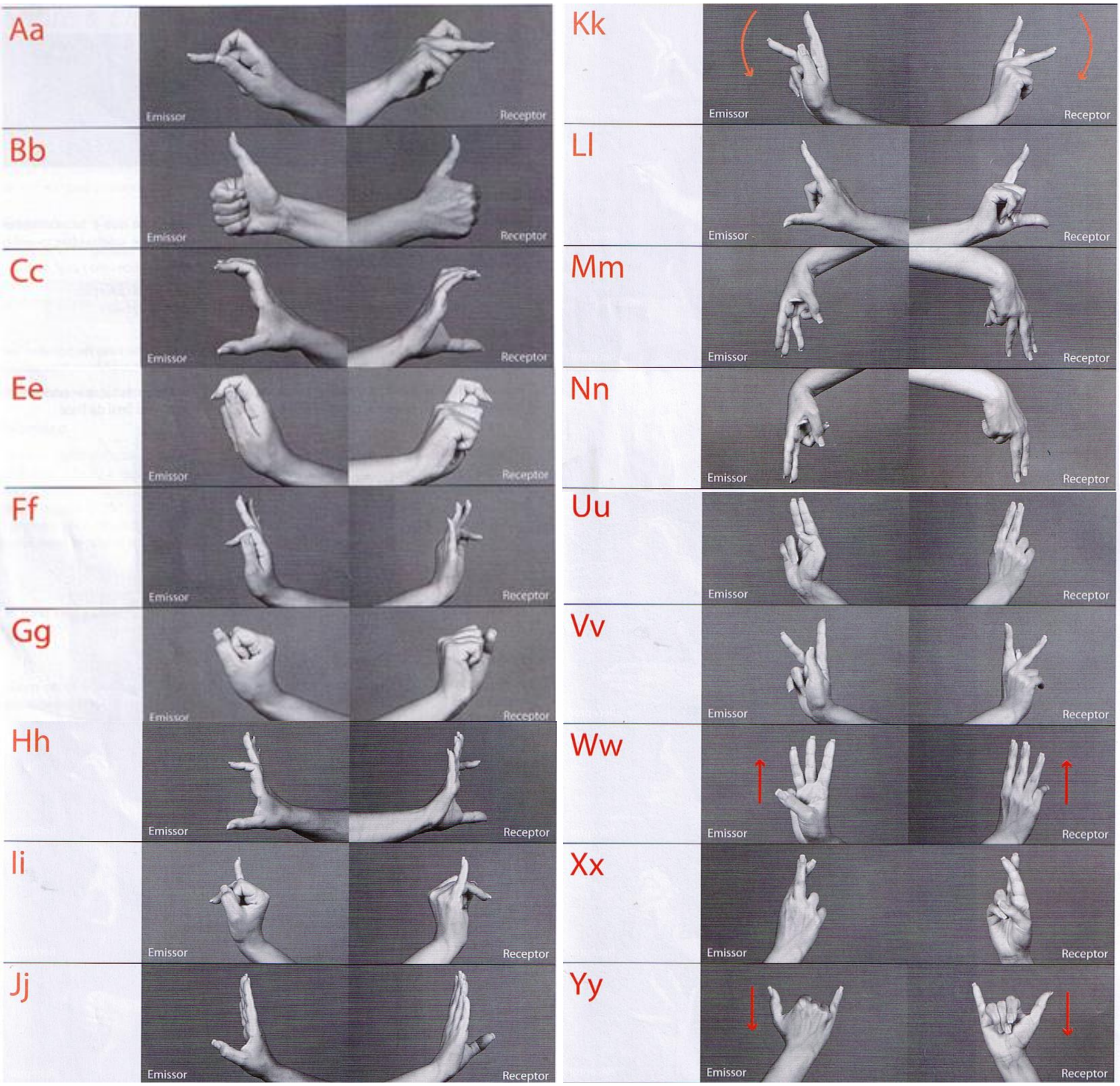
Appendix

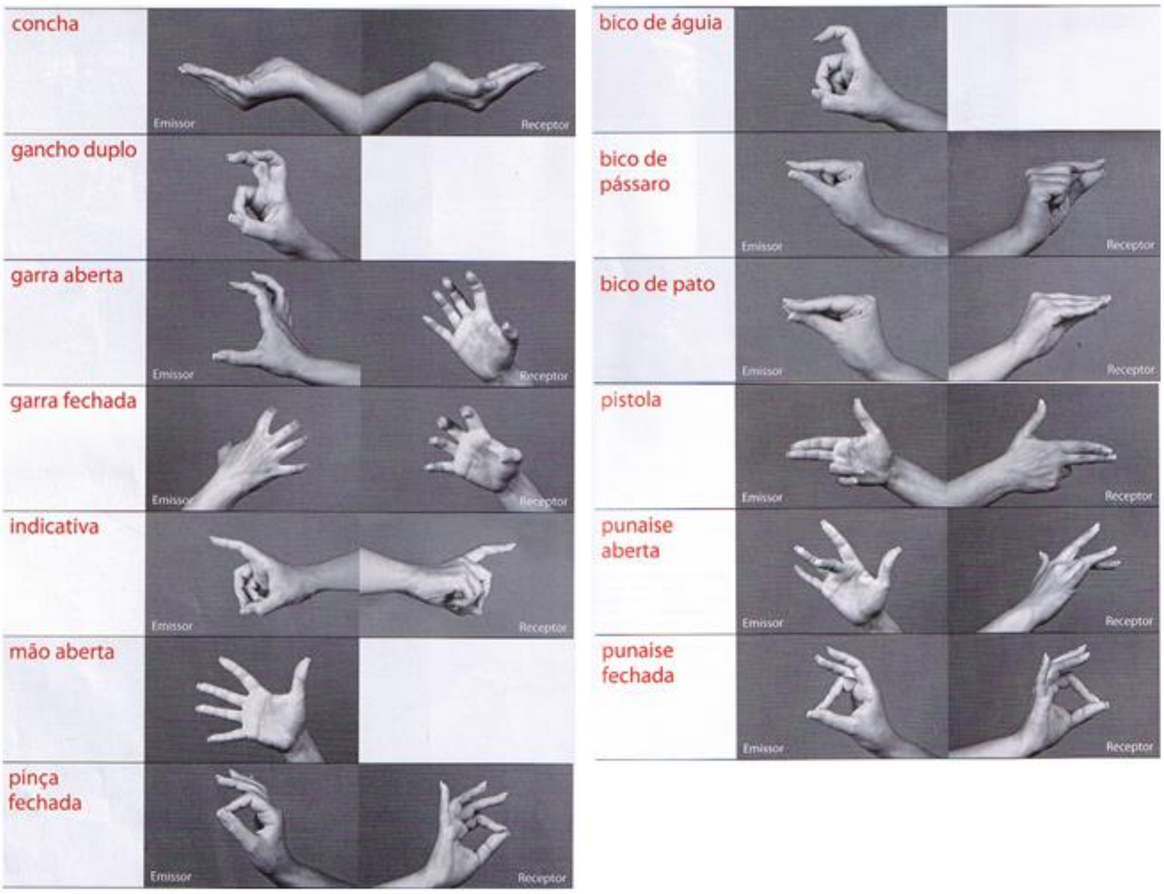
A.CROSS MATRIX FOR 52 POSTURES CLASSIFIER

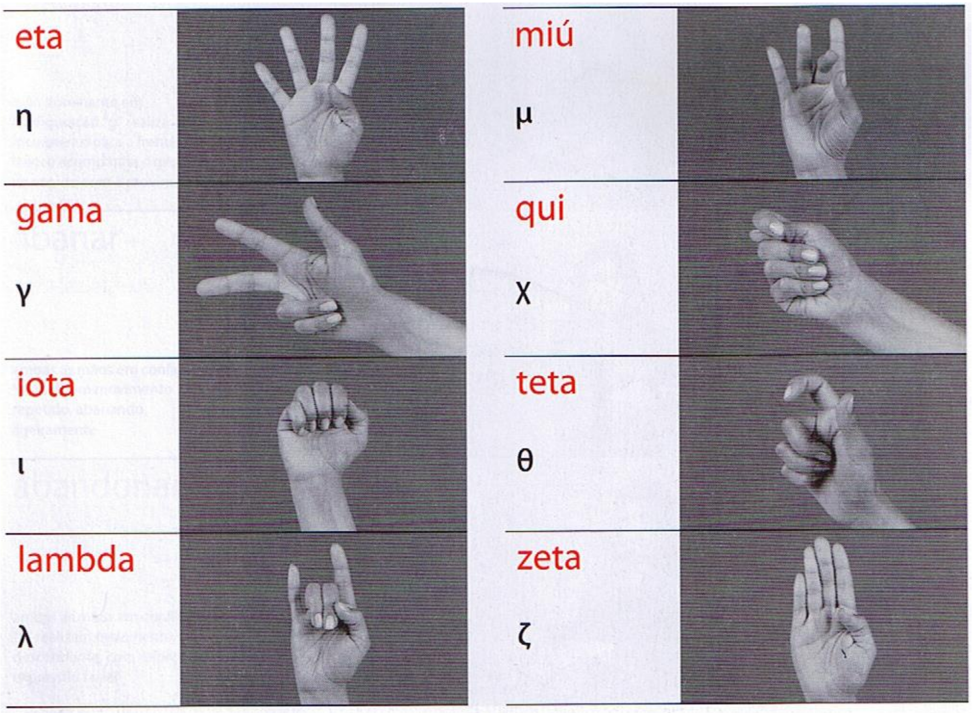
	a	b	c	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	1	2	3	4	5	6	7	8	9	bicoaguia	bicopassaro	bicopato	concha	ganchoduplo	garraaberta	garrafechada	indicativa	maoaberta	pincafechada	pistola	punaiseaberta	punaisefechada	eta	gama	teta	zeta	lambda	iota	Hit	Total	HIT %
a	40	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	41	0,9756
b	0	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	43	47	0,9149
c	0	0	45	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	45	50	0,9000
e	0	0	0	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	53	0,8679
f	0	0	0	0	49	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	49	50	0,9800
g	0	0	0	0	0	38	0	0	0	0	0	0	0	0	0	1	0	8	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	38	51	0,7451
h	0	0	1	0	0	0	45	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	45	50	0,9000
i	0	0	0	1	0	0	0	44	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2	0	44	53	0,8302
j	0	0	0	0	0	0	1	0	44	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	44	50	0,8800
k	0	0	0	0	0	0	0	0	0	46	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	46	50	0,9200
l	0	0	0	0	0	0	0	0	0	0	49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	49	50	0,9800
m	0	0	0	0	0	0	0	0	0	0	0	55	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	55	57	0,9649
n	0	0	0	0	0	0	0	0	0	0	0	3	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	49	0,9388
o	0	0	0	0	0	1	0	0	0	0	0	0	0	48	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	48	55	0,8727
p	0	0	0	0	0	0	0	0	0	0	0	0	0	1	53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	53	57	0,9298
q	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	49	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	49	51	0,9608
r	0	0	1	0	1	0	4	0	0	0	0	0	0	0	0	0	46	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	46	54	0,8519
s	0	0	0	1	0	7	0	1	0	0	0	0	0	0	0	3	0	30	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	45	0,6667
t	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	37	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	37	48	0,7708
u	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	56	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	56	62	0,9032
v	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	51	0	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	51	60	0,8500
w	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	1	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	2	0	0	43	53	0,8113
x	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	52	55	0,9455
y	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	47	49	0,9592
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	47	50	0,9400
2	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	41	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	3	0	0	0	2	0	0	0	0	41	53	0,7736
3	0	0	0	8	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	30	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	46	0,6522
4	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	45	50	0,9000
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	52	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	52	55	0,9455
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42	46	0,9130
7	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	55	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	55	59	0,9322
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	50	53	0,9434
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	58	60	0,9667
bicoaguia	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	3	0	0	0	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	53	0,8679
bicopassaro	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	48	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	52	0,9231
bicopato	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50	51	0,9804
concha	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	52	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	52	54	0,9630
ganchoduplo	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	47	0	0	0	0	0	0	0	0	0	0	0	4	0	0	47	56	0,8393
garraaberta	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	2	0	0	0	0	1	0	3	0	0	0	0	0	48	54	0,8889
garrafechada	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	45	0	0	0	0	0	2	0	0	0	0	0	0	45	51	0,8824
indicativa	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	42	1	0	0	0	0	0	0	0	0	0	0	42	46	0,9130
maoaberta	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	0	0	2	0	2	0	1	0	0	0	48	57	0,8421
pincafechada	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	56	0	0	0	0	0	0	0	0	0	56	58	0,9655
pistola	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	44	0	0	0	9	0	0	0	0	44	58	0,7586
punaiseaberta	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	45	4	1	0	0	0	0	0	45	54	0,8333
punaisefechada	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	2	38	6	0	0	0	0	0	38	49	0,7755
eta	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	4	1	0	1	2	30	0	0	4	1	0	30	49	0,6122
gama	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	43	0	0	0	0	43	53	0,8113
teta	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	2	1	0	0	0	0	0	0	39	0	1	0	39	47	0,8298
zeta	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	5	1	0	32	0	0	32	49	0,6531
lambda	0	0	0	0	0	0	0	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	1	1	1	0	0	42	0	42	52	0,8077
iota	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	46	48	0,9583
																																																					2363	2703	0,8729

Figure 34 - Using histogram intersection kernel with 32x32 feature vector size, depth information, and the full normalization process (with stretch method)

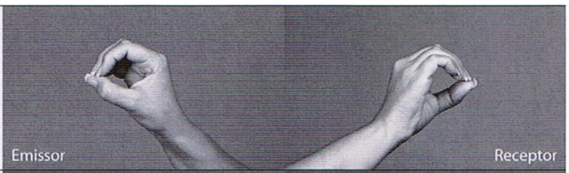
B. Hand Configurations (Bela Baltazar, 2010)



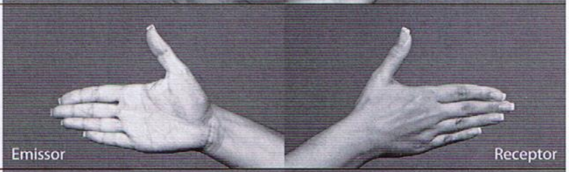

concha
Emissor
Receptor
gancho duplo
garra aberta
Emissor
Receptor
garra fechada
Emissor
Receptor
indicativa
Emissor
Receptor
mão aberta
pinça fechada
Emissor
Receptor
bico de águia
bico de pássaro
Emissor
Receptor
bico de pato
Emissor
Receptor
pistola
Emissor
Receptor
punaise aberta
Emissor
Receptor
punaise fechada
Emissor
Receptor


eta
η
gama
γ
iota
ι
lambda
λ
miú
μ
qui
χ
teta
θ
zeta
ζ

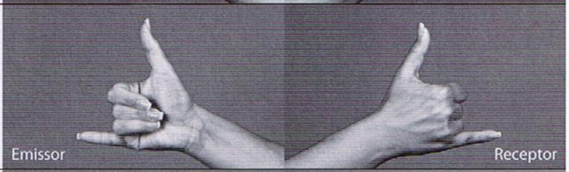
0


Emissor
Receptor

1


Emissor
Receptor

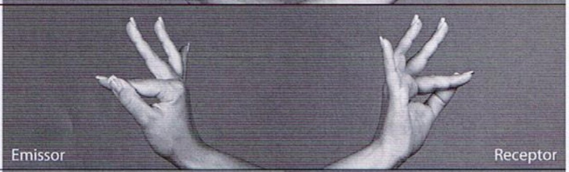
2


Emissor
Receptor

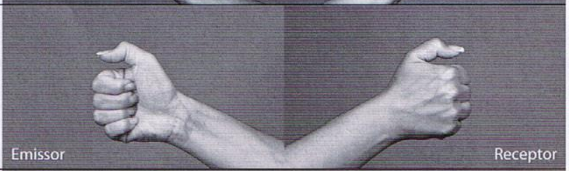
3

Emissor
Receptor

4


Emissor
Receptor

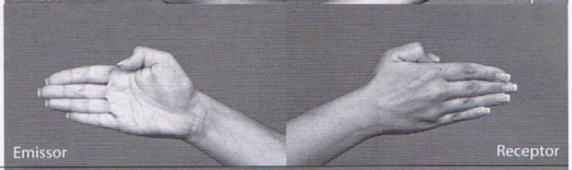
5


Emissor
Receptor

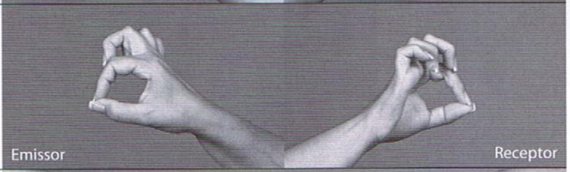
6

Emissor
Receptor

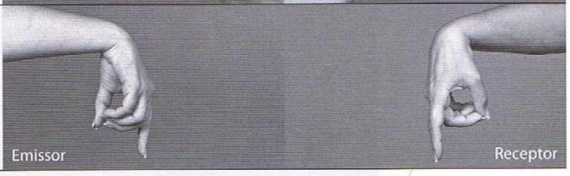
7


Emissor
Receptor

8


Emissor
Receptor

9


Emissor
Receptor