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 couldn’t 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 em 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 sistema desenvolvido 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 situações foram testadas: reconhecimento simples da configuração da mão; reconhecimento da configuração da mão em sinais; reconhecimento de sinais usando somente informação dos percursos das mãos; reconhecimento de sinais com o percurso das mãos e configuração das mãos. A primeira e terceira experiencias foram realizadas de forma a conferir o método de extração de *features*, enquanto a segunda e quarta experiencies foram conduzidas de forma a adaptar os primeiros sistemas ao problema real do reconhecimento de sinais em LGP.

A primeira e segunda experiência obtiveram taxas de acerto de 87.4% e 64.2% respetivamente enquanto as experiências respetivos ao reconhecimento de sinais obtiveram taxas de 91.6% para a experiência contendo só o movimento, e 81.3% com o movimento e a configuração das mãos.

Abstract

Just like 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 SLs and other languages is that the first one is signed and the second one is spoken, meaning that SL is a visual language. SL are the most common type of language among deaf people since no sense of hearing is required to understand it.

In this way, to build a bridge and ease the communication between deaf (and hard-of-hearing) people and people not familiarized with SL is the main motivation of this dissertation. The purposed system has as main features not being intrusive, discarding the usage of glove like devices or a setup with multiple cameras. This is achieved using the Kinect One sensor from Microsoft. Using a single device, it is possible to 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 movement information and sign recognition using movement and hand posture information. The first and third classes of experiments, were conducted, in order to confirm feature extraction method’s eligibility while the second and fourth experiments were conducted to address the problem. Recognition rates reached 87.4% and 64.2% for the first and second experiments, respectively. In the experiments concerning signs, recognition rates of 91.6% for movement data only, and 81.3% for both movement and hand configuration data were achieved.

CONTENTS

List of figures

Acronyms and Abbreviations

ASL	– American Sign Language
GUI	– Graphical User Interface
GR	– Gesture Recognition
ME	– Movement Epenthesis
NMF	– Non-manual features
PSL	– Portuguese Sign Language
SL	– Sign Language
SLR	– Sign Language Recognition
SVO	– Subject-Verb-Object
SOV	– Subject-Object-Verb
2D	– Two Dimensional
3D	– Three Dimensional

1 Introduction

1.1 Sign Language

Just like 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 SLs and other languages is that the first one is signed and the second one is spoken, meaning that SL is a **visual** language. SL are the most common type of language among deaf people since no sense of hearing is required to understand it.

Up until the late 1960s, SL were not considered real languages, being many times 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 can 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 novel 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 research about the SL linguistics.

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 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

However, according to (Bela Baltazar, 2010) in the PSL a sign is composed by 5 features being the first 3 what compose the base of any sign: **hands configuration**, **place of articulation**, **hands orientation,** and the other 2: **facial expression** and **body movement**, with equal importance and which 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”),* 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

To build a bridge and ease the communication between deaf (and hard-of-hearing) people and people not familiarized with SL is the main objective and a big motivational tool itself for the making of this dissertation, being this quest a few years old with only recent and significant breakthroughs (Capilla, 2012b).

The objective of this dissertation work goes through develop tools to collect and manage gesture data, build a training and testing database of captured gestures and develop and extend tools, algorithms and techniques for PSL recognition.

1.3 Problem Description

In Gesture Recognition (GR), one of the most difficult challenges is to turn the sensed and acquired gesture raw data into something meaningful, for example, in the context of a gesture control 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 above, 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 work with full proficiency and good results: **hand configuration, orientation of the hand, position, motion, contact point, plane** and **non-manual features.**

The position, motion, plane and the contact point of the sign are strictly connected and can be implicitly observed by analysing the hand path, which is the movement performed by the hands.

Most of the works, and usually with very interesting results (Almeida, 2011; Capilla, 2012a; Chai et al., 2013; Vogler, Metaxas, 1999), focus solemnly 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. Other projects 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 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 merge both hand depth information with the hands path to recognize and distinguish isolated signs, using only depth information in the context of PSL.

1.4 Thesis Hypothesis

- **H1 -** In the first hypothesis, it is stated that it is possible to extend and adapt state-of-the-art work in automatic sign recognition and in PSL recognition

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

(Almeida, 2011) by distinguishing signs in which the hands perform the same movement, but the hand configuration is different between those signs in PSL.

- **H2** - the second hypothesis is that an approach that uses both hand configuration classification and hand position information can outperform a system based only on the hand position, according to the problem of the first hypothesis using only depth information.

1.5 Objectives

The description of Portuguese Sign Language has its specific meanings and symbols, which differs from others. 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. It is the first objective to show that this research, is pairing with the peers in the literature, for simple and limited problems of Automatic Sign Language Recognition, with a specific application to PSL.

From the hypothesis described above**,** the following objectives were enunciated:

O1 – Posture Recognition System - specify, develop and test a system that uses fully 3D data structures to define, describe, record and classify the depth data of the hands and use that data in an efficient and effective manner, for real-time automatic posture recognition for PSL hand configurations.

O2 – Movement Recognition System - specify, develop and test a system that uses joint positions from both hand to define, describe, record and classify a sign in PSL, for real-time automatic sign recognition.

O3 – Sign Recognition System – specify, develop and test a system that, using the systems developed in O1 and O2, defines, describes, records and classifies signs in PSL using depth images from hands and depth joints. Comparing the results of O3 and O2 addresses specifically the second hypothesis (**H2**).

O4 – Software Application – develop and test a software application that, using systems O1, O2 and O3, lets the user record and view data from and for the hand configurations and signs systems in a useful way.

This work has the aim to collaborate with the research in order to increase the social inclusion of more than 100.000 hearing impaired people in Portugal (Bela Baltazar, 2010). Yet, are not objectives of this thesis to propose a final and unique solution to the immense problem that SLR presents.

1.6 Document Structure

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

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

Chapter 3: In this chapter, the details of the proposed system architecture and the application Graphical User Interface (GUI) are presented, its design and implementation, and the description of the components for the developed system. It is also presented the data collection methods and properties, details about the Corpus used and recording specifications.

Chapter 4: This chapter presents the results of the methods used for the Posture recognition and for the Sign recognition processes. The comparison between other works is also included in this chapter.

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

2 State of Art

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 capture. 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. Decide which sign/gesture represents, or even convey meaning to entire “sentences” is the result of the classification part.

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's derived from (Kadhim Shubber, 2013). Used worldwide by a multitude of individuals, being them people from the deaf communities and their teachers, or people associated with them, such as family or friends, have their citizens often segregated from the rest of the society by the difficulties in the communication with the rest of the people (Almeida, 2011).

Typing and/or writing in Portuguese, or any other written language, isn't straight forward for deaf and hard-of-hearing people. For those who have been deaf their whole lives it's akin to learning a new language.

Currently it is not possible for deaf and hard-of-hearing people to communicate with each other 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, it has been explored for many years, is still a challenging problem for real practice. A more cohesive and robust approach was developed by Microsoft (Chai et al., 2013) recently for the problem of the recognition of SL.

Despite that fact, in the particular case of Portuguese Sign Language remains unknown an efficient system to perform automatic recognition of PSL gestures and communication.

The Kinect, by being generated within a gaming purpose, rapidly saw its original purpose to be adapted to various usages because of its low-cost as depth sensor, being this activities very distinct, 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 it is being analysed depth information and not only colour information. The older version of the Kinect device had most of its problems confined with the low resolution available, turning the recognition process a bit harder (Khoshelham, Elberink, 2012). This process is expected to suffer a substantial change with the Kinect expected to be used in this project, the Kinect One sensor.

By using the segmentation from one posture to another and combining also the trajectory of the gesture (Chai et al., 2013),it is possible to use machine learning technology and pattern recognition technology to make the final decision of what's 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 that feature extraction was almost inexistent and the measurements were directly used as features.

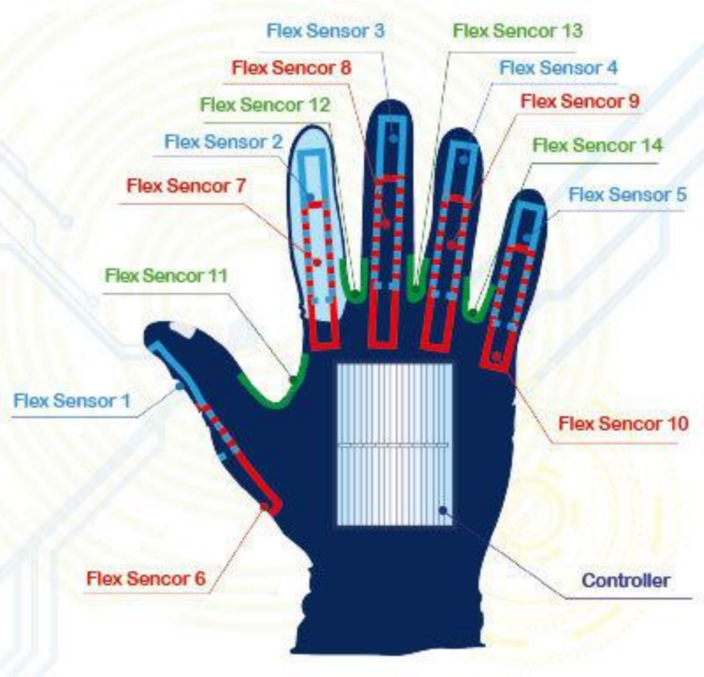


Figure 1 – Data glove example. Usually, this glove devices feature precise data about the hands parts positioning in 3d space and also including accelerometers, giving other information like velocity/acceleration, etc.

These gloves 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 contrary to the data extracted from the gloves systems that were concise and accurate compared to all 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 on simple video camera.

- Gloves systems can be used regardless of the environment, whether complex backgrounds or signer garment.

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, it was beginning to be acknowledge that the hand tracking stage of the system does not attempt to produce a fine-grain description of hand shape, 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 this systems to solve the hands segmentation issue there were needed algorithms such as “skin detection algorithms” or other methods to segment the hand. Many works were done with this approach or similarly, ranging from (Freeman, Roth, 1995; Parish et al., 1990) to (Pashaloudi, Margaritis, 2002b; Wilson, Bobick, 2000) and (Wang, Quattoni, 2006; Yang et al., 2010). Other approaches to solve this problem were 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 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.

Plus the above 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, 1998).

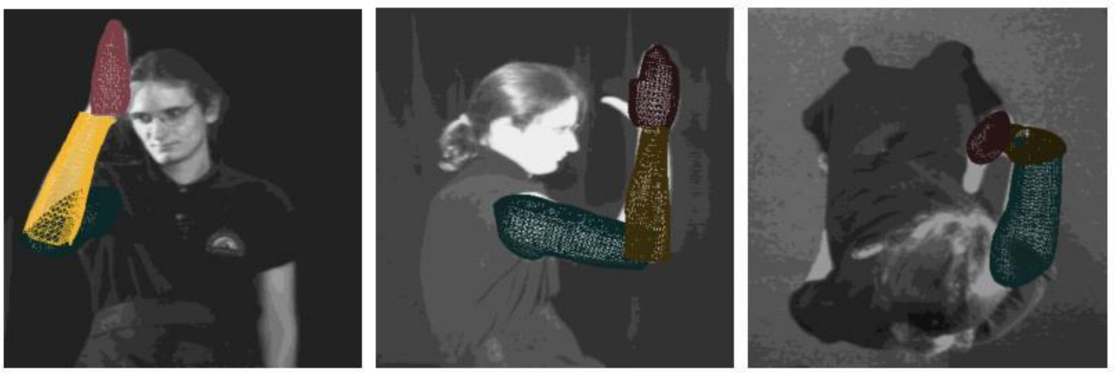


Figure 2 – 3 axis camera system (Vogler, Metaxas, 1998)

This systems don't give much flexibility to "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, requiring for that 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 this 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, while 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 problem, 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, easing this way 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 public in the last September. It features a RGB camera that outputs 1920x1080 pixels of colour data, 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 resolution on the depth image, the new sensor achieves has an improved accuracy on the depth values, as can be seen on Figure 3.

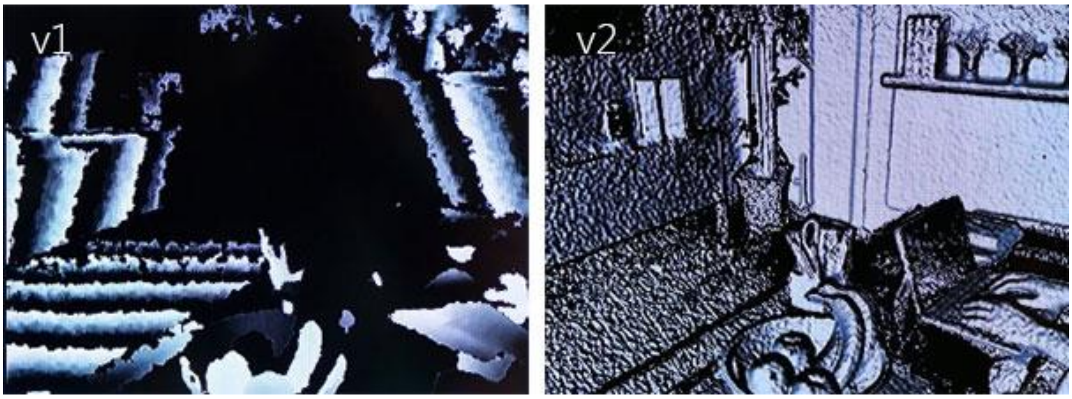


Figure 3 – Kinect versions depth comparison. The left image corresponds to the first version of the Kinect sensor, released in 2008, while in the right it is the same scene with the new Kinect One sensor. (Microsoft, 2014a)

2.3 Analysis and Classification Methods

Due to the nature of the problem of SLR, soon became usual to establish the comparison between the speech recognition and because 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, making the use of HMMs plausible for modelling both processes. Both systems have to consider context and 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, hence the sound, of following ones. This doesn’t exactly adds 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, movement 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).

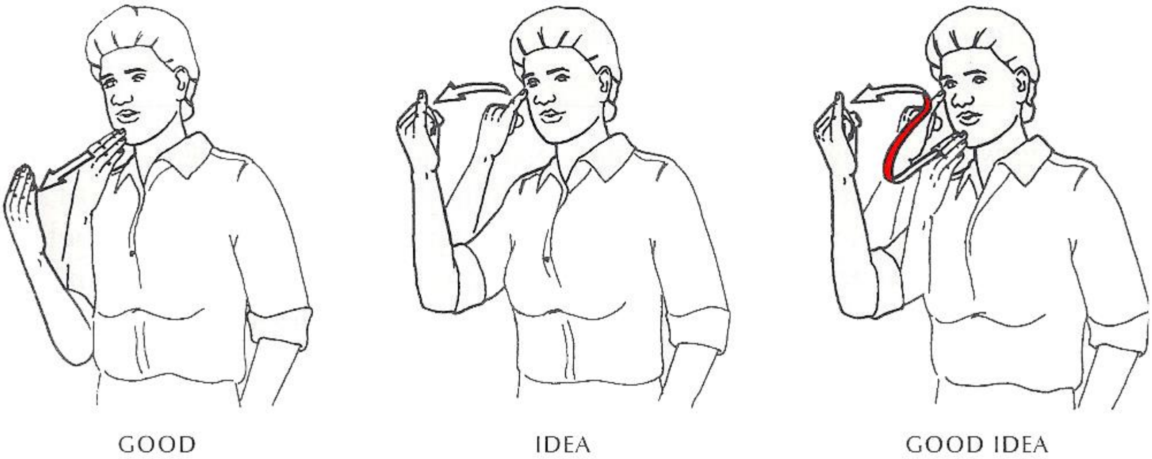


Figure 4 – The red path in the left image represents the movement epenthesis

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) didn’t focus on the typical finger signing usually focused till then and instead, focused on gestures, which represent whole words, since real SL conversations usually can only proceed at the normal pace of spoken conversations due to this 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) they improved his 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 set were required.

Pashaloudi and Margaritis (Pashaloudi, Margaritis, 2002a) 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, in this work it was 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, 2012b), 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 and 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 HMMs 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-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 Chinese Sign Language words. In their database, each word was recorded by 5 times.

2.4 Summary

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

After the data collection explained, it was enunciated 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 GR.

3 PhySaLiS

In this chapter, the developed system, referred as PhySaLiS (Portuguese Sign Language System), is presented. The system requirements derived from the objectives are enunciated in the first section, being followed by the description of the system architecture and the application GUI in the second section. The following section, 3.3 describe the Corpus used for both postures and signs systems and the specifications of the data collection regarding the recordings such as number of signers, repetition of words, etc. The section 3.4 describes the pre-processing done to the raw data before the feature extraction is performed. 3.5 explains the methods used for the feature extraction process in the hands configurations and in the movement data, and finally, section 3.6 explains the classifier creation methods and specifications.

3.1 System Requirements

To achieve the objectives purposed for this thesis, multiple requirements were set accordingly to each of the objectives. 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/Apply a technique to extract and normalize hands information from the depth image to be used for the posture system	O1,O3	Completed
SR2	Develop/Apply a technique to extract and normalize hands information from the joints information to be used for the sign system	O2,O3	Completed
SR3	Collect hand configurations performed by multiple users, using depth data, allowing the creation of a structured dataset for training and testing	O1, O3	Completed

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 sign recognition task	O2, O3	Completed
SR6	Develop a classification technique appropriate for the posture recognition task	O1, O3	Completed
SR7	Use a previously collected dataset to train and test the 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 Isolated Gestures, and to do that, a two layer system architecture is proposed.

Regardless of not being immediately visible in the diagram (Figure 5 – 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.

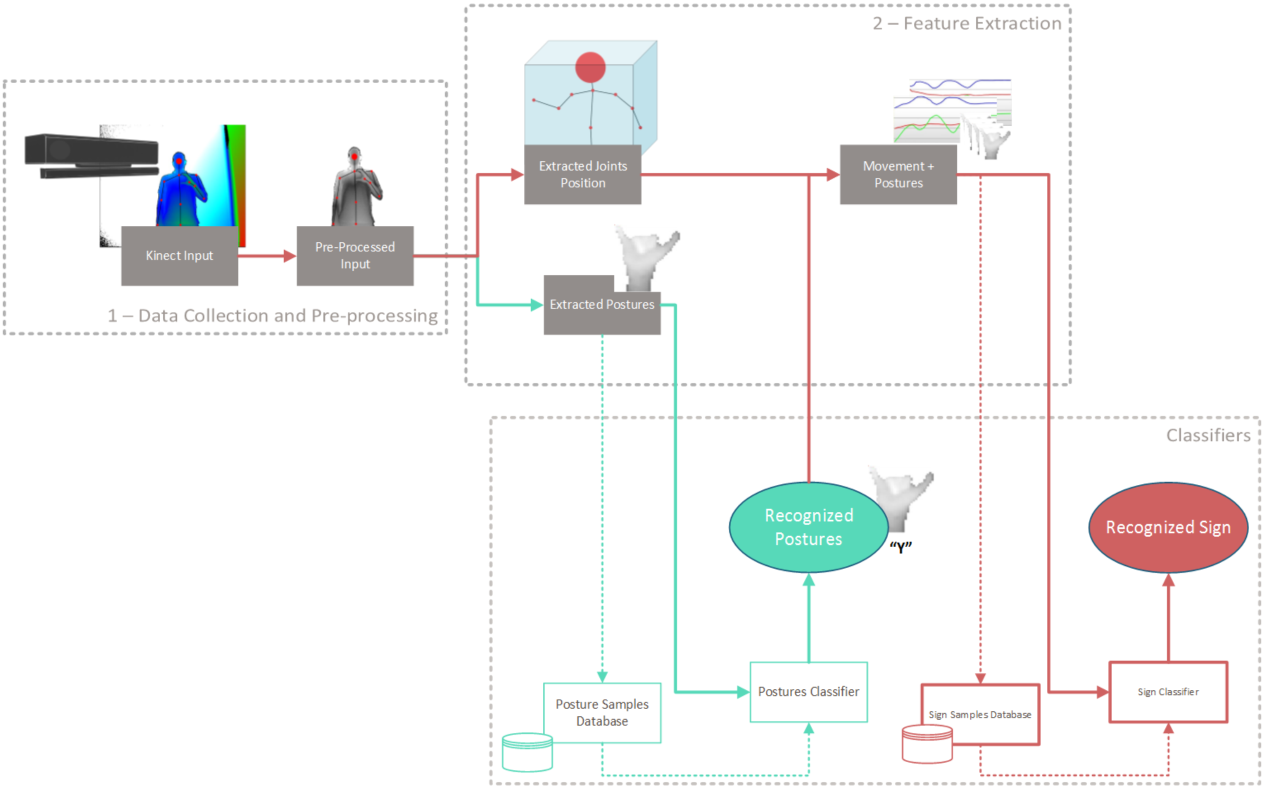


Figure 5 –System Architecture. The first module comprises the data collection and pre processing methods. The second one handles Feature Extraction, while the last one includes all the classifiers work, since creation to testing and usage.

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

The Pre-processing module takes care of the Depth Data and Joint information 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, be that the hand size or the arm size, or the application of Erode and Dilate filters to the depth stream (for noisier environments).

The second module, “**Feature Extraction”,** adds the functionality needed for the feature extraction and processing. This feature extraction will provide information for real-time automatic GR and provides the training samples set, a critical data source for the GR process, which will be stored in a database.

Like in (Almeida, 2011) **postures** refer to static hand configurations and **signs** to 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 of a set of posture acquiring functions to get and convert posture dimensions, 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* data normalization.

Finally, the **“Classifiers**” module handles the management of the collected data and the creation of the system classifiers. It allows the creation of both Support Vector Machines and Hidden Markov Models 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.

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.

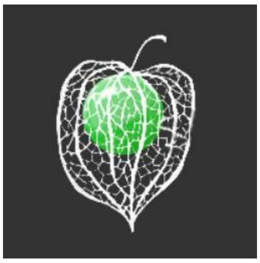
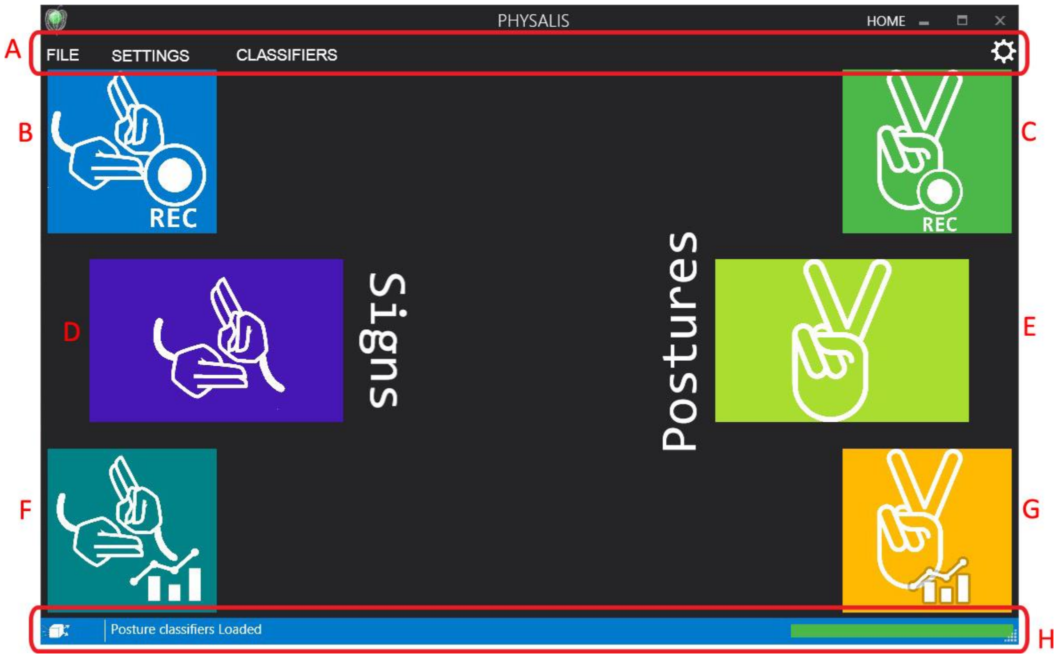


Figure 6 – PhySaLiS logo

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.

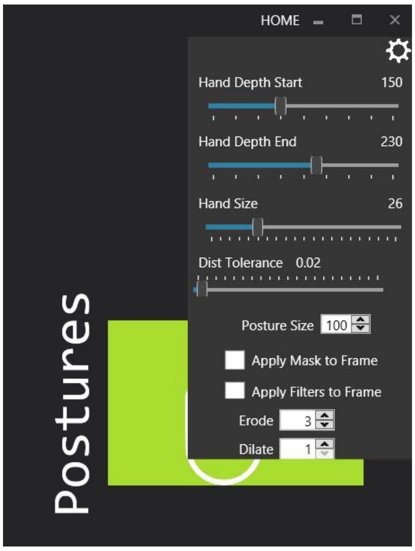
The application was developed in the *Metro* style, being composed by simple lines and effects. The "Home" screen is as follows:



Posture classifiers Loaded

Figure 7 – PhySaLiS home screen. At the centre left are the buttons to access the functionalities concerning signs, at the right, functionalities concerning postures.

In this screen it is possible to go to any of the major functionalities of the system. These main functionalities 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 and E).



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

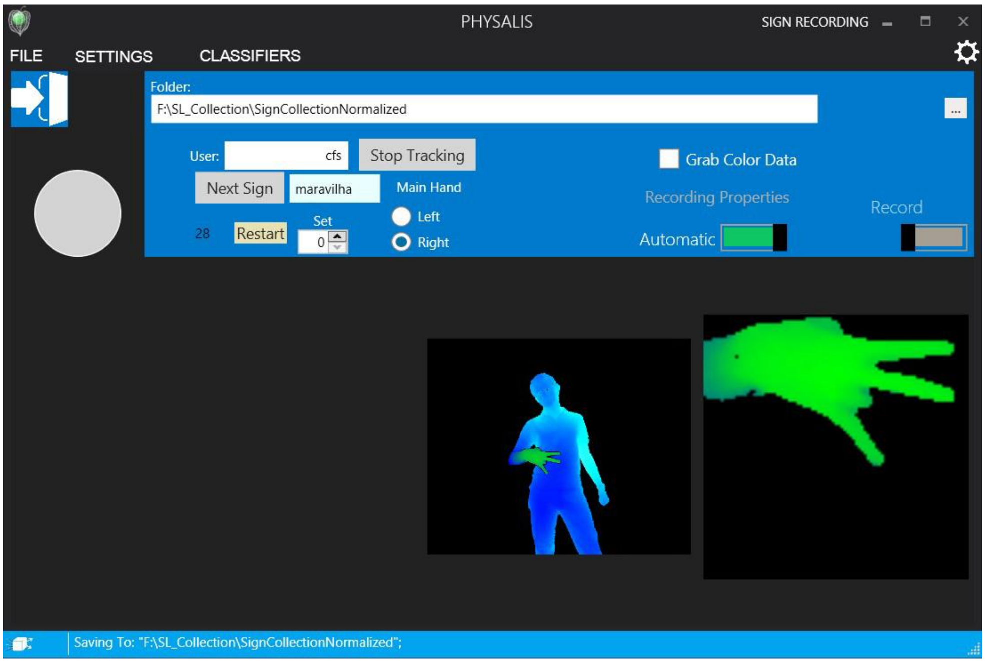
Figure 8 –Data collection options. These options relate to normalizations done in both postures and signs systems. The hand depth start, end and size define the bounding box that extracts the hand depth image. The distance Tolerance to detect if the signer is moving or not.

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

In the bottom of the application there is a status bar that shows information regarding operations that are done by the system and other status messages, represented with label H (Figure 7).

3.2.1.1 *Data Recording Screens*

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



Folder: F:\SL_Collection\SignCollectionNormalized

User: cfs

Next Sign maravilha

28 Set 0

Main Hand
○ Left
● Right

□ Grab Color Data

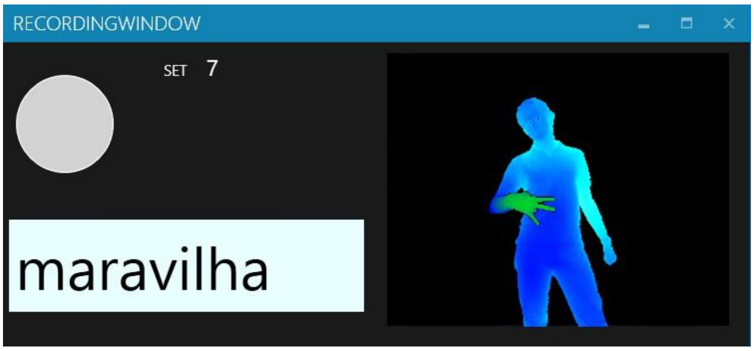
Automatic

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

Figure 9 – PhySaLiS sign recording screen. In this screen the user can create a new signer folder in which he can record new instances of signs. For the recordings, the user can define which hand behaves as the main hand, the sign and set to record and if the sign is recorded automatically or manually.

In B - sign recording screen (Figure 9) - it is possible to choose where to save the data and initiate the recording giving a name for the signer (each signer goes in distinct folders). After given the name for the signer folder, it is possible to record the signs either automatically, being the “start recording” and “stop recording” signals given by estimating movement from the hand joints. For the automatic method, it is required that the signer is in a standing pose (Figure 16) before the start signal, otherwise the recording won’t activate. This way all signs start from the same place/ pose. For the manual mode, a simple recording button is available, so the user can start and stop the recording.

The user should also select which sign and which instance of the sign is being recorded so the recording has the right label for the classifiers. This mode also shows another smaller window to pass to another screen, or projector, to help in the recording process, being the signer able to see this "*helper*" window in the other screen. It contains simple information as the sign to perform, the information weather the system is recording, ready to record or idle, and the depth stream visualization.

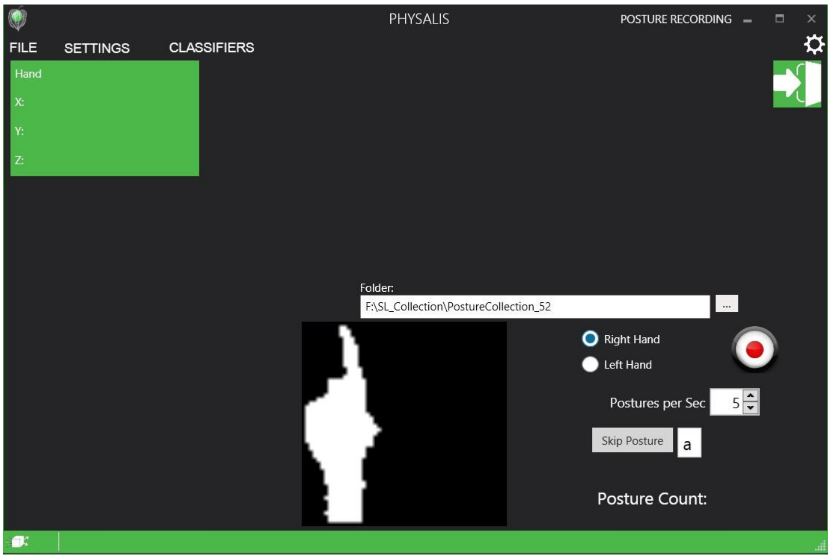


SET 7

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) – as 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 done, this way it is possible to save multiple instances of the



Hand
X:
Y:
Z:

Folder: F:\SL_Collection\PostureCollection_52

● Right Hand
○ Left Hand

Postures per Sec 5

a

Posture Count:

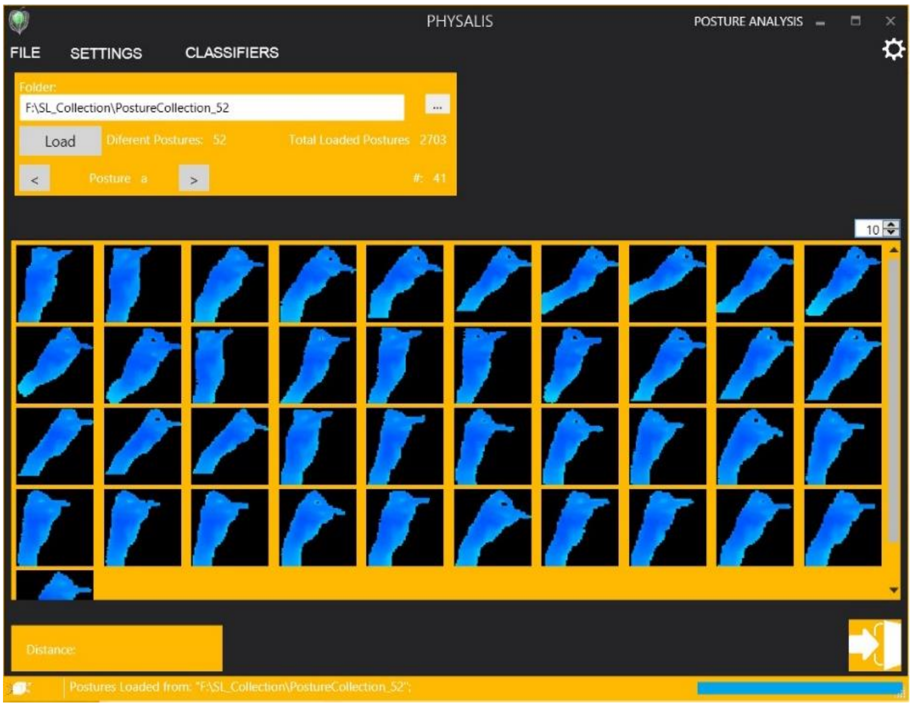
Figure 11 – PhySaLiS posture recording screen. In this the user can select which hand to track to perform de recording, select which posture to record and how many images per second are captured.

posture much faster (with small variations or not). By default, the right hand is recorded, but it is choosable which hand to record, either left or right.

3.2.1.2 Data Analysis Screens

F and G (Figure 7) give the user access for both Postures and Signs analysis screen. These screens serve the purpose of inspecting the 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

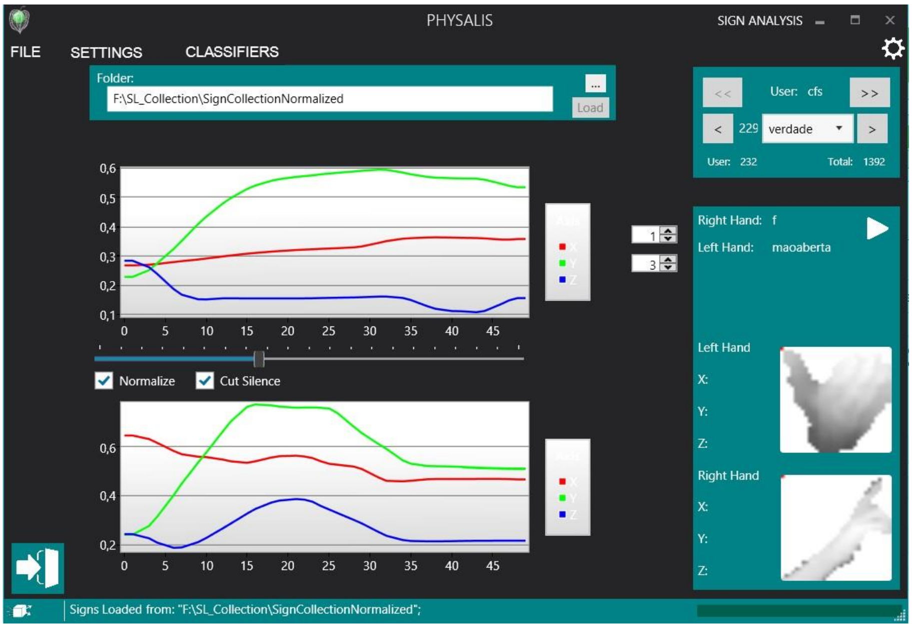
10

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 Postures Analysis Screen (Figure 12), each instance relate to an image of the hand, being each class composed by multiple images. In signs, each of the instances is composed by a movement and two images per frame, that correspond to the hand depth images, and it is possible to navigate through 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

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

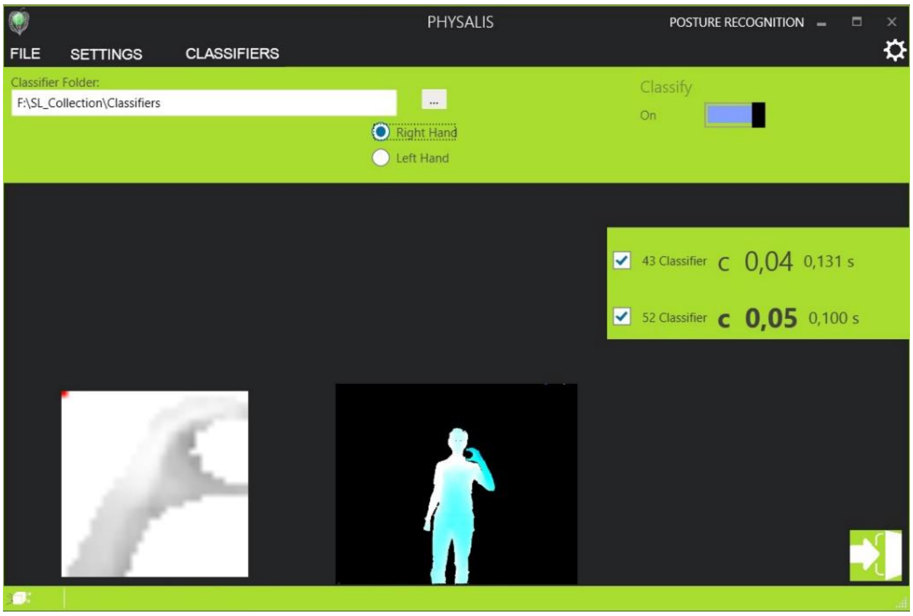
Figure 13 – PhySaLiS sign analysis screen. This screen gives visual representation of the data acquired and treated concerning signs. It is also possible to view any sign for any user recorded.

In the movement analysis screen (Figure 13) it is possible to navigate between all the instances of movements previously recorded and loaded in the system. It is also possible to observe the label of each hand for the current movement and for the current frame (classified with the 43 postures classifier), and to 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 D and E of the home screen (Figure 7) lets the user 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. The result shown in the classification is the classified hand configuration that most occurred in the last 10 recorded frames.



Classifier Folder: F:\SL_Collection\Classifiers

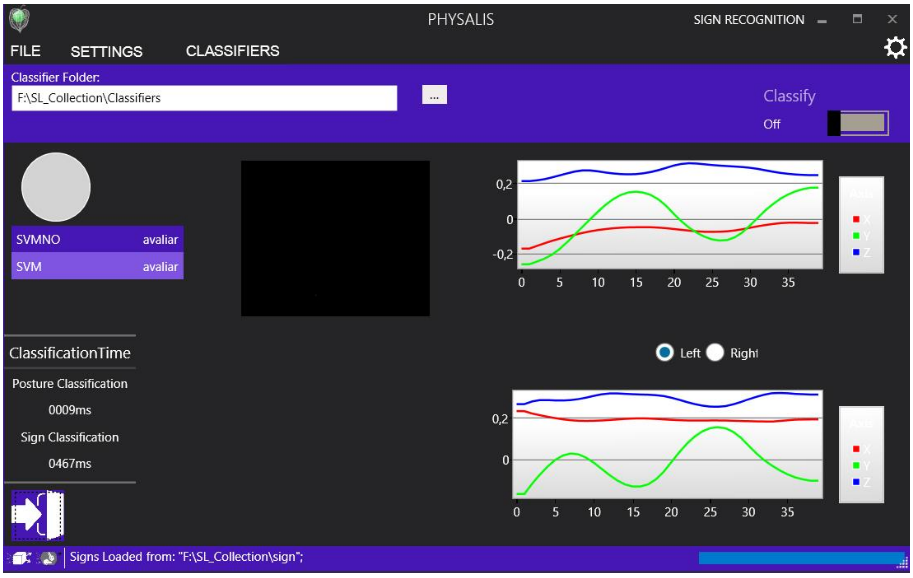
● Right Hand
○ Left Hand

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 postures with either left or right hand and to practically experiment the posture classifiers developed with the system.

Similarly, in the Sign Recognition Screen (Figure 15), the user should choose the main hand of the signer, and both 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

● Left ○ Right

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 main hand and to practically experiment both sign classifiers developed with the system. At the right it is possible to choose which hand is the main hand and to observe the chart of the performed sign.

3.3 Data Collection

3.3.1 Setup

To collect the data for this work a setup with Kinect One sensor was used which had the following constraints:

1. No direct sunlight in the room in which the recordings were 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, 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;
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 had to be calibrated regarding the signer size, task done by a simple tool developed for the purpose, which, capturing the signer in a standing pose, estimated the arm length. After having the arm span information, the collection proceeded by simply making the selected sign. All signs started to be recorded automatically from the standing pose (Figure 16), being the start signal given by an estimated amount of movement, and the same happened to the stopping signal.

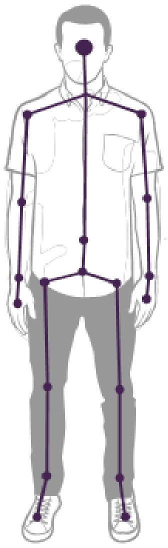


Figure 16 – Example of a standing pose (Microsoft, 2014b). *The back and legs stay straight, both arms fall along the torso and the head points forward.*

3.3.2 Corpora

An important aspect for any study involving human participants is to obtain the approval of a regulated ethics committee. As such, a detailed description of the data collections and associated studies/experiments were submitted for ethics approval. 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 features and that they could quit at any time. The experiments here reported have been evaluated and approved by the ethics committee of the University Institute of Lisbon (ISCTE-IUL) regulated by the dispatch nº7095/2011.[JF1].

Signs

The Corpus, or vocabulary, used in the system concerning **signs** was chosen to better address the problem at hands and to better illustrate the purpose of this work. 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 the 29 words were:

- Concerning individual signs :
    A. The "auxiliary" hand behaves similarly to the main hand (mirror);
    B. Only the main hand moves (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, 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 enunciate the problem this thesis tries to address.

The signs used were:

Table 3 – List of signs selected to compose the 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, being the first one done above the chest and 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*)

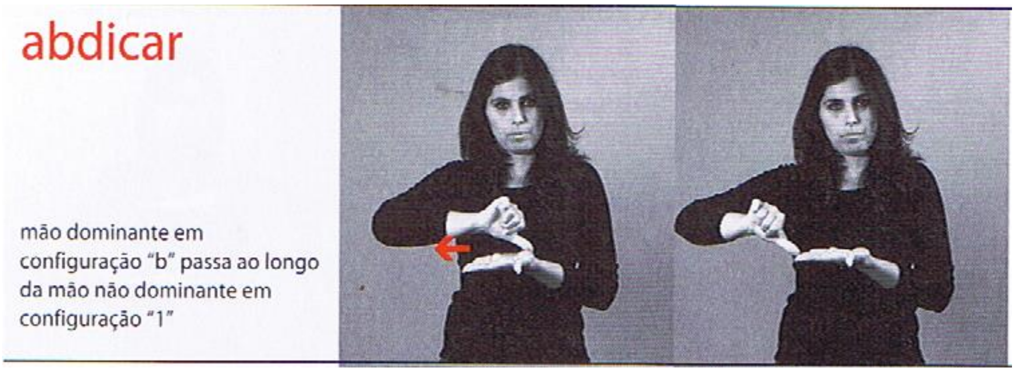
The signers were previously briefed about the collection method and no signer had any background on PSL, having them to learn each of the performed signs prior to the data collection.

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, and for each signer, eight repetitions of each of the 29 signs were collected.

Postures

For the hand configuration sub-system two different recordings and consequently two classifiers were done. For the first recording, 52 different postures were recorded. Each of those postures were recorded around 50 times with some small variations, such as small hand orientation or fingers angles and was recorded with 2 signers. Those variations were minimal in order not to change the posture. Basically, the postures done were in the form of finger spelling, 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.

The second recording for the hands configuration system was the one used in the sign recognition and was composed by 43 postures. For this recording, 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 signs. One example of this property is the word “*abdicar” – abdicate:* 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’”

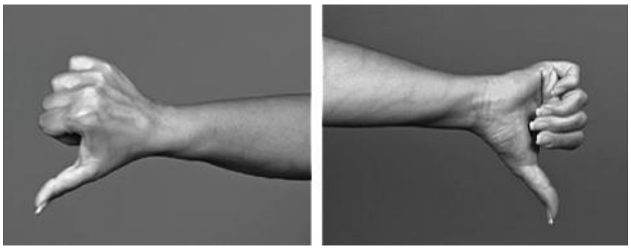


Figure 18 - Hand configuration "q" (Bela Baltazar, 2010). *On the left is the view of the signer, and on the left is the view of the "receiver"*

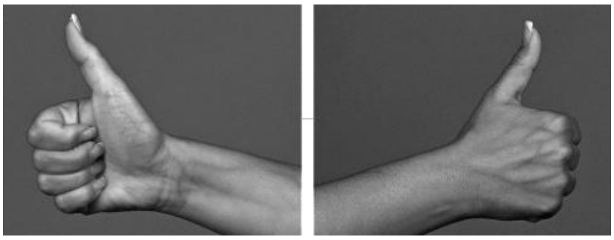


Figure 19 - Hand configuration "b" (Bela Baltazar, 2010) . *On the left is the view of the signer, and on the left is the view of the "receiver"*

For both signs and hand configurations data collections, the above setup conditions were used.

3.4 Pre-processing

Concerning the sensor data acquisition, the data is acquired via de Kinect One Sensor described above. From this sensor, only the Depth Stream/ output and the derived Body Stream is used in this system.

Generally, one of the first crucial steps in computer vision systems is the Segmentation. In this system, the first segmentation to be done is the Background removal, segmenting this way the background from the user, or signer.

3.4.1 Background Removal

Consider $D_1, D_2, \dots, D_{30}$ as a sequence of $512\ x\ 424$ of depth pixel values (16 bits) for the time instants $1, 2 \dots 30,$ respectively and MD the target mask to be created.

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

$$MD = D_1$$

- For the instants *n* such as $1 < n < 30$:
  - D_n As being 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 MD represented by dn_{xy} and md_{xy} respectively.
  - $If\ dn_{xy} < md_{xy}\ then\ md_{xy} = dn_{xy}$

At the end of the 30 instants, MD will be the depth image containing the minimum depth values (or closest) observed during those frames. Assuming that the sensor only “sees” the scenario, meaning this that the user must 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 5cm.
  - Considering D_n as being the depth map for the instant $n > 30$.
  - $If\ d_{xy} \geq (md_{xy} - \lambda)\ then\ d_{xy} = 0$

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

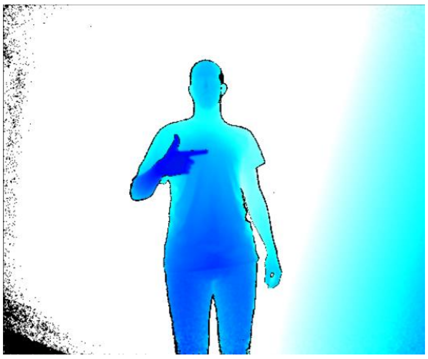


Figure 20 – Kinect sensor original depth input. The colour in the depth images is simply a form of representation, since the depth images contain only one value per pixel, corresponding to the distance of that point in space to the sensor in millimetres.

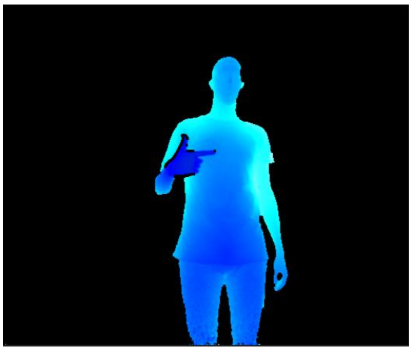


Figure 21 – Kinect sensor depth input after applied the background subtraction. With background subtraction - the body silhouette becomes the only visible object as it is the only object that moves in the sensor field of view.

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 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 depth image 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 in the Body structure and the depth image comes from the previous step, Background Removal.

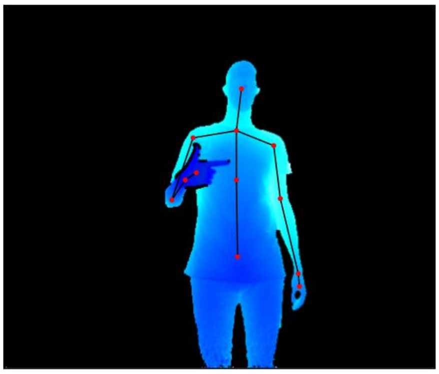


Figure 22 – Both depth and body inputs from Kinect in one frame. The red dots are the joints given by the Kinect. Each joint is represented by a point in 3D space.

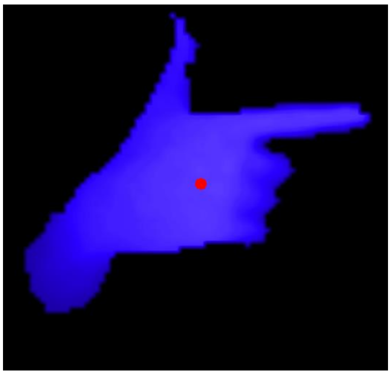


Figure 23 – Extracted hand depth image width the hand joint depicted as a red dot

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 (camera space) to the depth space (depth image space), mapping real world coordinates into pixels from the depth image.

3.5 Feature Extraction

The features to be passed for the classifier come from the hands 3D coordinates (joint position) and from the hand image.

Being the sign composed by many instants, and having in each instant both hand positions and both hands images, some normalizations are needed before passing the data onto the classifier, whether for training or recognizing purposes. These normalizations occur in two forms:

- The hands positions normalization for each frame, which results in the **Hands Paths**.
- The **hands depth images normalization**, which will be used in a first classifier that result in a **posture label** in each instant for each hand, representing the hands configurations for that instant.

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 reference the body **Spine Centre** joint depicted in (Figure 24).

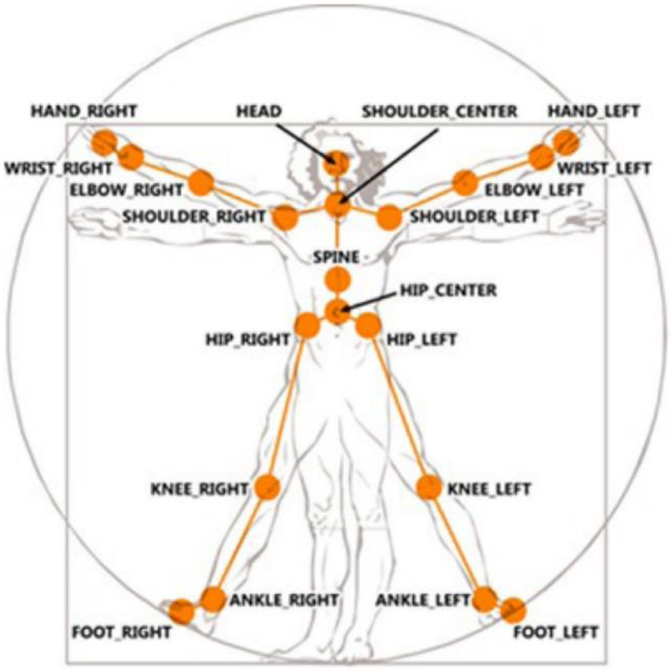


Figure 24 – Kinect body joints. The Spine Centre joint used in the normalization method is the joint above the "HIP_CENTER" and below the "SHOULDER_CENTER"

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

In this raw data, it is possible to see that the only the coordinates X and Y are according to the spine centre joint. This brings an obvious problem that is, if from recording to recording, the signer is at distinct distance from the sensor, the hand z value will vary greatly. In order to bring down the hands to the same reference, 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.

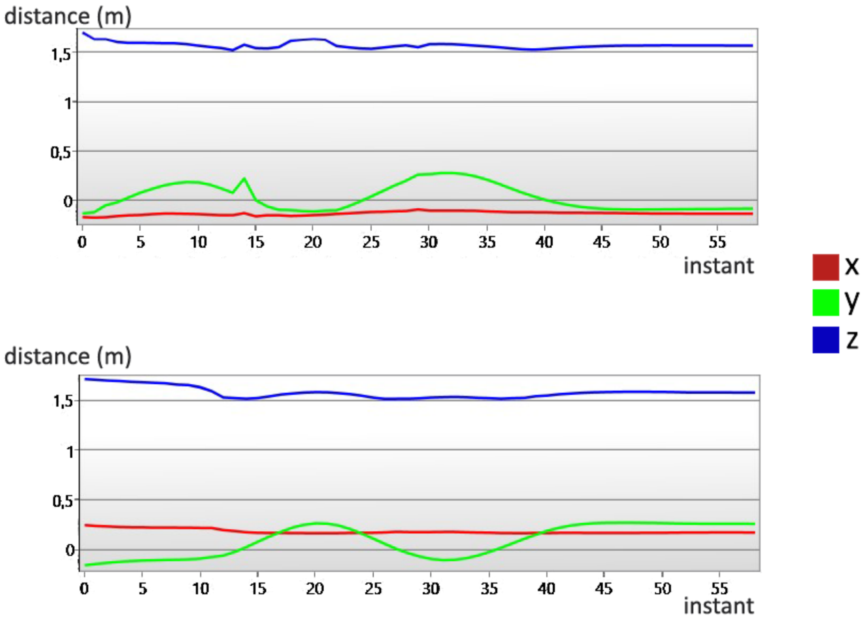


Figure 25 – Raw Hands data of an instance of the sign "avaliar”. The top chart corresponds to the left hand path while the bottom one is the right hand path chart. Each frame/instant contains 3 values, being them the coordinates X, Y and Z.

Being $H_{(x,y,z)}$ any of the hands joints Point given by the sensor and $SpineC_{(x,y,z)}$ the Spine Centre Point, the first 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.

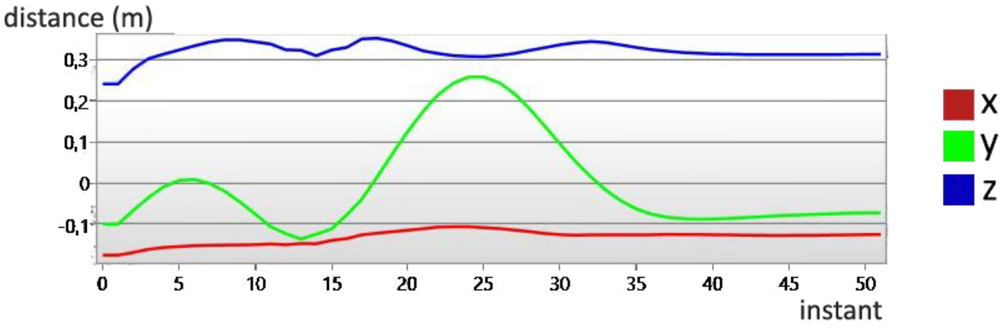


Figure 26 – Hand data of the same instance of the sign "avaliar” shown in Figure 25 after the first step of normalization. Only the left hand chart is shown. Each frame/instant contains 3 values, being them the coordinates X, Y and Z.

After the classifier was created to test this approach, it was noted that signers with distinct heights, hence distinct arm span for that matter, had distinct results. As the previous normalization step did nothing to ease this problem, another method was needed. This issue was diminished 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 in the system that takes into account the joints from the hand to the shoulder and calculates the total distance between the joints.

This is the same as creating a virtual box around the signer, which varies with the signer arm span, and for that, firstly it is needed to define the boundaries of said box for each instant:

$$Min_x = SpineC_x - AS\ , Max_x = SpineC_x + AS$$

$$Min_y = SpineC_y - AS\ , Max_y = SpineC_y + AS$$

$$Min_z = AS\ , \qquad Max_z = -AS$$

Where AS is the Arm Size value

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

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

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

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 space.

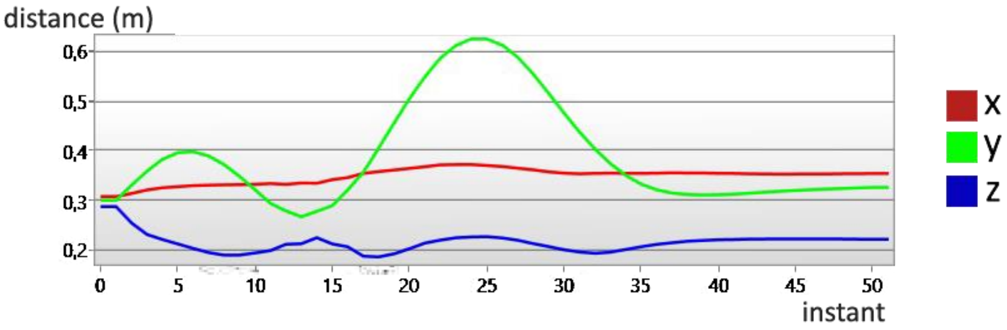


Figure 27 – Hand data of the same instance of the sign "avaliar" shown in Figure 25 after the second step of normalization. Only the left hand chart is shown. Each frame/instant contains 3 values, being them the coordinates X, Y and Z.

Once that the recording of the movement starts and stop automatically, being the start signal given by the start of movement, and the end by an estimate of movement which might not be precise due to noise on the joint detection by the sensor, the next step is to remove the frames at the end of the recorded movement with irrelevant information. This noisy frames are recorded due to the system recording the sign automatically with an estimate of movement. To compensate for this noise, some frames of the end part of the movement. By observing both hands movements, starting from the end of the movement, all 6 coordinates (3 coordinates from each hand) are observed in a window of 3 frames. While all the coordinates, in this 3 frame window have variations lower to a fixed threshold, then the frame is eliminated from the movement. It is possible to see, comparing Figure 27 and Figure 28, that after this method is applied, from the final to near the frame 40, the frames were eliminated.

The last step to create the feature vector is to normalize the movements' size, that is, make all movements have the same number of instants. A movement is created by all the positions from both hands along the size, in the form of an array per each of the axis (x, y and z) for each of the hands, giving a total of six arrays to describe a movement.

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

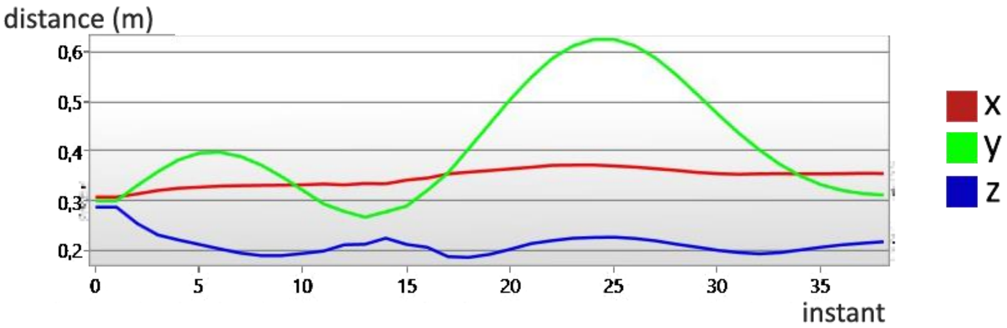


Figure 28 – Hand data of the same instance of the sign "avaliar" shown in Figure 25 after the third step of normalization. This third step removed information at the end of the sign, in which the hands are halted, hence considered "silence". Only the left hand chart is shown. Each frame/instant contains 3 values, being them the coordinates X, Y and Z.

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

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

In the end of the process, the movement graphic looks like:

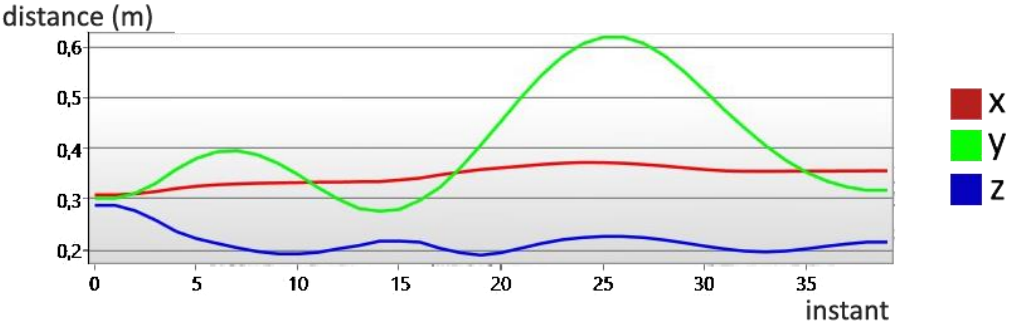


Figure 29 – Hand data of the same instance of the sign "avaliar" shown in Figure 25 after the final step of normalization. This fourth step normalizes all signs to the same length. Only the left hand chart is shown. Each frame/instant contains 3 values, being them the coordinates X, Y and Z for each hand.

3.5.2 Hands Depth Images Normalization

The original Hand image concerning the depth stream, given in the subsection Hands Segmentation that can be 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 needed for the sign, and as in a written language, there are signers who are right handed while others are left handed. Because of the previously stated, it is needed that the system can compare both left and right images as equals.
b) Signer distance to sensor - since the hand depth image is taken from the total depth image of the Kinect output, in which objects closer to the sensor have values closer to 0, and objects farther have increasingly bigger 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 less image proportion than a bigger hand, resulting in images with bigger areas without information for smaller hands;

To address the problem (a), when the hand depth image was correspondent to the right hand, it was only needed to flip or mirroring the image by the vertical axis, being this way, left and right hand images equivalent.

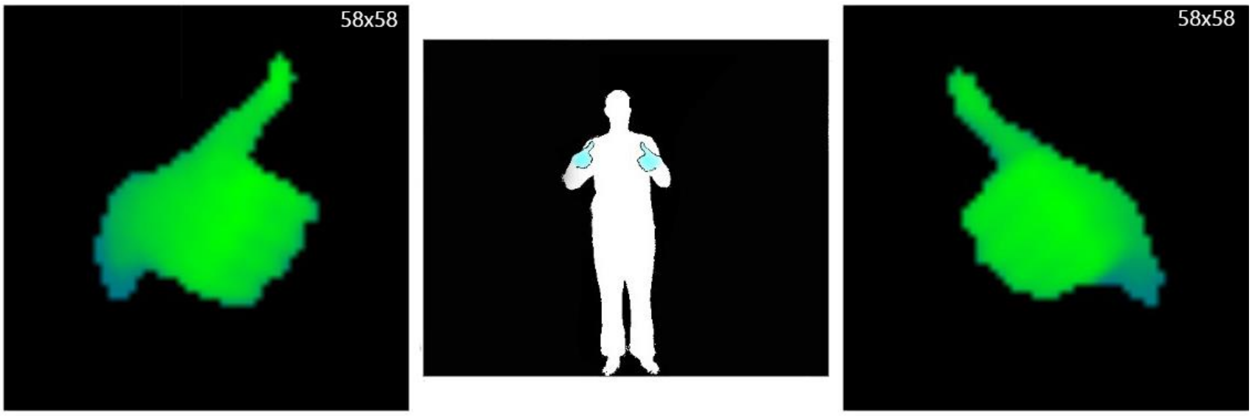


Figure 30 – Depth data input for the both hands. At the left side is the left hand and at the right side the right hand. The middle image is the original depth input after applying the background extraction. At the top right corners it is possible to see the original size of the hand image. The colour 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 is only 4.5 m, hence 4500.

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

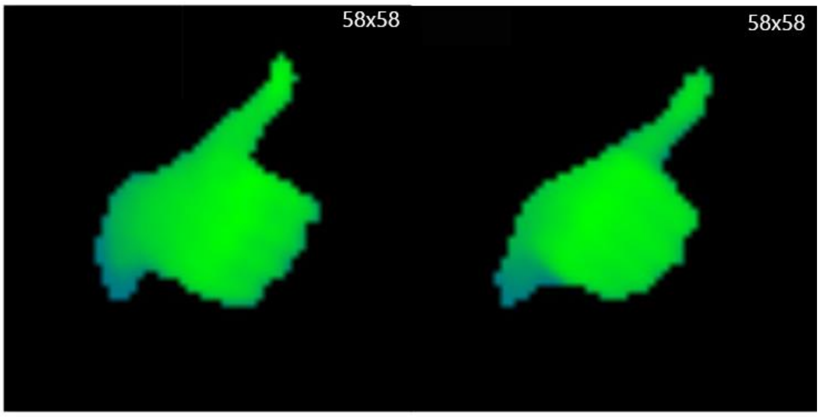


Figure 31- After mirroring one of the hands, the images become very similar. At the top right corners it is possible to see the original size of the hand image. The colour representation is merely visual since the input for depth values varies from 0 (the sensor) to the max range the sensor can infer depth (accurately is only 4.5 m, hence 4500.

For problem (b), the solution to eliminate the variable distances of the hand to the sensor was to assume that the closest value of the image, hence the one with the lower value that was different from 0, had 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 as 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 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 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 ratio of the hand portion in the image (called “Stretch Ratio”), and another method that does not conserve the ratio of the hand (simply called “Stretch”). Figure 32 shows the results from the first method – “Stretch Ratio”.

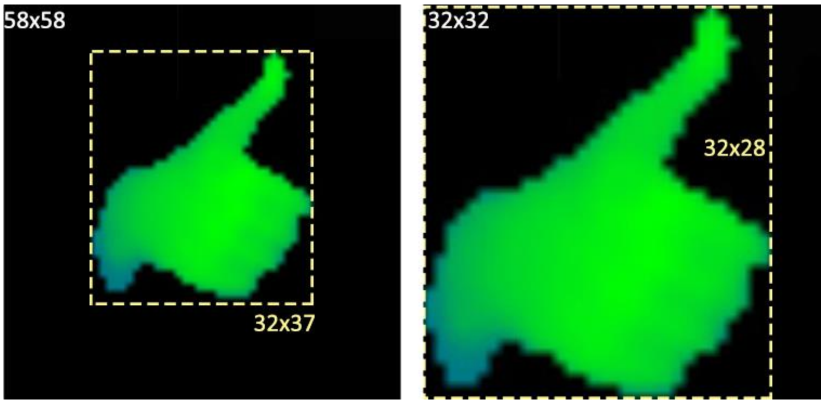


Figure 32 – At the left side is the hand depth image before applying any stretch method while in the right side is the result of the “Stretch Ratio” method. In both images, in white is the total size of the image while in yellow is the size of the square that contains the hand part in the image. The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

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

The other tested method, not conserving the ratio of the hand portion of the image, distorts the original image, in some cases in an insignificant amount. Despite not seeming a natural approach to the problem, tests were conducted to verify this approach. Figure 33 show the result the “Stretch” method.

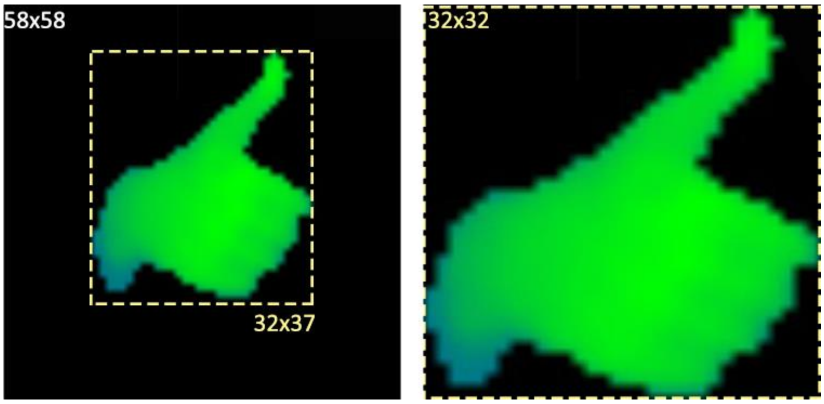


Figure 33 - At the left side is the hand depth image before applying any stretch method while in the right side is the result of the “Stretch” method. In both images, in white is the total size of the image while in yellow is the size of the square that contains the hand part in the image. The size is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

Similarly to what happened in the previous method, the pixel values are calculated by linear interpolation according by horizontal lines.

3.6 Classifiers

This section specifies the methods that have the role of creating and managing the posture and sign database and sample data (known as the training set) for the **learning** and **classification** processes.

To be able to classify recorded postures and signs, the system must first be trained to create those same classifiers. This training is accomplished in both scenarios using SVM on the collected dataset.

SVM can only solve binary problems, however, several approaches have been suggested to perform multi-class classification using SVM. In this thesis, it is used 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 should 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 SVM 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 posture classifier was created to verify the capture normalizations method.
- The second one was created to merge with the sign classifier.

The first classifier had a dataset composed by 52 different postures, where this postures had minor variations, while the second one had 43 different postures to address the problem of hand configurations varying position and palm orientation in signs, as described in subchapter Corpora. 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 above), 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 hand configuration recognized.

To train and test the classifier, a k-fold cross-validation method was applied. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. A k=10 value was used and despite the subsamples being randomly generated, it is assured that each subsamples has various instances from each hand configuration.

In each fold of the algorithm, of the 10 subsamples, 1 subsample is used as the validation data for testing the model, and the remaining 9 subsamples are used as training data. This process is repeated then 9 more times (performing a total of 10 folds), with each of the 10 subsamples being used exactly once as validation data. The results from the folds are 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 validation exactly once. No further testing was done with data not used in training neither in validation, mainly due to the low amount of data available.

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), creating the respective classifiers and saving 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 as well as the training and validation accuracies for each of the eight folds.

3.6.2 Signs Classifier

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

- The first sign classifier relies solely on the hand path, hence, the movement component of the sign, or the gesture part.
- The second sign classifier uses the hand path combined with the hands configuration.

For each sign on the dataset, the feature vector is an array with a fixed number of positions, being this the chosen normalized size for a sign (normalized method explained above).

Each position of this vector contains, for the first case of sign classifier, 6 doubles, corresponding the first three to the hands coordinates, X, Y and Z in that order, for the right hand, 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 values, being the first three the hands coordinates for the right hand, the fourth value corresponding to the label for the hand configuration of the right hand (recognized with the posture classifier with 43 classes described above), and the four remaining values having the same scheme as the first four values but for the left hand.

The posture labels, values in the fourth and eight positions of the feature vector, are normalized to have a similar range as the other features. The other features, X, Y and Z range from [0...1] for all 3 coordinates. The normalization done to the original posture label value, which originally vary from [0...42], is dividing the value by 43. This way, the posture label value given to the feature vector has the range [0...1].

To test the classifiers, in both cases, 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 signers times 6 signers). Despite the subsamples being randomly generated, it is assured that each subsamples has 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 classifiers.

Although it could be possible to achieve better results with a thorough exploration of the tested kernels parameters and even other kernels or machine learning techniques, considering the available time for realization of this thesis and considering that its main aim is not to explore the machine learning field, each kernel was used with the default values of the framework implementations.

4 Results and Evaluation

4.1 Posture Recognition

After collecting the data and creating the classifiers for the 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 with other works. There were created classifiers with the depth information, and others with binary information (shape of the hand). Variations in the hand depth image normalization process were also tested, being the first classifiers tested without addressing the problem of the signer hand size (showed in 3.5.2), while in the others, the full normalization process was included. And finally, variations in the feature vector size were also tested, with feature vector sizes of 64(images with 8x8), 256(16x16), 1024 (32x32) and 4096(64x64). Since no significant differences obtained for the last three sizes, but since the size of 32x32 obtained usually higher accuracy, this was the selected feature size.

Table 5 – Postures classifiers results for the data with 52 postures and with the 3 distinct normalization methods. The kernel used was the Histogram Intersection. Underlined are depicted the similar results from different feature sized vectors (32x32 and 64x64). The Training and Validation values are averaged accuracy values from all folds.

		Normalization Method					
Features		No Stretch		Stretch Ratio		Stretch	
Type	**Size**	Training	Validation	Training	Validation	Training	Validation
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	0,874
	64x64	0,942	0,809	-	-	0,970	0,871

From analysing the values, firstly, it is possible to see that it is best to use the depth information than the binary information in this problem. This was expected since depth features contain more information than the binary ones. The feature size selected to be used in this problem, in this conditions, was the 32x32 feature vector size. The processing workload needed to use the feature vector of 64x64 doesn’t increment the results in a statistically relevant amount, once that the posture system is to be used within the sign recognition system, having to classify both hands in each of the sign instants. In top of that, and because of the data collection method and sensor, the original input concerning the hands, has sizes varying from 130x130 pixels to 40x40 if 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 $64x64$ is used, the original Kinect data for the hand will be smaller therefore the feature vector must be scaled up.

With the “Stretch” method applied, the performance of the classifiers increased in an average of 9% in comparison to the “No Stretch” classifiers, and an average of 9% for the “Stretch Ratio” classifiers.

Analysing the data in Table 5, we can see that the best classifier uses 32x32 features with depth data, and the stretch method on the normalization of the hand depth image. In appendix 0 is the cross matrix table for the chosen classifier. It is possible to see that were 340 false positives out of 2703 recognitions, 30 postures had a recognition rate above the average 87.28% while the standard deviation was 9.12%.

Comparing with similar works, (Almeida, 2011) made use of Kinect v1 using only depth information. It achieved a 100% recognition rate on the 26 letters from the PSL alphabet (against this example’s 52 postures) using a Skeletal-based Template Matching adaptation. His data set contained only one user, and the testing was done with the same user present in the data set, but with new recordings. 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 he used SVM with a Gaussian Kernel, using an estimated parameter δ. Using also depth information, but instead of the Kinect, 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 testing what was the best method to use, there was created and tested the second classifier with the dataset containing 43 postures.

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 Validation values are averaged accuracy values from all folds.

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

Being this dataset composed by less postures (43) than the previous one, but having a lot more variation concerning the hand orientation, its accuracy dropped considerably comparing to 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 hands in signs, and with a set of 46 hand postures, it achieved a validation 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, this is the only suitable classifier to be used to recognize hand configurations with the motion of the hands in the middle of the signs.

4.2 Sign Recognition

For the sign recognition system, that was the main aim of this work, there were created 2 classifiers as well. In both cases, despite the 9 kernels were tested, as explained in 3.6.2, only the 3 best results will be discussed.

For the first classifier, that only concerned the moving part of the sign, that is, the hand path, the results were as follows:

Table 7 – Testing 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 Validation values are averaged accuracy values from all folds.

		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 for the movements were tested, being the movements of the dataset normalized to that size with the method described in 3.5.1.Analysing the data in Table 7, it is possible to see that the best classifier

Table 8 – Confusion matrix created in the classifier testing phase using the cross validation method. An average recognition rate of 91.59% was achieved for the selected 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**

uses features resized to 40 frames and the used kernel was the *Laplacian*. In is the confusion matrix table for the chosen classifier.

It is possible to see that were 117 false positives out of 1392 recognitions, 10 postures had a recognition rate below the average 91.59% while the standard deviation was 8.93%. In Table 3, presenting the corpora, were also depicted signs with similar or with the same movement. As it was expected, the signs that had lower recognition rate were precisely those that have the same movement, that are the pairs “escrever” and “graxa”, “balança” and “avaliar”, “discutir” and “guerra” and lastely “tubo(fino)” and “tubo(médio)”. Also, signs with similar movement as “eclipse” and “morrer” and “testemunha” and “verdade”, that have small differences only in the positioning of the sign, verified a lower accuracy rate. The classifier was also able to distinguish signs with a significant positioning difference but with the same movement, as the case of the pair “mesa” and “balcão”. That the sole difference is that the second one is performed bellow the chest, and the second one above the chest. Yet, despite “balcão” having the similarity with “mesa”, it also shares a similar positioning with the pair of signs “tubo”.

Comparing again with (Almeida, 2011), that achieved a 100% recognition rate on the 10 signs from the PSL alphabet (against this example’s 29 signs) using an algorithm of 3D Path Analysis. His data set contained only one user, and the testing was done with the same user present in the data set testing the system on the fly. In Rui’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. He achieved a recognition rate of 55.24% for 13 signs of LIBRAS. Souza’s system was 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 solution 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 labels in each frame were used as features. 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.

		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

Comparing this approach, that uses both movement and hand configurations information, with the previous approach, that uses only movements, there is a decrease of 10.3% comparing the best classifiers from both cases. After evaluating and validating the performance of the system using only the movement information a possible assumption is that this decreasing in the accuracy is due to the hand labels, that, despite normalized to fit the feature vector, introduce instability to the dataset. In the confusion matrix for this second sign classifier is possible to see that no sign increased recognition. Besides introducing error in the signs that were not supposed to benefit from this approach (signs that had no similar nor equal movement), the hand labels weren't helpful in distinguishing the pair of signs (Table 3) that should actually benefit from it.

Table 10 – Confusion matrix created in the sign classifier testing phase using the cross validation method. This classifier used both movement and configuration information. An average recognition rate of 81.32% was achieved for the selected 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
tubo(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 is (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 using a SVM with a Quadratic kernel for classifying the hand configuration and Hidden Conditional Random Fields to merge all the information. In his work, there was a substantial increase of the accuracy when comparing the systems without the hands information (55.24%) and this approach.

5 Conclusions and Future Work

5.1 Conclusions

In this work, there were presented, detailed and conducted experiments in the problem of sign language recognition in the context of Portuguese sign language using Support Vector Machines and the Kinect One sensor. First there were detailed the main components of the Portuguese Sign Language and there were specified the main problems this work proposed to tackle. In the development of the system, it was presented the two step signs classification proposed in the architecture. This system was created using only depth information retrieved with the Kinect One sensor.

The first hypothesis (**H1**) suggested that was possible to extend current works on PSL to distinguish a specific set of properties in signs: signs with the same movement but different hand configurations. While proved in (Gineke, Reinders, 2010; Souza, Pizzolato, 2013) that hand information was crucial for the sign recognition process and that it could increase the sign recognition accuracy, this was not verified in this system. Using only depth information, Souza and Pizzolato used hand depth images deriving also hands and facial orientation to achieve those conclusions. It can be concluded that analysing both hands at each frame with a single image is not a viable approach because of the posture classifier used. This system was able to correctly distinguish the paired signs of the classes 2 and 3 proposed in 3.3.2 because of the analysis of the hand path, but failed to distinguish pairs of the type 1 because of the hand configuration classification. Despite this work not being able to distinguish said classes of signs, it extended PSL state-of-art by using 53 postures of PSL and 29 signs, improving the previous 26 postures and 10 signs from (Almeida, 2011), researching and implementing some of the suggested future work, such as handling multiple signers. It was also able to verify that the approach to the movement part can be taken into consideration for other works, and for future work, by, as previously mentioned, being able to distinguish those specific classes of signs.

5.2 Future Work

As a future work of this research, it would be necessary to implement a better system to classify hand configurations. Other approaches to be implemented and tested could be: coupling other information with the hand depth image, such as hand orientation; use an appearance based model of the hand for each frame or a set of frames; try to assert the hand configuration analysing a set of frames for the movement or even the whole movement, instead of classifying many instants. Other future work, to improve the case of other signs must go through detecting facial expressions, either similarly to what was done in (von Agris et al., 2008) or merging this system with the ViKi (Visual Kinect) developed by Hélder Abreu (Abreu, 2014) with interesting results in the analysis of the lips. Analysing other body parts and movements is also a crucial step for a Sign Language system being able to identify all classes of signs. Concerning the machine learning, other approaches should be tested, either by a thorough investigation and manipulation in the kernels used or in new kernels, or by using other machine learning techniques such as HMM or even Neural Networks. A final improvement after the previous mentioned ones implemented, should be addressing the continuous sign language recognition problem.

Bibliography

Abreu, Hélder (2014): „Visual Speech Recognition for European Portuguese“.

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“. In:

Andersen, MR; Jensen, T; Lisouski, P (2012): „Kinect depth sensor evaluation for computer vision applications“. In:

Athitsos, Vassilis; Neidle, Carol; Sclaroff, Stan (2010): „Large lexicon project: American Sign Language video corpus and sign language indexing/retrieval algorithms“. In: *… of Sign Language …*., pp. 2–5.

Bela Baltazar, Ana (2010): *Dicionário de Língua Gestual Portuguesa*. Porto Editora.

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.

Brashear, Helene; Starner, Thad; Lukowicz, Paul; et al. (2003): „Using multiple sensors for mobile sign language recognition“. In:

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

Capilla, DM (2012b): „Sign Language Translator using Microsoft Kinect XBOX 360 TM“. In:

Chai, Xiujuan; Li, Guang; Lin, Yushun; et al. (2013): „Sign Language Recognition and Translation with Kinect“. In: *vipl.ict.ac.cn*.

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: *Automatic Face and Gesture Recognition, …*. — ISBN: 0769516025

Fang, Gaolin; Gao, Wen; Zhao, Debin (2004): „Large vocabulary sign language recognition based on fuzzy decision trees“. In: *… and Cybernetics, Part A: Systems and …*. 34 (3), pp. 305–314.

Freeman, WT; Roth, Michal (1995): „Orientation histograms for hand gesture recognition“. In: *… Face and Gesture Recognition*.

Gineke, A; Reinders, MJT (2010): „Influence of handshape information on automatic sign language recognition“. In: *Gesture in Embodied …*.

Hong, Seok-ju; Setiawan, Nurul Arif Setiawan; Lee, Chil-woo (2007): „Real-Time Vision Based Gesture Recognition for Human-Robot Interaction“. In:

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 on the Integration of Gesture in Language …*.

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*.

McGuire, RM (2004): „Towards a one-way american sign language translator“. In: *Automatic Face and …*.

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: 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 (2002a): „Hidden Marcov Models for sign Language Recognition: a Review“. In: *In2nd HellenicConference on …*. (April), pp. 343–354.

Pashaloudi, Vassilia; Margaritis, Konstantinos (2002b): „Hidden markov models for greek sign language recognition“. In: *… of 2nd WSEAS International Conference on …*.

Segen, Jakub; Kumar, Senthil (1999): „Shadow gestures: 3D hand pose estimation using a single camera“. In: *Computer Vision and Pattern Recognition, …*. 00 (c).

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:

Starner, Thad; Pentland, Alex (1995): „Visual Recognition of American Sign Language Using Hidden Markov Models.“. In:

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: *Computer Vision …*.

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: *Systems, Man, and Cybernetics, 1997. …*., pp. 156–161.

Vogler, Christian; Metaxas, Dimitris (1998): „ASL recognition based on a coupling between HMMs and 3D motion analysis“. In: *Computer Vision, 1998. Sixth …*., pp. 363–369.

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

Wang, SB; Quattoni, A (2006): „Hidden conditional random fields for gesture recognition“. In: *… Pattern Recognition, …*.

Wilson, AD; Bobick, AF (2000): „Realtime online adaptive gesture recognition“. In: *… Recognition, 2000. Proceedings. 15th …*. (505).

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: *Proc. IFAC/IFIP/IFORS/IEA symposium on analysis, …*.

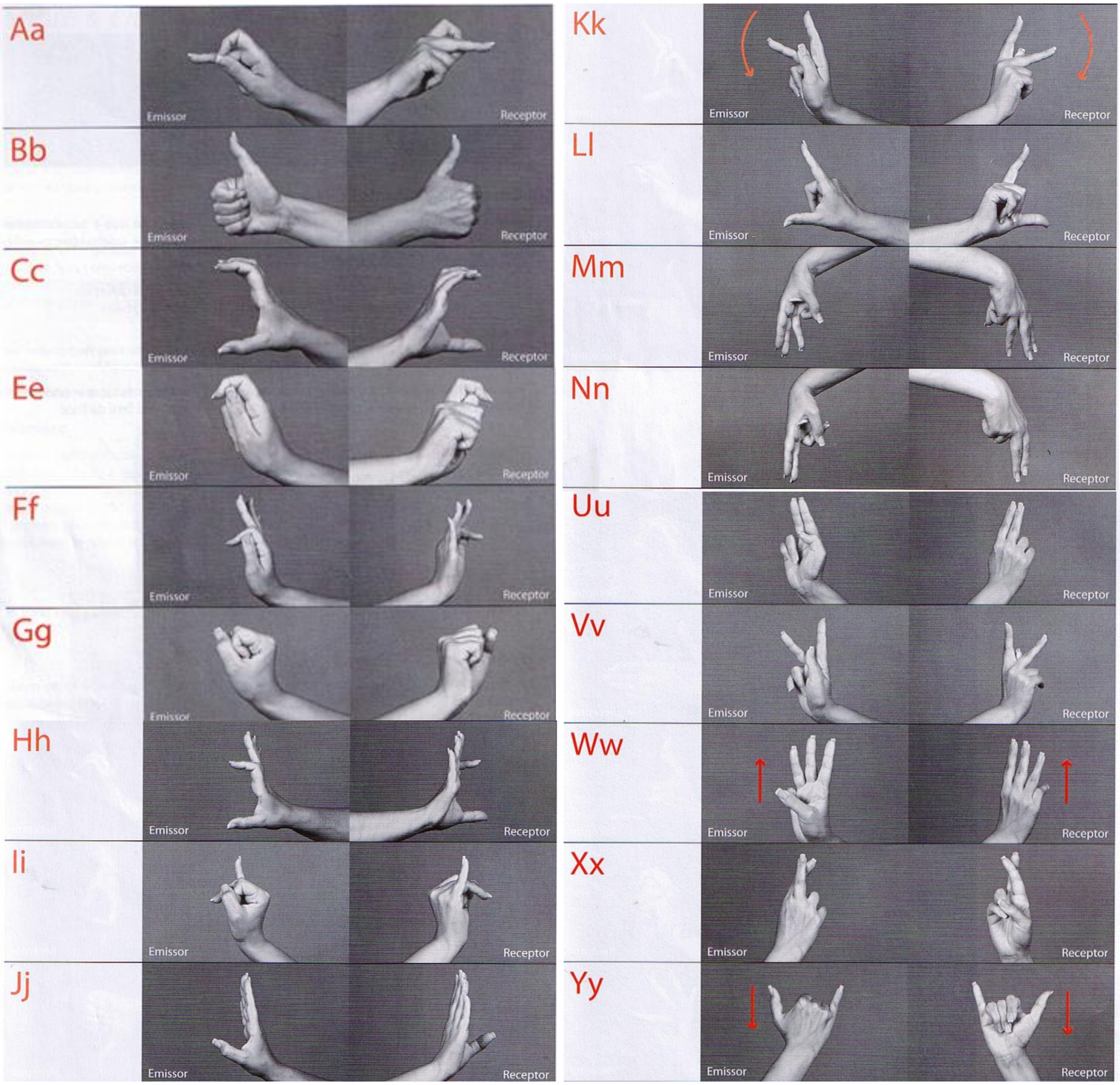
Appendix

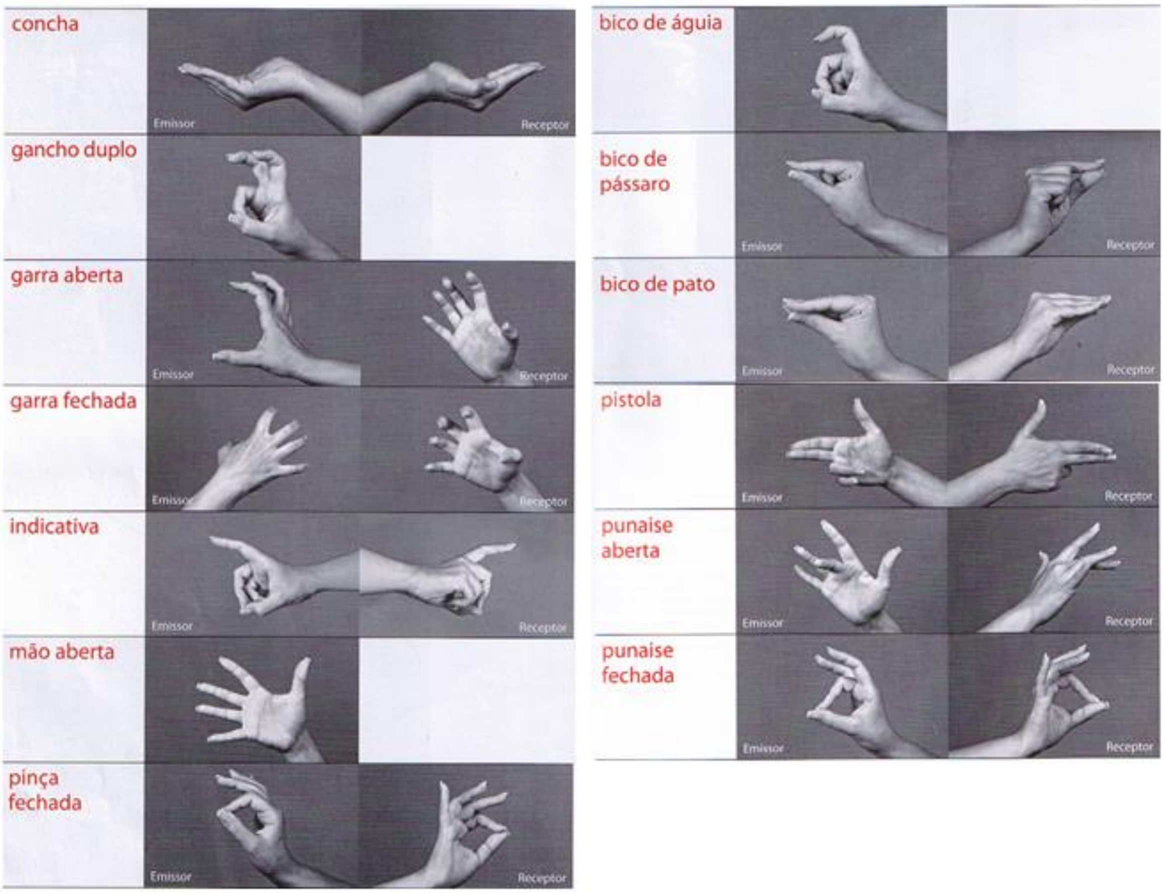
A.CROSS MATRIX FOR 52 POSTURES CLASSIFIER

	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

gancho duplo

garra aberta

garra fechada

indicativa

mão aberta

pinça fechada

bico de águia

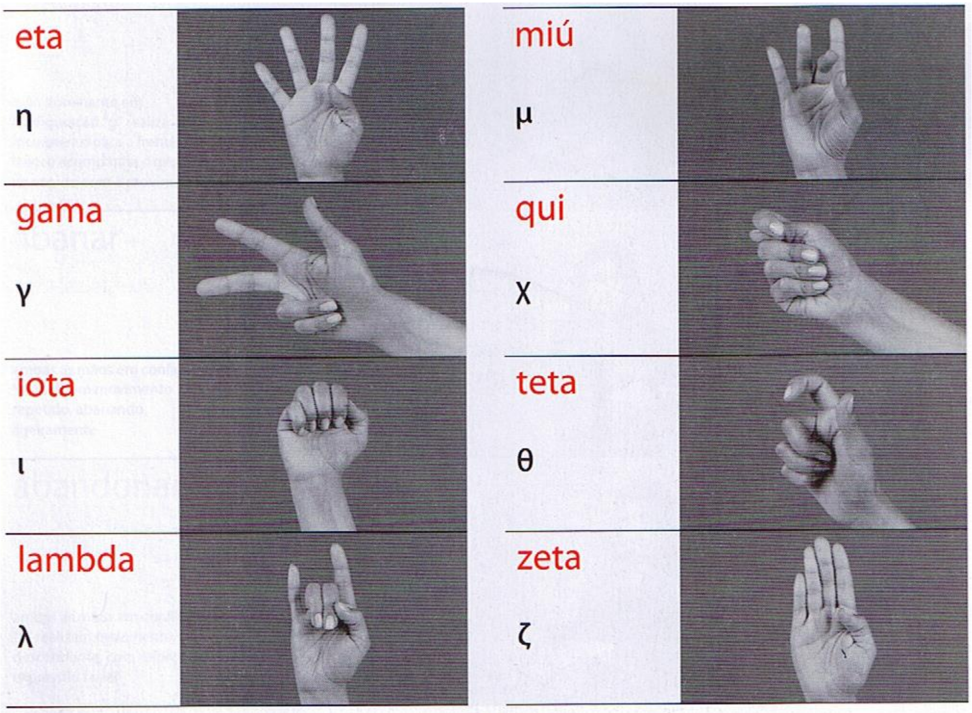
bico de pássaro

bico de pato

pistola

punaise aberta

punaise fechada



eta

η

gama

γ

iota

ι

lambda

λ

miú

μ

qui

χ

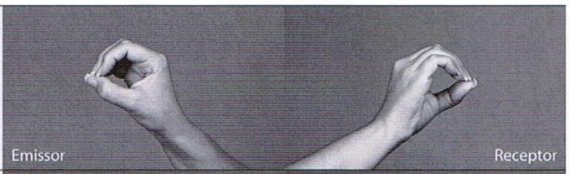
teta

θ

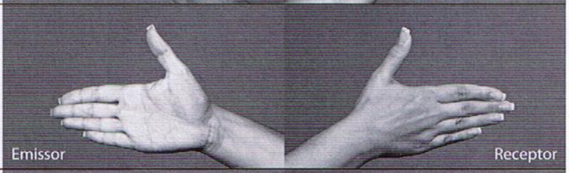
zeta

ζ

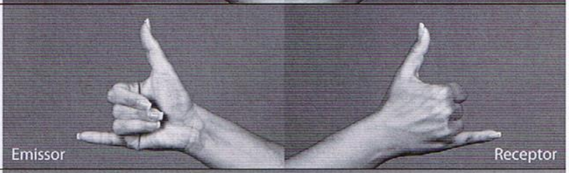
0


Emissor
Receptor

1


Emissor
Receptor

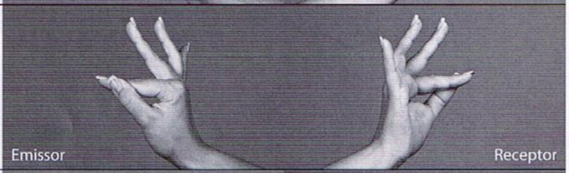
2


Emissor
Receptor

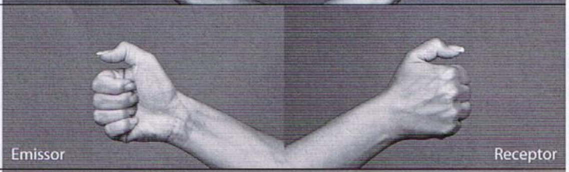
3

Emissor
Receptor

4


Emissor
Receptor

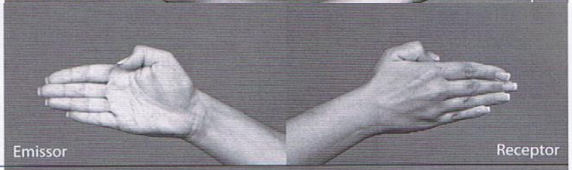
5


Emissor
Receptor

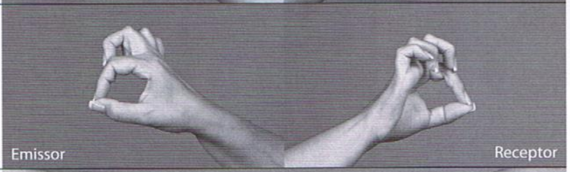
6

Emissor
Receptor

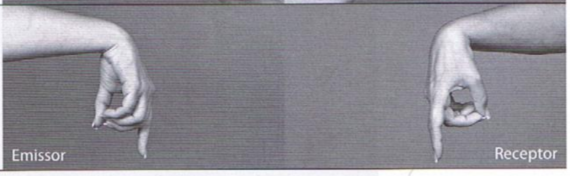
7


Emissor
Receptor

8


Emissor
Receptor

9


Emissor
Receptor