FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Universidade do Porto
Faculdade de Engenharia
FEUP

# Analysis of Expressiveness of Portuguese Sign Language Speakers

**Maria Inês Coutinho Vigário Rodrigues**

MASTER THESIS

Integrated Master in Bioengineering

Supervisor: Luis Filipe Pinto de Almeida Teixeira (PhD)

Co-supervisor: Eduardo José Marques Pereira (Eng.)

June 2014

# Abstract

Nowadays, there are several communication gaps that isolate deaf people in several social activities. The breakthroughs made so far in the area of reading nonverbal behaviour are crucial and constitute a strong motivation to direct research efforts towards studying this population.

This master thesis is part of a unique project that intends to study the differences among two populations: deaf people and hearing people. The ultimate goal is to move towards a better understanding of emotional and behaviour patterns, through face and body expressions analysis, in both deaf and hearing people. The hereby presented work, in particular, is a first step towards that ultimate goal in which the expressiveness of gestures in Portuguese Sign Language speakers is accessed in several different ways.

One of the main contributions of this work to the scientific community, that served as support to the work itself, was the construction of a video database of a very specific context (duo-interaction between deaf and hearing people). This database served as starting point for the development of a feature-based computer vision framework with the purpose of finding answers to the following problems: differentiate the deaf population from the hearing one, identifying based on body expressiveness different conversation topics, trying to identify through feature analysis different levels of mastery in Portuguese Sign Language. The approaches followed to answer these research questions were based on the use of machine learning classifiers in supervised and unsupervised ways. The results of the classification revealed that it was possible with the features considered to distinguish deaf from hearing people and also to distinguish different conversation topics featured in the videos of the database. The attempt of stratifying the subjects considered in levels of expertise of sign language was not achieved with the same success due to the abstractness of this issue which may be dependent on several other factor that are exterior to our approach such as the life experience of each subject.

The final outcomes of this study are very interesting and encouraging to carry on with further research on: discovering and studying the main differences of the considered populations and also to explore their body and facial expressions on several scenarios as well as their interactions.

# Acknowledgements

This dissertation would never be complete if I did not have the chance to acknowledge to all people that supported me throughout this time. For all of them I leave here before anything else a sincere thanks.

To Luis Teixeira, my supervisor, for all the accompaniment and wise advice during the planning of the work and also in the process of thinking about the results and other approaches; to Eduardo Marques, my co-supervisor, for his tireless patience with me to explain everything any time I needed and for giving me so many valuable advices during the process of this work; to INESC TEC for having such a friendly environment and amazing and helpful people in special to VCMI group (Visual Computing and Machine Intelligence) for being so welcoming and promoter of the mutual aid among its members.

Some people external to this project were also very helpful and without their efforts this work would not have been so successful: to Ana and Paula who are LGP experts and gave us several advices and helped us finding the volunteer population for our database, to the socio-psychologists from the Faculdade de Psicologia e Ciências da Educação da Universidade do Porto who helped us defining all the framework of our database, to Agrupamento de Escolas Eugénio de Andrade, Escola EB2/3 de Paranhos for providing the venue for acquisition of the videos of our database and finally to Stephano Piana from all his help with the EyesWeb software.

Besides all the people that contributed directly to the contents of this work a lot of people in my life helped finding the strengths throughout the hardest moments of this process. Moments to relieve stress, give a good laugh, get together for playing some songs or simply go out for coffee. For all this I want to give my deepest thanks to my closest friends who were always supportive and always had the right words in the right moments. I mean to thank equally to me friends that do not have the chance to see my everyday but whom I know that never forget about me.

I would also like to thank two entities that made my five years of college the best years of my life which are Tuna Feminina de Engenharia and Metal e Bio where I learned that in life everything is possible as long as we never lose track of what really matters and what makes us happy. For giving me some of the most amazing moments of my life and showing me that in five years we can find friendship for life.

Finally, and most important, I would like to thank my dear family specially to my mother Adelaide, my father António and my sister Ana who gave my unconditional support all my life. To my mother for being so kind and loving when I needed and also tough at the same time I want to thank since I know we are the same; to my father for his never ending patience and loving everyday of my life and also for picking me up late ours while I was working everytime I needed; to both my parents together for being so amazing and worrying so much about me and my future; to my oldest sister who helps me more than anyone else and whom I love and miss everyday.

Inês Vigário

*"And so with the sunshine and the great bursts of leaves growing on the trees, just as things grow in fast movies, I had that familiar conviction that life was beginning over again with the summer."*

F. Scott Fitzgerald, "The Great Gatsby"

*To my family and dearest friends,*

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| CM | Confusion Matrix |
| CR | Correct Rate |
| $k$-NN | $k$-Nearest Neighbours |
| HBA | Human Behaviour Analysis |
| LGP | Língua Gestual Portuguesa |
| MHI | Motion History Image |
| PCA | Principal Component Analysis |
| PCH | Pixel Change History |
| PSE | Pixel Signal Energy |
| QoM | Quantity of Motion |
| ROI | Region of Interest |
| SVM | Support Vector Machines |

# Chapter 1

# Introduction

After decades of neglect (due to the supremacy of the learning theory and then the cognitive revolution), the research on emotion, behaviour and expressiveness analysis was leveraged by the important work on facial expression by Tomkins [1][2] and continued by Ekman [3][4]. The power of nonverbal behaviour in emotional, or simply relational episodes, became a central issue in most psychology textbooks as these started to be invaded by photos with prototypical expressions and simple emotions[5]. In 1992 appeared the concept of basic or fundamental emotions. It was created by theorists in the area of behaviour analysis during the course of their works. Although researchers considered that human behaviour can be segmented in a set of fundamental emotions there was no unanimity in which ones these are [6]. This topic remains currently a doubt among people who dedicate themselves to analysing human behaviour.

Nowadays, the panorama regarding the nonverbal concomitants of emotional experiences has changed drastically. Every year, more than 50 books and papers are published featuring nonverbal channels of expressive communication (included in a wide range of different areas of expertise). The channels considered are mainly facial expression, gestures, gaze, vocal quality, paralinguistic features, posture and body position, head nods, among others [5]. Even though much of present-day expressive research is carried out with paper-and-pencil assessment of verbal reports of expressive content, the computer vision field is trying to emerge as a relevant tool for achieving nonverbal sensitivity in an computational and automatic way.

Body gestures serve as main communicative function and contain substantial affective and cognitive information that help us emphasize certain parts of our speech and pass on expressive content. If we think about spoken language, we can realise how disfluent it can be, full of false starts, hesitations, and speech errors. On the other hand, gestures are most of the times faithful to the speaker's communicative intention [7]. The interpersonal socio-emotional interaction and the recognition of a person's affective state are vital for communication. Concerning the computer vision field, the usage of automatic tools for extraction body features allows a better understanding of the human behaviour. This ability is named nonverbal sensitivity or emotional intelligence (defined as the ability to encode or express, and to decode or understand nonverbal cues [8]). Decoding and studying expressiveness is useful for several applications, in which the analysis of behavioural video is included as a way to describe the social interactions between two people [9].

This first chapter will focus on the main motivations and objectives for the development of this MSc thesis, as well as presenting the contributions given to the scientific community. A structure of the dissertation is also presented.

## 1.1 Motivation

Automated visual analysis of behaviour provides tools for the construction of intelligent computer vision systems. The idea of reaching nonverbal sensitivity through computational models is very appealing since this may help us understanding how the human visual system interprets all the sensory events in the environment and how it relates those events.

Sign language speakers experience their languages very passionately. This may be explained by the fact that language plays a crucial role in the construction of a community and in by fact that it is a clear mark of belonging [10]. Emotion recognition from body language and its implications to the social adjustment of a sign language speaker are, therefore, very important issues. Sign language expressions are composed of manual (hand gestures) and non-manual components (facial expressions, head motion, pose and body movements). Some expressions are performed only using hand gestures whereas some change the meaning when a facial and body expressions accompany hand gestures [11]. As it will be described on section 1.2, this study will focus on evaluating the differences between deaf and hearing people, in terms of expressive patterns through body gesture analysis. This study aims to be a preparatory work for future and more ambitious project that will intend to reduce the major gaps that nowadays prevents deaf people from interacting easily with other people in society.

In particular we will focus on doing preliminary evaluations of the motion expressiveness of subjects using LGP (from the Portuguese denomination Língua Gestual Portuguesa which can be translated to Portuguese Sign Language). The nature of this work constitutes a very important step, with relevant conclusions, towards a final goal of diminishing the difficulties that the deaf community has when it comes to interacting with the hearing community.

## 1.2 Objectives

The expressiveness analysis method to be developed in this master thesis will be focused on Portuguese Sign Language speakers. The main objectives that will guide this study were defined taking into consideration the motivations mentioned before and also the barriers faced by this population.

- Study the expressiveness of the human body in a feature-based approach.
- Build an unique video database of a very specific context (duo-interaction between deaf and hearing people) that encompasses technical annotations.
- Identify and test different methods to extract features, and analyse and classify their changes through temporal domain.
- Differentiate the deaf population from the hearing one.
- Identify based on body expressiveness different conversation topics.
- Identify through feature analysis different levels of mastery in Portuguese Sign Language.

## 1.3 Contributions

The proposed work had following main contributions:

- The creation of a database featuring video data of dialogues between deaf and hearing people, with manual annotations of body parts.
- Proof of the differences in the motion expressiveness between deaf and hearing people speaking LGP.

- The ability to distinguish different conversational moments through a classification method may be a valuable insight for other studies that mean to assess human behaviour for example.

## 1.4 Structure of the Dissertation

Besides this introductory chapter, this dissertation includes 4 other chapters. In Chapter 2 is presented a review about the state-of-the-art research works regarding their application on this specific problem with major focus on the works for motion expressiveness. Chapter 3 describes the methods used to accomplish the proposed objectives for this dissertation whose results and discussion are explored in Chapter 4. Finally, Chapter 5 is a conclusion to the presented dissertation, in which are also presented some cues regarding the future work for this project.

# Chapter 2

# Literature Review

Detecting and interpreting temporal patterns of nonverbal behavioural cues in a given context is a natural and often unconscious process for humans. However, this still remains a rather difficult task for computer vision systems [12]. It would be of great value if machines were able to achieve this kind of sensitivity and play a helpful role in several social situations like sensing for example agreement or disagreement among a group of people arguing about a certain topic.

The best way of achieving this goal is to successfully carry the inherent knowledge that humans have on this matter to computer systems, allowing these systems to sense activities and social relationships. The fusion of several elements such as motion, appearance, shape etc. is the cue to solving this fundamental, yet challenging, research topic that has driven the efforts of many researchers. Message production and processing, relational communication, social interaction and networks, deception and impression management, and emotional expression are the main applications for nonverbal sensitivity in computer systems [13].

The key aspects that one needs to take into account, when it comes to achieving nonverbal sensitivity based on body expressiveness, will be explored in this chapter, as well as the most relevant state-of-the-art studies. Aspects related to body gesture expressiveness and emotion will be addressed since those are the main focuses of this project targeting sign language speakers. Although this is not included on the framework of this study in particular, some aspects related to the detection and classification of emotions using human gesture will also be discussed since that is a relevant issue and goal for a continuation of the study that will be hereby presented.

## 2.1 Understanding Behaviour

Human Behaviour Analysis (HBA) is increasingly more of interest for computer vision and artificial intelligence researchers [14]. During any human interaction, there are many relations between human body parts and the surrounding environment. The key for understanding behaviour is then to analyse human interactions based on well defined existent relationships [15]. The main challenges in HBA are related to its: complexity and uncertainty. To overcome these challenges, visual analysis of behaviour focuses on three essential functions [16]:

- Representation and modelling: To extract and encode visual information from imagery data in a more concise form that also captures intrinsic characteristics of objects of interest;
- Detection and classification: To discover and search for salient, perhaps also unique, characteristics of certain object behaviour patterns from a large quantity of visual observations, and to discriminate them against known categories of semantic and meaningful interpretation;
- Prediction and association: To forecast future events based on the past and current interpretation of behaviour patterns, and to achieve object identification through behavioural expectation and trend.

The last item mentioned may constitute a major difficulty since, developing learning strategies (in the computer vision field) requires a deep knowledge in psychology of emotions in order to build behaviour models and patterns for extrapolation to several types of analyses.

### 2.1.1 Behaviour in Context

As mentioned previously, the analysis of human behaviour may be useful for several applications. Understanding the context of a visual environment is essential to properly interpret behaviour since the context will be distinct for each application. If we think about it, this concept of context evaluation is inherent to the human condition once we have a constant necessity of adjusting ourselves to the situation we are experiencing [17]. Computer vision research on HBA includes a broad range of studies on developing computer systems and models to achieve nonverbal sensitivity in different contexts and through different channels (such as face, voice, gait and body gesture).

#### 2.1.1.1 Human Body as a Vehicle of Expression

Body expression is an inherent feature of human behaviour. Many artists use it in exaggerated ways to express their ideas to the public in a nonverbal way. If they are expressive enough, the message they are trying to pass on is received by that public. An example of expression of the six emotions and also a neutral state through body expression is shown in Figure 2.1.

Whole-body expressions provide information about the emotional state of the producer, but also signal his action intentions. For example, a surprised body expression can signal the appearance of a new element in a scene but also give information on how the subject will deal with that surprise. The work in body expression was initiated in 1872 by Darwin who described in detail the body expressions of many different emotions [18]. Recent psychology studies have pursued Darwin's preliminary work: de Gelder [19] described the stimulus set of whole body expressions termed bodily expressive action stimulus test (BEAST), providing validation data through the creation of a database composed of 254 whole body expressions from 46 actors expressing 4 emotions; Van den Stock and Righart [20] investigated whole-body expressions of emotions in three different experiments from which realised the importance of emotional whole-body expressions in communication either when viewed on their own or in combination with facial expressions and emotional voices; Kleinsmith et. al propose a method that automatically recognizes affective states and affective dimensions from non-acted body postures and use observers to establish ground truth labels. The review paper by Kleinsmith and Bianchi-Berthouze [21] provides an overview, gathering and evaluation of the main state-of-the-art studies on body expressions as a communication channel.

As stated in Chapter 1, one of the key objectives of this work is the construction of a complete database, from which several behaviour-related features can be extracted. Researchers like the ones that developed the studies discussed so far, which are inserted on the context of computer vision for expressiveness and emotion analysis, would certainly benefit from a database containing this type of information regarding


**Figure 2.1:** Emotions displayed by body language in front of a uniform background. Extracted from [22].

the evaluation of body gestures and their expressiveness. Several state-of-the-art studies are working on extracting expressive cues from human gesture proving that these are valuable indicators of a person's profile and emotional state. In this sense, the dataset created for this study intends to serve in the future as a support to extract information on interaction between people (deaf and hearing) in order to infer if there are social difficulties in terms of communication and also try to detect different levels of empathy.

#### 2.1.1.2 Body Gesture

Gesture (Figure 2.2) is the use of motion of the limbs or body as a means of expressing/communicating of an intention or feeling [7]. This human communication channel is responsible for passing several signals to the context we are inserted in. Research on expression recognition in psychology and computer vision fields used to focus exclusively on photographs or video sequences of facial expressions, since this was considered to be the only powerful channel of nonverbal communication. More recently, researchers realised the amount of expressive information that body gestures may transmit in a more subtle way [23, 24, 25]. The recognition of human gesture has already shown its value in areas such as: behavior understanding, human-machine interaction, machine control, surveillance among others [26]. Studies showed that even in the absence of facial and vocal cues, it is possible to identify basic emotions signalled by static body postures [27], arm movement [25] and whole body movement [27, 28, 25]. Thus, gestures serve an important communicative function in face-to-face communication since they often occur in conjunction with speech. According to Cassell [29], the fact that *gestural errors* are extremely rare demonstrates how essential their nature is for accurate communication. If we think about spoken language, we can realise how disfluent it can be, full of false starts, hesitations, and speech errors. On the other hand gestures are almost always faithful to the speaker's communicative intention. Five different types of gestures may be defined [7]:

- Gesticulation: movements of the hands and arms that accompany speech;
- Language-like gestures: gesticulation with the intention of replacing a particular spoken word or phrase;
- Pantomimes: gestures that depict objects or actions, with or without accompanying speech;

- Emblems: familiar gestures such as, thumbs up, and assorted rude gestures (often culturally specific);
- Sign languages: linguistic systems which are well defined.

Gesture recognition (a relevant to be analysed in the case of sign language) can be defined as the process by which a human observer or a machine identifies the body gestures made by a subject. If we think about which types of gestures would be more easily recognized by computer systems, we would come to the conclusion those would be emblems or even sign language since they tend to be less ambiguous, less natural and more likely to be learned. Emblematic gestures carry more clear semantic meaning. Several steps can be identified that are transversal to most gesture recognition systems [7]:

1. Sensing human position, configuration, and movement in the scene using cameras and computer vision techniques;
2. Preprocessing;
3. Gesture modelling and representation;
4. Feature extraction and gesture analysis;
5. Gesture recognition and classification.

**Figure 2.2:** Examples of body gesture from FABO database. Extracted from [16].

In gesture recognition we find image-based and video based approaches encountered in literature. The first one was exploited by Coulson [30] who attributes six universal emotions to a total of 176 computer-generated mannequin figures produced from descriptions of postural expressions of emotion. Static gesture or pose recognition can be accomplished by a straightforward implementation using template matching, geometric feature classification, neural networks, or other standard pattern recognition techniques such as parametric eigenspace to classify pose [31, 32, 33]. Video-based approaches (more relevant to this MSc study) are challenging due to spatio-temporal variations and endpoint localization issues. These critical issues, although challenging, are vital for an accurate gesture perception making the video-based approaches more suitable than image-based ones. An effective gesture recognition system was introduced by Ravindra de Silva et al. [34] in which Discriminant Analysis was used to build affective posture predictive models and to measure the saliency of the proposed set of posture features in discriminating between 4 basic states: angry, fear, happy, and sad. Li and and Greenspan [35] also worked on segmentation and recognition of dynamic gestures reaching an effective model based on Dynamic Programming. In the work proposed by

Piana et al. [25] anger, disgust, fear, happiness, sadness and surprise are six states evaluated through the extraction of features such as kinectic energy, quantity of motion, barycenter tracking etc. This work is performed in a controled environment for monitor the behaviour of autistic children while playing with an interactive serious game.

#### 2.1.1.3 Expressive Gesture

The concept of gesture leads us to think about a set of temporal/body features responsible for conveying expressiveness. The scientific community is gaining interest in studying the ways of modelling and communicating expressive content in non-verbal interaction. Two performing arts in which this increased interest is evident are: the study of movements in music performances (evaluating the conductor)or full-body movements to evaluate the performance of a dancer [36]. In this particular context, gesture is considered to contain and convey information related to the emotional/affective domain, which is different from the traditional meaning we give to language gestures. In this case gesture is the responsible for conveying expressive content, i.e., relevant information that is suggestive of a certain state of emotion [36].

**Figure 2.3:** Image recordings of piano performances (left and top views) used for studying expressive gesture. Extracted from [36].

For the particular case of the proposed work it is important to assimilate that gestures and the way they are performed work as an identity of a subject. The same action may be executed in several different ways depending on the executant. It is not rare that we are able to recognize a person by the way they walk and, it is also possible to infer emotional states by the way that person is moving (the field of gait analysis)[37]. This perspective enables us to classify walking as an expressive gesture [26]. Several everyday actions may constitute expressive gesture: Pollick [38] investigated expressive content of actions such as knocking or drinking, Heloir and Gibet [39] work on the identification and representation of the variations induced by style for the synthesis of realistic and convincing expressive gesture sequences in sign language speakers.

It is possible to identify the main challenges concerning the study of expressive gestures [40]:

- Define and characterize the expressiveness and variability in human movement. This expressiveness is considered at all levels of gesture generation, and involves both a semantic dimension (from actions that convey a specific meaning to sign languages that imply the linguistic aspects of phonetics, phonology, prosody, etc.), and an expressive dimension induced by intentional variations or emotional states of the actor, and results in variations in the produced signals.

- Explore new motion representation spaces that reflect the expressiveness and variability contained in the data. This implies reducing the complexity of the high-dimensional motion data by proposing different embeddings for these data. Such embeddings should enable to characterize and parametrise specific action sequences, and give rise to original approaches for recognition, or generation of new behaviours, inspired by sensorimotor biological processes.
- Link the different levels of representation, from narrative scenarios through structural patterns of actions, to continuous streams of motion data. More precisely, the aim is to extract structural patterns from data and to understand how these discrete patterns influence the synthesis of gestures while preserving the semantics of actions as well as subtle expressive variations.
- Defining evaluation protocols that are necessary for evaluating the different hypothesis and models that are constructed at all the levels of the perception-production loop.

Automatic analysis and synthesis of expressive gesture can open novel scenarios in the field of interactive multimedia. Computational models of expressiveness in human gestures can contribute to new paradigms for the design of interactive systems, improved presence and physicality in the interaction

## 2.2 Body Expressiveness Representation

In order to be successful interpreting the expressiveness of a human movement/gesture automatically, it is crucial to use precise models that can capture that expressiveness. When performing this task, it is important to consider not only the activity (scene) we are studying but the whole shared environment in which the activity is impregnated in. Thus, modelling this activity is concerned not only with modelling action performed by different objects in isolation, but also the interactions and causal relationships among these actions [41]. For a computer to extract a behavioural model from a set of data it is necessary to choose the essential elements to be accounted, firstly when it comes to the suitable representation mechanism. In this work two types of behaviour representation strategies will be addressed: trajectory-based and pixel-based [16].

### 2.2.1 Trajectory-based Representations

Trajectory-based representation is a type of object-based representation which consists of describing in detail objects in a space over time based on the assumption that individual objects can be segmented reasonably well in a visual scene [16].

Trajectory-based representation is a method implemented for systems using passive or active markers which yield a high contrast in the images and provide a robust representation [42]. It aims to construct object-centred spatio-temporal trajectories centred on each object's boundary box or shape structure. The object trajectory is the history of the motion in a visual scene [16]. A trajectory of a certain object is computed by associating that object in consecutive frames using motion tracking [43]. To achieve this tracking it is necessary to apprehend some object appearance attributes like colour, shape or texture so that it is possible to establish inter-frame correspondence over time. In a trajectory-based representation a single trajectory track should be associated to an object of interest (see Figure 2.4).

Unfortunately, image quality issues make this task hard (due to noise and crowded environments). Some techniques have been developed to cope with occlusions and lighting changes [44, 45], and by adopting a tracking-by-detection strategy [46, 47]. Despite these scientific efforts, problems remain unsolved on how

**Figure 2.4:** Set of trajectories extracted for objects of interest, in this case several body parts. Extracted from [36].

to best combine useful information sources and filter out unreliable information, so that object tracking is not lost over time and that false positives do not occur.

Many human action recognition studies benefit from this approach [48, 49, 50] using it as a starting point for behaviour analysis. The main problems encountered by the authors of these studies that in some ways compromise the results are described bellow [16]:

- Lack of details: in a cluttered public scene we may have low fidelity visual detail, which prevents us from extracting sufficient image features.
- Severe occlusion: losing the points we are trying to track over time may happen if the objects are inserted in a large space shared with other objects where inter-object occlusion may occur. This leads to the attainment of discontinuous trajectories and inconsistent labelling in object association.
- Lack of context: an isolated trajectory of an object does not always capture distinctively the information that one may be looking for. Interpreting the behaviour of an object in an unconstrained environment analysing only its trajectories appears to be insufficient in several cases.

### 2.2.2 Pixel-based Representations

This second strategy neglects individual object entities, targeting its focus to all relevant pixels found on the image, colour and intensity gradient information. It is extremely useful since it is computationally simple and appears to be beneficial for representing object activities in cluttered or crowded scenes. It is most commonly used when an exact distinction between what constitutes foreground and background is wanted, for example in a busy urban scene with different elements such as people (foreground) and buildings, trees and permanent objects (background) [51]. Effective techniques have been implemented from a short block of input frames by casting the problem as an exercise for optimal labelling. These are useful approaches of background approximation that assign a set of labels or pointers for each pixel in the background image (Figure 2.5) such as: Combinatorial Optimization, Minimum Cut/Maximum Flow and Alpha Expansion [52].

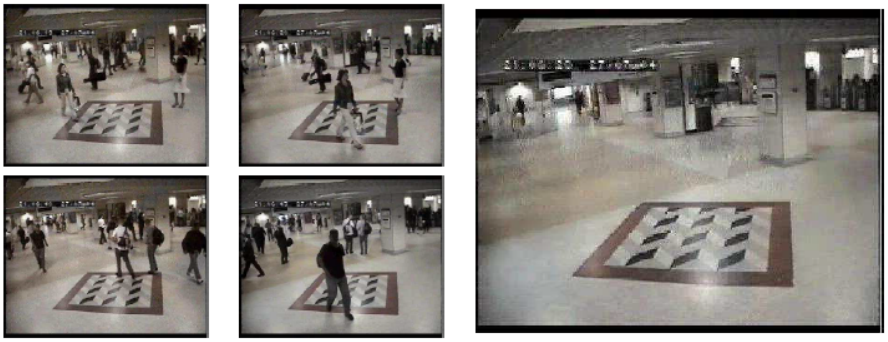

**Figure 2.5:** This image displays an effective background subtraction. Left:4 of the 20 input frames. Right: extracted background. Extracted from [51].

Although this pixel-based approach is considered a simple one, it can be of much value for modelling and understanding actions and activities [16]. After foreground pixel detection using for example an adaptive Gaussian mixture background model [53] more complete pixel-based representations may be obtained such as for example:

- The Motion History Image (MHI) which is used to detect visual changes in images (motion) by keeping a history of the changes which decays over time [54]. It is a widely used algorithm for its simple implementation and easy visualization. Looking at Figure 2.6 it is intuitive to perceive several aspects such as the direction of the motion, the current position of the object moving and its past localizations on the scenario. Computer vision techniques such as gesture recognition are accomplished in a feature-based statistical framework [55].
- Pixel Signal Energy (PSE) which uses temporal filters to measure the average magnitude of pixel-wise temporal energy over a backward window (its size determines the number of frames (history) needed to be stored) [56]. It allows one to extract reliable temporal change at individual pixels. Furthermore, pixel energy constitutes a good measure for exploiting synchrony in pixel-events. Following approaches like this one of detecting synchronous pixel events addresses the limitation of the short-sighted view of single pixels providing a more flexible framework for capturing global events as opposed to related spatio-temporal motion-energy measures [56].
- Pixel Change History (PCH) combines the two previous methods measuring the multi-scale temporal changes at each pixel [57]. A PCH image becomes a MHI image if its accumulation factor is set to one. This method can capture a zero-order pixel-level change, i.e., the mean magnitude of change over time [58].

If the assessment of an object's spatio-temporal information (particularly facial and body features motion) is intended, many researchers used optical-flow techniques [59, 60, 61, 62]. An optical flow field is usually a dense map of imagery displacement vectors that estimate pixel-wise apparent motion between consecutive image frames over time [16]. When it comes to facial and body expression recognition for example, a sequence of images contains much more information than a single image as it is visible on Figure 2.7, where an example of source vectors is shown for the detection of the wave of the subject's arms [62]. Flow vectors are computed by image grids instead of per pixel which happens to be computationally expensive. Thus, similarly to foreground pixel-based features, the optical flow-based representation avoids tracking individual objects.

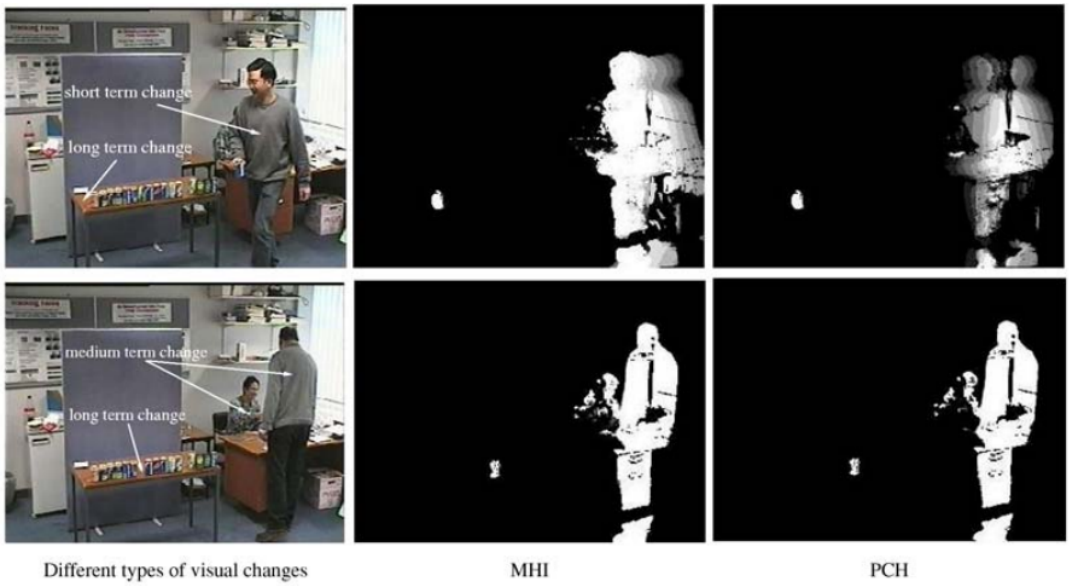


**Figure 2.6:** Examples of Motion History Images and Pixel Change History images for different types of visual changes. Extracted from [58].

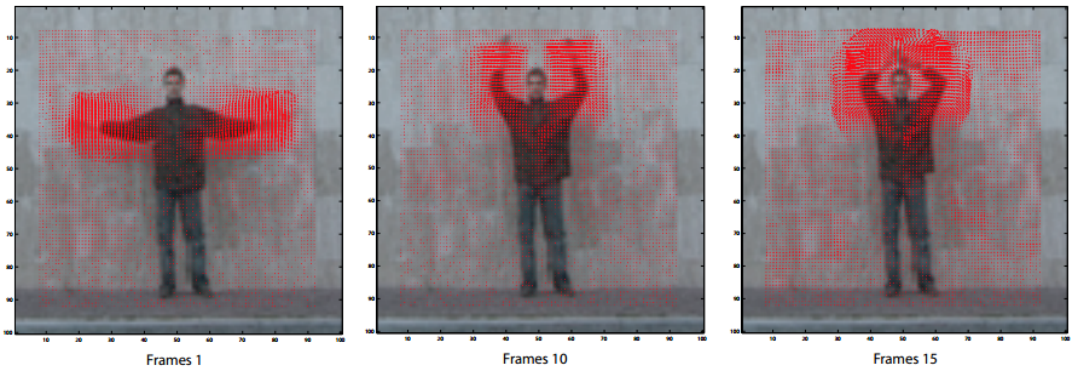


**Figure 2.7:** Optical flow of arm-waving action at frames 1, 10 and 15. Extracted from [62].

## 2.3 Classification Techniques

In order to obtain a system for automatic analysis of behaviour patterns there is necessity to learn statistical models in which the structure and the most suitable parameters are inferred [16]. In the case of a graphical model, its structure refers to (1) the number of hidden states of each variable in a model and (2) the factorisation of the model state space in order to find a suitable graph topology [16].

The development of robust pattern classifiers from a limited training set of observations (i.e., feature vectors) has long been one of the most relevant and challenging tasks in machine learning and statistical pattern recognition [63]. The main of goal of analysing in this chapter machine learning techniques is getting some insight regarding some of the methods that can lead us to answer the research questions stated in Chapter 1. Some strategies of *Supervised* and *Unsupervised* machine learning will be addressed as a subject of interest of the presented work. Some of the goals of this work are to answer a set a research question for which classification solutions will be used. Supervised techniques will be used to distinguish

deaf from hearing people and to identify different conversation topics in dialogues while Unsupervised techniques will be suitable on trying to identify different levels of mastery in LGP.

### 2.3.1 Supervised Classification

In the context of fully supervised classification it is assumed that the instances for each pattern class are well defined *a priori*. Any given pattern is uniquely associated with a corresponding target label, therefore this strategy assumes that labelled examples of behaviour categories are available. Categories are established on the first instance and several examples of each category must be created in order to obtain ground truths of the classification of each pattern category. This set of examples is obtained by intervention of human experts. Their work is to label the training set correctly in order to create a supervised training set [64].

Data annotation is an additional, expensive, and error-prone preparation process. The number of behaviour classes may be incredibly large meaning that listing them can be practically impossible. In areas such as bioinformatics, speech processing, or affective computing, exact class labels may not even be explicitly observable. Although annotating data might be extremely difficult and time consuming (or, sometimes, even impossible), supervised learning is still by far the most used approach to machine learning and pattern recognition systems [64].

#### 2.3.1.1 *k*-Nearest Neighbours (kNN)

This is a widely used algorithm in the field of Pattern Recognition with the function of performing classification, regression, missing data imputation or interpolation and density estimation. The idea of *k*-NN is to categorize query points based on their distance to points in a training dataset [65]. So, having as a starting point a set X of n points (training dataset) and a distance function (that may be Euclidean, Hamming, Jaccard, among others) it is possible to find with this algorithm the k closest points in X to a query point in Y (test dataset) [65].

This is a very simple algorithm commonly used in learning frameworks as a first approach to classification problems. Its simplicity makes it possible to compare the obtained results with *k*-NN with other more complex classification techniques [66].

#### 2.3.1.2 Support Vector Machines

Support Vector Machines (SVM) are a set of supervised learning methods that analyse data meant for recognising patterns for further classification and regression. The SVM algorithm as it is used nowadays was published by Vapnik and Cortes in 1995 [67]. It was primarily developed for two-group classification. The general conceptual idea is the following: having a set of training sample with specific labels assigned it is possible to create a non-probabilistic binary linear model to classify unknown samples (testing set).

SVMs are based on the concept of decision planes that define boundaries to make decision among sets of date. Generally, SVMs classify data by finding the best hyperplane to separate all data points of one class from another. The optimized hyperplane is the one that sets the widest margin between the two classes. The support vectors, as seen in Figure 2.8 surrounded by circles, are the points closest to the hyperplane [68].

In the case that the two classes are not linearly separable the hyperplane chosen (in this case denominated by soft margin) will be the one that finds a good balance between the number of points that are possible to separate and the margin between them. In this case it is impossible to avoid some mis-classification. Some binary classifications do not have a simple hyperplane as a useful separating criterion. For those

cases, a more complex classifier, allowing more general boundaries allowing more general boundaries between classes, may be more appropriate [69].

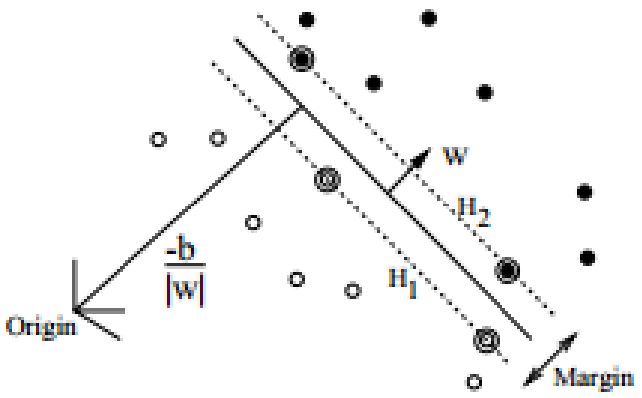


**Figure 2.8:** Linear separating hyperplanes for the separable case. The support vectors are circled. Extracted from [68]

SVMs are used for several application. Osuna et al.[70] used it for detecting frontal human faces on real images, Sharma[71] used it for handwritten digit recognition and Ward et al. [72] to predict the secondary structure of proteins, being these three examples of the most common applications of this method.

#### 2.3.1.3 Cross-validation: evaluating a classifier's performance

Creating models to predict data classification and using the same data both for creating the model and testing the prediction function is a methodological mistake: in this case, a model would just repeat the labels of the samples that it has just seen. The prediction would have a perfect score but its purpose of predicting labels of yet-unseen data would fail (phenomenon called overfitting). Cross-validation is a tool meant for evaluating the performance of a system and its accuracy. Its usage helps avoiding the above mentioned overfitting of data since it provides a much more accurate picture of your system's true accuracy [73]. In general terms with cross-validation we split our data into a training set and a smaller validation set, and then train on the training set and use the validation set to measure accuracy.

An important way of assuring that our data is not biased in some kind way we should take into consideration the following aspects [74]:

- Pick our validation data randomly from our existing collection data. This way you assure that the chosen set is diverse.
- The effectiveness of the performed cross-validation depends on how representative our dataset is of the range of possible inputs we want to see

Cross-validation techniques comprise several different ways of selecting the data to both train and test a model [74]:

1. Exhaustive cross-validation: in this case there is the leave-p-out cross-validation and leave-one-out cross-validation which is a particular case of the first one. The names of these techniques are quite

suggesting: they involve using p (or one) observations as the testing set and the remaining ones as training.

2. Non-exhaustive cross-validation: for this method we present $k$-fold cross-validation (see Figure 2.9) and 2-fold cross-validation as a particularisation of the first one.

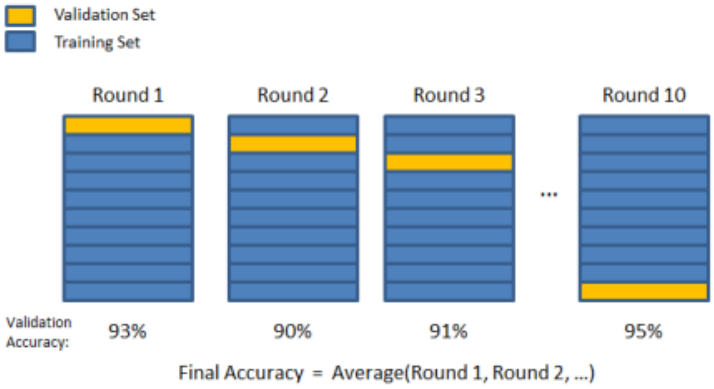


**Figure 2.9:** Schematisation of a $k$-fold cross-validation procedure. Extracted from [74].

### 2.3.2 Unsupervised Classification

Contrarily to Supervised classification, Unsupervised processes do not assume any pior knowledge of the existent pattern classes. So, these are procedures that use unlabelled data in its classification process. Why are unsupervised techniques useful[16]?

- Labelling large collections of data is costly;
- In some applications, the pattern characteristics can change over time. Unsupervised procedures can handle these situations.
- In some situations, unsupervised learning can provide insight into the structure of the data that helps in designing a classifier.

#### 2.3.2.1 $k$-Means Clustering

$k$-means (proposed by MacQueen in 1967[75]) is an iterative algorithm with the objective of classifying a set of data into a set number of clusters. A loose definition of clustering could be ”the process of organizing objects into groups whose members are similar in some way”. A cluster is therefore a collection of objects which are ”similar” to each other and are ”dissimilar” to the objects belonging to other clusters. In the case of this method the number of clusters is an input and the general procedure is the following:

1. Define k centroids (preferably placed as far from each other as possible), one for each cluster.;
2. Calculate the centroids of the clusters;
3. Associate each data point to the cluster with the nearest centroid;
4. Steps 2 and 3 are repeated until no change is observed in the clusters’ centroids.

One thing that it is unknown is the number of iterations that will be necessary to converge the data so, this is a process that may vary in terms of time of computation. The objective of the $k$-means algorithm is

to minimize the within cluster variability. The objective function (which is to be minimized) is the sum of square distances (errors) between each point and its assigned cluster center:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} = \|x_i^{(j)} - C_j\|^2 \tag{2.1}$$

where $\|x_i^{(j)} - C_j\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre $C_j$, is an indicator of the distance of the n data points from their respective cluster centres.

A drawback of this method is that it does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centres. The *k*-means algorithm is usually run multiple times in order to reduce this effect. *k*-means is widely used in the computer vision field namely: for image segmentation[76], feature learning[77] and behaviour pattern's analysis.[78].

#### 2.3.2.2 Agglomerative Hierarchical Clustering

Hierarchical clustering algorithms are approaches that can be either top-down or bottom-up. Bottom-up algorithms treat each sample as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all samples. Agglomerative hierarchical clustering is then a bottom-up clustering method in which clusters have sub-clusters, which in turn have sub-clusters, etc. Agglomerative hierarchical clustering starts with every single object (sample) in a single cluster. Then, in each iteration, it merges (or agglomerates as its name says) the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster. The final cluster shall have the following characteristics[79]:

- Clusters generated in early stages are nested in those generated in later stages.
- Clusters with different sizes in the tree can be valuable for discovery.

A final Matrix Tree Plot visually demonstrates the hierarchy within the final clusters, where each merger is represented by a binary tree. Summing up the general process is [79]:

1. Assign every sample to a separate cluster;
2. Assess all pair-wise distances between the clusters;
3. Build a distance matrix using the computed distance values;
4. Look for the pair of clusters with the smaller distance. Remove that pair from the matrix and merge it;
5. Evaluate all distances from this new cluster to all other clusters, and update the matrix.
6. Iterate until the distance matrix is reduced to a single element.

This s a technique used for example to improve other clustering algorithms [80], widely used for organising species taxonomy [81] or learning human behaviour [82]. These examples of applications trust on this algorithm's main advantages of being able to produce an ordering of the objects, which may be informative for data display and the fact that smaller clusters are generated, which may be helpful for discovery [79].

# Chapter 3

# Methodology

In the current chapter the methods that were combined to meet the objectives proposed for this work will be explained in detail. A newly created database is presented as well as the motivation behind its creation. The process and tools used for annotating parts of the database are also described. Several trajectory and pixel-based approaches are explored to be used as features of body expressiveness. Lastly it will be explained the supervised and unsupervised approaches used to perform classification of the considered data and find answers for our research questions.

## 3.1 LGP Database

### 3.1.1 Database Acquisition

To perform learning techniques there is the need to collect a set of data for studying. For this reason, the acquisition of a complete database containing people from the two populations, deaf and hearing people, took place for the purpose of this study and upcoming ones. The aim of the dataset that will be presented (still in development) is to enable the possibility of performing studies that analyse dialogue relationships (from sociological, psychological and technical perspectives) between two individuals in a relaxed environment. The conversation scenarios were recorded in a video format. The conversation scenarios defined by socio-psychologists from the Faculdade de Psicologia e Ciências da Educação da Universidade do Porto were the following:

1. Conversation between two deaf people;
2. Conversation between two hearing people;
3. Conversation between a deaf and a hearing person.

In order to execute these scenarios, some requirements in terms of the population were defined by the same team of socio-psychologists:

1. The population should be composed by at least six hearing and six deaf people (sign language speakers) from the same gender;
2. The individuals should know each other *a priori* and have some kind of affinity.
3. The acquisition should take place in a venue that was familiar to all subjects.

Table 3.1 displays the population that volunteered for the creation of this database. The contact with the volunteers was obtained by a partnership with the Agrupamento de Escolas Eugénio de Andrade, Escola

EB2/3 de Paranhos. Escola EB2/3 de Paranhos was the selected venue for the acquisition to take place because of the mentioned partnership and also because it was a familiar environment for the volunteer population.

**Table 3.1:** Volunteer population for the creation of the database. All subjects are females between the ages of 27 and 39.

| Deaf People | Gender | Age | Hearing People | Gender | Age |
|---|---|---|---|---|---|
| D1 | Female | 38 | H1 | Female | 38 |
| D2 | Female | 35 | H2 | Female | 27 |
| D3 | Female | 36 | H3 | Female | 30 |
| D4 | Female | 39 | H4 | Female | 29 |
| D5 | Female | 31 | H5 | Female | 30 |
| D6 | Female | 39 | H6 | Female | 37 |
| - | - | - | H7 | Female | 37 |

Since the focus of this database is to enable the analysis of behaviour/expressiveness, a set of conversation topics, that would awaken certain emotions in the individuals, was defined by the same socio-psychologists. Those topics were defined in a staggered way so that the discussion would generate emotions of increasing intensity in the actors of the conversation. The topics were chosen assuming that a dialogue would occur between a pair of subjects and that both would intervene actively. To build a framework for the videos' acquisition, four different conversation topics were defined belonging to two-fold moments: positive (1 and 2) and negative (3 and 4). The topics should be discussed in the following order:

1. **Talk about happy moments:**
    - Describe several happy moments that happened throughout the person's life;
    - Explain why those moments were happy;
    - Explain their context and circumstance;
    - Mention and explain some happy moments outside of the actor's personal life.
2. **Talk about people with which the actor has a strong love or friendship bond:**
    - Describe those people;
    - Explain the reason why the strong bond exists;
    - Explain the importance of those people in a personal point of view.
3. **Talk about sad moments**:
    - Describe several sad moments that happened throughout the person's life;
    - Explain why those moments were sad;
    - Explain their context and circumstance;
    - Mention and explain some sad moments outside of the actor's personal life.
4. **Talk about situations that awaken anger/indignation/injustice**:
    - Explain why those situations cause anger/indignation/injustice;
    - Explain how those situation may affect personal life on the short and long run;
    - Explain their context and circumstance;

- Mention a few typical situations of that kind.

Figure 3.1 sums up the structure of the database in terms of content. The volunteer subjects were coupled so that the conversation scenarios were covered. Each conversation between a pair of subjects was designated as a session having been a total of 9 sessions included so far in the database (whole organization presented on Table 3.2).

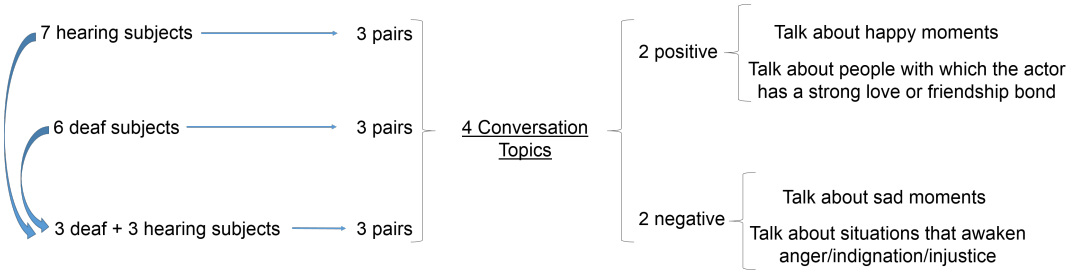


**Figure 3.1:** Diagram of the database structure. This diagram shows the number of pairs available and the conversation topics considered.

**Table 3.2:** Sessions comprising the different pairs of subjects featured in the database.

| Session | Subject |
|---|---|
| 01 | D3 and D2 |
| 02 | H3 and H5 |
| 03 | H1 and H4 |
| 04 | D1 and H6 |
| 05 | H6 and H2 |
| 06 | D6 and H2 |
| 07 | D5 and D6 |
| 08 | D6 and D4 |
| 09 | D5 and H7 |

Figure 3.2 represents the set used to record the videos. IP0, IP1, IP2 and IP3 represent the cameras used and K a Microsoft Kinect. These were all placed in strategic locations (at a height of 2.58 meters) for the best capture possible. The location of the cameras was hidden from the subjects so that the dialogues and reactions would not be affected by the awareness of the presence of the cameras. Two chairs were centred in the room in a way that was propitious for the dialogue in terms of proximity and comfort and for the video acquisition. All distances are presented in meters. Looking at Figure 3.3 it is possible to visually understand the perspective captured by each of the cameras. The same time frame is shown for each camera.

Prior to the start of the acquisitions, the subjects had time to read a script with the topics they would have to talk about, so that they could prepare and remember about some of the situations that they are asked about. Upon completion of each conversation moment, the pair of participants would leave the room, close the door, and re-enter to start the next moment.

At the end of the acquisitions, all subjects filled a questionnaire (previously formulated by the socio-psychologists) in which demographic and opinion enquiries were included. Two versions of the questionnaire were used, one for the deaf subjects and another for the hearing ones (Figure A.5 and A.6 in the Appendix).

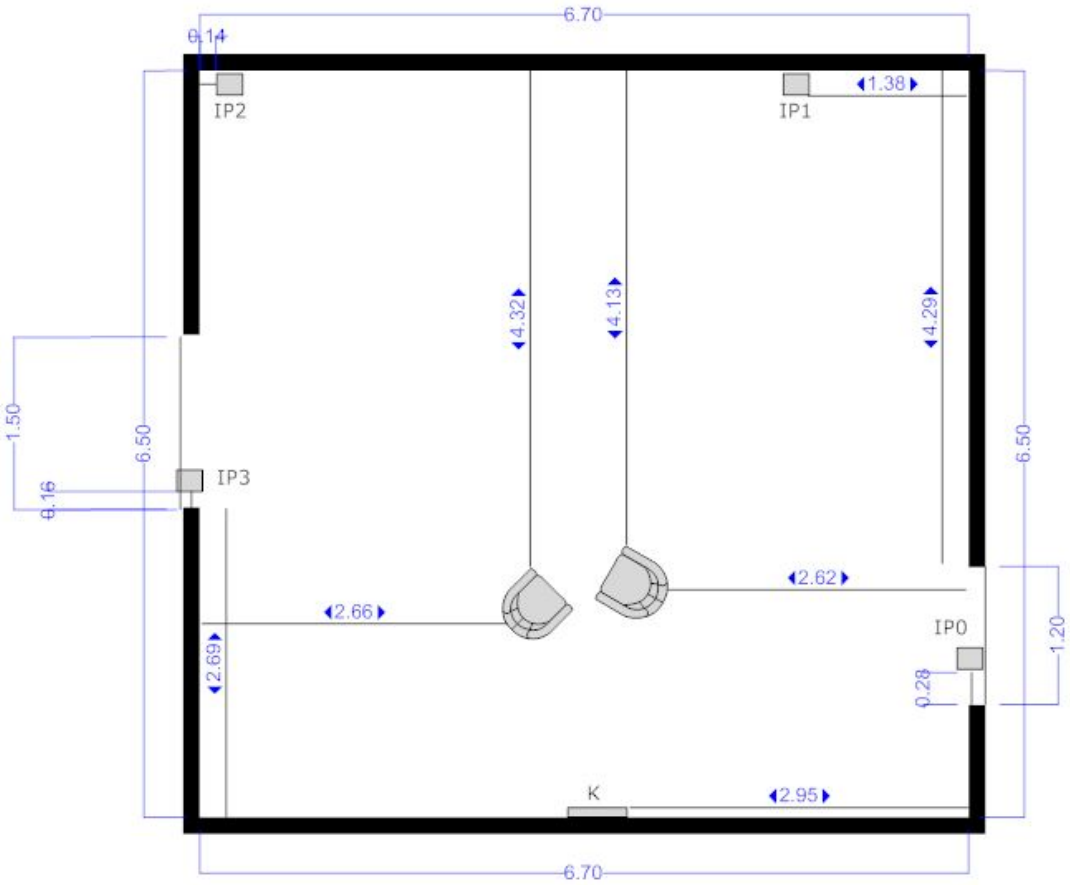


**Figure 3.2:** Sketch of the database acquisition set. The 4 different cameras used are IP0, IP1, IP2 and IP3. A Microsoft Kinect is represented with a K.

### 3.1.2 Manual Annotation

The objective of creating this dataset was to provide the scientific community with a complete set of material to study motion patterns, sign language gestures (and interpreting their semantics), emotions transmitted through facial and body expressions among other aspects.

Annotation is a methodology for adding information to a document at some level: a word or phrase, paragraph or section or the entire document. This information is ”meta-data”, that is, data about other data. One of the desired features for this database under construction was to complete its dual interaction videos with manual annotations regarding sociological, psychological and technical aspects. In the present work technical aspects regarding the motion of different body parts were covered in terms of manual annotation. For this purpose it was necessary to choose an adequate annotation tool among all the ones made available on-line.

VIPER-GT [1] is a tool developed by the Laboratory for Language And Media Processing (LAMP) at the University of Maryland. It consists of a Java graphical user interface that enables the process of authoring ground truth annotation of video data. It is designed to allow frame-by-frame markup of video metadata stored in the VIPER format. It encompasses different attributes which are divided in two categories: text (string based) and spatial (shapes). The idea of this first annotation phase was to create spatial information

[1] http://viper-toolkit.sourceforge.net/


**Figure 3.3:** A frame of a dialogue moment captured from the different cameras P0, P1, P2 and P3. The same time frame is displayed from the perspective of the four cameras.

on the different body parts movement so, only spatial attributes were used. The different shapes made available by VIPER-GT for spatial annotation are: bboxes, oboxes, ellipses, open and closed polygons, points and circles. The body parts annotated for the videos and shapes used were:

- **Head, nose, mouth and trunk:** using oboxes which are oriented rectangles. The string representation of an obox (and the codification that will appear in the xml metadata file) is x y h w and o where x and y are the coordinates of the top left corner of the box, h is the height, w is the width and o the orientation in degrees counter-clockwise. The coordinates in VIPER-GP count down and to the right from the top left corner.
- **Elbows, eyes and hands:** using ellipses which act exactly as oboxes in terms of string representation since it represents an obox bounding the ellipse. All these body parts were annotated for both left and right side.

The process of choosing these specific spatial attributes and descriptors for each body part took into consideration the similarities terms of shape to the actual body part. So, the fact that this tool allows one to create and edit shapes for all frames in video data is valuable for the annotation purpose of this project. Another valuable feature of this tool is the possibility of performing interpolation. Interpolation fills all values of all spatial attributes of all chosen object descriptors for all frames in the open range are overridden with values generated by an algorithm of linear interpolation between key frames predefined. An example of an annotated sequence of frames for the videos of the database is presented in Figure 3.4.

The annotations added so far to the database were not used in the present work directly but are a valuable contribution to the content of the database and the impact that it can have in the scientific community. It is well known that performing manual annotation is a time consuming task, and for this reason, only a part of the dataset was annotated precluding its use as a tool in this study (as a feature on the learning/classification framework). Excerpts (of approximately 15 seconds) from the videos of session 04 were the samples annotated during this study.

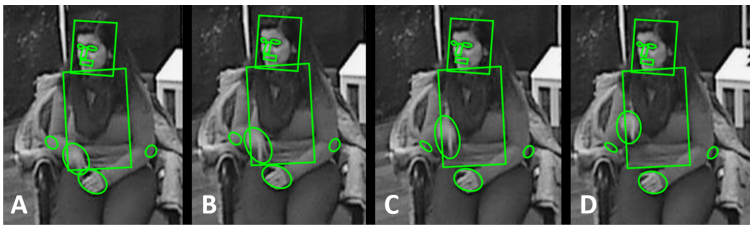


**Figure 3.4:** Sequence of annotated frames. Trunk, nose, head and mouth are annotated using oboxes and eyes, hands and elbows using ellipses.

### 3.1.3 Data Selection for Present Work

Taking into account the objectives of this work, came up the necessity of narrowing the number of videos of the database. This work is oriented to the study of sign language speakers and, for this reason, the sessions in which two hearing people are having a conversation had to be discarded since they are not using sign language. The filtered database used for building the framework for this study was: sessions 04, 06, 07, 08 and 09. Session 01, featuring subjects D3 and D2, was also discarded since the position of the subjects on the set was different (they ware standing) so, it would not be possible to use those subjects as a comparison with the other subjects.

Since the analysis of the subjects' motion was going to be performed in a frontal view, the data considered was only from cameras IP1 and IP2.

## 3.2 Tools and Libraries

In this section will be described the main functionalities of some tools and libraries that were used in this work in the precesses of feature extraction, feature selection and learning stage. The origin and purposes of each tool will be detailed.

### 3.2.1 OpenCV

OpenCV[2] was started at Intel in 1999 for the purposes of accelerating research in the development of applications of computer vision (for commercial purposes) in the world and, for Intel, creating a demand for ever more powerful computers by such applications. This open-source computer-vision library with over 500 functions can greatly simplify computer-vision programming. It includes advanced capabilities - face detection, face tracking, face recognition, Kalman filtering, and a variety of artificial-intelligence (AI) methods in ready-to-use form. In addition, it provides many basic computer-vision algorithms via its

[2]http://opencv.org/

lower-level APIs [83]. One of its main advantages is the fact that it is multi-platform framework: supports Windows, Linux and most recently Mac OS X.

Since the feature extraction of this work was made in **MATLAB r2011a** it was necessary to use the integration of the OpenCV computer vision library with MATLAB. This is made using a MEX interface (which stands for MATLAB Executable and provides an interface between MATLAB or GNU Octave and subroutines written in C, C++ or Fortran. ) allowing one to perform algorithm development, data analysis, and numerical computation in MATLAB. Basically when compiled, MEX files are dynamically loaded and allow non-MATLAB code to be invoked from within MATLAB as if it was a built-in function. Some features that will be presented below were obtained using functions from the Motion Analysis and Object Tracking [3] library included in OpenCV.

### 3.2.2 EyesWeb

Eyesweb[4] is a non-profit open software platform developed by Lab. InfoMus - DIST - University of Genoa, Italy. In short it is a Microsoft Windows based tool for creating interactive digital multimedia applications through visual programming. This tool enables the possibility to form digital sound and images in real-time, through the use of various Human-Computer Interactions (HCI). These interactions include, but are not limited to: Object Identification, Segmentation and Recognition, Face Recognition, Gesture Recognition and Motion Tracking. The mentioned functionalities are extremely valuable for a work such as the one that it is being hereby described. This tool's blocks are based on the OpenCV library. The EyesWeb Development Environment to build patches (composition of several Eyesweb blocks with a specific function) is shown Figure A.1 included in the Appendix.

### 3.2.3 Weka

Weka [5] is a suite of machine learning algorithms for data mining tasks written in Java. It is an open source software available under the GNU General Public License. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. On the presented framework Weka was used for feature selection since it is shown in literature to be a valuable tool for this type of task [84, 85].

### 3.2.4 CLUTO - A Clustering Toolkit

CLUTO[6] is a software package for clustering low and high dimensional datasets and for analysing the characteristics of various clusters. CLUTO offers three different types of clustering algorithms based on different paradigms: the *partitional*, *aglomerative* and *graphpartitioning*. A great advantage of the algorithms featuring in CLUTO comparing to other techniques is that CLUTO deals with the clustering problem as an optimization process which seeks to maximize or minimize a particular clustering criterion function defined either globally or locally over the entire clustering solution space. CLUTO provides a total of seven different criterion functions that can be used to drive both partitional and agglomerative clustering algorithms, that are described and analyzed in [86, 87].

[3] http://docs.opencv.org/modules/video/doc/motion_ analysis_ and_ object_ tracking.html
[4] http://www.infomus.org/eyesweb_ita.php
[5] http://www.cs.waikato.ac.nz/ml/weka/
[6] http://glaros.dtc.umn.edu/gkhome/views/cluto

A feature that CLUTO possesses that was considered to be very valuable and one of the main reasons for choosing this tool is the fact that it provides tools that analyse the discovered clusters, understand the relations between the objects assigned to each cluster and the relations between the different clusters. The visualisation tools of the outcomes of the clustering solution's statistics are also user friendly. CLUTO's algorithms have been optimized for operating on very large datasets both in terms of the number of objects as well as the number of dimensions. These algorithms can quickly cluster datasets with several tens of thousands objects and several thousands of dimensions.

CLUTO's distribution consists of both stand-alone programs (vcluster and scluster) for clustering and analyzing these clusters, as well as, a library via which an application program can access directly the various clustering and analysis algorithms implemented in CLUTO.

## 3.3 Feature Construction

This section introduces the various aspects related to the feature extraction process in this work that aimed to enable unsupervised and supervised machine learning techniques. When using machine learning techniques it is common to represent data by a fixed number of features (attributes) which can be binary, categorical or continuous. Finding a suitable data representation is not an easy task being strongly related to the measurements that are possible to perform with the available data [88].

Before building the feature vector the following statistics were measured for each video: frames per second (fps), total number of frames and duration. A continuous sampling was performed so that in the end were obtained videos with the same number of frames in total (120) and also the same fps. As a consequence, and depending on each video's fps, some videos underwent a downsample and other an upsampling. Table A.1 in the Appendix sums up the definitive number of samples obtained for each subject in each conversation topic. From now on these samples will be refereed as **miniclips**. Finally it was established for each video a region of interest (ROI) so that only the area bounding the subject was considered (in order to avoid noise interference in the subsequent analysis).

### 3.3.1 Motion History Image

The Motion History Image (MHI) is a static image template where pixel intensity is a function of the recency of the motion in a sequence. Basically in an MHI, pixel intensity is a function of the motion history at that location, where brighter values correspond to more recent motion. One of the advantages of the MHI representation is that a wide range of temporal motion events may be encoded in a single frame, and in this way, the MHI spans the time scale of human gestures. The MHI $H_\tau\ (x,y,t)$ can be computed from an update function $\psi(x,y,t)$ [89]:

$$H_\tau(x,y,t) = \begin{cases} \tau & if\,\psi(x,y,t) = 1 \\ max(0, H_\tau(x,y,t-1) - \delta) & otherwise \end{cases} \tag{3.1}$$

Here, $x$, $y$ and t show the position and time, $\psi(x,y,t)$ signals object's presence (or motion) in the current video image, the duration $\tau$ decides the temporal extent of the movement (e.g., in terms of frames), and $\delta$ is the decay parameter.

The final value to be included in the feature vector that will characterize each miniclip is the mean value of the **Quantity of Motion (QoM)** in each frame being the QoM the weighted sum of all the pixels of the motion history image: a miniclip in which wider movements are performed will have a higher QoM. The

standard deviation of the QoM of all the frames of a certain miniclip will also be an entry of the final feature vector. The miniclips in which the mean QoM was too low were suppressed.

### 3.3.2 Motion Gradient

The motion gradient is a feature that is obtained by the computation of the gradient of the MHI. Using this feature it is possible to get direction vectors pointing in the direction of the movement of a silhouette. The computation of this gradient is obtained by performing the convolution with separate Sobel filters in the X and Y directions [90]. Mathematically, the Sobel filter uses two 3×3 matrices that are convoluted with the image to compute approximations of the image derivative. One matrix is used for horizontal variations and another for vertical[90].

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \qquad G_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{3.2}$$

Orientation is then computed the following way:

$$orientation(\phi) = arctan\left(\frac{\frac{\partial MHI}{\partial y}}{\frac{\partial MHI}{\partial x}}\right) \tag{3.3}$$

The value to be included in the feature vector is also computed weighting for each frame the orientation obtained on each pixel by the value of the MHI on that pixel attributing this way more importance to the orientation associated to the pixels with a more recent motion. Mean and standard deviation values (regarding all video frames) of the weighted orientation entered the final feature vector of each miniclip.

### 3.3.3 Motiongrams

One of the features explained previously was the QoM. It is a rough estimation of the quantity of movement, and does not tell anything about the location of the movement inside the frame. This makes it difficult to know where the movement happened in the frame, for example whether movement was happening in the head, torso or feet. A solution to complement this idea would be to create a display that could visualise the QoM as well as the distribution of QoM in time and space. In this work it is proposed to combine motiongrams with MHI, in order to overcome these problems. The idea of the motiongram representation emerged from an analogy with the widely used waveform displays and spectrograms for efficient visualisation of some important features of audio material.

The motiongram is an approach in which motion is measured by summing up the active pixels in a motion image and plotting the value over time. The motiongram image obtained will give us the notion of the overall motion qualities in a video preserving some of the spatial information of where in the image/body the motion has occurred [91]. For our miniclips, two types of motiongram were obtained: a vertical one (for which the MHI intensity over all rows in a column was assessed) and a horizontal one (for which the MHI intensity over all columns in a row was assessed). The motiongram approach is based on collapsing a matrix of size M×N into two different ones of M×F (in which M is the number of rows and F the number of frames in a miniclip) and N×F (in which N is the number of rows and F the number of frames) for horizontal and vertical mostiongrams respectively. Figure 3.5 shows an overview of the processing of a acquisition of a video motiongram.

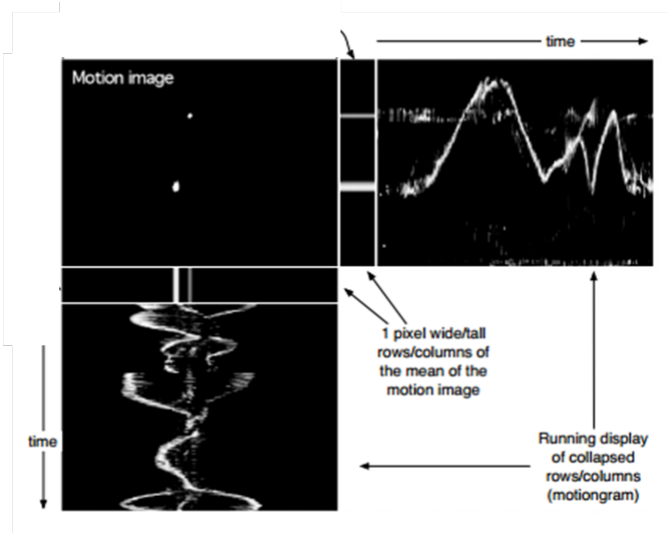


**Figure 3.5:** An overview of the process of creating a motiongram, showing the motion image, and the running motiongrams. Extracted from [91].

The features used (for each miniclip) were histograms in which every bin contained the normalized sum of the motiongram information over all frames for each column and row for vertical and horizontal motiongrams respectively.

### 3.3.4 Body Part Tracking

Figure 3.6 shows the tasks performed using the blocks available in Eyesweb. Firstly, the head, right arm and left arm of the subject in focus in a certain miniclip were tracked by a tracking algorithm (the patch used may be seen in the Appendix on Figure A.2). Tracking is never an easy task having most of the times to be executed in a semi-automatic way. In this case a simple tracking was used by an algorithm that from a given starting point performs template matching (through a specific kernel) between the current frame and the previous one. A distance threshold was defined so that if that value was exceeded the tracking point would be ”lost” and the tracking reinitialised with a new pointed manually inserted. With the tracking of these three points it was possible to extract the following relevant features:

**Occupation Rate**: the idea was to use a specific block in Eyesweb (Position Dependent Potentials) which, using a grid, divides an image (frame) in cells and uses it as input for a model that compute on the indexes of the cells the percentage of occupation of a certain point being tracked in a video. The percentage of occupation in each cell is updated over time (for the whole duration of the video). The feature extracted was the mean and standard deviation occupation rate of all cells in a certain video.

**Kinematic Features**: from the tracked points referred above it was possible, using two blocks available in Eyesweb (Figure A.4), to extract trajectories from a set of points and to compute from those trajectories kinematic features. From the trajectories (of head, right arm and left arm) both X and Y components of velocity, acceleration and movement direction were extracted and mean and standard deviation values were calculated for each miniclip and added to the feature vector.

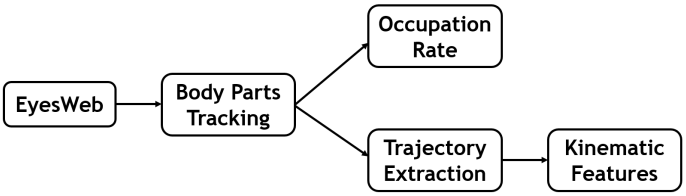


**Figure 3.6:** Schema of relationships of the features extracted using Eyesweb.

## 3.4 Feature Selection

In this section another problem in the process of extracting features is approached: the selection of the most relevant and informative features.

In machine learning, variable selection is used to find the subset of the available inputs that accurately predicts the output and remove the ones that are redundant and/or irrelevant. The objectives of variable selection are to improve predictive performance of the model, provide faster and more cost-effective predictors, and provide a better understanding of the underlying process that generated the data. Two different methods available in Weka (3.2.3) were used for feature selection.

**Information Gain Attribute Evaluation** evaluates the worth of an attribute by measuring the information gain with respect to the class (works for multi-class problems). It is widely used for standard feature selection method but has the disadvantage that does not take into account feature interaction [92]. With this method it is possible to get a ranked list of the most predictive features according to its Information Gain score.

**ReliefF Attribute Evaluation** implements the Relief method (proposed by Kira and Rendell in 1992 [93]) to apply multivariate relevance criteria to rank individual features according to their relevance in the presence of others. The Relief algorithm uses an approach based on the $K$-Nearest Neighbour algorithm. To evaluate the index, there is the need to firstly identify in the original feature space, for each example, the K closest examples of the same class(nearest hits) and the $k$ closest examples of a different class (nearest misses). Then, in projection on a given feature, the sum of the distances between the examples and their nearest misses is compared to the sum of distances to their nearest hits. Using the ratio of these two quantities it is possible to create an index independent of feature scale variations. The Relief method works for multi-class problems [88].

## 3.5 Classification Methods

The learning stage of the work being presented is essential since the different learning methods that will be addressed intend to answer the research questions enunciated as objectives of this work. A learning mechanism is all about finding patterns in the available data, being a pattern a process or event that can be given a name. The components of a pattern recognition system are presented in Figure 3.7 from which the components of pre-processing and feature extraction were already approached. In this section, the task of finding patterns that belong to the same class in our data will be addressed. That may be achieved by using classifiers (decision rules) which decide about the pattern based on an observation. Considering the components in the figure, the *feature extraction* aims to create discriminative features good for classification, the *teacher* (present in the case of supervised learning) provides information about the pattern, the *classifier* assigns classes to observations and the *learning algorithm* sets the pattern recognition for training examples.

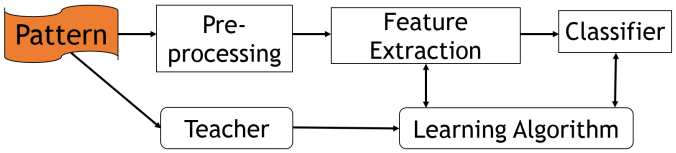


**Figure 3.7:** Components of a pattern recognition system .

### 3.5.1 Principal Component Analysis

Principal Component Analysis (PCA) is a useful statistical technique that has found application in fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension. PCA is a way of finding out which features are important for best describing the variance in a data set. It is most often used for reducing the dimensionality of a large data set so that it becomes more practical to apply machine learning where the original data are inherently high dimensional. This step may be included in the *Feature Extraction* component in Figure 3.7 being an important stage prior to classification.

PCA was applied for the final feature matrix so that the features would undergo an orthogonal transformation to be converted to a set of different features called the principal components of our data. The main steps of the PCA are the following [94]

1. **Obtain a feature matrix**: this is a M×N matrix in which M are the observations and N the number of features.
2. **Subtract the mean**: for PCA to work properly, there is the need to subtract the mean from each of the data features. This way the mean value of all observations will be set to zero for each feature.
3. **Calculate the covariance matrix**:a covariance matrix is a matrix whose element in the $i$, $j$ position is the covariance between the ith and jth elements of a vector (being covariance the measure of how much the features vary from the mean with respect to each other). A covariance matrix is always computed to compare two different features so, if we have a data set with more than 2 features (which is our case), there is more than one covariance measurement that can be calculated. In fact for an $n$-feature dataset can be calculated $\frac{n!}{(n-2)!\times 2}$ different covariance matrices. In general a $n$-features dataset, will originate a matrix that has $n$ rows and columns (so is square) and each entry in the matrix is the result of calculating the covariance between two separate features.
4. **Calculate the eigenvectors and eigenvalues of the covariance matrix**: let $A$ be an $n \times n$ matrix. The number $\lambda$ is an eigenvalue of $A$ if there exists a non-zero vector $v$ such that $Av = \lambda v$ In this case, vector $v$ is called an eigenvector of $A$ corresponding to $\lambda$. To the mentioned matrix of $n \times n$, $n$ eigenvectors are attributed. By this process of taking the eigenvectors of the covariance matrix, we have been able to extract lines that characterise the data.
5. **Choosing components**: in this step is where the notion of reduced dimensionality appears. Looking at the eigenvectors and eigenvalues it is possible to notice quite different eigenvalues. The main conclusion of this process is that the **Principal Components** of a feature vector will have higher eigenvalues. Having the components in order of significance, one can choose to ignore the components of lesser significance. Information is obviously lost however, if the eigenvalues are small, not much is lost. To sum up if some components are left out, the final data set will have less dimensions than the original.

### 3.5.2 Distinguishing Deaf from Hearing People

The task of distinguish the deaf population from the hearing one is one of the research questions of this work since spotting the aspects that differentiate these two populations may lead to finding the answers to remove the gaps between them. The assessment performed here is a first step towards that goal.

The *k*-Nearest Neighbours was one of the methods used for classification. The theory behind this classification algorithm (which belongs to a set of techniques called Instance Based Learning [95]) was previously explored in the 2.3.1.1. It is a very simple algorithm used frequently as a first approach to classification problems. This happens since *k*-NN is considered a *non-parametric lazy* algorithm meaning that it does not make any assumptions on the underlying data distribution and it does not use the training data points to do any generalization, reducing the training phase to a minimum. Recalling the method behind this algorithm: it starts by extending the local region around a data point until the $k^{th}$ nearest neighbour is found. For nominal data (which is the considered case), an object is classified by a majority vote scheme, with the object being assigned to the class most common amongst its *k*-nearest neighbours. There is a phenomenon, commonly referred as *Curse of Dimensionality*, strongly affects the performance of *k*-NN classifier. It states that in high dimensional spaces distances between nearest and farthest points from query points become almost equal. Therefore, nearest neighbour calculations cannot discriminate candidate points [96]. *k*-NN may also ”break down” when the data contains irrelevant/noisy features. To mitigate the occurrence of this phenomenon PCA was applied to the feature vectors.

As well as *k*-NN the SVM algorithm was already explained in 2.3.1.2. The approach for the distinction between deaf and hearing people was a regular SVM binary classification in which a support vector machine was trained, and then cross validation was performed on the classifier. The trained model was then used to classify (predict) new data (test data). The kernel function used was linear and several scaling factors (C parameter for the standard C-SVM formulation) were tested in order to infer which one would produce better results.

#### 3.5.2.1 Grouping the Miniclips

***k*-Nearest Neighbours** and **Support Vector Machines** were two classifiers used for pattern recognition in this work. Logically, for this problem the two classes were known *a priori* for each miniclip, if the performer was a deaf or a hearing person. Table 3.3 shows the different ways in which the miniclips were grouped. The intention was to compare the classification performances for the different groups (miniclips from each topic and also all topics grouped). Both algorithms were used with cross-validation (with the number of folds dependent on the number of samples per grouping) in order to avoid overfitting (phenomenon explained in 2.3.1.3). The statistical measures used to evaluate the performance of these two classifiers were the Confusion Matrix, Correct Rate, Recall and Precision (explained in detail on the Results and Discussion chapter.

**Table 3.3:** Sample grouping used for classification of subjects regarding their deaf or hearing condition.

| Grouping | Number of Samples |
|---|---|
| Topic 1 | 234 |
| Topic 2 | 296 |
| Topic 3 | 361 |
| Topic 4 | 375 |
| All Topics | 1266 |

### 3.5.3 Distinguishing Different Conversation Topics

Regarding the differentiation of the conversation topics (recalling that there are two with positive connotation and other two with negative)spotting the differences in terms of expressiveness among the four topics was the primary reason why these specific moments were included as a feature of the database described above. The same classifiers detailed on the previous subsection, *k*-NN and SVM, were used to approach this problem. *k*-NN was used in the same fashion but, on the other hand, as stated before, SVM is typically used for binary classification however, in this case we had, for the classification of the conversation topics, four different classes which leads to using multiclass SVM (with cross-validation). This problem was generalized with the approaches by Allwein et al. and Cardoso and Cardoso [97, 98]. In this work, and because the different categories are ordered, were performed three binary classifications, Topic1/Topic2, Topic2/Topic3 and Topic3/Topic4. With this vision of the multiclass classification problem it is possible to simplify the algorithm suggested in [99]. In this case a linear discriminator of all classes was necessary so we chose a weighted sum of the individual features and the bias to differentiate classes:

$$y(x) = x^T w + b \tag{3.4}$$

#### 3.5.3.1 Grouping the Miniclips

The performance evaluation was made using the same statistics used for differentiating deaf from hearing people. The groupings of miniclips used were the ones shown on Table 3.4.

**Table 3.4:** Sample grouping used for classification of the conversation topics.

| Grouping | Number of Samples |
|---|---|
| Subject D1 | 158 |
| Subject D4 | 73 |
| Subject D5 | 328 |
| Subject D6 | 265 |
| Subject H2 | 97 |
| Subject H6 | 157 |
| Subject H7 | 188 |
| All Subjects | 1266 |

### 3.5.4 Identifying Levels of Mastery in Portuguese Sign Language

For the purpose of distinguishing different levels of expertise in LGP it was used an agglomerative hierarchical clustering method to the feature data. To identify groups of similar feature values clustering procedures use distance measures to group data points in a way that provides minimal inner-cluster distances and maximal inter-cluster distances [100]. Contrarily to supervised techniques classification with unsupervised clustering is used if the groups are not known in advance which is the case our data for this purpose. The possibility of our miniclips containing information regarding this question was not known *a priori* meaning that we were not aware if it was going to be found an answer to this research question with the support of our data. This makes this question undoubtedly very abstract but also a valuable addition to the overall framework of this study.

CLUTO was the tool used to solve this problem. The clustering methods available treat each object as a vector in a high-dimensional space, and compute the clustering solutions using one of five different

approaches. Four of these approaches are partitional in nature, whereas the fifth approach is agglomerative (the one used). One of the drawbacks of the agglomerative approach is the requirement for the number of clusters, K, to be specified before the algorithm is applied. In order to overcome this issue, using the questionnaires that all subjects filled, it was decided to define 3 and 4 levels of expertise in LGP based on the score of each subject's answers to the following questions: number of years that was familiarised with sign language and the current profession. The combined score of the two answers was also used. Table 3.5 shows the organisation of the subjects in levels regarding their answers to the two mentioned questions. The weighted combination of the scores of each answer (0.5 for each) originated 4 different levels.

**Table 3.5:** Division of our population regarding the number of years in contact with the LGP and current job. for classification purposes.

| Level | Year Range | Subjects | Level | Current Profession | Subjects |
|---|---|---|---|---|---|
| 1 | 7-15 | D1/H2/H7 | 1 | Non Related to LGP | D4 |
| 2 | 16-25 | H6/D5/D6 | 2 | Speech Therapist | H7 |
| 3 | 26-35 | D4 | 3 | LGP Interpretation | H2 |
| - | - | | 4 | LGP Teaching | D1/D5/D6/H6 |

Without applying prior knowledge about classes, the clustering process aimed at discovering the inner structure of the data and possibly grouping the subjects in relevant clusters.

For the different clusterings solutions performed the quality was measured by using two different metrics that look at the class labels of the documents assigned to each cluster. The first metric is the widely used entropy measure that looks are how the various classes of documents are distributed within each cluster, and the second measure is the purity that measures the extend to which each cluster contained documents from primarily one class.

Given a particular cluster $S_r$ of size $n_r$ , its entropy is defined by:

$$E(S_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r} \tag{3.5}$$

where $q$ consists in the number of classes in the dataset, and $n_r^i$ is the number of samples of the $i$th class that were assigned to the $r$th cluster. Equation 3.5 measures the entropy of a cluster individually however, so that we can have information about the entropy of the entire clustering, a sum of the individual cluster entropies weighted according to the cluster size is more informative:

$$Entropy = \sum_{r=1}^{k}\frac{n_r}{n}E(S_r) \tag{3.6}$$

Ideally a clustering solution will have an entropy of zero. This would reveal that each cluster would contain only samples from a single class. In a similar fashion the purity of this same cluster is computed as follows:

$$P(S_r) = \frac{1}{n_r}max(n_r^i) \tag{3.7}$$

which is the ratio between the overall cluster size that the largest class of documents assigned to that cluster represents. The overall purity in given by:

$$Purity = \sum_{r=1}^{k}\frac{n_r}{n}P(S_r) \tag{3.8}$$

Concerning these two measures it is vital to refer that they intend to give us an indication of the clustering method since small entropy values and large purity values are indicators of a good clustering solution.

**Summing up** the main aspects explored in this chapter it is important to highlight the definition of the video database with conversational scenarios with both deaf and hearing populations as well as the main features (different camera views, technical annotations and conversation topics) included. The contents of this database were filtered for the purpose of this study and that data was used for feature extraction using different pixel and trajectory based techniques. Finally the supervised ($k$-NN and binary and multiclass SVM) and unsupervised (agglomerative hierarchical clustering) classification methodologies used to answer the research questions of this work stated in Chapter 1 ware detailed and its use explained for our context.

# Chapter 4

# Results and Discussion

In this chapter the main results of the described methods are presented as well as a relevant discussion. This chapter is divided regarding the main research questions to be answered, comprising as well relevant results that lead to the answer of those questions: differentiating deaf and hearing people, identifying different conversation topics (positive and negative connotation) and group the evaluated population in different levels of mastery in Portuguese Sign Language.

## 4.1 Feature Construction

The first clarification to be made is to reveal how the full feature vector (whose features were detailed in the previous chapter) was build. Table 4.1 show the appearance of the feature vector in which all features were compiled. The explanation of the name of each feature is the following:

- **mean_ QoM_ MHI**: mean value of the QoM (using the MHI) of all frames in a miniclip.
- **std_ QoM_ MHI**: standard deviation of the value of the QoM (using the MHI) of all frames in a miniclip.
- **mean_ Gradient**: mean value of the gradient of the MHI of all frames in a miniclip.
- **std_ Gradient**: standard deviation value of the gradient of the MHI of all frames in a miniclip.
- **horizontal_ motiongram**: histogram of the horizontal motiongram of a miniclip.
- **vertical_ motiongram**: histogram of the vertical motiongram of a miniclip.
- **mean_ TF**: mean values of the three trajectory features considered (velocity, acceleration and direction) for the whole duration of each miniclip (the three features considered separately despite appearing together in this representation). Values computed for the head, right arm and left arm.
- **std_ TF**: standard deviation values of the three trajectory features considered (velocity, acceleration and direction) for the whole duration of each miniclip (the three features considered separately despite appearing together in this representation). The trajectory features were computed for the three body parts tracked: head, right arm and left arm.

The full matrix of our miniclips had these attributes for all video samples. In the rest of this section will be given some relevant details regarding the implementation of some of the features used.

**Table 4.1:** Tabular schematisation of the structure of the feature vector. The full feature matrix M×N matrix in which M is the number of miniclips and N the number of features.

| mean_ QoM_ MHI | std_ QoM_ MHI | mean_ Gradient | std_ Gradient |
|---|---|---|---|
| vertical motiongram | horizontal motiongram | mean_ TF_ Head | std_ TF_ Head |
| mean_ TF_ RightArm | std_ TF_ RightArm | mean_ TF_ LeftArm | std_ TF_ LeftArm |

## 4.2 Implementation Details

The **MHI** feature was calculated using the OpenCV function *updateMotionHistory* from the Motion Analysis and Object Tracking library. To estimate the MHI of a frame a set of parameters needs to be inputted to the function:

1. silhouette: Silhouette mask that has non-zero pixels where the motion occurs.
2. mhi: Motion history image that is updated by the function (single-channel, 32-bit floating-point).
3. timestamp: current frame.
4. duration: maximal duration of the motion track in frames.

In order to obtain the MHI it was necessary to calculate motion masks (silhouette). These masks were obtained by subtracting consecutive grayscale video frames enabling the extraction of moving elements from static irrelevant background. The image obtained after subtracting two consecutive frames was afterwards binarised using the Otsu method. The duration parameter was empirically inferred. It corresponds to the maximal time that the motion history will be saved meaning that when that duration is reached the gray pixels corresponding to the oldest timestamp are converted to black pixels. This inference was done by testing different durations for the several miniclips to be evaluated. A single value (duration of 10 frames) was chosen for all miniclips so that the superimposition of different movements (in different directions for example) was reduced to a minimum. Figure 4.1 shows a sequence MHI images. The fact that the subjects are considerably distant from the camera, diminishes the quality of pixel-based approaches since the resolution is not the desired to extract the maximum detail from the frames. The presence of several isolated non-zero pixels actually confirms that fact. On the other hand the conceptual idea of the MHI is captured being clearly visible the history of the arms' motion kept on the pixel intensity.

The **Motion Gradient** was also calculated with the support of the OpenCV Motion Analysis and Object Tracking library. The function used was *calcMotionGradient* that gives as output the gradient mask and orientation (see 3.3) with the following parameters as input:

1. mhi: motion history single-channel floating-point image.
2. delta1: minimal (or maximal) allowed difference between MHI values within a pixel neighbourhood.
3. delta2: Maximal (or minimal) allowed difference between MHI values within a pixel neighbourhood. That is, the function finds the minimum $m(x,y)$ and maximum $M(x,y)$ MHI values over a 3 x 3 neighbourhood of each pixel and marks the motion orientation at $(x,y)$ as valid only if $min(delta1, delta2) <= M(x,y) - m(x,y) <= max(delta1, delta2)$

The delta parameters were chosen empirically, i.e., several values were tested until the best combination (that originated the best results) was found.

Lastly some details regarding the implementation of the **Motiongrams** will be addressed (see Figure 4.2 for an example of vertical and horizontal motiongrams). As explained before, for each video was

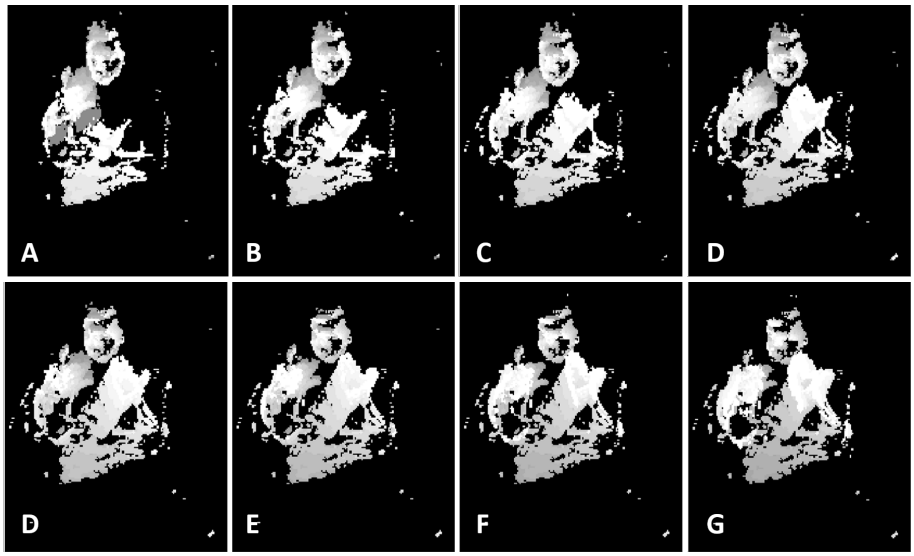


**Figure 4.1:** Sequence of MHI images obtained from a sequence of frames of a miniclip. Higher values of pixel intensity (brighter) represent pixels in which motion occurred more recently.

selected a ROI that would incude the subject in analyses. For this reason all obtained miniclips had different resolutions. Since the histograms of the motiongrams are features whose size (in terms of bins) is directly related to the resolution (number of rows and columns of an image), a mathematical standardisation had to be performed.

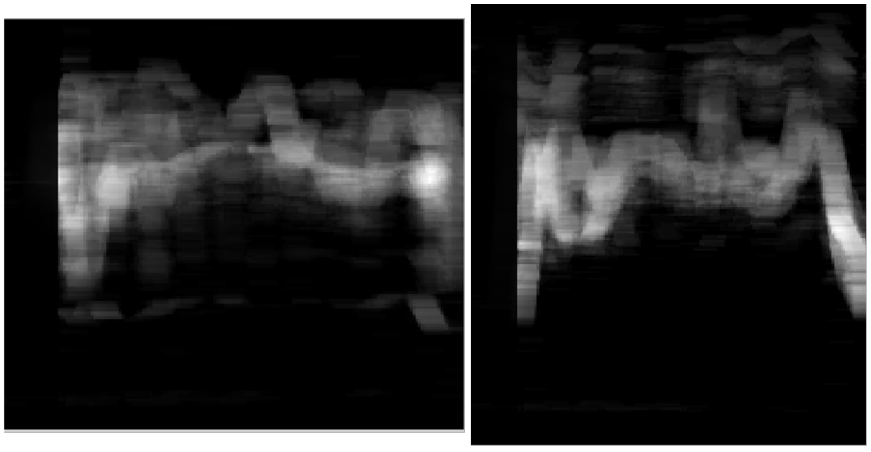

**Figure 4.2:** Examples of motiongrams of a miniclip before size reduction. Left: horizontal motiongram. Right: vertical motiongram.

Figure 4.3 shows a flowchart detailing the manipulation that was performed to the histograms of every miniclip to reduce their size to 114 bins in the one concerning the vertical motiongram and to 162 concerning the horizontal motiongram. These two values are the minimum values found for width and height respectively in all the miniclips. This simple computation allowed the vectors on the feature matrix to have the same dimension.

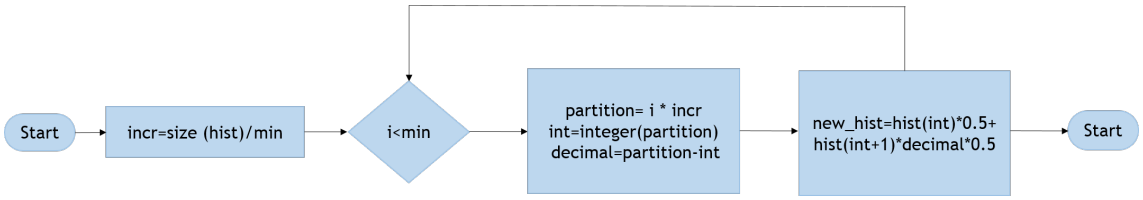


**Figure 4.3:** Fluxogram representing the size reduction of the histograms.hist is the original histogram; size(hist) represents the size (bins) of the histogram being considered; min is the the dimension (rows) of the smallest histogram from all videos (114 for vertical and 162 for the horizontal); integer(partition) is the integer part of the value of partition; decimal is the decimal part of the value of partition.

## 4.3 Feature Selection

The feature reduction is one of the key processes for knowledge acquisition. Being the considered feature data multidimensional, large in size and being intended for classification, if we do not evaluate the relevancy of each feature the classification may end up with wrong results and occupying resources especially in terms of time. The main goal is to discard the redundant and inconsistent features that affect the classification. Figure 4.4 and 4.5 show the results of the feature evaluation using the **Information Gain** and **ReliefF** methods (which were used in Weka with the default parameters). In terms of revealing the features that are most discriminative both methods are concordant leading use to conclude that the motiongrams (vertical and horizontal) are the ones in which more significant differences are observed between the miniclips used, in other words, the variance observed among all samples is higher.

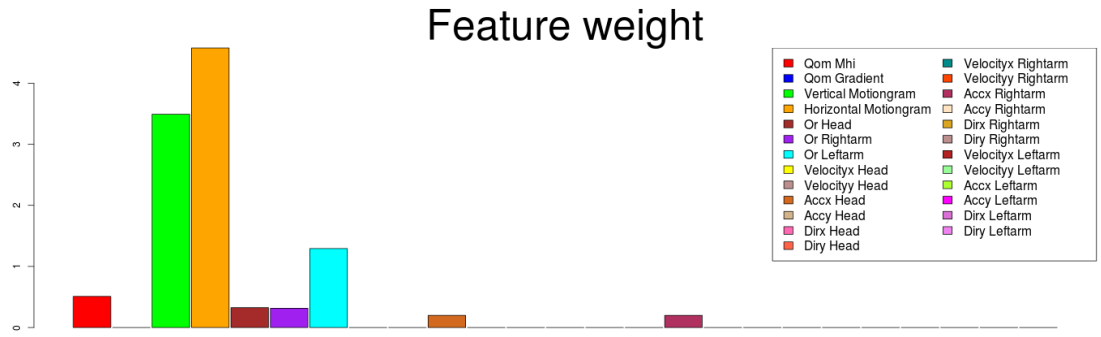


**Figure 4.4:** Feature selection performed in Weka using the Information Gain method. Vertical and horizontal motiongrams display the highest weight.

This outcome of the feature selection was somewhat expected, at least regarding the power of the motiongrams. This feature makes a spatio-temporal representation of the motion on the miniclips which is a thorough approach. They are running displays that make it possible to see both the location and level of movement of a video sequence over time. Although our miniclips were affected by noise from background pixels, Jensenius [101] states that the motiongram implementation is capable of visualising the main motion features even with quite drastic changes of for example inversion, colour, filtering, background, lighting, clothing, video size and compression. The trajectory features were also expected to be relevant and able to differentiate the miniclips however their weight in both charts is not significant. The tracking method and the process of extraction of features from trajectories made available in EyesWeb was most likely not

effective or not adequate for our videos. In the future, different trajectory-based approaches should be tested.

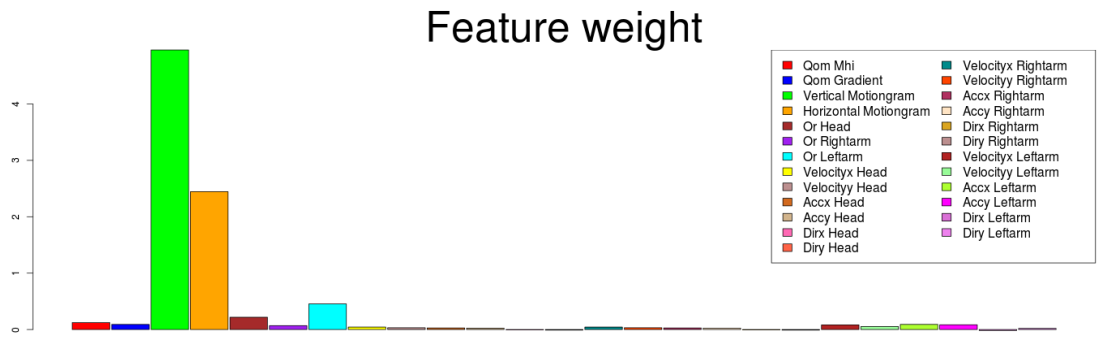


**Figure 4.5:** Feature selection performed in Weka using the ReliefF method. Vertical and horizontal motiongrams display the highest weight.

## 4.4 Classification Analysis

From this point on, the actual results and achievements of this research work will be presented. The main measures extracted from the classifiers used to answer the research questions will be presented in a tabular form for a better understanding of the performance of each solution. The discussion and main inferences regarding the several results are also stated. Taking into consideration the task of selecting the most relevant features only the vertical and horizontal motiongrams were used to generate the upcoming results.

### 4.4.1 Distinguishing Deaf from Hearing People

Distinguishing deaf from hearing people in terms of expressiveness when it comes to using LGP is unquestionably the main enquiry to be answered by this work. As presented on the motivation section (1.1) finding the differences between these two populations in terms of expressiveness is a step in the process of dealing with the gaps that separate these populations.

In order to attempt on finding differences among the miniclips of deaf and hearing people two different classifiers were used: ***k*-Nearest Neighbours** and **Support Vector Machines**. The measures used to evaluate these tests were:

- **Confusion Matrix (CM)**: contains information about actual and predicted classifications done by a classification system. It can give us several information regarding how many samples from each class were correctly classified or not (true positives, true negatives, false positives and false negatives). The result tables present the **Worst CM** of each miniclips' grouping and also the **Cumulated CM** which is the sum of the CM that result from the several fold of the cross-validation process. In this work all CM were normalized using the total amount of samples of each class. For this reason every cell in a line of a CM represents a percentage of samples classified as a certain class. This favours a simpler analysis and comparison of the results.
- **Correct Rate (CR)**: percentage of correctly classified samples.
- **Recall**: fraction of retrieved instances that are relevant (also called Sensitivity).
- **Precision**: fraction of retrieved instances that are relevant (also called Positive Predictive Value).

#### 4.4.1.1 *k*-Nearest Neighbours

As mentioned on section 3.5.2 the performance of *k*-NN algorithm is frequently affected by problems related to the elevated number of dimensions of the considered data. Recent research indicates that the number of dimensions alone does not necessarily result in problems finding the nearest neighbours [102], since relevant additional dimensions can also increase the contrast. The difficulties only arise when irrelevant dimensions reduce the contrast on the features off the samples considered. To test if irrelevant dimensions were present and also to reduce the computation cost in terms of time, the Principal Component Analysis method was applied to the feature matrix. This way, we wanted to find if applying dimensionality reduction the signal-to-noise ratio is high enough to use that transformation used by PCA to represent the data. The results of the *k*-NN with the different data groupings (miniclips from each topic separately and all topics together) are shown on Table 4.2.

**Table 4.2:** *k*-NN classification results obtained for the distinction between deaf and hearing people after PCA.

| Topic | Worst CM | | Accumulated CM | | CR | Recall | Precision |
|---|---|---|---|---|---|---|---|
| Topic1 | 0.75 | 0.25 | 0.82 | 0.18 | 0.7568 | 0.7989 | 0.8158 |
| | 0.42 | 0.58 | 0.34 | 0.66 | | | |
| Topic2 | 0.69 | 0.31 | 0.75 | 0.25 | 0.6911 | 0.7371 | 0.7531 |
| | 0.50 | 0.50 | 0.40 | 0.60 | | | |
| Topic3 | 0.77 | 0.23 | 0.82 | 0.18 | 0.7854 | 0.8492 | 0.8252 |
| | 0.36 | 0.64 | 0.29 | 0.71 | | | |
| Topic4 | 0.75 | 0.25 | 0.80 | 0.20 | 0.7357 | 0.8296 | 0.7988 |
| | 0.54 | 0.46 | 0.43 | 0.57 | | | |
| All Topics | 0.71 | 0.29 | 0.74 | 0.26 | 0.6609 | 0.7440 | 0.7374 |
| | 0.53 | 0.47 | 0.49 | 0.51 | | | |

Instead of analysing the results obtained by the *k*-NN with PCA process separately, the *k*-NN was computed without submitting the feature matrix to a prior PCA so that the effects (concerning the results) of the dimensionality reduction could be evaluated. The results are visible on Table 4.3.

**Table 4.3:** *k*-NN classification results obtained for the distinction between deaf and hearing people before PCA.

| Topic | Worst CM | | Accumulated CM | | CR | Recall | Precision |
|---|---|---|---|---|---|---|---|
| Topic 1 | 0.97 | 0.03 | 0.97 | 0.03 | 0.9892 | 0.9886 | 0.9943 |
| | 0.04 | 0.96 | 0.05 | 0.95 | | | |
| Topic 2 | 0.93 | 0.07 | 0.95 | 0.05 | 0.9540 | 0.9767 | 0.9496 |
| | 0.05 | 0.95 | 0.04 | 0.96 | | | |
| Topic 3 | 0.97 | 0.03 | 0.98 | 0.02 | 0.9723 | 0.9760 | 0.9812 |
| | 0.05 | 0.95 | 0.04 | 0.96 | | | |
| Topic 4 | 0.96 | 0.04 | 0.98 | 0.02 | 0.9751 | 0.9850 | 0.9794 |
| | 0.04 | 0.96 | 0.03 | 0.97 | | | |
| All Topics | 0.96 | 0.04 | 0.97 | 0.03 | 0.96280 | 0.9743 | 0.9688 |
| | 0.05 | 0.95 | 0.05 | 0.95 | | | |

From the selected statistics the ones that indicate in a more direct way the level of quality of the performance of the classification are the correct rate, the recall and the Precision. In the case that the classifier

does not make mistakes, correct rate = precision = recall = 1. But, in the real world, this task is almost impossible to achieve. Since Recall measures the completeness, or sensitivity, of a classifier, Precision, the exactness of a classifier, and the Correct Rate obviously how many samples are correctly classified these are reliable indicators to be followed.

Taking this into consideration and comparing these three variables for all groupings, on Tables 4.2 and 4.3 it is clear that the performance of the *k*-NN classifier is better when the PCA is not applied. This may happen for a number of reasons and it is not linear that PCA should maintain the integrity of the information contained on the original feature matrix. To avoid this loss of information different normalizations such as Z-Score or Min-Max for example could be used before using PCA but those approaches would go beyond the main purposes of this work (which is sustained on more high-level objectives) [103].

With this revealing result in relation to the effect of PCA on the data considered, it was decided not to apply PCA for the remaining of the tests to be performed. One drawback of this decision could be the computation time of the tests however, the difference on the running times was not considered significant so, not applying dimensionality reduction was not detrimental to the progress of the work.

Looking in isolation at the *k*-NN results the performance for all the different miniclips' groupings is very good being the highest values verified for the grouping of Topic 1. The grouping was made this way so that it would be possible to compare the performance of the classifier for a different number and combination of samples. Topic 1 grouping is by indication of Table 3.3 the one with less samples. This could point that the classifier is more accurate with less samples to impair its judgement. This may be a plausible inference although the performance differences are not that disparate what can lead one to conclude that the classifier performs very accurately in general terms. The CMs show concordant results among all the groupings being the misclassification rate never higher than 5%. No major differences are observed between the worst CM and the accumulated one which confirms the good performance values.

#### 4.4.1.2 Support Vector Machines

The same test was performed using binary SVM to identify the two classes: deaf and hearing. The intention of using a different classifier for the same miniclip groupings and classes was to evaluate the differences in terms of performance on a more complex classifier, less influenced by dimensionality problems. A linear kernel function was used, since it demonstrated to be the one originating the best results. Several C-SVM values were used as input (ranging from $2^{-2}$ to $2^{8}$) being chosen the one that optimized the performance values (Correct Rate, Recall and Precision).

**Table 4.4:** SVM classification results obtained for the distinction between deaf and hearing people.

| Topic | Worst CM | | Accumulated CM | | CR | Recall | Precision |
|---|---|---|---|---|---|---|---|
| Topic 1 | 0.95 | 0.05 | 0.97 | 0.03 | 0.9880 | 0.9866 | 0.9785 |
| | 0.00 | 1.00 | 0.01 | 0.99 | | | |
| Topic 2 | 0.94 | 0.06 | 0.98 | 0.02 | 0.9831 | 0.9830 | 0.9818 |
| | 0.00 | 1.00 | 0.00 | 1.00 | | | |
| Topic 3 | 0.97 | 0.03 | 0.99 | 0.01 | 0.9859 | 0.9918 | 0.9902 |
| | 0.06 | 0.94 | 0.01 | 0.99 | | | |
| Topic 4 | 1.00 | 0.00 | 1.00 | 0.00 | 0.9863 | 0.9852 | 0.9898 |
| | 0.06 | 0.94 | 0.03 | 0.97 | | | |
| All Topics | 0.94 | 0.06 | 0.98 | 0.02 | 0.9753 | 0.9667 | 0.9610 |
| | 0.18 | 0.82 | 0.04 | 0.96 | | | |

The results obtained for the SVM classifier corroborate the ones obtained with $k$-NN being the performance slightly improved (comparing Correct Rate columns). For the SVM, contrarily to $k$-NN, the best performance is observed for Topic 3. Regarding the CMs it is observed a balance in the misclassification rates of the classes which leads one to conclude that the classifier is accurate for both classes. Accumulated CM are more a less equivalent for all the miniclip groupings being the highest misclassification rates observed for the grouping of all topics (where more samples can confuse the classifier).

Summing up the outcomes of the experiences made to identify miniclips of deaf and hearing people it is reasonable to say that the information contained in the videos of our database is rich and the feature approach made to extract that information was appropriated. This inference can be made since the classifiers were able to undergo a training stage with an outcome model that can accurately classify testing samples. With this promising results, one of the research questions proposed on the initial stages of this project was answered successfully. The LGP experts consulted stated that in order to distinguish if a subject using LGP is deaf or hearing we should focus mainly on evaluating facial features. With this result we prove that the body features are also very descriptive. In the future, and as a continuation of this study, facial features should also be evaluated to find out if body and facial features can complement each other.

### 4.4.2 Distinguishing Different Conversation Topics

This subsection addresses the analysis of the results regarding the identification of the miniclips from the different conversation topics. When the database was created it was planned in the definition of four different moments was something used to make it more robust and complete. It is important to recall that a ”script” with directions regarding what should be the content of each acquisition moment was given to the subjects so that they could prepare to take part in the dialogue. Since the subjects were speaking LGP, only a Portuguese Sign Language speaker could evaluate the contents of the videos and verify if the four distinct moments were in fact present. So, this supervised evaluation was done for some of the videos confirming that the subjects were demonstrating different levels of expressiveness and emotion accordingly to the current discussion topic.

The remaining videos were also analysed without expert supervision in order to extract some conclusions based on the queues given previously by the experts. It was possible to deduce that, regarding the presence of four different conversational moments, the framework of the database being developed is accurate and valuable.

To find out if it was possible with the selected features to discriminate the expressiveness of the subjects on the miniclips of those moments, $k$-NN and SVM were also the supervised learning algorithms used.

#### 4.4.2.1 $k$-Nearest Neighbours

Table 4.5 shows the results for the $k$-NN test performed for the different groupings of miniclips. The measures presented are the same as for the tests on Section 4.4.1. Looking at the results and comparing primarily the performances of the different groups (evaluated by the Correct Rate, recall and precision) it is possible to state that the performance of the $k$-NN classifier is very good since the Correct Rates for all groups are around 90%. With this grouping structure one can infer from these results that the performance

of the classifier is not compromised by the fact that all subjects (CR = 0,9285, Recall = 0,9105 and Precision = 0,9349) are grouped together.

These results confirm that the feature selection process effectuated chose in fact features that enable the classification process to differentiate our miniclips and that the miniclips from the different topics are actually different in terms of motion expressiveness. The values observed for recall and precision also corroborate this conclusion since by definition, recall tells us from all the positive examples existent (from a certain class), what fraction did the classifier pick up while, precision tells us from all the examples the classifier labelled as positive (for a certain class), what fraction were correct. The values observed for these two measures are also very good. Looking at the CM for all groupings it is possible to observe a pattern for the majority of the groupings which is that Topics 1 and 2 display the highest misclassification rates. This is not a strange occurrence since if we look at what the subjects were supposed to talk about, for example in Topic 1 we can conclude that the contents are not that different from the ones of Topic 2. This may be an explanation for observed occurrence.

#### 4.4.2.2 Support Vector Machines

On Table 4.6 are shown the results for the SVM test performed for the different groupings of miniclips. Similarly to what was done to answer the question of the previous section, the SVM classifier was used here to corroborate the results of the *k*-NN classifier. In this case, we used a multiclass SVM approach, since were used miniclips from 4 different moments. The values of the performance measures considered are somewhat different from the ones of the *k*-NN. While on the *k*-NN the performance results are fairly homogeneous among the different groupings, this is not observed for the SVM. Considering the CM of this classification a particular phenomenon is observed, which is the fact that the misclassifications happen mainly in adjacent classes which is consequence of the fact that the classes are considered in an ordered way (as mentioned in Section 3.5.3). Due to this fact, the classifier used in case of doubt decides of the label to assigned between adjacent classes. Contrarily to the *k*-NN classifier for the same problem Topics 1 and 2 do not display the highest rates of misclassification which is due to the nature of the algorithm used and the above mentioned fact (classes are considered in an ordered fashion).

Although SVM is a classifier that usually performs better than *k*-NN, this test was used to confirm the results of the previous one. This confirmation was achieved however, the results were not improved comparing to the *k*-NN ones. To overcome this issue, further parameter refinement should be done on the multiclass SVM classifier. Nevertheless, the results are concordant with the *k*-NN which allows one to conclude that the classification solution performs well revealing also that the motiongrams are descriptive and discriminative of the miniclips.

For this research question there is a concordance between the results of the *k*-NN and the SVM classifiers. In the case of the SVM the worst performance is observed for the grouping of all topics which is expected since a larger number of samples complicates the task of the classifier. However, this phenomenon is not observed for the *k*-NN in which no major differences are observed among all the miniclip groupings.

**Table 4.5:** *k*-NN classification results obtained for the distinction the different conversation topics.

| Subject | Worst CM | | | | Accumulated CM | | | | CR | Recall | Precision |
|---|---|---|---|---|---|---|---|---|---|---|---|
| D1 | 0.77 | 0.06 | 0.03 | 0.13 | 0.86 | 0.06 | 0.01 | 0.07 | 0.9143 | 0.8597 | 0.9074 |
| | 0.03 | 0.90 | 0.03 | 0.05 | 0.02 | 0.92 | 0.02 | 0.04 | | | |
| | 0.02 | 0.00 | 0.95 | 0.03 | 0.01 | 0.02 | 0.94 | 0.03 | | | |
| | 0.10 | 0.07 | 0.00 | 0.83 | 0.04 | 0.03 | 0.01 | 0.92 | | | |
| D4 | 0.94 | 0.06 | 0.00 | 0.00 | 0.87 | 0.08 | 0.06 | 0.00 | 0.9639 | 0.9481 | 0.9182 |
| | 0.33 | 0.67 | 0.00 | 0.00 | 0.02 | 0.88 | 0.02 | 0.08 | | | |
| | 0.00 | 0.00 | 1.00 | 0.00 | 0.02 | 0.00 | 0.97 | 0.01 | | | |
| | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.02 | 0.00 | 0.97 | | | |
| D5 | 0.81 | 0.10 | 0.07 | 0.02 | 0.87 | 0.08 | 0.06 | 0.00 | 0.9331 | 0.8660 | 0.9314 |
| | 0.03 | 0.90 | 0.01 | 0.06 | 0.02 | 0.88 | 0.02 | 0.08 | | | |
| | 0.00 | 0.00 | 0.99 | 0.01 | 0.02 | 0.00 | 0.97 | 0.01 | | | |
| | 0.00 | 0.03 | 0.00 | 0.97 | 0.00 | 0.02 | 0.00 | 0.97 | | | |
| D6 | 0.89 | 0.01 | 0.04 | 0.06 | 0.95 | 0.01 | 0.03 | 0.01 | 0.9363 | 0.9515 | 0.9637 |
| | 0.00 | 0.81 | 0.08 | 0.11 | 0.03 | 0.87 | 0.06 | 0.05 | | | |
| | 0.00 | 0.00 | 0.97 | 0.03 | 0.00 | 0.01 | 0.97 | 0.02 | | | |
| | 0.04 | 0.04 | 0.01 | 0.92 | 0.01 | 0.03 | 0.03 | 0.92 | | | |
| H2 | 0.81 | 0.11 | 0.06 | 0.03 | 0.93 | 0.04 | 0.02 | 0.00 | 0.9166 | 0.9356 | 0.9826 |
| | 0.00 | 1.00 | 0.00 | 0.00 | 0.02 | 0.94 | 0.03 | 0.00 | | | |
| | 0.00 | 0.00 | 0.96 | 0.04 | 0.00 | 0.05 | 0.91 | 0.04 | | | |
| | 0.00 | 0.04 | 0.04 | 0.92 | 0.00 | 0.05 | 0.06 | 0.89 | | | |
| H6 | 0.72 | 0.25 | 0.03 | 0.00 | 0.84 | 0.15 | 0.01 | 0.00 | 0.8976 | 0.8394 | 0.8962 |
| | 0.11 | 0.80 | 0.09 | 0.00 | 0.07 | 0.86 | 0.07 | 0.01 | | | |
| | 0.02 | 0.06 | 0.92 | 0.00 | 0.01 | 0.05 | 0.92 | 0.01 | | | |
| | 0.00 | 0.00 | 0.03 | 0.97 | 0.00 | 0.01 | 0.03 | 0.96 | | | |
| H7 | 0.84 | 0.13 | 0.00 | 0.03 | 0.94 | 0.04 | 0.00 | 0.01 | 0.9698 | 0.9458 | 0.9999 |
| | 0.00 | 0.96 | 0.02 | 0.02 | 0.00 | 0.95 | 0.03 | 0.02 | | | |
| | 0.00 | 0.04 | 0.96 | 0.00 | 0.00 | 0.01 | 0.99 | 0.00 | | | |
| | 0.00 | 0.04 | 0.00 | 0.96 | 0.00 | 0.01 | 0.00 | 0.98 | | | |
| All Subjects | 0.90 | 0.03 | 0.04 | 0.03 | 0.91 | 0.03 | 0.03 | 0.03 | 0.9285 | 0.9105 | 0.9349 |
| | 0.03 | 0.91 | 0.02 | 0.04 | 0.03 | 0.91 | 0.02 | 0.04 | | | |
| | 0.01 | 0.01 | 0.96 | 0.03 | 0.01 | 0.01 | 0.96 | 0.03 | | | |
| | 0.01 | 0.03 | 0.03 | 0.93 | 0.01 | 0.04 | 0.03 | 0.92 | | | |

**Table 4.6:** SVM classification results obtained for the distinction the different conversation topics.

| Subject | Worst CM | | | | Accumulated CM | | | | CR | Recall | Precision |
|---|---|---|---|---|---|---|---|---|---|---|---|
| D1 | 0.80 | 0.20 | 0.00 | 0.00 | 0.83 | 0.17 | 0.00 | 0.00 | 0.9172 | 0.9215 | 0.9024 |
| | 0.08 | 0.92 | 0.00 | 0.00 | 0.08 | 0.93 | 0.00 | 0.00 | | | |
| | 0.00 | 0.05 | 0.95 | 0.00 | 0.00 | 0.02 | 0.98 | 0.00 | | | |
| | 0.00 | 0.00 | 0.10 | 0.90 | 0.00 | 0.00 | 0.13 | 0.87 | | | |
| D4 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.9861 | 0.9886 | 0.9792 |
| | 0.17 | 0.83 | 0.00 | 0.00 | 0.09 | 0.91 | 0.00 | 0.00 | | | |
| | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | | | |
| | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 | | | |
| D5 | 0.88 | 0.13 | 0.00 | 0.00 | 0.82 | 0.18 | 0.00 | 0.00 | 0.9104 | 0.9037 | 0.8963 |
| | 0.15 | 0.85 | 0.00 | 0.00 | 0.08 | 0.88 | 0.04 | 0.00 | | | |
| | 0.00 | 0.08 | 0.85 | 0.08 | 0.00 | 0.05 | 0.92 | 0.03 | | | |
| | 0.00 | 0.00 | 0.05 | 0.95 | 0.00 | 0.00 | 0.04 | 0.96 | | | |
| D6 | 0.85 | 0.15 | 0.00 | 0.00 | 0.92 | 0.08 | 0.00 | 0.00 | 0.9059 | 0.8964 | 0.8995 |
| | 0.00 | 0.71 | 0.29 | 0.00 | 0.03 | 0.86 | 0.11 | 0.00 | | | |
| | 0.00 | 0.13 | 0.81 | 0.06 | 0.00 | 0.06 | 0.88 | 0.05 | | | |
| | 0.00 | 0.00 | 0.12 | 0.88 | 0.00 | 0.00 | 0.07 | 0.93 | | | |
| H2 | 0.93 | 0.07 | 0.00 | 0.00 | 0.97 | 0.03 | 0.00 | 0.00 | 0.8656 | 0.8677 | 0.8531 |
| | 0.22 | 0.67 | 0.11 | 0.00 | 0.11 | 0.74 | 0.16 | 0.00 | | | |
| | 0.00 | 0.15 | 0.85 | 0.00 | 0.00 | 0.12 | 0.88 | 0.00 | | | |
| | 0.00 | 0.00 | 0.17 | 0.83 | 0.00 | 0.00 | 0.17 | 0.83 | | | |
| H6 | 0.90 | 0.10 | 0.00 | 0.00 | 0.97 | 0.03 | 0.00 | 0.00 | 0.9808 | 0.9848 | 0.9771 |
| | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | | | |
| | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | | | |
| | 0.00 | 0.00 | 0.09 | 0.91 | 0.00 | 0.00 | 0.06 | 0.94 | | | |
| H7 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.9788 | 0.9841 | 0.9817 |
| | 0.00 | 0.83 | 0.17 | 0.00 | 0.00 | 0.93 | 0.07 | 0.00 | | | |
| | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | | | |
| | 0.00 | 0.00 | 0.05 | 0.95 | 0.00 | 0.00 | 0.02 | 0.98 | | | |
| All Subjects | 0.50 | 0.50 | 0.00 | 0.00 | 0.70 | 0.29 | 0.01 | 0.00 | 0.7880 | 0.8211 | 0.7788 |
| | 0.08 | 0.58 | 0.33 | 0.00 | 0.04 | 0.76 | 0.19 | 0.01 | | | |
| | 0.00 | 0.14 | 0.86 | 0.00 | 0.01 | 0.08 | 0.86 | 0.06 | | | |
| | 0.00 | 0.00 | 0.40 | 0.60 | 0.00 | 0.00 | 0.20 | 0.80 | | | |

### 4.4.3 Identifying Levels of Mastery in Portuguese Sign language

The first analysis that should be done is to the parameters used to build a classifier in CLUTO since this way we can understand the intention behind the solution used for clustering. The idea was to create an agglomerative hierarchical tree to build a hierarchy of clusters so, the following parameters were considered:

- **nfeatures**: number of descriptive to features to display for each cluster.
- **showsummaries**: used to discovered clusters and identify relations among the set of most descriptive features of each cluster..

- **showtree**: used so that a hierarchical agglomerative tree is used. This tree considers all clusters and the relations among them.
- **fulltree**: used to have subdivisions of the hierarchical agglomerative tree created by the previous parameter.
- **labeltree**: label the nodes of the tree with the most relevant set of features.
- **clmethod**: used to set the clustering method as agglomerative.
- **crfun**: selects the particular clustering criterion function to be used in finding the clusters.
- **sim**: selects the similarity function to be used for cluster formation.
- **zscores**: used to analyse the samples in each cluster and compute the $z$-score of its similarity to the other samples in that cluster and also in other clusters.

Further details on this algorithm's parameters and methods can be found on CLUTO's official manual [100]. The selection of these parameters was done after several others were tested being these the ones that generated better results. Considering the extended number of parameters of this algorithm it is possible to benefit from a wide-range of information and statistics about the clusters found. Basically the intention is to measure the quality of the clustering process through: the quality of each cluster (indicated by the criterion function that it uses) and the degree of similarity between elements in the same cluster.

Setting the parameter clmethod to *bagglo* determines that the methodology of the algorithm will be an *agglomerative* approach. This means that the algorithm will find the clusters by partitioning the entire dataset into either a predetermined or an automatically derived number of clusters. In our case as mentioned in Section 3.5.4 the number of clusters were in fact known *a priori* to be 3 in the case that we wanted to group our subjects by the number of years familiarised with LGP or 4 for the current profession or the combined score of the answer of the two questions. Tables 4.7, 4.8 and 4.9 show the statistics that evaluate the clusterings performed. The clustering scores in Tables 4.8 and 4.9 are actually the same since the number of clusters defined as input was four in both cases. The portion of the tables that varies is the last four columns where a class-distribution table (or CM) is shown giving information about how the different classes are distributed in each one of the clusters (different scores obtained by the answers to the considered question). Looking at this class-distribution table, can be a first analysis to determine the quality of the different clusters.

**Table 4.7:** Clustering statistics and performance measures obtained by the agglomerative clustering solution. The classes considered are regarding the number of years in contact with LGP.

| cid | Size | ISim | ISdev | ESim | ESdev | Entpy | Purty | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 70 | 0.669 | 0.075 | 0.332 | 0.079 | 0.625 | 0.557 | 0.44 | 0.56 | 0.00 |
| 2 | 167 | 0.606 | 0.105 | 0.477 | 0.117 | 0.596 | 0.707 | 0.71 | 0.28 | 0.01 |
| 3 | 1029 | 0.597 | 0.112 | 0.437 | 0.12 | 0.75 | 0.646 | 0.28 | 0.65 | 0.07 |

**Table 4.8:** Clustering statistics and performance measures obtained by the agglomerative clustering solution. The classes considered are regarding the current profession of the subject.

| cid | Size | ISim | ISdev | ESim | ESdev | Entpy | Purty | 4 | 3 | 1 | 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 70 | 0.669 | 0.075 | 0.332 | 0.079 | 0.584 | 0.6 | 0.60 | 0.36 | 0.00 | 0.04 |
| 2 | 167 | 0.606 | 0.105 | 0.477 | 0.117 | 0.529 | 0.695 | 0.69 | 0.28 | 0.01 | 0.02 |
| 3 | 144 | 0.709 | 0.078 | 0.559 | 0.087 | 0.669 | 0.674 | 0.67 | 0.12 | 0.18 | 0.03 |
| 4 | 885 | 0.602 | 0.111 | 0.481 | 0.127 | 0.536 | 0.739 | 0.74 | 0.01 | 0.05 | 0.20 |

**Table 4.9:** Clustering statistics and performance measures obtained by the agglomerative clustering solution. The classes considered are the number of years in contact with LGP combined with the current profession.

| cid | Size | ISim | ISdev | ESim | ESdev | Entpy | Purty | 3 | 4 | 2 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 70 | 0.669 | 0.075 | 0.332 | 0.079 | 0.695 | 0.557 | 0.04 | 0.56 | 0.36 | 0.04 |
| 2 | 167 | 0.606 | 0.105 | 0.477 | 0.117 | 0.831 | 0.413 | 0.41 | 0.28 | 0.29 | 0.02 |
| 3 | 144 | 0.709 | 0.078 | 0.559 | 0.087 | 0.845 | 0.424 | 0.25 | 0.42 | 0.30 | 0.03 |
| 4 | 885 | 0.602 | 0.111 | 0.481 | 0.127 | 0.66 | 0.682 | 0.06 | 0.68 | 0.06 | 0.20 |

Before going through the results themselves it is necessary to clarify the meaning of all the statistics presented on the remaining columns of the tables above. The column labelled as *cid* indicates the id of the cluster, *Size* the number of samples that belong to each cluster, *ISim* and *ISdev* represent both the average and standard deviation in terms of similarity between each cluster (internal similarities) whereas *ESim* and *ESdev* represent the same statistics but for similarity of the objects of each cluster and the rest of the objects (external similarities). It is important to note that the discovered clusters are ordered in increasing order, meaning that clusters that are tight and far away from the rest of the objects have smaller cid values.

This research question of inferring the different levels of mastery in Portuguese Sign Language was undoubtedly the most abstract one that we proposed ourselves to answer. The intention was to discover if the features captured for the miniclips of each subject would contain hidden information regarding the expertise of each subject. Trying to find if the miniclips can be grouped according to these criteria is the logical approach.

The simple statistics reported help one finding out if the attempt of grouping our data is reasonable or not. These statistics have to do with the quality of each cluster as measured by the criterion function used and the similarity between the objects in each cluster.

For a better understanding of the results, the tables will be analysed separately:

The first clustering data presented (Table 4.7 in which with a 3-way clustering is intended to group our samples so that it is possible to fit the data in 3 classes (in which the levels of mastery in LGP are defined based on the number of years the subjects know that language.

Resorting to the opinion of the LGP experts this indicator alone might not be a clear of how experienced a subject is in terms of LGP. This happens since many factors (regardless of the years in contact with the language) may influence this: for example if a subject has to use sign language on a daily basis or not, if he has parents who are deaf making most likely that language to be the one that the subject considers to be his mother tongue, whether he was born hearing or deaf and in the case of being deaf if the deafness was acquired or from birth among other facts. As it is visible all the above mentioned factors show that the years in contact with the language are not an unequivocal indicator since the life experience of the subject may overcome that issue. The overall entropy and purity of this clustering are (computed from the values on Table 4.7 using Equations 3.6 and 3.8) **0.723** and **0.649** respectively. These values reveal that the clustering was performed poorly (the various miniclips are not distributed within each cluster in an organised way) since the entropy is over 0.7. Purity which is supposed to be as close as 1.0 as possible is quite low which indicates us that the samples in each cluster are not as homogeneous as desired. Analysing the diagonal of the CM we have the information about the percentage of samples that were we classified correctly. See that class 3, although it is the one with more elements (see Size column) it is the cluster with less samples assigned to it. Classes 1 and 2 also display a large percentage of misclassification.

When performing the clustering (4-way) to be fitted for the levels in terms of professional occupation of the subjects, the overall entropy and purity of the solution were the best of the three analysis performed (**0.553** and **0.718**).

From the questions featured on the questionnaires (Figure A.5 and A.6) that were common to the questionnaires of both deaf and hearing this one regarding the current professional occupation of the subjects was the one considered by LGP to possibly be more discriminative when it comes to the expertise on this language. This is shown by an improvement on the values of entropy and purity obtained for this solution. The CM of this clustering process is an indicator as well of its best performance comparing to the one analysed previously. The larger values in its diagonal reveal that the overall misclassification is lower, being observed a higher rate of well classified samples comparing to the misclassified ones for the first class.

The last table shows the results of the same 4-way clustering process but using a different class distribution. Here the values used as classes on the previous tests are combined to get a new set of classes (weighted sum of scores). With this fusion of classes it was intended to get classes that would concatenate more information about the expertise of each subject. The less promising results of entropy and purity (**0.706** and **0.611**) tell us that combining the classes of the first test with the ones from the second impair the performance of the clustering solution being the organisation of the miniclips in clusters less organized. The CM of this last table reveals that the samples are mainly distributed for the first three classes. The low value of entropy and high of entropy are confirmed by the observation of a large number of misclassifications for all classes (specially for the first class). Agglutinating more information it was expected a more detailed description of our subjects however, this was not observed. To overcome this issue maybe a different type of combination of the information could be tested.

In this way the clustering solution was best fitted for the case when the subjects are grouped by their current profession which leads us to conclude that this question featured on the questionnaires is the one that most suitably describes the subjects' expertise in LGP.

In this chapter the results of the methods applied to meet the goals of this work were presented. The first relevant result was the outcome of the feature selection process that revealed that the Motiongrams were the features with more ability to differentiate the contents of the miniclips considered from all the subjects. For this reason these features were used for the classification processes that followed. The results of the classification used to answer the research questions of this work revealed that it was possible with the features considered to distinguish deaf from hearing people and also to distinguish different conversation topics featured in the videos of the database constructed. The attempt of stratifying the subjects considered in levels of expertise in LGP through the usage of an agglomerative clustering algorithm was not achieved with the same success since this is more abstract issue which may not be clearly identified just by motion features being most likely dependent on several other factor that are exterior to our approach and particular from each subject's life experience and contact with the LGP.

# Chapter 5

# Conclusions and Future Work

Deaf people who are sign language speakers are individuals considered to be extremely expressive. The lack of ability to use spoken language causes these subjects to rely much more on other means to convey their expressive intentions and emotions. Automated visual analysis of behaviour provides tools for the construction of intelligent computer vision systems. The idea of reaching nonverbal sensitivity through computational models is very appealing since, this may help one understanding how the human visual system interprets all the sensory events in the environment and how it relates those events. This study is framed in this mindset focusing on automated visual analysis of expressiveness of LGP speakers using computer vision techniques.

With this work important breakthroughs were achieved regarding the analysis of the expressiveness of LGP speakers. In the chapter of literature review, existing studies that analyse body gestures were studied. Algorithms already available in literature that constitute powerful classification tools to understand the problems faced in this study were analysed. These constituted valuable guidances for the development of the framework designed to capture the the body expressiveness of the subjects considered. The studies presented for behaviour representation and emotion analysis supported by body gestures were also analysed once they are integrated in the long-term goals of the project in which this study is inserted.

The main goals of this work, defined *a priori*, were to answer a series of research questions through the use of classification techniques supported by the extraction of features from videos of a database build for the purpose of this study. The database, which is inserted in a very specific context (duo-interaction between deaf and hearing people), is meant to be made available to the scientific community. For this study in particular, its contents allowed us to extract relevant conclusions regarding the considered population so, we consider that it can also valuable for other studies among the scientific community.

Several features were extracted from whole body motions such as: MHI, Motion Gradient, Motiongrams and features related to trajectories (velocity, acceleration and direction). Through feature selection processes the Motiongrams displayed the best ability to differentiate the videos considered. So, these were the features selected to take part in the classification processes.

Regarding the research questions, it was possible to distinguish with success the deaf and the hearing populations using $k$-NN and SVM classifiers that showed concordant results. This achievement is considered a major breakthrough since a human observer struggles to make this distinction easily, at least if

he considers body features alone, so, with this work we proved that computer vision technique using supervised classification techniques are capable of making this distinction accurately. Different conversation topics were included as a feature of the database so, another task proposed for this work was to distinguish those different moments represented in different videos. Using again a $k$-NN classifier and a multiclass version of the SMV classifier these conversation topics were distinguished correctly with a very low error rate. Concerning the last research question it was known that distinguishing different levels of mastery in LGP would be a arduous task since the support used (to divide the subjects in levels) was information contained in questionnaires answered by the subjects. The classification was done by an agglomerative clustering algorithm (unsupervised classification) and was not very successful in stratifying the considered subjects in levels. The conclusion extracted from this test is that identifying this characteristic of each subject is not linear being dependent on a numerous amount of factors that may not be considered by the questions featured in the questionnaires. Further work on this matter is necessary to identify the more relevant indicators of the expertise of a person in LGP: maybe a multi-variable regression for which we would need to define other questionnaires and tests with combinations of demographic, academical and other data that could evaluate more accurately the expertise of a subject. A larger population would also be beneficial.

Summing up, the main breakthrough done with this work was that contrarily to what was expected body gestures alone contain a great amount of expressive content that allowed us to identify valuable aspects in the videos made available by the construction of the database.

As said before, this study is a preliminary work encompassed within a broader project which aims to integrate facial expressions and body gestures for further behavioural analysis in terms of emotions (targeting the deaf and hearing community). With this mindset some improvements to the developed work in this master thesis can be pointed. Regarding the motiongrams, which are the dominant and more discriminative features of the videos used, it is possible to further investigate their potentialities in order to extract for example informations for tracking of different elements. Regarding the acquisition of the videos it is also important to re-evaluate the way the subjects' conversations are acquired since the data available so far is not adequate to analyse for example facial features which require more detail that can only be obtained by increasing the proximity of the subjects to the cameras.

The continuation of this work intends to go in the direction of finding emotional and behaviour patterns through face and body expressions analysis, of deaf and hearing people. New techniques and approaches shall need to be reviewed and tested for the future work of this project since this is an very ambitious goal. It is ambitious but has an extremely valuable purpose of finding solutions for the gaps that between deaf and hearing people.

# References


[1] S. S. Tomkins. *Affect Imagery Consciousness: Volume I The Positive Affects*. New York Springer Publishing, 1962.

[2] S. Tomkins. *Affect Imagery Consciousness: Volume II: The Negative Affects*. Springer Series. Springer Publishing Company, 1963.

[3] Paul Ekman and Wallace V. Friesen. *Unmasking The Face*. Prentice-Hall, 1975.

[4] Paul Ekman and Wallace V. Friesen. *Facial action coding system:A technique for the measurement of facial movement.* Consulting Psychologists Press., 1978.

[5] J. Harrigan, R. Rosenthal, and K. Scherer. *New Handbook of Methods in Nonverbal Behavior Research*. Series in affective science. OUP Oxford, 2008.

[6] Paul Ekman. Are there basic emotions. *Psycological Review*, 99:550–553, 1992.

[7] Hatice Gunes, Massimo Piccardi, and Tony Jan. Face and body gesture recognition for a vision-based multimodal analyzer. In *Proceedings of the Pan-Sydney Area Workshop on Visual Information Processing*, VIP '05, pages 19–28, Darlinghurst, Australia, Australia, 2004. Australian Computer Society, Inc.

[8] DoryA. Schachner, PhillipR. Shaver, and Mario Mikulincer. Patterns of nonverbal behavior and sensivity in the context of attachment relations. *Journal of Nonverbal Behavior*, 29(3):141–169, 2005.

[9] L. Riek, Afzal S., and Robinson P. Affect decoding measures and human-computer interaction. *In Proceedings of Measuring Behaviour*, 2008.

[10] Josep M. Nadal, Pilar Monreal, and Santiago Perera. Emotion and linguistic diversity. *Procedia - Social and Behavioral Sciences*, 82(0):614 – 620, 2013. <ce:title>World Conference on Psychology and Sociology 2012</ce:title>.

[11] I. Ari and L. Akarun. Facial feature tracking and expression recognition for sign language. In *Signal Processing and Communications Applications Conference, 2009. SIU 2009. IEEE 17th*, pages 229–232, 2009.

[12] Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. *Image and Vision Computing*, 31(2):203 – 221, 2013. <ce:title>Affect Analysis In Continuous Input</ce:title>.

[13] Dimitris Metaxas and Shaoting Zhang. A review of motion analysis methods for human nonverbal communication computing. *Image and Vision Computing*, 31(6 - 7):421 – 433, 2013. <ce:title>Machine learning in motion analysis: New advances</ce:title>.

[14] Alexandros Andre Chaaraoui, Pau Climent-Perez, and Francisco Florez-Revuelta. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, 39(12):10873 – 10888, 2012.

[15] Kamrad Khoshhal Roudposhti and Jorge Dias. Probabilistic human interaction understanding: Exploring relationship between human body motion and the environmental context. *Pattern Recognition Letters*, 34(7):820 – 830, 2013. <ce:title>Scene Understanding and Behaviour Analysis</ce:title>.

[16] S. Gong and T. Xiang. *Visual Analysis of Behaviour: From Pixels to Semantics*. SpringerLink : Bücher. Springer, 2011.

[17] Arnold Wiliem, Vamsi Madasu, Wageeh Boles, and Prasad Yarlagadda. A suspicious behaviour detection using a context space model for smart surveillance systems. *Computer Vision and Image Understanding*, 116(2):194 – 209, 2012.

[18] Charles Darwin. *The expression of the emotions in man and animals*. London: John Murray, 1872.

[19] Beatrice de Gelder and Jan Van den Stock. The bodily expressive action stimulus test (beast). construction and validation of a stimulus basis for measuring perception of whole body expression of emotions. *Frontiers in Psychology*, 2:181, 2011.

[20] Jan Van den Stock and Ruthger Righart. Body expressions influence recognition of emotions in the face and voice. *Emotion*, 3:487–494, 2007.

[21] A. Kleinsmith and N. Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *Affective Computing, IEEE Transactions on*, 4(1):15–33, 2013.

[22] Konrad Schindler, Luc Van Gool, and Beatrice de Gelder. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural Networks*, 21(9):1238 – 1246, 2008.

[23] Anthony P. Atkinson, Mary L. Tunstall, and Winand H. Dittrich. Evidence for distinct contributions of form and motion information to the recognition of emotions from body gestures. *Cognition*, 104(1):59 – 72, 2007.

[24] Harald G. Wallbott. Bodily expression of emotion. *European Journal of Social Psychology*, 28(6):879–896, 1998.

[25] Stefano Piana, Alessandra Staglianó, and Antonio Camurri ANDFrancesca Odone. A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition. *IDGEI International Workshop*, 2013.

[26] Bon-Woo Hwang, Sungmin Kim, and Seong-Whan Lee. 2d and 3d full-body gesture database for analyzing daily human gestures. In De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang, editors, *Advances in Intelligent Computing*, volume 3644 of *Lecture Notes in Computer Science*, pages 611–620. Springer Berlin Heidelberg, 2005.

[27] A.P. Atkinson, W.H. Dittrich, A.J. Gemmell, and A.W. Young. Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception*, 33:717–746, 2004.

[28] Chikahito Nakajima, Massimiliano Pontil, Bernd Heisele, and Tomaso Poggio. Full-body person recognition system. *Pattern Recognition*, 36(9):1997–2006, 2003.

[29] Justine Cassell. A framework for gesture generation and interpretation. In *Computer Vision in Human-Machine Interaction*, pages 191–215. Cambridge University Press, 2000.

[30] Mark Coulson. Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonver*, 22(2):117–139, 2004.

[31] H. Ohno and M. Yamamoto. Gesture recognition using character recognition techniques on two-dimensional eigenspace. *Proceedings of ICCV*, pages 151–156, 1999.

[32] R. Cutler and M. Turk. View-based interpretation of real-time optical flow for gesture recognition. *Proceedings of IEEE International Conference of Automatic Face & Gesture Recognition*, pages 416–421, 1998.

[33] P. Peixoto, J. Gonçalves, and H. Araújo. Real-time gesture recognition system based on contour signatures. *ICPR'2002 –16th International Conference on Pattern Recognition*, pages 11–15, 2002.

[34] P. De Silva, M. Osano, A. Marasinghe, and A.P. Madurapperuma. Towards recognizing emotion with affective dimensions through body gestures. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 269–274, 2006.

[35] Hong Li and Michael Greenspan. Model-based segmentation and recognition of dynamic gestures in continuous video streams. *Pattern Recognition*, 44(8):1614 – 1628, 2011.

[36] Antonio Camurri, Barbara Mazzarino, Matteo Ricchetti, Renee Timmers, and Gualtiero Volpe. Multimodal analysis of expressive gesture in music and dance performances. In Antonio Camurri and Gualtiero Volpe, editors, *Gesture-Based Communication in Human-Computer Interaction*, volume 2915 of *Lecture Notes in Computer Science*, pages 20–39. Springer Berlin Heidelberg, 2004.

[37] Yuichi Kobayashi. The emotion sign: Human motion analysis classifying specific emotion. *JCP*, 3(9):20–28, 2008.

[38] Frank E Pollick, Helena M Paterson, Armin Bruderlin, and Anthony J Sanford. Perceiving affect from arm movement. *Cognition*, 82(2):B51 – B61, 2001.

[39] Alexis Heloir and Sylvie Gibet. A qualitative and quantitative characterisation of style in sign language gestures. In Miguel Sales Dias, Sylvie Gibet, MarceloM. Wanderley, and Rafael Bastos, editors, *Gesture-Based Human-Computer Interaction and Simulation*, volume 5085 of *Lecture Notes in Computer Science*, pages 122–133. Springer Berlin Heidelberg, 2009.

[40] Miguel Sales Dias, Sylvie Gibet, Marcelo Wanderley, and R. Bastos. *Gesture-Based Human-Computer Interaction and Simulation, Proceedings of Gesture Workshop 2007*. Lecture Notes in Computer Science. Springer, December 2009.

[41] Tao Xiang and Shaogang Gong. Beyond tracking: Modelling activity and understanding behaviour. *International Journal of Computer Vision*, 67(1):21–51, 2006.

[42] Thomas B. Moeslund and Fredrik Bajers. Computer vision-based human motion capture - a survey, 1999.

[43] Heng Wang, A. Klaser, C. Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176, June 2011.

[44] S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking group of people. *Computer Vision and Image Understanding*, 80:42–56, 2000.

[45] I. Haritaoglu, D. Harwood, and L.S. Davis. W4: real-time surveillance of people and their activities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):809–830, Aug 2000.

[46] Bo Wu and Ram Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, 2007.

[47] M.D. Breitenstein, H. Grabner, and L. Van Gool. Hunting nessie - real-time abnormality detection from webcams. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1243–1250, Sept 2009.

[48] Antonios Oikonomopoulos, Ioannis Patras, Maja Pantic, and Nikos Paragios. Trajectory-based representation of human actions. In *Proceedings of the ICMI 2006 and IJCAI 2007 International Conference on Artifical Intelligence for Human Computing*, ICMI'06/IJCAI'07, pages 133–154, Berlin, Heidelberg, 2007. Springer-Verlag.

[49] Yu-Gang Jiang, Qi Dai, Xiangyang Xue, Wei Liu, and Chong-Wah Ngo. Trajectory-based modeling of human actions with motion reference points. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision ECCV 2012*, volume 7576 of *Lecture Notes in Computer Science*, pages 425–438. Springer Berlin Heidelberg, 2012.

[50] Imran Saleemi, Khurram Shafique, and Mubarak Shah. Probabilistic modeling of scene dynamics for applications in visual surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1472–1485, 2009.

[51] David Mark Russel and Shaogang Gong. Minimum cuts of a time-varying background. *British Machine Vision Association*, pages 809–818, 2006.

[52] Scott Cohen. Background estimation as a labeling problem. In *Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2*, ICCV '05, pages 1034–1041, Washington, DC, USA, 2005. IEEE Computer Society.

[53] Chris Stauffer and W.E.L. Grimson. Learning patterns of activity using real-time tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):747–757, 2000.

[54] A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257–267, Mar 2001.

[55] R. Venkatesh Babu and K.R. Ramakrishnan. Recognition of human actions using motion history information extracted from the compressed video. *Image and Vision Computing*, 22(8):597 – 607, 2004.

[56] Jeffrey Ng and Shaogang Gong. Learning pixel-wise signal energy for understanding semantics. In *In Proc. BMVC*, pages 695–704. Press, 2001.

[57] Tao Xiang, Shaogang Gong, and Dennis Parkinson. Autonomous visual events detection and classification without explicit object-centred segmentation and tracking. *In British Machine Vision Conference*, pages 233–242, 2002.

[58] Tao Xiang and Shaogang Gong. Beyond tracking: modelling activity and understanding behaviour. *International Journal of Computer Vision*, 67:2006, 2006.

[59] T. Gautama and M.M. Van Hulle. A phase-based approach to the estimation of the optical flow field using spatial filtering. *Neural Networks, IEEE Transactions on*, 13(5):1127–1136, 2002.

[60] Ahmad R. Naghsh Nilchi and Mohammad Roshanzamir . An efficient algorithm for motion detection based facial expression recognition using optical flow. *International Journal of Engineering and Applied Sciences*, 14:318–323, 2006.

[61] Y. Yacoob and L.S. Davis. Recognizing human facial expressions from long image sequences using optical flow. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(6):636–642, 1996.

[62] Ali S and Shah M. Human action recognition in videos using kinematic features and multiple instance learning.

[63] A.K. Jain, R. P W Duin, and Jianchang Mao. Statistical pattern recognition: a review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):4–37, 2000.

[64] Friedhelm Schwenker and Edmondo Trentin. Pattern classification and clustering: A review of partially supervised learning approaches. *Pattern Recognition Letters*, 37(0):4 – 14, 2014. <ce:title id=>Partially Supervised Learning for Pattern Recognition</ce:title>.

[65] Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Softw.*, 3(3):209–226, September 1977.

[66] Gongde Guo, Hui Wang, David A. Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In *CoopIS/DOA/ODBASE'03*, pages 986–996, 2003.

[67] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September 1995.

[68] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, June 1998.

[69] J.F. Pinto da Costa, R. Sousa, and J.S. Cardoso. An all-at-once unimodal svm approach for ordinal classification. In *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, pages 59–64, Dec 2010.

[70] Edgar Osuna, Robert Freund, and Federico Girosi. Support vector machines: Training and applications. Technical report, Cambridge, MA, USA, 1997.

[71] Anshuman Sharma. Handwritten digit recognition using support vector machine. *CoRR*, abs/1203.3847, 2012.

[72] J. J. Ward, L. J. McGuffin, B. F. Buxton, and D. T. Jones. Secondary structure prediction with support vector machines. *Bioinformatics*, 19(13):1650–1655, 2003.

[73] Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-validation. In LING LIU and M.TAMER Ã–ZSU, editors, *Encyclopedia of Database Systems*, pages 532–538. Springer US, 2009.

[74] Chris McCormick. K-fold cross-validation, with matlab code. Computer Vision and Machine Learning Projects and Tutorials, July 2013.

[75] J. MacQueen. Some methods for classification and analysis of multivariate observations, 1967.

[76] Mei Yeen Choong, Wei Leong Khong, Wei Yeang Kow, L. Angeline, and K.T.K. Teo. Graph-based image segmentation using k-means clustering and normalised cuts. In *Computational Intelligence, Communication Systems and Networks (CICSyN), 2012 Fourth International Conference on*, pages 307–312, July 2012.

[77] Adam Coates and Andrewy Ng. Learning feature representations with k-means. In *Neural Networks: Tricks of the Trade*. Springer Berlin Heidelberg, 2012.

[78] Elke Braun, Bart Geurten, and Martin Egelhaaf. Identifying prototypical components in behaviour using clustering algorithms. *PLOS ONE*, 5(2), 2010.

[79] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[80] Jianjun Cheng, Xiaoyun Chen, Haijuan Yang, and Mingwei Leng. An enhanced k-means algorithm using agglomerative hierarchical clustering strategy. In *Automatic Control and Artificial Intelligence (ACAI 2012), International Conference on*, pages 407–410, March 2012.

[81] Xiang Li, Huaimin Wang, Gang Yin, Tao Wang, Cheng Yang, Yue Yu, and Dengqing Tang. Inducing taxonomy from tags: An agglomerative hierarchical clustering framework. In Shuigeng Zhou, Songmao Zhang, and George Karypis, editors, *ADMA*, volume 7713 of *Lecture Notes in Computer Science*, pages 64–77. Springer, 2012.

[82] Sang-Wan Lee, Yong Soo Kim, and Zeungnam Bien. Learning human behavior patterns for proactive service system: Agglomerative fuzzy clustering-based fuzzy-state q-learning. In *Computational Intelligence for Modelling Control Automation, 2008 International Conference on*, pages 362–367, Dec 2008.

[83] Shervin Emami and Valentin Suciu. Facial recognition using opencv. *Journal of Mobile, Embedded and Distributed Systems*, 4(1), 2012.

[84] Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.*, 5:1205–1224, December 2004.

[85] Yang Li, Bin-Xing Fang, You Chen, and Li Guo. A lightweight intrusion detection model based on feature selection and maximum entropy model. In *Communication Technology, 2006. ICCT '06. International Conference on*, pages 1–4, Nov 2006.

[86] Ying Zhao and George Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, pages 515–524, New York, NY, USA, 2002. ACM.

[87] Ying Zhao and George Karypis. Criterion functions for document clustering: Experiments and analysis. Technical report, 2002.

[88] Isabelle Guyon and André Elisseeff. An introduction to feature extraction. In Isabelle Guyon, Masoud Nikravesh, Steve Gunn, and LotfiA. Zadeh, editors, *Feature Extraction*, volume 207 of *Studies in Fuzziness and Soft Computing*, pages 1–25. Springer Berlin Heidelberg, 2006.

[89] Md. Atiqur Rahman Ahad, J. K. Tan, H. Kim, and S. Ishikawa. Motion history image: Its variants and applications. *Mach. Vision Appl.*, 23(2):255–281, March 2012.

[90] G.R. Bradski and J. Davis. Motion segmentation and pose recognition with motion history gradients. In *Applications of Computer Vision, 2000, Fifth IEEE Workshop on.*, pages 238–244, 2000.

[91] Alexander Refsum Jensenius. Using motiongrams in the study of musical gestures. In *Proceedings of the International Computer Music Conference*, pages 499–502, New Orleans, LA, 2006. Tulane University.

[92] OpenTox. Information gain attribute evaluation@ONLINE, June 2014.

[93] Kenji Kira and Larry A. Rendell. A practical approach to feature selection. In *Proceedings of the Ninth International Workshop on Machine Learning*, ML92, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc.

[94] Lindsay I Smith. A tutorial on principal components analysis. Technical report, Cornell University, USA, February 26 2002.

[95] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. *Mach. Learn.*, 6(1):37–66, January 1991.

[96] N. Kouiroukidis and G. Evangelidis. The effects of dimensionality curse in high dimensional knn search. In *Informatics (PCI), 2011 15th Panhellenic Conference on*, pages 41–45, Sept 2011.

[97] Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *J. Mach. Learn. Res.*, 1:113–141, September 2001.

[98] Jaime S. Cardoso and Maria J. Cardoso. Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. *Artif. Intell. Med.*, 40(2):115–126, June 2007.

[99] Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. *The Annals of Statistics*, 26(2):451–471, 04 1998.

[100] George Karypis. *Cluto: A Clustering Toolkit*. University of Minnesota, Department of Computer Science, November 2003.

[101] Alexander Refsum Jensenius. Evaluating how different video features influence the visual quality of resultant motiongrams. In *Proceedings of the 9th Sound and Music Computing Conference*, pages 467–472, Copenhagen, 2012.

[102] Michael E. Houle, Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Can shared-neighbor distances defeat the curse of dimensionality? In *Proceedings of the 22Nd International Conference on Scientific and Statistical Database Management*, SSDBM'10, pages 482–500, Berlin, Heidelberg, 2010. Springer-Verlag.

[103] Derong Liu, Huaguang Zhang, Marios M. Polycarpou, Cesare Alippi, and Haibo He, editors. *Advances in Neural Networks - ISNN 2011 - 8th International Symposium on Neural Networks, ISNN 2011, Guilin, China, May 29-June 1, 2011, Proceedings, Part II*, volume 6676 of *Lecture Notes in Computer Science*. Springer, 2011.

# Appendix A

# Appendix

Here in the Appendix are attached some auxiliary informations to the methodologies described on Chapter 4.

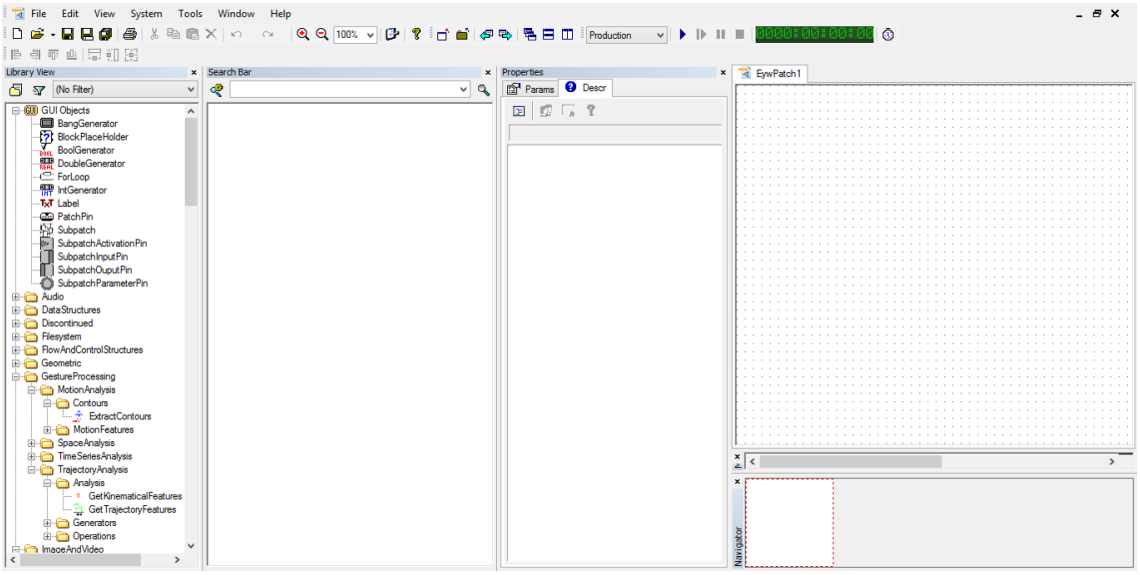


**Figure A.1:** Eyesweb Development Environment.

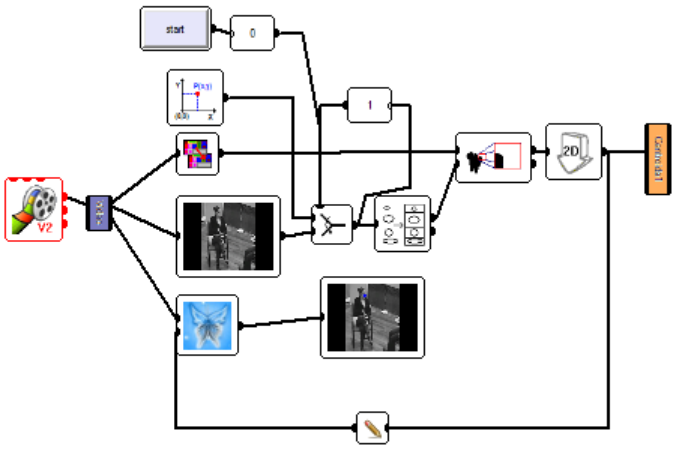


**Figure A.2:** Eyesweb patch used for body parts tracking. An initial coordinate was given for the centroid position of each body part to initiate the tracking.

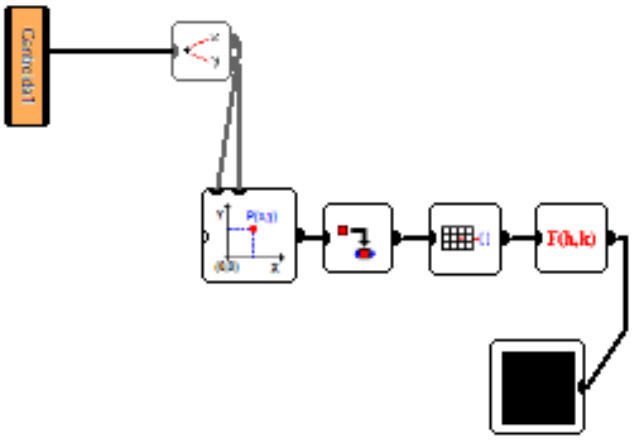


**Figure A.3:** Eyesweb patch used for computation of the Occupation Rate of each body part's centroid.

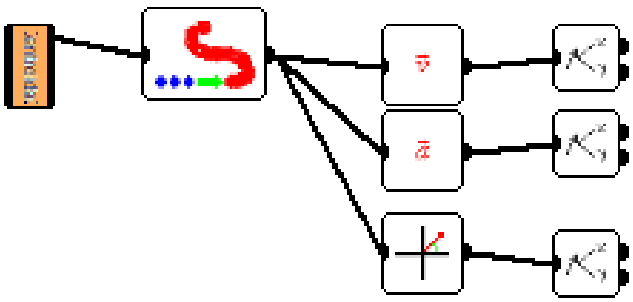

**Figure A.4:** Eyesweb patch used for computation of kinematic features from trajectories: velocity, acceleration and direction.

As seguintes afirmações por favor, indique em que medida concorda com cada uma das seguintes afirmações.

No espaço anterior a cada afirmação, escreva o número (1, 2, 3, 4, 5, 6, 7, 8, ou 9) que indica em que medida as seguintes afirmações descrevem a sua opinião. Quando responder, por favor procure utilizar todos os números desta escala (1 a 9). Não há respostas corretas ou erradas a estas questões.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Discordo totalmente 1 | 2 | Discordo moderadamente 3 | 4 | Neutro 5 | 6 | Concordo moderadamente 7 | 8 | Concordo totalmente 9 |

_____ Considero importante a formação profissional dos surdos

_____A comunidade ouvinte não conhece a comunidade surda.

_____A comunidade ouvinte não se preocupa com a comunidade surda.

_____Existe igualdade de oportunidades educativas/profissionais para surdos e ouvintes.

_____A comunidade ouvinte descrimina a comunidade surda.

_____Existe preconceito em relação à pessoa surda.

_____A Língua Gestual Portuguesa deveria ser lecionada para os alunos em geral.

_____Os surdos e os ouvintes têm os mesmos direitos.

_____Considero que os surdos devem frequentar as mesmas escolas que ouvintes.

**INSTRUÇÕES**: Por favor, indique os seguintes dados demográficos.

1. Nome _____.
2. Idade _____.

3. Sexo:
   _____ Masculino
   _____ Feminino

4. Qual a sua naturalidade?
   _____ Portuguesa
   _____Outra

5. Qual o grau de escolaridade que concluiu?

_____ Sem escolaridade
_____ Ensino primário
_____ Ensino básico
_____ Ensino secundário
_____ Ensino superior
_____ Ensino profissional
_____Outro. Qual?_____.

6. Qual a sua situação atual perante o emprego?

_____Empregado

_____Desempregado

_____Reformado.

_____Estudante.

7. Qual a sua profissão?_____.

8. Conhece a Língua Gestual Portuguesa?

_____Não.

_____Sim.

Se sim, há quanto tempo? _____.

9. Qual a sua formação ao nível da Língua Gestual Portuguesa (LGP)?

____docência de língua gestual portuguesa.

____interpretação de língua gestual portuguesa.

____cursos básicos de língua gestual portuguesa (por ex. terapia da fala).

____não tem.

10. Tem conhecimento da comunidade surda ou de associações representativas das pessoas surdas?

_____Não.

_____Sim, mas poucas.

_____Sim, muitas.

11. Tem amigos surdos?

_____Não.

_____Sim, mas poucos.

_____Sim, muitos.

12. Nasceu surdo?

_____Sim.

_____Não.

Se não, com que idade ficou surdo(a)?_____. E qual a razão?_____.

13. Qual o seu tipo de surdez?

_____.

14. Os seus pais são ouvintes ou surdos?

_____Ouvintes

_____Surdos

_____Ouvinte/Surdo.

15. Tem mais familiares com surdez?

_____Não.

_____Sim.

16. Usa prótese auditiva?

_____Sim.

_____Não, nunca.

_____Não, mas já usei.

17. Usa implante?

_____Não.

_____Sim.

***Agradecemos a sua participação.***

**Figure A.5:** Questionnaire used for deaf subjects containing demographic and opinion questions. Portuguese version of the questionnaire.

As seguintes afirmações por favor, indique em que medida concorda com cada uma das seguintes afirmações.

No espaço anterior a cada afirmação, escreva o número (1, 2, 3, 4, 5, 6, 7, 8, ou 9) que indica em que medida as seguintes afirmações descrevem a sua opinião. Quando responder, por favor procure utilizar todos os números desta escala (1 a 9). Não há respostas corretas ou erradas a estas questões.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Discordo totalmente 1 | 2 | Discordo moderadamente 3 | 4 | Neutro 5 | 6 | Concordo moderadamente 7 | 8 | Concordo totalmente 9 |

_____ Considero importante a formação profissional dos surdos

_____A comunidade ouvinte não conhece a comunidade surda.

_____A comunidade ouvinte não se preocupa com a comunidade surda.

_____Existe igualdade de oportunidades educativas/profissionais para surdos e ouvintes.

_____A comunidade ouvinte descrimina a comunidade surda.

_____Existe preconceito em relação à pessoa surda.

_____A Língua Gestual Portuguesa deveria ser lecionada para os alunos em geral.

_____Os surdos e os ouvintes têm os mesmos direitos.

_____Considero que os surdos devem frequentar as mesmas escolas que ouvintes.

**INSTRUÇÕES**: Por favor, indique os seguintes dados demográficos.

1. Nome _____.
2. Idade _____.
3. Sexo:
   _____ Masculino
   _____ Feminino
4. Qual a sua naturalidade?
   _____ Portuguesa
   _____Outra
5. Qual o grau de escolaridade que concluiu?

5. Qual o grau de escolaridade que concluiu?

_____ Sem escolaridade
_____ Ensino primário
_____ Ensino básico
_____ Ensino secundário
_____ Ensino superior
_____ Ensino profissional
_____Outro. Qual?_____.

6. Qual a sua situação atual perante o emprego?

_____Empregado

_____Desempregado

_____Reformado.

_____Estudante.

7. Qual a sua profissão?_____.

8. Conhece a Língua Gestual Portuguesa?

_____Não.

_____Sim.

Se sim, há quanto tempo? _____.

9. Qual a sua formação ao nível da Língua Gestual Portuguesa (LGP)?

____docência de língua gestual portuguesa.

____interpretação de língua gestual portuguesa.

____cursos básicos de língua gestual portuguesa (por ex. terapia da fala).

____não tem.

10. Tem conhecimento da comunidade surda ou de associações representativas das pessoas surdas?

_____Não.

_____Sim, mas poucas.

_____Sim, muitas.

11. Tem amigos surdos?

_____Não.

_____Sim, mas poucos.

_____Sim, muitos.

***Agradecemos a sua participação.***

**Figure A.6:** Questionnaire used for hearing subjects containing demographic and opinion questions. Portuguese version of the questionnaire.

| Session | Subject | Topic | Number of Samples |
|---|---|---|---|
| 04 | H6 | 1 | 33 |
| 04 | H6 | 2 | 43 |
| 04 | H6 | 3 | 62 |
| 04 | H6 | 4 | 44 |
| 04 | D1 | 1 | 33 |
| 04 | D1 | 2 | 42 |
| 04 | D1 | 3 | 60 |
| 04 | D1 | 4 | 33 |
| 06 | H2 | 1 | 32 |
| 06 | H2 | 2 | 20 |
| 06 | H2 | 3 | 42 |
| 06 | H2 | 4 | 35 |
| 06 | D6 | 1 | 31 |
| 06 | D6 | 2 | 19 |
| 06 | D6 | 3 | 42 |
| 06 | D6 | 4 | 35 |
| 07 | D5 | 1 | 26 |
| 07 | D5 | 2 | 22 |
| 07 | D5 | 3 | 35 |
| 07 | D5 | 4 | 37 |
| 07 | D6 | 1 | 26 |
| 07 | D6 | 2 | 21 |
| 07 | D6 | 3 | 34 |
| 07 | D6 | 4 | 37 |
| 08 | D4 | 1 | 21 |
| 08 | D4 | 2 | 10 |
| 08 | D4 | 3 | 19 |
| 08 | D4 | 4 | 23 |
| 08 | D6 | 1 | 19 |
| 08 | D6 | 2 | 10 |
| 08 | D6 | 3 | 19 |
| 08 | D6 | 4 | 37 |
| 09 | D5 | 1 | 25 |
| 09 | D5 | 2 | 58 |
| 09 | D5 | 3 | 53 |
| 09 | D5 | 4 | 89 |
| 09 | H7 | 1 | 26 |
| 09 | H7 | 2 | 58 |
| 09 | H7 | 3 | 53 |
| 09 | H7 | 4 | 64 |

**Table A.1:** Samples considered for subject in each session through each conversation topic.