Inf Syst Front (2017) 19:1243–1259
DOI 10.1007/s10796-017-9765-z

Grammatical facial expression recognition in sign language discourse: a study at the syntax level

Fernando A. Freitas[1] · **Sarajane M. Peres**[1] · **Clodoaldo A. M. Lima**[1] · **Felipe V. Barbosa**[2]

Published online: 30 May 2017
© Springer Science+Business Media New York 2017

Abstract Facial Expression Recognition is an already well-developed research area, mainly due to its applicability in the construction of different system types. Facial expressions are especially important in the area which relates to the construction of discourses through sign language. Sign languages are visual-spatial languages that are not assisted by voice intonation. Therefore, they use facial expressions to support the manifestation of prosody aspects and some grammatical constructions. Such expressions are called Grammatical Facial Expressions (GFEs) and they are present at sign language morphological and syntactic levels. GFEs stand out in automated recognition processes for sign languages, as they help removing ambiguity among signals, and they also contribute to compose the semantic meaning of discourse. This paper aims to present a study which applies inductive reasoning to recognize patterns, as a way to study the problem involving the automated recognition of GFEs at the discourse syntactic level in the Libras Sign Language (Brazilian Sign Language). In this study, sensor Microsoft Kinect was used to capture three-dimensional points in the faces of subjects who were fluent in sign language, generating a corpus of Libras phrases, which comprised different syntactic constructions. This corpus was analyzed through classifiers that were implemented through neural network Multilayer Perceptron, and then a series of experiments was conducted. The experiments allowed investigating: the recognition complexity that is inherent to each of the GFEs that are present in the corpus; the use suitability of different vector representations, considering descriptive characteristics that are based on coordinates of points in three dimensions, distances and angles therefrom; the need for using time data regarding the execution of expressions during speech; and particularities that are connected to data labeling and the evaluation of classifying models in the context of a sign language.

Keywords Grammatical facial expression · Sign language · Pattern recognition · Machine learning

✉ Sarajane M. Peres
sarajane@usp.br

Fernando A. Freitas
fernandokorban@usp.br

Clodoaldo A. M. Lima
c.lima@usp.br

Felipe V. Barbosa
felipebarbosa@usp.br

[1] School of Arts, Sciences and Humanities, University of Sao Paulo, Sao Paulo, Brazil

[2] Faculty of Philosophy, Languages and Literature, and Human Sciences, University of Sao Paulo, Sao Paulo, Brazil

1 Introduction

The automated recognition of facial expressions attracts the attention from researchers and from the industry, mainly due to its application potential. The analysis of affective facial expressions (Ekman and Friesen 1977; Ekman 1978; Whissell 1989; Jack et al. 2014), for example, opens up possibilities for the development of systems which make use of the analysis of different types of emotions which are expressed during speech (Petridis and Pantic 2011), or while shopping (Popa et al. 2010); for the synthesis of expressions (Yang and Chiang 2008) and sonification (Valenti et al. 2008); or also on systems that use human behavior analysis to monitor people (Yong et al. 2011; Joho et al. 2009; Zhao

et al. 2011) or to support everyday decision-making, such as in a smart house (Yu et al. 2012).

The affective facial expressions are naturally incorporated to human behavior, and, many times, they support discourse construction in the context of both oral languages and sign languages. However, sign languages are visual-spatial languages, and they cannot count on the acoustic support of voice intonation when discourse is built. Due to that, facial expressions are granted new functions, presenting themselves as an important element to assist the manifestation of prosody aspects and some grammatical functions. In that context, facial expressions are then called Grammatical Facial Expressions (GFEs) and they are present at sign language morphological and syntactic levels. GFEs, at a morphological level, provide adjectivization and modify the semantics of signs in sign languages when they are being performed. They can act in an isolate manner, through alterations which are only performed in the facial expression while a sign is being executed, or by supporting the adjectivization of signs through alterations which are performed in the facial expression along with modifications of other lexical units. Regarding the GFEs at the syntax level, the one which is studied in this paper, the collaboration is provided in the composition of the meaning and cohesion of what is being signed, expressing, for example, denials, statements, questions, and relative expressions. The absence of GFEs during discourse in sign languages may represent a lack of fluency and/or some other linguistic problem – which may have arisen when the sign language was being learned, or which may be due to linguistic problems at a neuro-linguistic level (Lillo-Martin and Quadros 2005).

The first studies which aimed at interpreting sign languages were focused on extracting characteristics mainly from hand signals, as observed on Kelly et al. (2009b), a study in which its authors only analyze head movements to support hand signal interpretation, not using information from facial expressions. Nonetheless, nowadays researchers also started considering non-hand signals when they wish to interpret a sign language. Through that, the study of sign language interpretation gained a multi-modal character, due to the consideration of the simultaneousness of its components.

However, as far as we know, the efforts to study the problem of GFE recognition are still elementary, as compared to the efforts taken into the study of affective facial expression recognition. Some of those studies were mainly developed for the American Sign Language (Nguyen and Ranganath 2008; Aran et al. 2009; Michael et al. 2009; Ding and Martinez 2010; Caridakis et al. 2011; Gibet et al. 2011; Kostakis et al. 2011; Yang and Lee 2011; Nguyen and Ranganath 2012; Kacorri 2013), and there are some initiatives for the German (von Agris et al. 2008), Czech (Krnoul et al. 2010; Hrúz et al. 2011), Irish (Kelly et al. 2009a, b) and Turkey (Ari et al. 2008) sign languages. Another aspect which hinders the development of applications from ones which have already been developed for the recognition of GFEs is that each sign language has its own set of GFEs, and each comprises different settings of face elements, which may or may not be accompanied by head movements. Thus, it is necessary to conduct specific studies for each sign language; as, although they may inherit knowledge from studies conducted for other languages, they need to have their particularities analyzed.

The study herein is the first one to be conducted in regards to recognizing GFEs in Libras Sign Language. This study is based on the construction of inductive models, by means of binary classifiers, in order to recognize GFEs within the execution of phrases in Libras. The inductive model was chosen as a way to provide an analysis which is originated from the observation of real GFE manifestations, and neural network Multilayer Perceptron model was selected as a tool for the construction of classifiers.

As this study was the first of its type to be conducted for Libras Sign Language, it was necessary to design and build a corpus of sign language phrases.[1] That corpus was built from the capturing of signals from 45 different phrases, in nine different syntactic contexts. Signals were performed by two signers who were fluent in Libas Sign Language, and capturing was implemented through sensor Microsoft Kinect along with Microsoft Face Tracking Software Development Kit for Kinect for Windows. Those resources allowed the acquisition of RGB (frames of a video) images and of 100 three-dimensional points which described the face of signers in each image. Sensor Kinect as the tool to capture data was mainly chosen due to its low cost and convenient use, factors which facilitate the construction of future systemic applications which may originate from this study. Besides those characteristics, the use of that sensor in similar studies from different areas also reinforces its adaptability to the task related to acquiring face points (Li et al. 2013; Takahashi et al. 2013; Yargic and Dogan 2013; Zhang et al. 2014).

The classifying models are present in the strategy adopted for this study, only as a tool to provide the means to analyze the following aspects: the complexity that is inherent to each of the GFEs that are present in the corpus of Libras Sign Language phrases; the use suitability of different vector representations, considering descriptive characteristics that are based on coordinates of points in three dimensions, distances and angles therefrom; the need for using time data regarding the execution of expressions during speech; and, finally, particularities that are connected

[1] Available at https://archive.ics.uci.edu/ml/datasets/Grammatical+Facial+Expressions – UCI Machine Learning Repository (Lichman 2013).

to data labeling and the evaluation of classifying models in the context of a sign language. The analysis of those aspects is the main goal of this study, and hence the research herein also supports refining the description of Libras Sign Language.

In order to present this paper contents, it was organized in the following sections: Section 2 shows a summary on the context of GFEs in Libras Sign Language and on Machine Learning, which is represented herein by artificial neural network Multilayer Perceptron; comments on related studies are outlined in Section 3; the task of recognizing GFEs is defined in Section 4, which also contains the corpus of phrases in Libras Sign Language, the preprocessing on corpus data, and the characteristic vectors which were created to present data to the classifying models; experiments with the obtained classifiers and results are presented in Section 5; and some general analyses are presented in Section 6. Finally, Section 7 summarizes the main conclusions of this study.

2 Background

This section is dedicated to providing information which is necessary to understand the study that is presented in this paper. GFEs are presented in the context of Libras Sign Language, as well as a summary of how they manifest themselves in terms of facial elements. Also, an overview of Machine Learning and Multilayer Perceptron is presented.

2.1 Sign languages and grammatical face expressions

The first studies on sign languages were conducted around 1960, by William C. Stokoe Jr. In his work Stokoe (1978), Stokoe identified three basic parameters of American Sign Language (ASL), and argued that they are present in all sign languages. Stokoe (1978) argues that the fundamental difference between sign languages and oral languages is related to the simultaneous structure upon which the elements composing language lexical units – the signs – are organized. According to the same author, signs can be seen as a set of minimal, non-holistic traces, which do not have meanings when considered isolatedly. Complementing Stokoe's studies, Battison (1974) and Aarons (1994) argued that the studying sign languages was more complex. Battison introduced the importance of observing orientation and direction aspects in the basic parameters, defending that those aspects constituted a new parameter in sign composition. And, from that study on non-manual traces (the minimal aspects of a sign), Aarons introduced the fifth basic parameter into the system: non-manual markers. Thus, the five basic parameters for the composition of a gestural system in a sign language are:

1. Handshape: the format the hand assumes when executing signals.
2. Location: the location where the sign is performed within the three-dimensional space, either "anchored" to the signer's body or performed in the neutral space in front of the addressee.
3. Movement: the movement performed by the hands within the three-dimensional space during signing.
4. Orientation: orientation assumed by the palm of the hand during sign production, and the direction towards which that signal is executed.
5. Non-manual markers: head position and movements, body position and movements, looks, and FEs.

Each of those parameters can be explored in a detailed manner, so the real understanding of all aspects involving language study and its particularities can be achieved. This paper presents a study on one of those parameters, the facial expressions, considering their grammatical functionalities within sign language discourse; that is, considering Grammatical Facial Expressions (GFEs).

The Grammatical Facial Expressions (GFEs) are within the context of Non-Manual Markers (NMM).[2] GFEs are related to specific constructions in sign language, both at the morphological and in the syntactical levels, and they are mandatory in certain contexts. They can, then, be used to modify signals, imposing a so-called alteration in some of its minimal traces, which leads the meaning of the phrase being said to be altered. The study in this paper is specifically focused on GFEs at their syntactic level.

At the syntactic level, GFEs are responsible for building questions and relative phrases, determining polarity (affirmative and negative phrases) and conditionals, and for building constructions with a topic and with a focus. It is interesting to notice that signing one of the mentioned constructions without the execution of facial expressions would make the respective phrase grammatically wrong. In the Libras Sign Language, GFEs are applied in the making of ten different types of syntactic constructions (Quadros and Karnopp 2004; Ferreira-Brito 1990):

– Interrogative expressions (*WH-questions*): through the use of interrogative expressions of types WHO, WHEN, WHY, HOW, WHERE. These expressions are characterized by a small raising of the head, accompanied by foldings in the forehead. For example: <WHAT JOHN PAY>(WH-question)[3]
– Interrogative (*y/n*): questions for which YES or NO are expected as answers. For example: <JOHN BUY

[2]NMMs are characterized by head positions and movements, body position and movements, looks, and FEs.

[3]Notation indicating that the interrogative facial expression (WH-question) was used in the whole phrase. Symbols <> mark the period in which that expression is executed.

CAR>(y/n). In the execution of this kind of sentence, the head is visually observed to be lowered, and eyebrows are observed to be raised.
- Interrogative (*doubt*): they express a kind of distrust. For example: <JOHN BATHROOM LOCKED>(doubt). In this expression, the lips are compressed, the eyes are more closed, foldings in the forehead are observed, and the shoulders are slightly tilted to the side or back.
- WHAT or WHO, appearing in subordinate sentences without the interrogative GFE, using markers which are specific to that phrase. For example: <I KNOW WHO STOLE> (affirmative).[4]
- Negative expressions: normally, those sentences have an explicit negative element, such as NO, NOTHING, NEVER, which may be incorporated to the signals or only expressed through non-manual markers (Arrotéia 2005). Those GFEs may be visually described through a horizontal movement of the head, or through movements in the outline of the mouth (lowering of corners), which are always associated to the lowering of eyebrows and lowering of head.
- Affirmative: sentences which express ideas or affirmative actions, using, for example, <I GO MALL>(affirmative). Those expressions are characterized by the vertical up-and-down movement of the head.
- Conditional expressions: phrases which establish a condition to perform something, for example, <IF RAIN>(condition)<I NOT GO PARTY> (negative). This GFE is characterized by the lowering of the head and raising of eyebrows during the execution of the condition, followed by another GFE, which may be negative or affirmative.
- Relative expressions: these are inclusions in a phrase in order to explain something, add information, or fit another relative question into what is being said. The interruption of the expression between the inserted information (raising of eyebrow) and the rest of the phrase characterizes this GFE, for example <GIRL FELL BYCICLE>(relative) SHE THERE HOSPITAL.
- Topics: a different way to organize discourse, for example, <FRUITS> (topic) I LIKE BANANA, i.e., the topic is FRUITS, and in this context, "I like bananas". Those expressions may be characterized in three different ways:
  1. raising of eyebrows, head tilted down and to the side, and eyes wide open (applied in this study);
  2. movement of the head down and forward; eyes wide open, followed by a large movement of the head back and to the side;
  3. head forward, slightly tilted up or down, mouth open with its upper part raised, eyebrow raised, and eyes wide open.
- Focus: phrases which introduce new information in discourse, which may: (a) establish a contrast; (b) inform another thing; (c) emphasize something. For example, if someone says that "Mary bought the car", and that information is wrong, the following speech may come right thereafter: NO <PAUL> (focuses) BOUGHT CAR. That GFE is characterized by the same GFE used in topic (a).

A summary on how facial elements are organized in the execution of each GFE is shown in Table 1. The following symbol scheme is used in this table: ↑ for an upward movement; ↓ for a downward movement; ↔ for a rightward or leftward movement; ↕ for an upward and downward movement; ∗ for compression; ⋄ for opening; ⊖ for withdrawal; ∩ for downward mouth corners.

The moment in which GFEs take place in the context of a sentence and the GFE used influence the meaning of a line in a discourse. Kacorri discusses this issue with details (Kacorri 2013). An affirmative sentence (*John likes Mary* – with the signs "JOHN", "LIKE" and "MARY"), for example, becomes an interrogative sentence (*Does John like Mary?*) if an interrogative(y/n) GFE is used together the same signs involved in the original sentence.

2.2 Machine learning

The ability for a computer program to use historical data in order to improve its performance in the resolution of a certain task is defined by Mitchell (1997) as Machine Learning. That context is related to the inductive learning, in which the system tries to generalize a solution through prior data, thus creating a model that is capable of explaining, under some aspect, the context that such data represent. In that process, an error measure is used in order to support the "learning" dynamics, as in each cycle of the process, the

Table 1 Grammatical Facial Expressions: mapping considering syntactic functions; description considering time–related and unrelated characteristics

Syntactic functions	Eyebrow	Eyes	Mouth	Head
Interrogative (WH-question)	↓			↑
Interrogative (y/n) / Conditional	↑			↓
Interrogative (doubt)	↓	∗	∗	⊖
Negative	↓		∩	↔
Affirmative				↕
Relative expressions	↑			
Topic / Focus	↑	⋄		↓

[4]It was not applied in this study.

performance of the program is measured in order to verify the need to alter its parameters, with aims at finding a better solution for the related task. Inductive learning may be reached through the application of supervised or unsupervised methods. In the first case, the methods consider a labeled data set, and the model's parameters are adjusted through the minimization of an objective function.

A well-known technique, which implements supervised inductive learning, is Multilayer Perceptron (MLP). MLPs were created from the basic concept that was introduced by Rosenblatt in 1958, called Perceptron, which is capable of solving simple problems with linearly separable patterns. MLP is a Perceptron network with a high connectivity degree, which may have one or more layers of internal neurons between the network input and output. Each neuron has a non-linear activation function, which must be differentiable in its whole domain. The potential of this type of neural network was highlighted after Rumelhart and McClelland's studies in 1986, with their backpropagation error algorithm, which is responsible for training and adjusting the weights of each neuron, according to the backpropagation of the error information that is found in each training iteration.

A Perceptron may be formally defined through Eq. 1, in which w is the weight vector that is applied in the neuron inputs; x is the input vector; b is a bias vector that can be positive or negative, and which is adjusted at each learning cycle along with the weight vectors; and φ is the activation function which will be responsible for a neuron response after an input. Thus, the equation which defines a Perceptron is

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right) = \varphi\left(\mathbf{w}^T\mathbf{x} + \mathbf{b}\right), \tag{1}$$

in which n is the input dimension which is provided to the neuron.

MLP networks are composed of interconnected neurons (Perceptrons), and they are generally organized as follows: an input layer that is responsible for receiving the information that is going to be processed by the neurons in a network; one or more hidden layers, which are the neuron groups between the input and the output layers; and, finally, the output layer, that is responsible for producing the network response. The information is progressively propagated in that kind of network; i.e., each neuron is connected to all other neurons in the following layer, with information being conveyed from layer to layer, without the presence of feedback. For that type of neural network, the use of a learning algorithm is made necessary, as it is responsible for adjusting the weights of neuron connections, and, consequently, for extracting the characteristics of a problem. Generally, that algorithm acts in two ways: in sequence (*online*), in which an error is analyzed at each iteration, or in *batch*, with errors being calculated only at the end of an epoch, i.e., only after all training data are passed on to the network.

The Backpropagation error algorithm is one of the most used algorithms in supervised learning neural network architectures, and its proposition by Werbos (1974) enabled great breakthroughs in the area of neural networks (Haykin 2009).

In summary, in the Backpropagation error algorithm, an error must be determined in an m iteration, which is assessed by subtracting the resulting value of iteration $y_j(m)$ from the desired value $d_j(m)$. The mean squared error of an interaction s, represented by ϵ_{mean}, will be responsible for representing function cost (Eq. 2), which, inversely, is the network performance that is measured by and activation function. That cost function represents the need for adjusting the neural network parameters in order to obtain learning (approximation) of a pattern.

$$\epsilon_{mean} = \frac{1}{M}\sum_{m=1}^{M} \epsilon(m) \tag{2}$$

The Backpropagation error algorithm applies a correction Δw_{ji} in synaptic weights, where i is the number of input values in a neuron and j is the neuron identifier. Correction is proportional to the partial derivative of the sum of errors that are found in the last layer $\epsilon(m)$ regarding Δw_{ji}, i.e., $\partial\epsilon(m)/\Delta w_{ji}$, which will define which rate will be applied to the weight Δw_{ji}, as the error $\epsilon(m)$ is modified.

3 Related works

The literature survey that is organized in this section regards studies that were conducted on facial expression recognition, considering the scope of sign languages. From the set of studies herein, the analyses that were conducted on that area are noticed to be mainly focused on supporting the recognition of signs from sign languages in multi-modal studies. Within multi-modal studies, the use of the information regarding facial expressions helps removing ambiguities and it increases the accuracy of strategies which aim to automate phases or tasks of a discourse recognition process in sign languages. No studies covering all facial expressions that are used in the context of a sign language were found. One of the most comprehensive studies is Kacorri's (2013). In that study, the author models sentences in American Sign Language (ASL), and shows how those expressions can take place during signing. The study by Nguyen and Ranganath (2012) presents an analysis of GFEs (in the scope of ASL) and is the study that is closest to the discussion presented herein. In the remaining related studies, analysis of GFEs were only conducted as an aid for the recognition of hand

signs under a perspective that aims its translation inside the discourse.

In Nguyen and Ranganath (2012), the strategy for recognizing the GFEs was designed from a manual marker in the image in order to outline and describe the facial expression, followed by the application of a prediction strategy that was based on the analysis of geometrical shapes. The authors tested their predictive model under two perspectives: (a) only considering a person as the source of the discourse under analysis (user-dependent approach); (b) considering more than one person as the source of the discourse under analysis (user-independent approach).

An analysis regarding the nature of analyzed data in correlated studies revealed that approaches which were based in data capturing via RGB-type cameras are still predominant. All studies used videos from that kind of camera to provide data for analysis. In regards to the extracting of descriptive characteristics, there are studies that extract them considering the whole face, such as in Ding and Martinez (2010), and authors which only analyze some regions of the face, such as Saeed (2010), who analyzes GFEs by only using the capturing of lip characteristics. The characteristics are extracted considering the time domain and the simultaneousness of expressions (Kostakis et al. 2011; Kacorri 2013), and considering different kinds of features, such as in Caridakis et al. (2011). Some of the related studies deal with the analysis taking into account the non-manual signs in a multi-modal analysis (Kelly et al. 2009a, b; Michael et al. 2009; Yang and Lee 2011; Nguyen and Ranganath 2012; von Agris et al. 2008; Krnoul et al. 2010; Ari et al. 2008).

Although little research effort has been employed in the analysis of GFEs, there is already a series of databases which are specifically prepared for that kind of study. Some of them are: UWB-07-SLR-P (Campr et al. 2008); database of National Center for Sign Language and Gesture Resources;[5] Boston University American Sign Language Linguistic Research Project database; a base with no name that was proposed by Aran et al. (2007). The use of those databases as reference data sets could be a good strategy for the comparison of studies conducted in that area. However, as all databases refer to videos, and, as far as it was possible to verify, their descriptive characteristics are not made available, their use would result in the construction of approaches which also need to deal with the image-processing problem.

Through the research on the recognition of GFEs, a clear concern is observed to exist with the problems related to time-dependence and occlusions. In the first case, the importance of time representations is connected to the fact that the execution of a facial expression is present in the signing of one or more signs. In turn, regarding the problem with occlusions, which is characterized by the loss of information that originates from the presence of a hand between the capture sensor and the face, it represents an analysis factor which is indispensable for the case of applications whose final objective is the automated recognition of signs, since the positioning of a hand next to the face is frequent in the construction of a discourse in sign languages.

[5] http://www.bu.edu/asllrp/cslgr/.

4 Recognition of grammatical facial expressions

The problem with the recognition of GFEs, implemented in this study, is defined in this section. Besides that, the corpus that was generated in order to allow for the induction of classifying models is described, as well as the procedures that were applied to adapt the data from the corpus for their use in the experiments. In this section, the different characteristics that are used to describe a GFE in the corpus are presented, as well as how GFEs are organized in a vector representation.

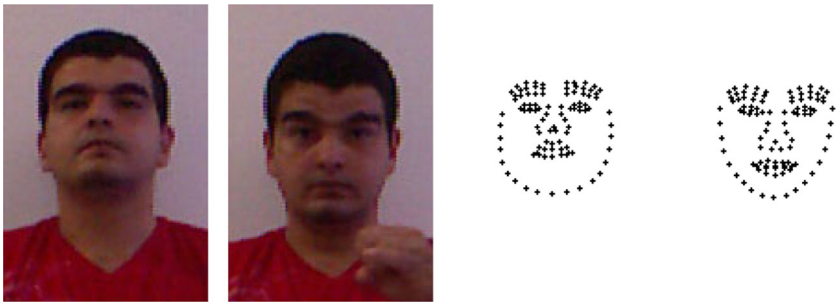
4.1 Definition of the problem

A facial expression $FE_i \in \{FE_1, FE_2, ...FE_n\}$ is the manner through which a set of points $P = \{p_1, p_2, ...p_n\}$, extracted from a human face, are arranged in the three-dimensional space. Those points have three coordinates (x, y, z); x is a pixel coordinate in the horizontal axis, y is a pixel coordinate in the vertical axis, and z is a depth coordinate which is provided in millimeters. Figure 1 shows a neutral face (a face which performs no GFEs) and an example of GFE used in Libras Sign Language, considering a real frame from a video where a GFE is being executed and the respective points (x, y) that were extracted from that face.

In order to define the basic classification problem that was used to perform the analyses herein proposed, consider a video as a sequence of frames $S = \{f_1; f_2; ...f_n\}$, of size n. A vector representation of that video is used as an input for a binary classification model, whose goal is to classify each of the frames as if they regard to the execution of a specific GFE. Upon finding the frames which regard to the execution of a GFE, the classifier marks the discourse segment where a syntactic modifier of a sentence occurs.

4.2 Corpus: grammatical facial expressions of Libras

The study herein has been conducted for the first time for the Libras Sign Language, and it was therefore necessary to produce a specific corpus for the study to be conducted. Although there are several videos which contain hundreds of people using Libras Sign Language in different contexts which are publicly available, the use of such videos would result in the need for processing and interpreting images,



(a) Neutral Face (b) GFE Conditional (c) Neutral Face Points (d) GFE Points

Fig. 1 Example of a neutral face and the execution of a GFE, and the respective points (x, y) that were extracted from those faces

inserting an additional phase into the composition of the approach.

Sensor Microsoft Kinect™[6] was used to capture the data which regard to Libras Sign Language speeches. Such choice was motivated by the possibility, observed in previous studies, to adjust the sensor for the task of capturing data. It was also motivated by the quality of the data that were obtained through its use, and also because it is an inexpensive sensor which is easily coupled to applications that may be implemented for real-life usage.

The data were obtained through the use of *Microsoft Face Tracking Software Development Kit for Kinect for Windows* (*Face Tracking SDK*)[7] – a mechanism that was especially designed to support the tracking of faces which are imaged by sensor Kinect. From the use of the different features that are available in that development package, it was possible to build an application that was capable of extracting face points from a person in front of a Kinect device. The developed application is capable of capturing around 27 frames per second in real time, of making an image available in each frame, and of making available the 100 points extracted from a face, for each of the frames.

Library *Face Tracking SDK* uses two methods for capturing the points of a face. The first one is called *StartTracking*, and the second one, *ContinueTracking*. The first method is used to start the process to obtain data. The second method, which is executed after the first one, uses information from the previous frame to identify the face points in the current frame. It is a point-prediction system which aids obtaining information which allows obtaining a higher number of frames per second. The use of *ContinueTracking* is optional. However, with the sole use of *StartTracking*, the capturing rate of frames per second is much lower, once it is not capable to use previous information to support the analysis of the current frame (the frame being captured). Due to its limited use of the information from the previous frame, *StartTracking* is not cost-effective (its cost is high and its performance is unsatisfactory) for the goals of this study, once the reconstruction of the frames that are captured in a video do not represent the contents in a fluent way.

Even though the data extracting tool is capable of providing several points of a face with an adequate precision for the recognition of GFE patterns, it has trouble dealing with occlusions. Naturally, a sign language is composed of signs which take the region of the head as the location. In those cases, invariably, the hands of a signer are placed between their face and the sensor which captures their face points. The capturing from sensor Kinect is hindered when there are occlusions. The sensor, along with *Face Tracking SDK* features, are capable of retrieving the capture quality when the occlusion ends. However, the frames that are captured while occlusions take place are not well represented.

For the construction of the corpus, five phrases involving each one of the GFEs of interest were chosen in this study (Table 2). The phrases were compose of signs which avoid face occlusions (by the hands), and they were captured from five different executions from two signers that were fluent in Libras Sign Language. The phrases were executed sequentially, five times, in a single data-capturing session. The data files were stored separately according to their syntactic context (each set of phrases with a kind of GFE), and they contain original images, the original images with the plotting of points (x, y), background images with the plotting of the points, and the respective videos. Coordinates (x, y, z) of 100 points in each frame are stored in text files with an index (time in milliseconds) which indicates which the corresponding image file is.

All data were labeled (1 – presence of GFEs or positive class, and 0 – no expressions or negative class) by two coders who were fluent in Libras Sign Language. The labeling provided by coder 1 was used with the data labels in the supervised learning that was applied in the implementation of GFE recognition using MLP. Table 3 describes the corpus, as built in terms of number of captured frames, number

[6] A device that is capable of capturing RGB images which hold depth information, and also of capturing acoustic information (http://msdn.microsoft.com/en-us/library/hh855347.aspx).

[7] http://msdn.microsoft.com/en-us/library/jj130970.aspx.

Table 2 Set of sentences that compose the corpus

WH-questions	Yes/no questions
When did Waine pay?	Did Waine buy a car?
Why did Waine pay?	Is this yours?
What is this?	Did you graduate?
How do you do that?	Do you like me?
Where do you live?	Do you go away?
Doubt questions	**Topics**
Did Waine buy A CAR?	University ... I study at "*anonymous*"!
Is this YOURS?	Fruits ... I like pineapple!
Did you GRADUATE?	My work ... I work with technology!
Do you like ME?	Computers ... I have a notebook!
Do you GO AWAY?	Sport ... I like volleyball!
Assertions	**Negatives**
I go!	I don't go!
I want it!	I didn't do anything!
I like it!	I never have been in jail!
I bought that!	I don't like it!
I work there!	I don't have that!
Conditional clauses	**Focus**
If rain, I don't go!	It was WAYNE who did that!
If you miss, you lose.	I like BLUE.
If you don't want, he wants.	It was Wayne who pay for it!
If you don't buy, he wants.	The bike is BROKEN.
If it's sunny, I go to the beach.	YOU are wrong.

Relative clauses

The girl who fell from bike? ... She is in the hospital!
The *"anonymous"* university? ... It is located in *"anonymous"*!
That enterprise? ... Its business is tecnology!
Waine, who is Lucas's friend, is graduated in Pedagogy!
Celi, the deaf school, is located in "*anonymous*"!

Table 3 Corpus description

GFE	Frames			Agreement
	+	−	Total	Coefficient
Interrogative (WH-question)	1158	1456	2614	0.83
Interrogative (y/n)	1247	1881	3128	0.67
Interrogative (doubt)	1271	1538	2809	0.77
Negative	1240	1466	2706	0.83
Affirmative	942	1194	2136	0.60
Conditional	1137	2804	3941	0.82
Relative Expression	1194	3040	4234	0.90
Topic	827	2794	3621	0.80
Focus	861	1886	2747	0.81

of positive frames, and number of negative frames for each kind of phrase and for each signer, and the total number of captured frames.

That table also shows the labeling agreement coefficients that were obtained by comparing the labeling between the two coders. The agreement measure that was used is Krippendorff's Alpha (Artstein and Poesio 2008). That coefficient ranges from −1 to 1. Negative values indicate that the labeling was random or of insufficient agreement; values between 0 and 0.2 indicate light agreement; between 0.2 and 0.4, fair agreement; between 0.4 and 0.6, moderate agreement; between 0.6 and 0.8, substantial agreement; and, finally, between 0.8 and 1, perfect agreement. Note in Table 3 the labeling agreement between two humans can only be considered as perfect in six cases. However, the measure also provides robustness for the labeling that was used to train the models in this study (the labeling provide by one of the human coders), once the smallest agreement (which would be the most problematic one) can be considered as of moderate/substantial confidence.

The presentation of that measure can help understand a difficulty that is inherent to the problem. In fact, there is a difficulty regarding the decision-making over the initial and final frames of a video snippet in which a GFE occurs; that is so because there is a moment of transition between a face that is neutral and a face which is set in a way to express a GFE, and the observable differences in the arrangement of face elements are very subtle. Besides that, a GFE is characterized by a set of different face arrangements. The subtlety of observable differences in a face between two frames and the joint analysis of the differences in all the elements of a face makes the process of human labeling difficult, making in inaccurate and variable among different coders.

4.3 Data pre-processing

The data that were obtained from the sensor are already provided in a state that is almost appropriate to be used in the analysis and recognition of patterns. Thus, only two simple strategies to pre-process data were applied: translation and normalization. They were required to reduce the influence of variations while locating the signer in axes x and y in regards to the positioning of the sensor; and to reduce the variations in axis z, which represents the distance between signer and sensor. The translation procedure was applied for all points that were captured through their shifting in regards to a point in space that was chosen to represent its origin $(0, 0, 0)$: the reference point that was chosen for translation was the nose tip. In that procedure, the average of all nose points (average x, average y) was subtracted from all remaining points in axes x and y. The average of the nose points in each signer was applied separately. In Fig. 2 it is

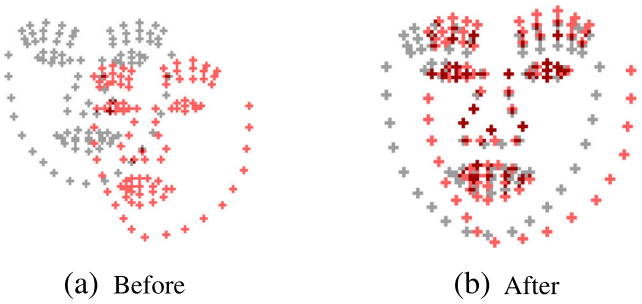


Fig. 2 Juxtaposition of the images of two signers. Signer 1 is in *black* and signer 2 is in *red*

Fig. 3 Final representation of the chosen points, with the watermarked image of all points at the *bottom*

possible to observe the effects of the application of such procedure.

The second procedure was normalization, whose objective was to minimize the influence of the distance between a signer and a sensor (axis z), and to also minimize the differences between the ability for each signer to move their face muscles. The procedure was applied on data which refer to the Euclidean distances and to the angles (which were calculated among each signer's face points). In that procedure, the relative values which refer to the variations of each distance and each angle were used, instead of the absolute values from such variations.

Sensor Kinect, combined to the application of *Face Tracking SDK* features, enables the capturing of 100 points of a face being sensed. However, many of those points have a high correlation, once they regard to the description of regions which are very close within the area where face elements (mouth, nose, eyes, eyebrows, and face outline) move. For that reason, also as a part of the pre-processing task, a study was conducted on the existing correlation among those points. In that study, the correlation among points was calculated considering the shifting they undergo when face elements move. The considered movements were all which may take place in the GFEs of Libras Sign Language. Three specific videos for the performance of that task were recorded, in a way to enable the capturing of the 100 face points, considering the movement of face and head elements, as specified below:

1. Eyebrow, mouth, and head movements.
2. Eyebrow and mouth movements.
3. Head movements.

Each video composed an experiment where the 100 points were analyzed according to the correlation that existed among them. At each experiment iteration, the highest correlation measure found indicated pairs of points to be replaced by their midpoints. That procedure was repeatedly executed while there was a correlation measure of 0.65 or higher, starting from a correlation at 1 and subtracting 0.005 from each iteration. However, replacing pairs of points with correlations which were lower than 0.97 lead to a mischaracterized representation of a face. For example, the grouping of points from left eyebrow with the points of the right eyebrow. Therefore, the correlation measure 0.97 was empirically determined as the minimum correlation to be considered for the replacement of pairs of points. The final set of points is composed of 8 points, as illustrated in Fig. 3a. From the set of 8 selected points for each frame, the two sets of measures were extracted to compose characteristics vectors: D, a set of distances, and A, a set of angles.

It is important to point out that the obtained set of points complies with Chang and Huang (2010) and Wang et al. (2010) which use the same points in their studies. Besides that, the final set of points is also similar to the one that was used in the studies in Yu et al. (2012) and Nguyen and Ranganath (2012) and Dahmane and Meunier (2012), which add some extra points between the mouth and the nose.

4.4 Characteristics vector for grammatical facial expressions

Aiming to explore different ways to organize the characteristics that were extracted from a face (coordinates of points, distances, and angles), some combinations of those characteristics were proposed in order to compose the characteristics vector (vector representation) to be used in the experiments of this study. Four types of characteristics vectors were considered:

– Vector XY_c: Vector with information regarding coordinates (x, y) of points.
– Vector XYZ_c: Vector with information regarding coordinates (x, y, z) of points.
– Vector XY_{da}: Vector with information (distances and angles) derived from coordinates (x, y) of each point.

Table 4 Description of the characteristics vectors

Type of vector	Information	Abbreviation
XY_c	coordinates	XY
XYZ_c	coordinates	XYZ
XY_{da}, $XY_{da}Z_c$	distances and angles	DA
	distances	D
	angles	A
reference point: eyes		
XY_{da}, $XY_{da}Z_c$	distances and angles	DAE
	distances	DE
	angles	AE
reference point: nose		
XY_{da}, $XY_{da}Z_c$	distance and angles	DAN
	distances	DN
	angles	AN
XY_{da}, $XY_{da}Z_c$		
	from literature	Q

Abbreviations: XY and XYZ: coordinates; D: distances; A: angles; E: eyes; N: nose; Q: reference in the literature (Yu et al. 2012; Chang and Huang 2010; Wang et al. 2010)

– Vector $XY_{da}Z_c$: Vector with information derived from coordinates (x, y) of each point, plus the depth information of each point; i.e., using the z coordinate of all points of a frame also in the composition of the vector.

Table 4 identifies each vector type, informs the characteristics that compose them, and attributes an abbreviation for used characteristics. The use of reference points includes distances and angles that were calculated between the points of a face and a fixed reference. In total, 22 characteristics vectors were created for this study.

Besides that, the use of a sliding window complements the representation of data, in a way to characterize the information on the movement of face elements in time (frame by frame). Parameter J defines the size of a "window" in frames, considering a frame sequence. In Table 5 the use of windows of three different sizes is explained.

For the case of windows of size $J = 1$, the problem is considered in its timeless aspect. In the case of the use of a representation with windows of size $J > 1$, it is necessary to define the "frame of interest", i.e., the frame which that window representation refers to. For this study, the first frame in the window was used as the "frame of interest".

Table 5 Examples of windows with three sizes: 1, 2 and 3

Size	Window 1	Window 2	...	Window m
1	$\{f_1\}$	$\{f_2\}$	...	$\{f_n\}$
2	$\{f_1; f_2\}$	$\{f_2; f_3\}$	...	$\{f_{n-1}; f_n\}$
3	$\{f_1; f_2; f_3\}$	$\{f_2; f_3; f_4\}$	...	$\{f_{n-2}; f_{n-1}; f_n\}$

5 Experiments and results

The experiments executed in this study aimed at providing means to analyze the complexity of the recognition of GFEs from Libras Sign Language. Therefore, that refers to the induction of a series of binary classifiers, which are implemented through MLPs that are trained with the Backpropagation error algorithm. The problem with the recognition of GFEs is modeled considering that:

– The information to be recognized comes from the signing of a phrase in Libras Sign Language, containing the use of a GFE of interest that was recorded in a video S, which must be seen as a sequence of frames $\{f_1, f_2, ..., f_n\}$.
– A vector representation V, containing information that was extracted from each of the frames in a video, is used as the input for a classifying model, and it may consider a window of frames, in order to enable a representation of time during which a movement takes place.
– The classifier analyzes the information regarding each of the frames, in order to decide whether it belongs or not to the video segment in which the GFE of interest occurs.
– The response Y from the classifier $\in \{+1, -1\}$. $+1$ means the frame belongs to the video segment in which the GFE occurs; and -1 means the frame does not belong to the video segment in which the GFE occurs.

Different MLP architectures were trained, considering:

– All the characteristics vectors possibilities that were presented in Section 4.4.
– Variations in window sizes, considering interval $[1, N_{max}]$, where N_{max} corresponds to half of the number of frames in the smallest video segment containing the GFE of interest.[8]
– Variations in the following neural network training parameters: backpropagation that was implemented with a descending gradient; number of neurons in the hidden layer (varying in $[\sqrt{d} - -5, \sqrt{d} + 5]$, where d

[8] This parameter assumes different values in each experiment, always considering the shortest time for the execution of an expression in the phrases. Therefore, it prevents a "window" from being large enough to contain frames which represent: non-expression – expression – non-expression.

is the size of the input vector); learning rate (varying in [1, 0.5, 0.25, ..., 0.0156]).
- Variations in the training and testing strategies, adopting signer-dependent and independent strategies, in a way that S_1 indicates that signing was performed by the first signer and S_2 indicates signing was performed by the second signer:
  - Exp. 1: training and validation with S_1 and test with S_1.
  - Exp. 2: training and validation with S_2 and test with S_2.
  - Exp. 3: training and validation with S_1 and test with S_2.
  - Exp. 4: training and validation with S_2 and test with S_1.
  - Exp. 5: training and validation with S_1 and S_2 and test with S_1 e S_2.
- Complete phrases were used for the training (two phrases), validation (one phrase), and testing (two phrases) of classifiers.
- The labeled corpus, considering the labeling provided by one of the human coders.

Table 6 Best results obtained with the MLP models for each of the GFEs

Exp.	F-score no windows	F-score with windows	J	F-score no windows	F-score with windows	J
	Interrog. (WH-question)			Interrog. (y/n)		
1	0.8578	0.8942	6	0.9179	0.9412	3
2	0.8640	0.8988	2	0.8349	0.9129	6
3	0.8320	0.8743	2	0.7788	0.8365	6
4	0.8871	0.8979	3	0.9132	0.9445	4
5	0.8287	0.8599	4	0.8341	0.8860	5
	Interrogative (doubt)			Negative		
1	0.9461	0.9607	4	0.9333	0.9582	3
2	0.9432	0.9700	11	0.6868	0.7269	6
3	0.8391	0.9052	5	0.6863	0.6760	6
4	0.8933	0.9228	3	0.8498	0.8806	5
5	0.9169	0.9452	5	0.7602	0.7830	4
	Affirmative			Conditional		
1	0.8409	0.8773	4	0.9459	0.9534	3
2	0.8333	0.8641	6	0.7930	0.8814	6
3	0.7469	0.7478	2	0.7530	0.7704	5
4	0.7945	0.8331	4	0.9271	0.9410	2
5	0.8057	0.8209	3	0.8357	0.8776	2
	Relative Expression			Topic		
1	0.9661	0.9680	2	0.9198	0.9544	5
2	0.9710	0.9759	2	0.8793	0.9322	3
3	0.8717	0.8653	2	0.8276	0.8953	5
4	0.9463	0.9579	3	0.8780	0.9233	4
5	0.8792	0.8973	3	0.8351	0.9164	6
	Focus					
1	0.9630	0.9836	6			
2	0.8952	0.9213	2			
3	0.8857	0.9022	2			
4	0.9120	0.9538	2			
5	0.8876	0.9321	2			

Results in terms of F-score. J: size of the window

Table 7 Analysis of the type of errors which were made by the best classifier for each of the GFEs

GFE	Total error	Bord. error	False−	False− (bord.)	Falso+	Falso+ (bord.)
Int. (WH)	34 (7%)	13 (38%)	21	4 (19%)	13	9 (69%)
Int. (y/n)	59 (4%)	26 (44%)	30	9 (30%)	29	17 (59%)
Int. (doubt)	14 (3%)	9 (64%)	4	1 (25%)	10	8 (80%)
Negative	17 (4%)	7 (42%)	7	5 (71%)	10	2 (20%)
Affirmative	33 (9%)	11 (33%)	21	3 (14%)	12	8 (67%)
Conditional	17 (3%)	10 (59%)	8	3 (38%)	9	7 (78%)
Relative Exp.	9 (1%)	5 (54%)	5	1 (20%)	4	4 (100%)
Topic	11 (2%)	6 (55%)	5	2 (40%)	6	4 (67%)
Focus	20 (4%)	20 (100%)	9	9 (100%)	11	11 (100%)

Based on above configurations, a series of classifying models was built and tested in each subset of experiments (Exp.1 to Exp. 5). A summary of results is presented in Tables 6 and 7. Table 6 presents the obtained results in regards to performance(F-score) of the best classifiers for each GFE type, considering representations with and without the use of sliding windows.

Table 7 presents the results obtained from the analysis of what kind of error the best classifiers obtained to each GFE. That information in the table is presented in terms of:

- the GFE being analyzed.
- Total error made by the classifier, in terms of the number of frames that were mistakenly classified, and the percentage it represents of the total number of frames that were presented in the test of the model.
- Borderline errors, in terms of the number of borderline frames that were mistakenly classified, and how much that represents in the total number of frames that were mistakenly classified in the test.
- Number of GFE frames that were mistakenly classified as "non-expression" (false negatives).
- Absolute and relative number of GFE frames that were mistakenly classified as "non-expression' within a borderline segment.
- Number of "non-expression" frames that were mistakenly classified as GFE (false positives).
- Absolute and relative number of "non-expression" frames that were mistakenly classified as GFE within a borderline segment.

The "borderline errors" are defined as classification errors that take place within the transition range, between the occurrence of the GFE under analysis and its "inexistence" (herein referred to as "non-expression" phase). In this study, a range of six frames was chosen as the borderline segment (or transition segment). That segment corresponds to three frames before the expression occurs and three frames thereafter.

As an example of the procedures that were adopted in order to analyze borderline errors, consider the sequence of labels that was assigned by the human labeler for a sequence of frames: 00001111. If, as a response from the neural model that was applied to the same frame sequence, 01111111 or 00000001 is obtained, three borderline errors will be identified for each case. The three first positive labels for the first sequence, and the three negative labels for the second sequence.

6 Analysis

Although only the results obtained with the best classifying models were listed for each of the GFEs, the total experiments and results obtained during this study development enabled the making of some comments. Before analyzing them, it is important to remember that the modeling herein allowed observing the results provided by binary classifiers, which were dedicated to the analysis and recognition of GFEs in an individual manner. The modeling allowed the construction of a recognizing model which was capable of identifying a GFE inside the execution of a sentence whose semantic goal was to communicate information which requires the presence of a special marker, which is built through a facial expression.

The first comment to be made regards to the ease of recognition that was observed for each GFE. Under a general analysis, it is plausible to say that the recognition of GFEs was successfully reached.

The classifying models for the interrogative GFE (WH-question) presented F-scores which varied by 0.03 in average. Considering the five best models that were generated for that expression, the number of borderline errors ranged from 29% to 51% of the total error. In turn, for the interrogative GFE (y/n), the results in terms of F-score presented higher variation, with results in 0.94 for experiments where

Fig. 4 Negative Grammatical Facial expression: comparison among signings

testing was conducted with signings of S_1 and 0.80 for experiments tested with signings of S_2. The speech of S_1 is more emphatic, with more marked GFEs, which makes the classifier's job easier. The borderline errors for that GFE followed a similar pattern, varying from 24% to 67% of the errors that were made by the classifier. The recognition of the interrogative GFE (doubt) was very well resolved by the classifier, indicating that the facial expressions that were involved in the construction of that GFE are better discriminated. The borderline errors are also numerous, reaching around 94% of total errors.

The difficulties which were found are more clearly concentrated in the negative GFE, where the lowest results were obtained, with models reaching F-scores in the vicinity of 0.60; and in the affirmative GFE, where the best results are weak as compared to the best results from the remaining GFEs (except the negative GFE). The borderline errors, in both cases, are lower than the other cases, in average. This fact confirms the difficulty in this recognition task, since the errors are concentrated in the middle of the GFE manifestation, i.e., when the GFE is really well-defined. One of the reasons for the difficulty with recognizing the negative GFEs lies in the fact their execution may assume different styles of facial element configuration (please see Fig. 4). Note that signer S_1 characterizes GFEs more intensely. Another reason is the fact that these GFEs reflect continuous head movements, which add an extra difficulty with interpreting the spatial relationships among the points which describe face elements.

Table 8 Characteristics which stood out in the best results

GFE	Type of vector	Information	Reference point.
Int. (WH-question)	$XY_{da}Z_c$	distance	W/eyes
Int. (y/n)	$XY_{da}Z_c/XY_{da}$	angles	nose
Int. (doubt)	XY_c/XYZ_c	coordinates	W
Negative	XY_{da}	angles	W
Affirmative	XY_{da}	distance	W
Conditional	XY_c/XY_{da}	distance	nose
Relative Expression	XY_{da}	distance/angles	W
Topic	$XY_{da}/XY_{da}Z_c$	distance	W/eyes/nose
Focus	XY_{da}	distance/angles	eyes/W

W: without reference point

The conditional GFE presents results which are similar to the interrogative GFE (y/n). That is expected given that both are characterized by the same configuration of facial elements, being differentiated due to the moment they are executed in a phrase. It is interesting to notice that more expressive borderline errors were observed for the conditional GFE (as compared to the interrogative GFE (y/n)), which illustrates that there are interpretation differences between those two GFEs, which are either expressed in the human labeling or in the classifier's decision margin.[9]

The classifier's performance for the relative GFE varies between F-scores of 0.78 to 0.97 for experiments where the signing of S_2 was applied in the classifier's test, and from

[9]There are no elements in this study to allow for evaluating whether this GFE is more difficult to be labeled by human labelers, or if it is difficult for a classifier to interpret the transition phase between non-expression – expression – non-expression.

Table 9 Improvement of results when windows are used

GFE	Highest alteration	Window in the highest alteration
Interrogative (WH-question)	0.0423	2
Interrogative (y/n)	0.0780	6
Interrogative (doubt)	0.0661	4
Negative	0.0401	6
Affirmative	0.0386	4
Conditional	0.0884	6
Relativa Expression	0.0181	3
Topic	0.0813	6
Focus	0.0445	2

0.86 to 0.96 where the signing of S_1 was applied in the classifier's test. The borderline errors are present, making up for 77% of the total error.

The results obtained for topic and focus GFEs are very similar. They presented little F-score variation for the five best models of each experiment, with F-scores of around 0.90. The borderline errors for the topic GFE ranged from 44% to 81%. In the case of the focus GFE, there was an experiment in which 100% of the errors were in borderline.

In regards to the study that was conducted on the different ways to represent face elements and their relationships, there were some highlights, i.e., some characteristics which frequently appeared in the best results. Table 8 summarizes the vector representation analysis, listing the most frequent characteristics in the five best recognition results of each of the GFEs.

The results obtained by the generated classifying models, also considering the five best models, are very similar when the same experiment and the same GFE are considered. However, it is possible to observe that, in a general way, the use of windows leads to improvements in recognition results. Even though, in some cases, the improvement has been very subtle. Table 9 shows a summary of the relative improvement in the best results, provided by the use of windows. Notice that an improvement that was higher than 8% existed in experiments involving conditional, topic, and interrogative (y/n) facial expressions, which are expressions that have similar facial modifications (raising of eyebrows and lowering of head).

From the analysis of borderline errors, it is interesting to notice they can be interpreted as the shifting of the start or end of a GFE. And, also, to the human eye, the small variations in position of face elements which occur between two frames may represent a complex analysis situation. Notice that, if borderline errors were admitted and could be disregarded because the human labeling if imprecise in such frames – the accuracy of classifiers could be considered as being higher. In this sense, Table 10 illustrates what was the accuracy of the classifiers which reached the highest F-scores for each GFE. Table 11 brings the same analysis considering F-score results of the fifth best model that was obtained for each GFE.

Finally, some additional comments may be made in regards the results that were obtained in the experiments:

1. There were no significant differences between using vector XY_{da} or vector $XY_{da}Z_c$ (this one with depth information, as it considers coordinate z).
2. There were no significant differences when using a part of the face (eyes or nose) as the reference.
3. Vectors XY_c and XYZ_c did not stand out in a great deal of facial expressions.
4. There was no standard size for windows, and some experiments were limited to the maximum window size that was established by the phrases in which a grammatical facial expression appeared quickly.

The use of neural models that were trained with learning rates under 0.3 with 13 and 14 neurons in the hidden layer is observed to be the most present within the set of models with the best F-scores. Besides that, the use of the training strategy whose stopping is determined by the validation error indicated that no more than 85 epochs are needed for

Table 10 Accuracy improvement that was represented by the acceptance of borderline errors, for experiments which reached the best results

GFE	F-score (original)	Total error (original)	Total error (final)
Interrogative (WH-questions)	0.8988	34 (7%)	21 (4%)
Interrogative (y/n)	0.9412	25 (5%)	19 (4%)
Interrogative (doubt)	0.9700	14 (3%)	5 (1%)
Negative	0.9582	17 (4%)	10 (2%)
Affirmative	0.8773	33 (9%)	22 (5%)
Conditional	0.9534	17 (3%)	7 (1%)
Relative Expression	0.9759	9 (1%)	4 (< 1%)
Topic	0.9544	11 (2%)	5 (< 1%)
Focus	0.9836	20 (4%)	0

Table 11 Accuracy improvement that was represented by the acceptance of borderline errors, for experiments which reached the lowest results among the five best ones, for each of the GFEs

GFE	F-score (original)	Total error (original)	Total error (final)
Interrogative (WH-question)	0.8341	206(13%)	96(6%)
Interrogative (y/n)	0.7788	326(19%)	186(11%)
Interrogative (doubt)	0.8391	290(19%)	208(14%)
Negative	0.6760	650(41%)	509(15%)
Affirmative	0.7469	265(25%)	173(12%)
Conditional	0.7530	307(15%)	170(11%)
Relative	0.8653	161(8%)	106(13%)
Topic	0.8276	155(8%)	35(4%)
Focus	0.8857	121(9%)	39(6%)

the model to reach a good generalization error. In a general way, the configuration which was used for the neural models is very simple, which shows that the technique itself that was used for the induction of models did not represent a difficulty in this study.

7 Conclusions

This study aimed at developing a set of models for the recognition of patterns which are capable of solving the problem regarding the recognition of facial expressions that are used in the context of Libras Sign Language, the Grammatical Facial Expressions, considering them at a syntactic level. The recognition models were built with Multilayer Perceptron neural networks that were trained to solve binary classification problems. Besides that, in order to investigate the influence of time information in the representation of data, frame windows were used as a way to represent such information in the experiments.

In order to reach that goal, data collecting procedures were performed through the use of Microsoft Kinect, and algorithms were developed in order to extract characteristics and conduct the experiments. In regards to the experiments and analysis of results, one may see this work as the investigation of five questions: dependency of users in the automated recognition of GFEs; characteristics which better represent the GFEs; the influence of time information in the solving of the problem; the influence of the information regarding depth in the problem; and use of a binary classification in the classification of GFEs.

The signing of signer 1 is naturally known to be more defined than the one of signer 2, which can be compared to the voice intonations of different oral speeches. Practically all instances of Experiment 4 (training with the second signer and testing with the first one) had higher results than intances of Experiment 3 (training with the first signer and testing with the second one), which demonstrates that, when training the neural network with more subtle expressions, its performance will be higher when it is tested with well-marked expressions. In Experiment 5 (when mixing different signers), intermediate results were observed.

In order to identify the best descriptors, combinations were made with some data (distances, angles, and information regarding depth), and tests were performed solely with data regarding (x, y) and (x, y, z). Another possibility that was envisioned was also the need to include a reference point in regards to distance and angle measures. For that, the tip of the nose and the average distance between the eyes were chosen. Eight points were chosen through the correlation study that was applied to the 100 points, in order to select the less correlated ones. In regards to which information should be used, the use of distances and angles bring better results than coordinates.

In regards to the use of depth information, the vectors which used the depth information did not stand out in relation to vectors which only used information derived from (x, y), showing that the use of Microsoft Kinect is not justified because it provides that kind of information, since there are other techniques, such as Active Shape Model (ASM) and Active Appearance Model (AAM), which perform the same task of making the information of points (x, y) available in pixels. However, a study to compare the accuracy of that capturing is needed in order to tell which solution would be more adequate in order to provide the information about the face.

In the study concerning time-dependence, the use of time information was concluded to positively influence virtually all experiments, with a window that varied proportionally to the execution time of expressions. Despite that finding, no uniform windows were observed for all expressions, but windows were observed to tend to have sizes between 3 and 6 frames.

Considering the results regarding facial expressions, they were satisfactory, and some difficulties were faced for some

expressions, with a special mention to the negative facial expression. However, a large number of experiments had F-scores over 80%, even when training the technique with a signer and testing it with the other one, or training and modeling with both signers, which demonstrates that the solution for the problem regarding a binary solution is feasible.

The studies in the sign language field are recent, and they were started in 1960 by Stokoe. In regards to the automated recognition of facial expressions in this scope, no studies about Libras Sign Language have been identified to have been conducted, and only a few focusing on sign languages were found. Therefore, this study presents an initial study on the automated recognition of facial expressions within the scope of such Sign Language.

Finally, it is important to mention that this study may be considered the initial effort to support the development of a comprehensive solution for the automated translation of a sign language, contributing to break the barriers of the communication between deaf and hearing communities, using science in order to build a tool which contributes to a more accessible world.

References

Aarons, D. (1994). *Aspects of the syntax of american sign language*. PhD thesis: Boston University Graduate School.

Aran, O., Ari, I., Guvensan, A., Haberdar, H., Kurr, Z., Turkmen, I., Uyar, A., & Akarun, L. (2007). A database of non-manual signs in turkish sign language. In *IEEE 15th signal processing and communications applications. SIU* (pp 1–4). IEEE.

Aran, O., Burger, T., Caplier, A., & Akarun, L. (2009). A belief-based sequential fusion approach for fusing manual signs and non-manual signals. *Pattern Recognition*, *42*(5), 812–822.

Ari, I., Uyar, A., & Akarun, L. (2008). Facial feature tracking and expression recognition for sign language. In *23rd International symposium on computer and information sciences. ISCIS'08* (pp 1–6). IEEE.

Arrotéia, J. (2005). O papel da marcaċão não-manual nas sentenċas negativas em língua de sinais brasileira (lsb). PhD thesis, Dissertaċão de Mestrado. Unversidade Estadual de Campinas.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

Battison, R. (1974). Phonological deletion in american sign language. *Sign language studies*, *5*, 1–14.

Campr, P., Hrúz, M., & Trojanová, J. (2008). Collection and preprocessing of czech sign language corpus for sign language recognition. In *Proceedings of the sixth international conference on language resources and evaluation*.

Caridakis, G., Asteriadis, S., & Karpouzis, K. (2011). Non-manual cues in automatic sign language recognition. In *Proceedings of the 4th international conference on pervasive technologies related to assistive environments* (pp 37–46). ACM.

Chang, C. Y., & Huang, Y.C. (2010). Personalized facial expression recognition in indoor environments. In *International joint conference on neural networks* (pp. 1–8). IEEE.

Dahmane, M., & Meunier, J. (2012). Sift-flow registration for facial expression analysis using gabor wavelets. In *11th International conference on information science, signal processing and their applications* (pp 175–180). IEEE.

Ding, L., & Martinez, A. M. (2010). Features versus context: an approach for precise and detailed detection and delineation of faces and facial features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(11), 2022–2038.

Ekman, P. (1978). Facial signs: facts, fantasies, and possibilities. *Sight, Sound, and Sense*, 124–156.

Ekman, P., & Friesen, W.V. (1977). Facial action coding system.

Ferreira-Brito, L. (1990). Uma abordagem fonológica dos sinais da lscb. *Espaċ,o: Informativo Técnico-Científico do INES*, 20–43.

Gibet, S., Courty, N., Duarte, K., & Naour, T. L. (2011). The signcom system for data-driven animation of interactive virtual signers: Methodology and evaluation. *ACM Transactions on Interactive Intelligent Systems*, *1*(1), 6.

Haykin, S. (2009). *Neural networks and learning machines* (Vol. 3). Pearson Education Upper Saddle River.

Hrúz, M., Trojanová, J., & ŻEleznỳ, M. (2011). Local binary pattern based features for sign language recognition. *Pattern Recognition and Image Analysis*, *21*(3), 398–401.

Jack, R. E., Garrod, O. G., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*, *24*(2), 187–192.

Joho, H., Jose, J. M., Valenti, R., & Sebe, N. (2009). Exploiting facial expressions for affective video summarisation. In *Proceedings of the ACM international conference on image and video retrieval* (p. 31). ACM.

Kacorri, H. (2013). Models of linguistic facial expressions for american sign language animation. *ACM SIGACCESS Accessibility and Computing (105)*, 19–23.

Kelly, D., Delannoy, J.R., McDonald, J., & Markham, C. (2009a). Incorporating facial features into a multi-channel gesture recognition system for the interpretation of irish sign language sequences. In *IEEE 12th international conference on computer vision workshops*. (pp. 1977–1984). IEEE.

Kelly, D., Reilly Delannoy, J., Mc Donald, J., & Markham, C. (2009b). A framework for continuous multimodal sign language recognition. In *International conference on multimodal interfaces* (pp. 351–358). ACM.

Kostakis, O., Papapetrou, P., & Hollmén, J. (2011). Distance measure for querying sequences of temporal intervals. In *Proceedings of the 4th international conference on pervasive technologies related to assistive environments* (pp. 40:1–40:8). ACM.

Krnoul, Z., Hruz, M., & Campr, P. (2010). Correlation analysis of facial features and sign gestures. In *IEEE 10th international conference on signal processing* (pp. 732–735). IEEE.

Li, D., Sun, C., Hu, F., Zang, D., Wang, L., & Zhang, M. (2013). Real-time performance-driven facial animation with 3ds max and kinect. In *2013 3rd International conference on consumer electronics, communications and networks* (pp. 473–476). IEEE.

Lichman, M. (2013). UCI machine learning repository. http://archive.ics.uci.edu/ml.

Lillo-Martin, D., & Quadros, R.Md. (2005). The acquisition of focus constructions in american sign language and língua brasileira de sinais. In *Language development conference on Boston university* (Vol. 29, pp. 365–375).

Michael, N., Metaxas, D., & Neidle, C. (2009). Spatial and temporal pyramids for grammatical expression recognition of american sign language. In *Proceedings of the 11th international ACM SIGACCESS conference on computers and accessibility* (pp. 75–82). ACM.

Mitchell, T. (1997). *Machine learning*. McGraw-Hill.

Nguyen, T.D., & Ranganath, S. (2008). Towards recognition of facial expressions in sign language: tracking facial features under occlusion. In *15th IEEE international conference on image processing. ICIP* (pp. 3228–3231). IEEE.

Nguyen, T. D., & Ranganath, S. (2012). Facial expressions in american sign language: tracking and recognition. *Pattern Recognition*, *45*(5), 1877–1891.

Petridis, S., & Pantic, M. (2011). Audiovisual discrimination between speech and laughter: why and when visual information might help. *IEEE Transactions on Multimedia*, *13*(2), 216–234.

Popa, M., Rothkrantz, L., & Wiggers, P. (2010). Products appreciation by facial expressions analysis. In: *Proceedings of the 11th international conference on computer systems and technologies and workshop for PhD students* (pp. 293–298). ACM.

Quadros, R.Md., & Karnopp, L.B. (2004). Língua de sinais brasileira: estudos lingüísticos (Vol. 1).

Saeed, U. (2010). Comparative analysis of lip features for person identification. In *Proceedings of the 8th international conference on frontiers of information technology* (pp. 20:1–20:6). ACM.

Stokoe, W. C. (1978). Sign language structure. *Annual Review of Anthropology*, *9*, 365–390.

Takahashi, M., Clippingdale, S., Okuda, M., Yamanouchi, Y., Naemura, M., & Shibata, M. (2013). An estimator for rating video contents on the basis of a viewer's behavior in typical home environments. In *2013 International conference on signal-image technology & internet-based systems* (pp. 6–13). IEEE.

Valenti, R., Jaimes, A., & Sebe, N. (2008). Facial expression recognition as a creative interface. In *Proceedings of the 13th international conference on intelligent user interfaces* (pp. pp 433–434). ACM.

von Agris, U., Knorr, M., & Kraiss, K.F. (2008). The significance of facial features for automatic sign language recognition. In *8th IEEE International conference on automatic face & gesture recognition. FG'08* (pp 1–6). IEEE.

Wang, H., Huang, H., Hu, Y., Anderson, M., Rollins, P., & Makedon, F. (2010). Emotion detection via discriminative kernel method. In *Proceedings of the 3rd international conference on pervasive technologies related to assistive environments* (pp. 7:1–7:7). ACM.

Werbos, P. (1974). Beyond regression: new tools for prediction and analysis in the behavioral sciences.

Whissell, C. (1989). The dictionary of affect in language. *Emotion: Theory, Research, and Experience*, *4*(113–131), 94.

Yang, C. K., & Chiang, W. T. (2008). An interactive facial expression generation system. *Multimedia Tools and Applications*, *40*(1), 41–60.

Yang, H. D., & Lee, S. W. (2011). Combination of manual and non-manual features for sign language recognition based on conditional random field and active appearance model. In *International conference on machine learning and cybernetics* (Vol. 4, pp. 1726–1731). IEEE.

Yargic, A., & Dogan, M. (2013). A lip reading application on ms kinect camera. In *2013 IEEE International symposium on innovations in intelligent systems and applications* (pp. 1–5). IEEE.

Yong, C.Y., Sudirman, R., & Chew, K.M. (2011). Facial expression monitoring system using pca-bayes classifier. In *International conference on future computer sciences and application* (pp. 187–191).

Yu, Y. C., You, S. D., & Tsai, D. R. (2012). Magic mirror table for social-emotion alleviation in the smart home. *IEEE Transactions on Consumer Electronics*, *58*(1), 126–131.

Zhang, D., Liu, X., Yan, N., Wang, L., Zhu, Y., & Chen, H. (2014). A multi-channel/multi-speaker articulatory database in mandarin for speech visualization. In *2014 9th International symposium on Chinese spoken language processing* (pp. 299–303). IEEE.

Zhao, S., Yao, H., Sun, X., Xu, P., Liu, X., & Ji, R. (2011). Video indexing and recommendation based on affective analysis of viewers. In Proceedings of the 19th ACM international conference on multimedia (pp. 1473–1476). ACM.

Fernando A. Freitas is researcher and professor at the Federal Institute of Education, Science and Technology of São Paulo, Brazil, since 2014. He works in under-graduation programs. He is Master in Information Systems (2014) at the University of São Paulo; Control and Automation Engineer (2012) at the Federal University of Itajubá, Brazil. His main research interests are computational intelligence, data mining, machine learning, pattern recognition, and assistive technology in the sign language area.

Sarajane M. Peres is researcher and professor at the University of São Paulo, Brazil, since 2007. She works in under-graduation and graduation Information Systems programs. She is PhD in Electric Engineering (2006) at the University of Campinas; Master in Manufacturing Engineering (1999) at the Federal University of Santa Catarina; Bachelor in Computer Science (1996) at the State University of Maringá, Brazil. She worked as researcher and professor at the State University of Western Paraná (1998–2005) and at State University of Maring (2005–2007), Brazil. She is the tutor of an Information Systems undergrad program, since 2010, in which works with junior scientific research and extension activities, under the Tutorial Education Program financed by the Ministry of Education, Brazil. Her main research interests are computational intelligence, data mining, machine learning, pattern recognition and human behavior modeling and analysis. She is leader of the Artificial Intelligence research group (GrIA), which is registered by CNPq - National Counsel of Technological and Scientific Development.

Clodoaldo A. M. Lima is researcher and professor at the University of São Paulo, Brazil, since 2010. He works in under-graduation and graduation Information Systems Programs. He is PhD and Master in Electric Engineering (2004/2000) at the University of Campinas; Bachelor in Electric Engineering (1997) at the State University of Juiz de Fora, Brazil. He worked as researcher and professor at the Mackenzie University (2007–2010), Brazil. His main research interests are signal processing - mainly biomedical signals -, machine learning - mainly kernel methods and committee machines -, biometric systems, non-parametric modelling and time series prediction.

Felipe V. Barbosa is researcher and professor at the University of São Paulo, Brazil, since 2010. He qualified initially as a Speech and Hearing Therapist at University of São Paulo (2001), completing his PhD on Deaf Studies at the same university in 2007. He has worked as sign language interpreter since 1993 and has worked clinically with deaf clients since 2001. Currently, he is leader in the research group on Sign Language and Cognition (LiSCo) at the Department of Linguistics. There are three major priorities in his work: description of Brazilian Sign Language, assessment of Brazilian Sign Language and atypical sign language.