Automatic segmentation of grammatical facial expressions in sign language: towards an inclusive communication experience

Maria Eduarda de A. Cardoso, Fernando de A. Freitas,
Felipe V. Barbosa, Clodoaldo A. de M. Lima, Sarajane M. Peres
University of São Paulo
Brazil
{felipebarbosa, c.lima, sarajane}@usp.br

Patrick C. K. Hung
Ontario Tech University
Canada
patrick.hung@uoit.ca

Abstract

Nowadays, natural language processing techniques enable the development of applications that promote communication between humans and between humans and machines. Although the technology related to automated oral communication is mature and affordable, there are currently no appropriate solutions for visual-spatial languages. In the scarce efforts to automatically process sign languages, studies on non-manual gestures are rare, making it difficult to properly interpret the speeches uttered in those languages. In this paper, we present a solution for the automatic segmentation of grammatical facial expressions in sign language. This is a low-cost computational solution designed to integrate a sign language processing framework that supports the development of simple but high value-added applications for the context of universal communication. Moreover, we present a discussion of the difficulties faced by this solution to guide future research in this area.

1. Introduction

Sign languages are recognized as natural languages and have received attention from several areas of knowledge, including Linguistics – which studies languages and their phenomena from a theoretical and applied standpoint – and Computer Science, especially Artificial Intelligence, Computer Vision and Graphic Processing – which use computational processes to manipulate language and its elements. Studies in these areas have provided advances for the deaf community in terms of accessibility, social and educational inclusion, and enhancement of language policies. Despite the relevance of this studies, advances in communication technology in the visual-spatial modality (including communication) are insignificant when compared with those for oral modality. Gadgets that interpret and synthesize oral languages, using oral communication to implement functionalities, have already reached the state of practice. These services range from help desk or personal assistants to companion robots and smart toys. This is a profitable segment in which little attention has been paid to accessibility or inclusion issues to provide universal design services [1]. As this segment grows based on resources that do not meet a portion of the population that has certain needs, the exclusion factor is aggravated. The exclusion of children, the elderly and people with disabilities might be the most impactful, and research and industry related to smart toys and companion robots do not seem to be progressing to reverse this picture.

Therefore, we should recognize that the technology available for visual-spatial language processing has not reached the maturity of existing technology for oral languages [2]. This makes it difficult to link visual-spatial modality of communication to electronic devices and digital applications, although there are some initiatives working to address it[1]. The difficulty of effectively processing sign languages comes from the poor formalization established for them and the need to recognize both linguistic aspects and complex patterns from elements used in their respective particular articulations.

The outstanding advances for this technology concern the processing of a finite set of signs, considering predominantly the elements formed by hand gestures (hand shapes and hand movements [3, 4, 2]), for interpretation or synthesis of the language. This technological apparatus can be incorporated into gadgets. However, to establish a proper communication experience, it is required to develop the processing of a linguistic element little explored by the computational linguistic area – the Grammatical Facial Expressions (GFE) [5, 6, 7, 8].

The research presented herein complements previous research efforts [9, 10] in which we propose to pro-

[1] Examples: UNI (https://bit.ly/2M9Ywo6), Mimix Sign Language Translator (https://bit.ly/2JGWslq), Hand Talk Translator (https://bit.ly/2Fg4kpI), SignAloud Gloves (https://bit.ly/2QoHQrp).

URI: https://hdl.handle.net/10125/63923
978-0-9981331-3-3

cess sign language based on a modular and extensible framework (Figure 1). The framework comprises a series of pattern recognition modules, each one specialized in recognizing one element of a sign language, and a grammar module responsible for combining the outputs of the former modules and providing the meaning of a composition of elements [11]. This modular approach should not be seen as a definite solution for the automatic processing of sign languages since relevant questions concerning a natural language are left out. However, modularity and finite vocabularies allow the development of low-complexity computational applications that contribute to the mitigation of the social and digital exclusion of the "sign language commonwealth" [12, 13, 14, 15].

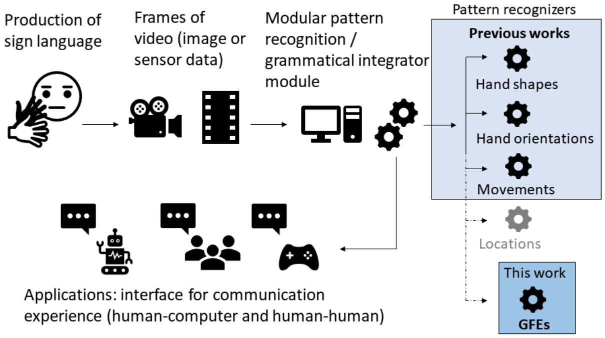

Production of sign language
Frames of video (image or sensor data)
Modular pattern recognition / grammatical integrator module
Pattern recognizers
Previous works
Hand shapes
Hand orientations
Movements
Locations
This work
GFEs
Applications: interface for communication experience (human-computer and human-human)

Figure 1. Basic ideas of a framework to provide sign language processing for building low-complexity applications. Dashed line: next step in the grammar module. Gray color: open issue.

In this paper, we explore the automatic segmentation of GFEs in phrases uttered in the Brazilian sign language (Libras). The task of segmentation concerns the labeling of video frames, which contain the production of sign language given by a person, according to the GFEs used in this language. In this work, this task was modeled as a multiclass classification problem solved with the Multilayer Perceptron (MLP) neural network[2]. Six GFEs, of the eight expressions used in Libras' syntactic constructs, compose the set of classes used in the classification problem. The remaining of this text is organized as follows: theoretical background (Section 2), which comprises basic concepts of GFEs, classification tasks and MLP neural networks; related work (Section 3); GFE segmentation problem definition (Section 4); experiments and results (Section 5); conclusions and next steps (Section 6).

2. Theoretical background

2.1. Facial expressions in sign languages

Facial expressions are part of human communication because they allow expressing emotions and transmitting affective information. However, in sign languages, facial expressions play an additional and essential role. They allow us to give meaning to what we say [11]; as a result, they are part of the construction of the discourse's syntax and semantics. In this process, they are called "grammatical facial expressions" (GFE). The GFEs are known as non-manual markers and are related to language-specific structures at the phonetic-phonological, morphological and syntactic levels [11, 19, 20, 21, 22]. At the phonetic-phonological level, GFEs have the role of differentiating those signs that have the same parameters in other constituent elements of the language (hand shapes, hand orientations, movements and location). At the morphological level, GFEs are applied as morphemic markers for purposes that include determining the intensity of an adjective or expressing superlative and comparative constructs. Finally, at the syntactic level, they determine interrogative, relative conditionals or topic phrases, mark focus or define polarities. In this paper, we are interested in processing the GFEs used at the syntactic level, since it could be pointed out as the most critical level in communication experiences.

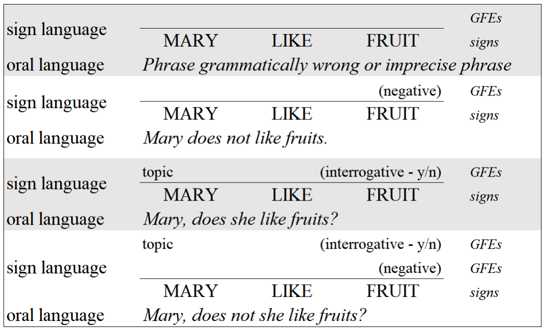
Figure 2 shows the construction of phrases using signs with and without GFEs. If the phrase is performed using only the three signs "Mary", "like" and "fruit", it would represent a phrase grammatically wrong or an imprecise affirmative phrase – "Mary likes fruits". On the other hand, if GFEs are used, the phrases assume different meanings, e.g., a negative phrase, an y/n-interrogative phrase with topic or an y/n-interrogative phrase with topic and (negative) polarization. Other types of phrases at the syntactical level used in Libras are: wh-questions, doubt-questions, conditional expressions, relative expressions and focus.

Figure 3 shows two frames extracted from the video containing the utterance of the phrases "I go!" (left) and "I don't go!" (right), using affirmative and negative GFEs, respectively. Note that the same hand gesture (same sign) was performed in both frames, and the GFEs imposed the desired meaning for the discourse.

2.2. The classification task and the MLP neural networks

In the data mining field, the term "classification" defines the task of finding a mapping between datapoints from a dataset to classes in a set of known classes [23].

[2] MLP-based hardware has been extensively explored in the specialized literature [16, 17]. The ease of embedding an MLP classifier in hardware, coupled with the efficiency of this neural network in classification recognition problems, motivates its application in this study. Hardware implementations contribute to improving security and privacy in gadgets, mainly in the toy computing contexts [18].



sign language				*GFEs*
	MARY	LIKE	FRUIT	*signs*
oral language	*Phrase grammatically wrong or imprecise phrase*			
sign language			(negative)	*GFEs*
	MARY	LIKE	FRUIT	*signs*
oral language	*Mary does not like fruits.*			
sign language	topic		(interrogative - y/n)	*GFEs*
	MARY	LIKE	FRUIT	*signs*
oral language	*Mary, does she like fruits?*			
	topic		(interrogative - y/n)	*GFEs*
sign language			(negative)	*GFEs*
	MARY	LIKE	FRUIT	*signs*
oral language	*Mary, does not she like fruits?*			

Figure 2. Phrases in sign language using GFEs

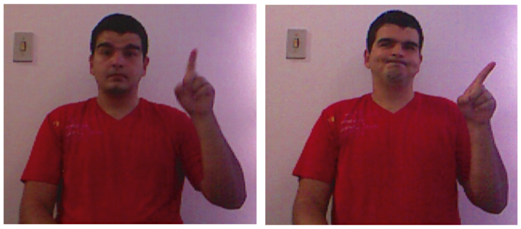


Figure 3. GFEs performed in an affirmative phrase (left) and in a negative phrase (right) – frames from the Grammatical Facial Expressions Dataset [10]

Formally, this mapping occurs between a set of input datapoints ($\overrightarrow{x} \in \Re^D$) to a finite set of labels (C_c), in which D is the dimension of the input vector and c is the cardinality of the set C. The mapping is modeled in terms of a function: $\Im : \Re^D \times W \rightarrow C$, where W is a space of parameters adjustable through a supervised learning algorithm. In order to model a classification problem, the dataset is required to be previously labeled considering each class in C as a possible label. After the mapping is built, it can be applied to classify unknown datapoints.

The Multilayer Perceptron (MLP) neural network comprises a set of sensory neurons that constitutes its input layer; one or more hidden layers of neurons that perform signal processing through non-linear activation functions (differentiable at all points); an output layer with neurons that perform signal processing through non-linear or linear activation functions (differentiable at all points); and two sets of synaptic weights (W_h and W_k) associated to the hidden and output neurons, respectively, and that are in charge of inhibiting or enhancing the neurons input signals [24].

The MLP is a feedforward neural network commonly trained with an error backpropagation algorithm, which is based on the error correction rule known as "generalized Delta rule" [25, 24]. To implement the training process, this neural network architecture uses two types of signals [25]. The former is the functional signal, which is propagated from the neurons of the input layer to the neurons of the output layer, in a forward propagation model. The latter is the error signal, which is generated in the output layer and is propagated back through the network. Both signals are produced by weighting the output activation functions with the synaptic weights. The MLP training algorithm optimizes the sets of synaptic weights to find values that fit the mapping that solves the classification problem. The optimization process runs until the classification error produced at the network output reaches a target minimum limit.

3. Related work

Studies that focus on understanding the role of GFEs in the context of sign languages have been presented in the specialized literature [5, 6, 7, 8]. These studies reinforce that GFEs need to be considered in automatic sign language processing, otherwise sign language-oriented applications may cause expectations breaches or misconceptions about the discourse at hand. The relevance of GFEs has been confirmed, for instance, by studies in language disorders which affect deaf people. In order to diagnose language disorders, Marshall et al. [5] used phrases in which the lack of facial expression would alter the respective meaning. In the experiment, facial expression omissions and meaning changes through lack of facial grammar were used as part of the indicators of language impairment. The studies carried out by Benitez-Quiroz et al. [6, 7] focused on analyzing the consistency of the GFEs characterization in terms of facial joints (or fiducial points) applied to perform them. Sequences of videos with discourse in sign language were annotated to create a linguistic model. The resulting model proved to exist discriminant characteristics for at least nine phrases classes, which also highlights the relevance of the GFEs in sign languages. The impact of GFEs was measured in an experiment carried out by Kacorri and Huenerfauth [8]. In this experiment, the sign language discourse was analyzed by inserting GFEs in digital animations. The discourse synthesized in the animations was interpreted by a group of deaf people. The authors observed that the absence of facial expressions and the low frequency of their use compromise the discourse comprehension significantly.

Because of the relevance of GFEs for sign languages, researchers have recently turned their attention to automatically interpret them through pattern recognition techniques. However, the need to intensify research efforts in this task is still clear, since most research in automatic sign language processing focuses on the recognition of manual gestures. For an overview of this research context, the reports organized in specialized literature reviews [3, 4, 2] may be useful.

With respect to works similar to the one discussed herein, we highlight initiatives related to other sign languages [26, 27, 28, 29] and those in which the efforts were applied in the Libras' context [10, 30, 31, 32, 33]. Liu et al. [26, 27] have explored pattern recognition in the context of (six) GFEs from the American sign language (ASL) through statistical models. These authors pay attention to eyebrow shapes and periodic head nods and shakes, since such characteristics are relevant components of many non-manual grammatical markers in ASL. This approach achieved accuracy rates of about 85% in the pattern recognition task. The GFEs in the British sign language were studied by Caridakis, Asteriadis and Karpouzis [28]. In this work, Elman networks were applied to analyze patterns in five GFEs and the accuracy rates varied between 66% and 100%, depending on the complexity of the test. Some gestures of Indian sign language, which involve the execution of GFEs, were studied by Kumar, Roy and Dogra [29], using multiple classifiers (Hidden Markov Models and Gaussian Mixture) combined through an Independent Bayesian Classifier Combination model. This strategy achieved accuracy rates between 87% and 93%.

In the context of Libras, all works that were found have used the same corpus of videos – the Grammatical Facial Expressions Dataset [10, 30, 34]. This corpus was designed to support studies related to binary segmentation – separating frames of video with a particular GFE from frames of video with other GFE or neutral expression. Therefore, in most of the works, the results were reported in terms of F-score. The works of the proponents of the corpus [10, 30] show results of GFE segmentation implemented by using MLP. In particular, Freitas et al. [30] describe an extensive setup of feature extraction and experiments in both contexts, speechmaker-dependent and speechmaker-independent. The F-score obtained in the experiments varied between 0.68 and 0.97, in speechmaker-dependent contexts, and between 0.67 and 0.95 in speechmaker-independent contexts. These authors have also been concerned with providing a detailed analysis of the complexity of GFEs and their impact on segmentation results. A strategy to estimate parameters in differential equation models, which describe non-linear trajectories of longitudinal data, is presented by Hu and Treinen [33]. The authors analyzed their strategy using the mouth movement of the Libras GFEs. Although this work did not aim to classify GFEs, it is interesting to show existing patterns in using GFEs. For instance, they concluded that, to analyze the mouthing behavior in a Libras phrase, it would be more informative to investigate the trajectory decay time than the trajectory cycle length. The last two related works are studies whose focus was to analyze the capacity of strategies while solving the classification task. Unlike Freitas et al. [10, 30] and our work, these authors are not concerned with the study of sign language *per se*. In the work presented by Uddin [31], Random Forests were applied, achieving F-scores between 0.97 and 1 for experiments including speechmakers in the training and test dataset. Bhuvan et al. [32] applied four techniques: MLP and Radial Base Function neural networks, Bayesian classifiers and Random Forests. Their best results achieved F-scores of 0.9 with MLP neural networks and Random Forests. However, the authors provided few details about the setup of the experiments.

4. Problem definition

The GFEs considered in this paper belong to a finite set of $n+1$ types of grammatical facial expressions $GFE = \{GFE_1, GFE_2, \ldots, GFE_n, GFE_{n+1}\}$, in which GFE_{n+1} is the neutral facial expression. An instance of a GFE is described by a set of m (x, y)-points $P = \{p_1,\ p_2,\ \ldots,\ p_m\}$, extracted from the human face and arranged in a bidimensional space (Figure 4). Next, let a video be a sequence of frames $S = \{q_1,\ q_2,\ \ldots,\ q_T\}$, in which T is the length of the sequence and q is a video frame containing the face image of a person performing a GFE while uttering a discourse in sign language. The aim of the classification model is to point out which GFE occurs in each video frame. Assuming that the classification model will perform its task successfully, sequences of video frames will be classified as containing a GFE_i, with $i = \{1, \ldots, n+1\}$, representing the solution to the GFE segmentation problem.

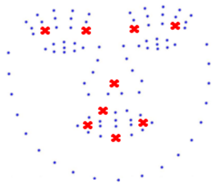


Figure 4. The 100 (x, y)-points extracted from the human face. The points marked with red crosses are the fiducial facial points used in this work.

A vector representation for the faces in the video frames is used as input for a multiclass classification model. This representation is based on the set of points P or on features derived from them. Thus, a face contained in a video frame is represented as a $(c{*}m)$-vector, i.e., the dimension of the input space depends on the number of characteristics (c) describing a point and the number of points (m) in P. For instance, as illustrated by the red crosses in Figure 4, if one considers nine (x, y)-points extracted from the face ($c = 2$ and $m = 9$), then the vector representing the face in the video frame q is $\overrightarrow{x_q} = \{x_1, y_1, x_2, y_2, \ldots, x_9, y_9\}$. In this case, the classification model works in a 18-dimension space.

The classifier models proposed herein consider the GFE segmentation problem under the temporal aspect. However, since traditional MLP does not have mechanisms for processing temporal information, it is necessary to embed it in the vector representation. For this, consider the sequence S as a time-varying signal whose variation relates to the movement of the facial fiducial points (e.g., eyebrows or lips movement) or to the movement of the head. The information contained in this signal can be incorporated into a vector representation through a windowing procedure. When applying this procedure, the vector representation for the frame q_t becomes the concatenation of some frames before and after it in the sequence S. Thus,

$$\overrightarrow{x^l_{q_t}} = \{x_{q_{t-\lfloor l/2 \rfloor}} \ldots, x_{q_{t-1}}, x_{q_t}, x_{q_{t+1}}, \ldots, x_{q_{t+\lfloor l/2 \rfloor}}\},$$

in which $\overrightarrow{x^l_{q_t}}$ is a vector representation for a windowed datapoint, each x_q is as defined before, x_{q_t} is the frame of interest[3] for which a class (GFE_i) will be associated, l is the size of the window and $t = \{\lfloor l/2 \rfloor \ldots, T - \lfloor l/2 \rfloor\}$ shows the position of x_q in S. The windowing procedure creates a $(l * c * m)$-input space.

5. Experiments and results

In this paper, the automatic segmentation of GFEs was modeled as a multiclass classification problem and an MLP was adopted to treat this problem. This strategy was tested through experiments established in speechmaker-dependent and speechmaker-independent contexts, following the strategy presented by Freitas et al. [30]. We created two problems with different levels of complexity: a multiclass problem that comprises two GFEs and the neutral expression, totaling three classes (*Experiment* $\#1$); and a multiclass problem that comprises six GFEs and the neutral expressions, totaling seven classes (*Experiment* $\#2$). The idea was to evaluate the suitability of using these classification models in the construction of applications that involve the recognition of several GFEs at once. In this section, we present the dataset used in the experiments, the preprocessing procedures applied to the data, the setup of the experiments and the results.

5.1. Grammatical facial expressions dataset

The Grammatical Facial Expressions dataset[4] [10, 30, 34] was used in the experiments reported herein. This dataset contains 18 videos referring to two speechmakers uttering phrases in Libras, using nine GFEs. Such videos were recorded with the Microsoft Kinect sensor, using a capture rate of approximately 30 frames per second. In each video, a speechmaker performs five repetitions of five phrases in Libras, which require the performing of at least one GFE among those used in that language. The information regarding facial expressions is stored through 100 spatial coordinates (x, y, z) of the face contour and face fiducial points located in the eyes, nose, eyebrows and mouth. In this dataset, each frame of video was manually labeled by a sign language expert. Labeling refers to the occurrence of GFEs and it can be used as ground truth for classification models evaluation. The occurrence of a GFE is labeled as a positive class, and the absence of it is labeled as a negative class. In this way, a sequence of frames with the label "1" shows a video segment in which a GFE is being used. Therefore, this is a dataset prepared for supporting experiments in binary segmenting problems (separation of the occurrence of a GFE from the occurrence of other GFE or the neutral expression)[5].

We adapted the original dataset to support experiments related to multiclass classification problems. Originally, the phrases related to each GFE formed a subset of data used to train a model whose aim was to find a specific GFE in a phrase. In Figure 5, the upper two squares illustrate two disjoint subsets of data used to train two classification models with specific purposes: a model to segment the affirmative GFE (square to the left in the figure); a model to segment the negative GFE (square to the right in the figure). In this figure, each subset of data comprises one type of phrase. Phrases A and B are explicitly illustrated. The phrases are composed of frames regarding neutral expressions (class $-$) and by frames regarding a specific GFE (class $+$). The square at the bottom of the figure refers to the adapted dataset. Both subsets of data used to support two binary classification problems were placed together to form a bigger subset of data suitable to train a multiclass classification model. In this case, the problem comprises three classes: neutral expression (class 0), GFE affirmative (class 1) and GFE negative (class 2).

The Grammatical Facial Expression dataset comprises nine GFEs. However, some of the phrases chosen to represent conditional expressions, relative expres-

[3] Representations with frame of interest at the beginning of the window or at the end of the window are also possible.

[4] https://archive.ics.uci.edu/ml/datasets/Grammatical+Facial+Expressions

[5] The translation of all phrases from Libras to English is presented by Freitas et al. [30].

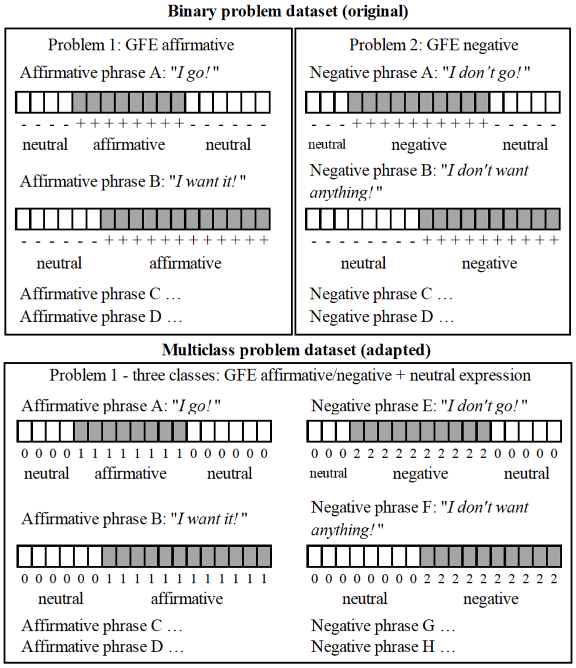


Figure 5. Subset of data used to build: two binary classification models (top) and a multiclass classification model (bottom)

sions, and topics also involved a second GFE in another segment of the utterance. A similar example to what occurs in the dataset is shown in the third phrase of Figure 2. The occurrence of the second GFE was labeled as a negative class. To also consider such phrases in the study presented herein, it would be necessary to re-label the entire dataset, since the labeling procedure is a subjective task and should be done following the same pattern for all frames (and phrases) [35, 36]. Because of this fact, such types of phrases have been excluded from the scope of this paper. Following what has been mentioned above, we have constructed 16 new subsets of data that involve the GFEs: wh-questions, y/n-questions, doubt questions, affirmative, negative and focus. Of these, 15 subsets comprise two-to-two combinations of six GFEs, and one subset comprises the combination of all GFEs.

5.2. Data preprocessing

Data from the Grammatical Facial Expressions dataset were made available in their raw state. Thus, we applied the following preprocessing procedures to adequate the data to the requirements of our study: normalization and displacement; attribute selection; feature extraction; vector representation and data windowing.

The normalization procedure adjusted the (x, y)-coordinates of each video frame within a unit hypercube. The displacement procedure centralized the coordinates in relation to the speechmaker's nose coordinates. We applied these procedures to scale down the effects of the variations (zoom in, zoom out and body displacements) of the speechmakers positions in relation to the sensor. The attribute selection concerns the choice of fiducial points to be used in the face's representation. We have used the same points used by Freitas et al. [30]. In this work, the authors analyzed the correlation between the 100 face fiducial points captured while the GFEs were performed. The results revealed the eight fiducial points needed to represent the variation in the facial movements that would be relevant in the GFEs discrimination: two points in each eyebrow and four points in the mouth (see the red points in Figure 4). We used the eight facial fiducial points and the Euclidean distances calculated between these points as descriptive features. Finally, we built two types of vector representation. The first type is composed of the (x, y)-coordinates of the eight facial fiducial points. The second type is composed of the distances calculated between such points. The temporal information was incorporated into the vector representation following the formalization presented in the Section 4.

5.3. Experiments setup

We carried out the experiments with two main purposes: (a) analyzing the suitability of the MLP for GFE automatic segmentation; and (b) studying the complexity of segmenting multiple GFEs.

We adopted the holdout method. In this method, the dataset must be split into two subsets, one for the model's training phase and one for the model's test phase. To build these subsets, the video frames referring to the five repetitions of three phrases related to the GFEs under analysis were assigned to the training set, and the video frames referring to the five repetitions of the two remaining phrases were used to compose the subset for testing. This division resulted in a relatively balanced training set compared to the test set. This is a good scenario for the experiments since the classification model has good learning conditions and more realistic test conditions. The ratio of the number of video frames in both subsets, to each GFE and each speechmaker, is shown in the Figure 6. The training and test subsets' information are presented separately. In this figure, the dark colors refer to the relative number of video frames pertaining to the GFE segments, and the light colors refer to the same idea for video frames with neutral expressions. In general terms, the speechmaker B utterances are longer than those produced by the speechmaker A. However, the former produced fewer video frames with neutral expressions (45% of frames from the speechmaker A are in GFEs segments; only 37% of frames from the speechmaker B compose the segments of GFEs). Except for the GFE wh-question case, the speechmaker B produced longer segments with GFEs, mainly when performing the GFEs for doubt-questions, negative phrases and to express focus.

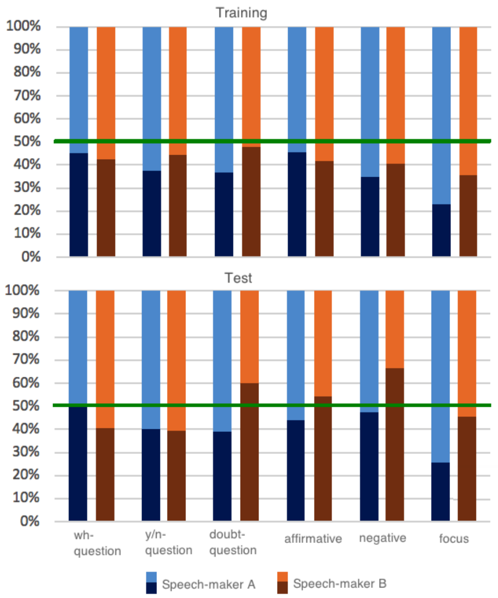


Figure 6. The ratio of the number of video frames in training and test subsets

The MLP architectures were built with one hidden layer, whose number of neurons was set to the mean between the number of neurons in the input layer and the number of neurons in the output layer [37]. The number of neurons in the input layer depends on the vector representation, the adoption of the windowing procedure and the size of the window when it is adopted. We adopted windows with three, five and nine frames. The number of neurons in the output layer was set equal to the number of classes: three in the *Experiment* #1 and seven in the *Experiment* #2. We trained and tested a set of MLPs using: 0.01, 0.1 and 0.5 for the learning rate, and 500 and $3,000$ for the number of epochs.

5.4. Results and discussion

Experiment $\#1$ addressed a multiclass classification problem related to two GFEs and neutral expressions. The classification models created under the conditions presented in the Section 5.3 produced similar results, except for those models trained with no-windowed data whose results were unsatisfactory. The results reported here were obtained considering: windows of size 3, rate of learning in 0.1, $3,000$ epochs of training, random initialization for the MLP synaptic weights and 10 runs for each combination of GFEs. Table 1 shows the results obtained with the two data representations, in the speechmaker-dependent context (training and test phases were carried out with data produced by the same speechmaker).

Table 1. Accuracy for multiclass classification, considering two GFEs and neutral expression and the speechmaker-dependent context (μ: mean; σ: standard deviation)

GFEs		coordinates		distances	
		μ	σ	μ	σ
	Speechmaker A				
affirmative	doubt-quest.	0.87	0.01	0.88	0.01
affirmative	focus	0.88	0.01	0,90	0.01
affirmative	negative	0.84	0.01	0.85	0.00
affirmative	wh-quest.	0.89	0.01	0.88	0.01
affirmative	y/n-quest.	0.89	0.01	0.88	0.01
doubt-quest.	focus	0.96	0.00	0.96	0.00
doubt-quest.	negative	0.93	0.00	0.92	0.01
doubt-quest.	wh-quest.	0.94	0.00	0.93	0.00
doubt-quest.	y/n-quest	0.94	0.00	0.94	0.00
focus	negative	0.95	0.00	0.95	0.01
focus	wh-quest.	0.95	0.00	0.96	0.00
focus	y/n-quest.	0.97	0.00	0.96	0.00
negative	wh-quest.	0.93	0.00	0.91	0.01
negative	y/n-quest.	0.93	0.00	0.91	0.01
wh-quest.	y/n-quest	0.95	0.01	0.95	0.01
	Global mean	0.92	0.04	0.92	0.04
	Speechmaker B				
affirmative	doubt-quest.	0.77	0.02	0.73	0.01
affirmative	focus	0.70	0.01	0.71	0.02
affirmative	negative	0.60	0.02	0.58	0.02
affirmative	wh-quest.	0.73	0.01	0.71	0.02
affirmative	y/n-quest.	0.81	0.01	0.78	0.01
doubt-quest.	focus	0.91	0.01	0.85	0.01
doubt-quest.	negative	0.80	0.01	0.74	0.02
doubt-quest.	wh-quest.	0.89	0.01	0.87	0.01
doubt-quest.	y/n-quest	0.81	0.01	0.88	0.01
focus	negative	0.72	0.03	0.63	0.01
focus	wh-quest.	0.87	0,00	0.87	0.00
focus	y/n-quest.	0.92	0.01	0.89	0.01
negative	wh-quest.	0.75	0.02	0.76	0.01
negative	y/n-quest.	0.82	0.03	0.82	0.02
wh-quest.	y/n-quest	0.88	0.02	0.90	0.01
	Global mean	0.80	0.09	0.78	0.01

The results listed in Table 1 show the difficulty faced by the multiclass classification models when analyzing the expressions performed by speechmaker B, as already observed in the tests with the binary classification problem presented by Freitas et al. [30]. According to these authors, the facial expressions performed by speechmaker A are more demarcated, i.e., their strengthened facial movements create a simpler decision surface for the classification models. Based on the experiments, it can be observed that the analysis of head movements presents greater complexity since the classifiers presented the lowest accuracy values when affirmative and negative GFEs were involved in the decision problem. This decrease in the classification model accuracy can be attributed to the smoothing of the movements produced by the offset preprocessing procedures. A method for nullifying shifts in body movements while preserving the intensity of head movements should be adopted to address this weakness in the proposed approach. Finally, the best results were obtained for problems involving the GFEs focus and y/n-question, in which there are upward eyebrow movements and head tilt down, and eye-opening in the case of focus. Furthermore, we observed that the GFE doubt-question, characterized by eyes and mouth contraction, also represents an easy problem for the classification model.

In tests concerning the speechmaker-independent context, the classification models are trained with phrases from one speechmaker and tested with phrases from the other. In these tests, the classification models lost performance. The best results are shown in Table 2. We reported only results with a mean accuracy greater than 0.6. The cases in which this lower limit was not reached were considered unsatisfactory. No test case reached this threshold for classification models trained with phrases from speechmaker A and tested with phrases from speechmaker B. As in the results previously discussed, in this case, the GFEs used in doubt-question, y/n-question and for expressing focus appear as representatives of an easier problem to solve.

Table 2. Accuracy for multiclass classification, considering two GFEs and neutral expression and the speechmaker-independent context (μ: mean; σ: standard deviation)

GFEs		coordinates		distances	
		μ	σ	μ	σ
speechmaker B (training) / speechmaker A (test)					
affirmative	doubt-quest.	0.66	0.11	0.65	0.04
affirmative	focus	0.65	0.00	0.65	0.01
doubt-quest.	focus	0.85	0.04	0.79	0.03
doubt-quest.	negative	0.66	0.06	0.70	0.02
doubt-quest.	wh-quest.	–	–	0.64	0.03
doubt-quest.	y/n-quest.	0.63	0.15	0.64	0.07
focus	wh-quest.	0.67	0.05	0.68	0.08
focus	y/n-quest.	0.70	0.05	0.72	0.07
wh-quest.	y/n-quest.	0.61	0.15	–	–

In summary, combinations of GFEs characterized by particular eye and mouth shapes and upward eyebrow movements represent easier pattern recognition problems. On the other hand, more difficult problems are those in which the fiducial points belonging to the eyes and mouth are neutral or those involving the recurrent movement of the head. The characterization of GFEs in terms of facial element shapes and head movements is presented in Table 3[6]. In this table, the first three lines are related to the most easily identified GFEs.

Table 3. GFEs characterization in terms of facial elements shapes and head movements

GFEs	eyebrows	head	eyes	mouth
doubt-question	$\downarrow$	$\ominus$	$*$	$*$
focus	$\uparrow$	$\downarrow$	$\diamond$	
y/n-question	$\uparrow$	$\downarrow$		
wh-question	$\downarrow$	$\uparrow$		
negative	$\downarrow$	$\leftrightarrow$		$\cap$
affirmative		$\updownarrow$		

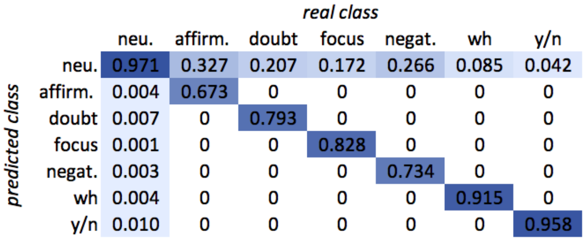
The ***experiment*** #2 addressed the six GFEs and the neutral expression. The classification models reached the mean accuracy of 0.91 and 0.75 for speechmakers A and B, respectively, in the speechmaker-dependent context. For both cases, the standard deviation was 0.01. These results show that the increase in the number of GFEs did not represent a significant increase in the problem's complexity seeing that, for the speechmaker A case, the obtained accuracy is close to the global mean accuracy obtained in *experiment* #1; although the accuracy for the speechmaker B has dropped compared to the respective global accuracy in *experiment* #1, the performance drop is small compared to the number of expressions that became part of the problem.

We can conduct a more detailed analysis of the complexity of segmenting the six GFEs by analyzing the classification errors. Confusion matrices support this type of analysis, and two examples obtained from *experiment* #2 are shown in Figures 7 and 8[7]. In such matrices, the errors are presented in relative values, i.e., for each GFE, the cells on the main diagonal refer to the percentage of correct classifications. In the first case (Figure 7), data from speechmaker A is under analysis. All error occurrences involve neutral expressions, and most of them are related to the indication of the neutral expression in a video frame in which a GFE occurs. For the second case (Figure 8), in which data from speechmaker B are considered, the phenomenon repeats for most errors. This phenomenon is partly related to the

[6] $\uparrow$ - upward movement; $\downarrow$ - downward movement; $\leftrightarrow$ - rightward and leftward movement; $\updownarrow$ - upward and downward movement; $*$ - compression; $\diamond$ - opening; $\ominus$ - withdrawal; $\cap$ - downward mouth corners.

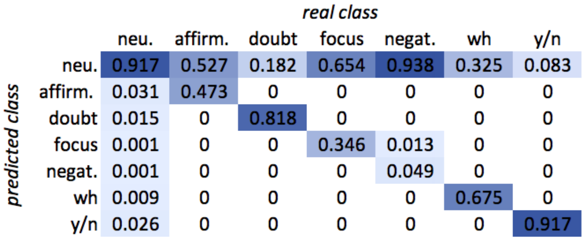
[7] The variation observed in the accuracy of the different runs performed for each set of classification models was small enough to support the generalization of the conclusions obtained from an example of a confusion matrix.

occurrence of errors in the transitions between the segments of GFEs and segments of neutral expressions. On average, 85% of the classification errors in the speechmaker A case occur in three frames that either precede or succeed a segment transition. For speechmaker B, on average, 50% of the errors occur under the same conditions. The exceptions occur in video frames containing the negative GFE, since part of them were classified as belonging to the class of GFE that express focus. The characterizations of such GFEs are different from each other (see Table 3), which raises the hypothesis that the problem is the representation of head movements. In addition, notice that the complexity associated with GFEs negative and affirmative is confirmed in this experiment. However, for GFEs that do not involve head recurrent movements, the results of the classification models varied similarly to the results observed in *experiment* $\#1$.



		real class						
		neu.	affirm.	doubt	focus	negat.	wh	y/n
predicted class	neu.	0.971	0.327	0.207	0.172	0.266	0.085	0.042
	affirm.	0.004	0.673	0	0	0	0	0
	doubt	0.007	0	0.793	0	0	0	0
	focus	0.001	0	0	0.828	0	0	0
	negat.	0.003	0	0	0	0.734	0	0
	wh	0.004	0	0	0	0	0.915	0
	y/n	0.010	0	0	0	0	0	0.958

Figure 7. Confusion matrix (with relative values): multiclass classification errors on data from the speechmaker A



		real class						
		neu.	affirm.	doubt	focus	negat.	wh	y/n
predicted class	neu.	0.917	0.527	0.182	0.654	0.938	0.325	0.083
	affirm.	0.031	0.473	0	0	0	0	0
	doubt	0.015	0	0.818	0	0	0	0
	focus	0.001	0	0	0.346	0.013	0	0
	negat.	0.001	0	0	0	0.049	0	0
	wh	0.009	0	0	0	0	0.675	0
	y/n	0.026	0	0	0	0	0	0.917

Figure 8. Confusion matrix (with relative values): multiclass classification errors on data from the speechmaker B

Although the results can still be improved, especially for the case of speechmaker B, they show the feasibility of solving the GFEs automatic segmentation problem in the speechmaker-dependent context. On the other hand, the generalization capability of speechmaker-independent models is unsatisfactory. The problem of generalization involving the particular articulation profile of different individuals has been studied in the area of gesture analysis also in other contexts [35, 38] and similar conclusions have been obtained.

6. Conclusion

In this paper, we presented an automatic solution for segmenting GFEs, modeled as a multiclass classification problem and treated by an MLP neural network. The experiments were performed for six types of GFEs considering the speechmaker-dependent and speechmaker-independent contexts. In the speechmaker-dependent context, our strategy presented promising results. Although, for speechmaker B, there may be sufficient misclassifications to suggest the impracticability of our approach, two aspects contradict this impression: (a) the numerous errors in transitions show that part of them refers to solutions that slightly shifted the GFEs segmentation and, up to a displacement limit, they do not represent errors for the segmentation task; and (b) by incorporating the segmentation responses to the grammar mentioned in the framework of Figure 1, part of them will be corrected since the grammar rules will prevent certain misclassifications from being accepted[8]. Due to this and the low computational cost involved in obtaining the solution, the proposed approach is feasible to be embedded in gadgets such as companion robots or smart toys. The weakness of our approach concerns the results in speechmaker-independent contexts. While it is possible to create custom solutions, this would not be the best option for large-scale applications. Ensembles of classifiers or mixtures of experts can be applied to overcome this difficulty without imposing high computational costs on the solution.

References

[1] C. Hitchcock and S. Stahl, "Assistive technology, universal design, universal design for learning: Improved learning opportunities," *J. of Special Education Technology*, vol. 18, no. 4, pp. 45–52, 2003.

[2] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *Int. J. of Machine Learning and Cybernetics*, vol. 10, pp. 131–153, Jan 2019.

[3] S. C. Ong and S. Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," *IEEE Trans. on Pattern Analysis & Machine Intelligence*, no. 6, pp. 873–891, 2005.

[4] D. H. Neiva and C. Zanchettin, "Gesture recognition: A review focusing on sign language in a mobile context," *Expert Systems with Applications*, vol. 103, pp. 159 – 183, 2018.

[5] C. Marshall, K. Mason, K. Rowley, R. Herman, J. Atkinson, B. Woll, and G. Morgan, "Sentence repetition in deaf children with specific language impairment in british sign language," *Language Learning and Development*, vol. 11, no. 3, pp. 237–251, 2015.

[8] Correction of similar classification errors were reported in the other modules of this framework [9].

[6] C. F. Benitez-Quiroz, K. Gökgöz, R. B. Wilbur, and A. M. Martinez, "Discriminant features and temporal structure of nonmanuals in american sign language," *PloS One*, vol. 9, no. 2, p. e86268, 2014.

[7] C. F. Benitez-Quiroz, R. B. Wilbur, and A. M. Martinez, "The not face: A grammaticalization of facial expressions of emotion," *Cognition*, vol. 150, pp. 77–84, 2016.

[8] H. Kacorri and M. Huenerfauth, "Continuous profile models in asl syntactic facial expression synthesis," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2084–2093, ACL, 2016.

[9] R. C. B. Madeo, S. M. Peres, C. A. M. Lima, and C. Boscarioli, "Hybrid architecture for gesture recognition: Integrating fuzzy-connectionist and heuristic classifiers using fuzzy syntactical strategy," in *Int. Joint Conf. on Neural Networks*, (Australia), pp. 1–8, June 2012.

[10] F. Freitas, S. Peres, C. A. Lima, and F. Barbosa, "Grammatical facial expressions recognition with machine learning," in *Proc. of 27th Florida Art. Intel. Research Society Conf.*, pp. 180–185, AAAI, 2014.

[11] W. Sandler, "Prosody and syntax in sign languages," *Trans. of the Philological Society*, vol. 108, no. 3, pp. 298–328, 2010.

[12] R. C. B. Madeo, S. M. Peres, H. H. Bíscaro, D. B. Dias, and C. Boscarioli, "A committee machine implementing the pattern recognition module for fingerspelling applications," in *Proc. of the ACM Symp. on Applied Computing*, (Switzerland), pp. 954–958, March 2010.

[13] C. R. de Souza and E. B. Pizzolato, "Sign language recognition with support vector machines and hidden conditional random fields: Going from fingerspelling to natural articulated words," in *Proc. of the 9th Int. Conf. on Machine Learning and Data Mining in Pattern Recognition*, (NY, USA), pp. 84–98, July 2013.

[14] P. Escudeiro, N. Escudeiro, M. Norberto, and J. Lopes, "Virtualsign game evaluation," in *Serious Games, Interaction and Simulation*, (Cham), pp. 117–124, Springer Int. Publishing, 2017.

[15] M. Zikky, R. Y. Hakkun, A. F. A. Rizqi, A. Hamid, and A. Basuki, "Development of educational game for recognizing indonesian sign language (SIBI) and breaking down the communication barrier with deaf people," in *Proc. of the 21st Int. Computer Science and Engineering Conf.*, pp. 83–88, 2018.

[16] A. W. Savich, M. Moussa, and S. Areibi, "The impact of arithmetic representation on implementing mlp-bp on fpgas: A study," *IEEE Trans. on Neural Networks*, vol. 18, no. 1, pp. 240–252, 2007.

[17] E. Ortigosa, A. Cañas, E. Ros, P. Ortigosa, S. Mota, and J. Díaz, "Hardware description of multi-layer perceptrons with different abstraction levels," *Microprocessors and Microsystems*, vol. 30, pp. 435–444, 11 2006.

[18] J. L. Espinosa-Aranda, N. Vállez, J. M. Rico-Saavedra, J. Parra-Patino, G. B. García, M. Sorci, D. Moloney, D. Pena, and O. Déniz-Suárez, "Smart doll: Emotion recognition using embedded deep learning," *Symmetry*, vol. 10, p. 387, 2018.

[19] R. M. de Quadros and L. B. Karnopp, *Língua de sinais brasileira: estudos linguísticos*. Artmed Editora, 2009. in Portuguese.

[20] L. Ferreira-Brito, "Uma abordagem fonológica dos sinais da lscb," *Informativo Técnico-Científico do INES, Rio de Janeiro*, vol. 1, no. 1, pp. 20–43, 1990. in Portuguese.

[21] O. Crasborn, "Nonmanual structures in sign language," *Encyclopedia of Language and Linguistics*, vol. 8, 2006.

[22] R. Pfau and J. Quer, *Nonmanuals: their grammatical and prosodic roles*, p. 381–402. Cambridge Language Surveys, Cambridge University Press, 2010.

[23] J. Han, M. Kamber, and J. Pei, *Data mining concepts and techniques*. Morgan Kaufmann Publishers, 3 ed., 2012.

[24] L. Fausett, *Fundamentals of neural networks: architectures, algorithms, and applications*. Prentice-Hall, 1994.

[25] S. S. Haykin, *Neural networks and learning machines*, vol. 3. Pearson Upper Saddle River, NJ, USA:, 2009.

[26] J. Liu, B. Liu, S. Zhang, F. Yang, P. Yang, D. N. Metaxas, and C. Neidle, "Recognizing eyebrow and periodic head gestures using CRFs for non-manual grammatical marker detection in ASL," in *10th Int. Conf. and Work. on Automatic Face and Gesture Recogn.*, pp. 1–6, IEEE, 2013.

[27] J. Liu, S. Y. F. Liu, B. Zhang, P. Yang, D. N. Metaxas, and C. Neidle, "Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions," *Image and Vision Computing*, vol. 32, no. 10, pp. 671–681, 2014.

[28] G. Caridakis, S. Asteriadis, and K. Karpouzis, "Nonmanual cues in automatic sign language recognition," *Personal and Ubiquitous Computing*, vol. 18, no. 1, pp. 37–46, 2014.

[29] P. Kumar, P. P. Roy, and D. P. Dogra, "Independent bayesian classifier combination based sign language recognition using facial expression," *Information Sciences*, vol. 428, pp. 30 – 48, 2018.

[30] F. A. Freitas, S. M. Peres, C. A. Lima, and F. V. Barbosa, "Grammatical facial expression recognition in sign language discourse: a study at the syntax level," *Information Systems Frontiers*, vol. 19, no. 6, pp. 1243–1259, 2017.

[31] M. T. Uddin, "An ada-random forests based grammatical facial expressions recognition approach," in *Int. Conf. on Informatics, Electronics & Vision)*, pp. 1–6, IEEE, 2015.

[32] M. Bhuvan, D. Rao, S. Jain, T. Ashwin, R. Guddetti, and S. Kulgod, "Detection and analysis model for grammatical facial expressions in sign language," in *Region 10 Symp. – "Technologies for Smart Cities"*, pp. 155–160, IEEE, 2016.

[33] Y. Hu and R. Treinen, "A one-step method for modelling longitudinal data with differential equations," *British J. of Mathematical and Statistical Psychology*, vol. 72, no. 1, pp. 38–60, 2018.

[34] M. Lichman, "UCI machine learning repository," 2013.

[35] R. C. B. Madeo, S. M. Peres, and C. A. M. Lima, "Gesture phase segmentation using Support Vector Machines," *Expert Systems with Applications*, vol. 56, pp. 100 – 115, 2016.

[36] N. T. Roman, P. Piwek, A. M. B. R. Carvalho, and A. R. Alvares, "Sentiment and behaviour annotation in a corpus of dialogue summaries," *J. of Universal Computer Science (Print)*, vol. 21, pp. 561–586, 2015.

[37] M. C. F. De Castro, F. C. C. De Castro, J. N. Amaral, and P. R. G. Franco, "A complex valued hebbian learning algorithm," in *IEEE Int. Joint Conf. on Neural Networks*, vol. 2, pp. 1235–1238, 1998.

[38] A. S. Ramakrishnan, "Segmentation of hand gestures using motion capture data," Master's thesis, University of California, 2011.