Recognition of affective and grammatical facial expressions: a study for Brazilian sign language

Emely Pujólli da Silva[1], Paula Dornhofer Paro Costa[1], Kate Mamhy Oliveira Kumada[2], José Mario De Martino[1], and Gabriela Araújo Florentino[3]

[1] University of Campinas, Campinas, Brazil
{paulad,martino}@unicamp.br
[2] Federal University of ABC, Santo André, Brazil
[3] Seli Institute, São Paulo, Brazil

Abstract. Individuals with hearing impairment typically face difficulties in communicating with hearing individuals and during the acquisition of reading and writing skills. Widely adopted by the deaf, Sign Language (SL) has a grammatical structure where facial expressions assume grammatical and affective functions, differentiate lexical items, participate in syntactic construction, and contribute to intensification processes. Automatic Sign Language Recognition (ASLR) technology supports the communication between deaf and hearing individuals, translating sign language gestures into written or spoken sentences of a target language. The recognition of facial expressions can improve ASLR accuracy rates. There are cases where the absence of a facial expression can create wrong translations, making them necessary for the understanding of sign language. This paper presents an approach to facial recognition for sign language. Brazilian Sign Language (Libras) is used as a case study. In our approach, we code Libras' facial expression using the Facial Action Coding System (FACS). In the paper, we evaluate two convolutional neural networks, a standard CNN and hybrid CNN+LSTM, for AU recognition. We evaluate the models on a challenging real-world video database of facial expressions in Libras. The results obtained were 0.87 f1-score average and indicated the potential of the system to recognize Libras' facial expressions.

Keywords: Facial action unit recognition, Grammatical facial expression, Affective facial expression, Sign language

1 Introduction

Sign Languages (SLs) are visuospatial linguistic systems structured on gestures that are adopted around the world by deaf people to communicate. Analogously to spoken languages, SLs emerged spontaneously, evolved naturally, reflecting the worldwide sociocultural differences and giving origin to a wide range of variations such as the British Sign Language (BSL), the American Sign Language (ASL), the Chinese Sign Language (CSL), the Brazilian Sign Language (Libras), among others.

Being a minority language in most territories, deaf individuals frequently need a sign language interpreter in their access to school and public services. In such scenarios, the absence of interpreters typically results in discouraging experiences, even in educational deficits, preventing the inclusion of deaf individuals in society. In other situations, such as health care, the need for a sign language interpreter can be embarrassing or a risk factor for urgent care.

Aiming to overcome existing obstacles in the communication between hearing and deaf people, in the last decade, many efforts have been dedicated to the development of Automatic Sign Language Recognition (ASLR) technology [41, 62, 44, 61]. ASLR systems recognize and translate sign language content in video into text. Optionally, the text output can be the input of a Text-To-Speech (TTS) synthesizer, resulting in a translation from source sign language to target spoken language.

Together with hand gestures and other non-manual markers, a key challenge in the development of ASLR technologies is the modeling and classification of facial expressions.

In SLs, more than communicating affective states, facial expressions represent morphemes, that fulfill syntactic and pragmatic functions. For this reason, the problem of recognizing facial expressions in SL context, can be considered more complex than the typical affective computing problem. In fact, facial expressions of emotions represent only a subset of common existing facial expressions in sign languages (Figure 1). Also, more recently, researchers have argued that the recognition of facial expressions can improve ASLR accuracy rates [55, 1].

The present work adopts a deep neural network architecture to recognize SL facial expressions coded as Action Units (AUs) of the Facial Action Coding System (FACS) [9]. Although some works implement AU recognition systems, their application unrelated to emotions is scarce [36]. Taking Brazilian Sign Language (Libras) as our case study, we coded Libras' facial expressions using FACS and we adapted existing network architectures to be more generic in the recognition of action units other than emotional AUs. While state-of-the-art AU classification works only handled eight to twenty AU labels [49, 27, 7, 69], our experiment considers 80 categories, derived from a comprehensive facial expression survey

Fig. 1. Examples of facial expressions in sign language that are not associated to the expression of emotions.

in Libras. Our resulting classification accuracy is competitive with the literature results and with improved generalization ability.

The paper is organized as follows. In Section 2, we discuss the state-of-the-art of Facial Expression Recognition (FER), and systems that are based on AU recognition. In Section 3, we summarize the role of facial expressions in Libras. Also, their FACS association is introduced. The methodology of our approach is presented in Section 4. The experimental results are shown in Section 5. We end by considering the implications of our findings in Section 6.

2 Related Work

Facial Expression Recognition (FER) has been studied for decades and many approaches have been proposed [10, 42]. There are two main approaches to FER. One considers the AUs as the features to be recognized in the face [3]. The second regards a set of prototypical facial expression of emotion defined by Paul Ekman (1993)[8], as the characteristics to be identified.

Due to the availability of facial expression information and data type from the affective perspective, it is more frequently encountered the second FER strand where only emotional labels are considered [35, 59, 40]. A large number of surveys in Emotional Facial Expression Recognition (EFER) have been published over the years [48, 25] and lately, well-design network architectures have achieved better accuracy and exceeded previous results [32, 64]. Such architectures are composed with decision tree, naive Bayes, multilayer neural networks and K-nearest neighbours, hidden Markov model (HMM), shallow networks, and deep neural networks [63, 64, 32].

Although many works adopt FACS as the visual appearance building blocks of emotion, their study unrelated to emotions is scarce [31]. Particularly in SL, there are not enough researches that associated non-manuals markers with FACS. Most of the works that carry a non-manual marker recognition scheme treats the facial expression by creating their classes of facial actions [60]. A comprehensive survey of AU analysis can be found in Martinez *et. al.* (2017) [36]. AU recognition can also be applied in intelligent vehicle systems that detect and recognize the facial motion and appearance changes occurring during drowsiness [56]. Moreover, the sensitivity of FACS to subtle expression differences shows its capability and application in the medical field as descriptive of characteristics of painful expressions [30], depression, or to examine evoked and posed facial expressions in schizophrenia patients [16].

Recently, the use of CNN-based representations has been adopted to model facial actions [6]. Walecki *et. al.* (2017) [57] proposed a convolutional neural network (CNN) model jointly with a Conditional Random Field structure to ease the inference. In Tran *et. al.* (2017)[29], the second layer is conditioned to the latent representation from the first layer, in other words, a two-layer latent space is learned. Although these models are accurate, they have slow performance. Also, like most works, these models generally need a feature extraction step [47]. In Li *et. al.* (2017)[26] the Visual Geometry Group (VGG) network is first used

with region learning and a fully connected long short-term memory (LSTM) network to obtained features in the task of AU detection, resulting in a complex system. Other recent works have attempted a hybrid approach, by combining AU relations through either generative approaches or discriminative approaches [18]. However, like most works they still try to learn a global representation from the input image [47]. In Chu *et. al.* (2019)[7], the presented hybrid network takes advantage of spatial CNNs, temporal LSTMs, and their fusions to achieve multi-label AU detection. The performance over 12 AUs from the BP4D dataset average F1 score of 82.5 with a 10-fold protocol.

The lack of sufficient details (e.g., parameter optimization strategy, preprocessing procedure, training, and testing protocols) makes it difficult to compare some different AU recognition methodologies, even on the same dataset. With the purpose of standardization for a fair comparison, challenges arose. Held for this purpose, FERA (Facial Expression Recognition and Analysis challenge) and EmotioNet challenges evaluated AUs recognition and discrete emotion recognition. The last EmotioNet 2017 challenge involved AUs occurrence detection for 11 AUs. The top algorithm used residual blocks and a sum of binary cross-entropy loss (PingAn-GammaLab) and achieved 94.46% accuracy [69]. The baseline results for the challenge was 80.7% accuracy using Kernel Subclass Discriminant Analysis (KSDA) [3].

The main novelty of our proposed approach resides in the joint region detection of AU and analysis of AU presence. In our framework, we perform a feature extraction stage and a task learning problem. The first architecture builds on a standard CNN. The second one was built upon an ensemble of CNN and LSTM. Differentiated from works where the whole face is used for AU intensity estimation and localization, our approach treats the problem by delimiting face regions in the presence or absence of AU to overcome the greater number of AU available on our sign language application.

3 Facial Expressions in Libras

Most SLs in the world can be considered understudied, meaning that aspects of their grammar and morphology are still undocumented or unknown. Also, compared to spoken languages, there is a lack of annotated SL corpora, a key input for supervised machine learning modeling.

In this scenario, a first contribution of the present work was, with the help of sign language linguists specialists, to conduct a detailed survey of existing facial expressions in Libras, and to code them using FACS.

In Libras, facial expressions that convey an idea of feeling and emotion are called Affective Facial Expressions (AFE). Affective facial expressions may start before a specific sign and end after the sentence has been completed. In other words, AFEs modulate the whole sentence, modifying the full meaning of a sequence of signs. AFEs are adopted, for example, when the signer communicates ideas sarcastically or when he/she is describing a sad event. A visual characteristic of AFEs is that they employ an integrated set of facial muscles.

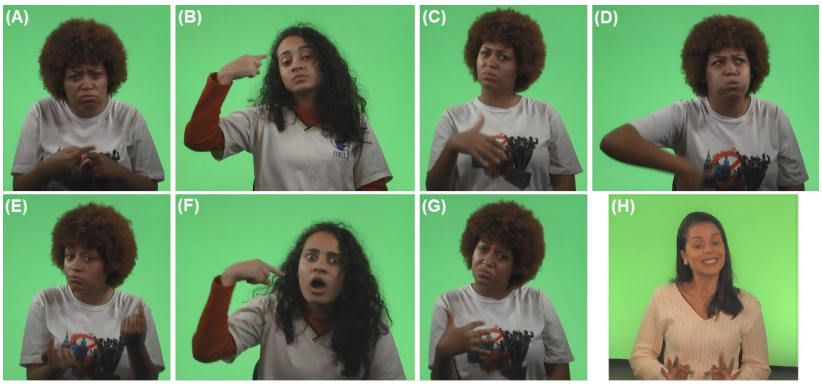

Fig. 2. In the performance of the signs in Libras, we can analyze the variation of the facial expressions by the images (A) and (E), where we have examples of grammatical facial expressions of sentence - GES. In the image (A), the interpreter is signing "why?" and in (E) she is signing "how?". In images (B) and (F), the signs "lawyer" and "crazy" are performed with the same manual gesture. Their difference is only based on the eyes, eyebrows action, and on the mouth open, which is an example of the grammatical expression of distinction - GED. Also, in images (C), (D), and (G), the interpreter performed the sign "expensive", "very expensive", and "little expensive", respectively. The intensity of the sign is displayed by the change in the facial expression, which passed from neutral (C) to frown and inflated cheeks (D), or to frown and crocked mouth down (G). Those are examples of grammatical facial expressions of intensity - GEI. In the last picture (H), the interpreter shows the sign "happy", which is accompanied by a characteristic affective facial expression - AFE.

Grammatical Facial Expressions (GFE) in Libras are expressions that typically occur at specific points of a sentence or are associated to a specific sign execution [14]. Observing the different properties of grammatical facial expressions we can categorize them into Grammatical Facial Expression for Sentence (GES), Grammatical Facial Expressions of Intensity (GEI), Grammatical Facial Expressions of Homonymy (GEH) and Grammatical Facial Expressions of Norm (GEN).

GES defines the type of sentence that is being signed [51]. Accordingly with the structure and information of the sentence, it can be classified into: WH-question (WH), Yes/No question (YN), Doubt question (DQ), Topic (T), Negative (N), Affirmative (A), Conditional clause (CC), Focus (F) and Relative clause (RC). In Libras, there are GES markers that are expressed by the face and head movements.

GEI differentiates the meaning of the sign assuming the role of a quantifier. For example, the same sign associated with the word "expensive" can have its meaning attenuated to "little expensive" or "very expensive", depending on the signer's facial expression. In Fig. 2 (C) , (D) and (G), we show frames of those signs.

Also, without its characteristic GEH a sign is incomplete and cannot be distinguished from other signs with the same manual signal. In other words, GEHs helps to define the meaning of a sign. For example, in the Fig. 2 (B) and (F), we have the representation of two signs with different meanings, in which the manual sign is the same, but the facial expression is different.

The facial expressions that are part of the signal by norm and whose function is to complete a manual signal, we define as the GEN. When a GEN sign is performed without the facial expression that defines it, the signal loses its meaning.

It is possible to notice that many of the non-manual articulators found in Libras are also used in other sign languages [51, 11]. For example, Yes/No questions in American SL (ASL) are associated with raised eyebrows, head tilted forward and widely-opened eyes, and WH-questions with furrowed eyebrows and head forward. Topics are described by raised eyebrows and head slightly back, and negations are expressed with a head shake[1, 28]. In the German SL (DGS), a change of head pose combined with the lifting of the eyebrows, corresponds to a subjunctive. Lip pattern, tongue, and cheeks that are not related to the articulation of words can provide information redundant to gesturing to support differentiation of similar signs [55]. Thus, these facial expressions function intersections reinforce the ability to generalize an application in Libras to other sign languages.

3.1 Coding Libras' facial expressions using FACS

The lack of pattern in the description of facial expressions in Libras and in SLs in general, becomes a major problem to implement a computational recognition problem. Our approach consisted of coding the Libras' facial expressions using the facial action coding system [9]. FACS describe face muscle variations through 52 action units (AUs), that can occur alone or in combination. Table 1 presents this novel coding of Libras' Facial Expressions using FACS.

Due to our limited access to examples of all Libras facial expression, we focus our model on the listed GES class. Likewise, for the AFE class, we adopt the prototypical seven basic emotions: happiness, sadness, surprise, fear, anger, disgust, contempt, as reported in the literature. This assumption is possible since studies show that basic emotions are used and recognized by the Deaf to convey affective states [43, 19, 23].

Note in Table 1, that the number of AUs that participate in the performance of the basic emotions (16) is at least two times lower than the number of AUs that are found in Libras facial expressions (39). Libras' facial expressions contain and transcend the regular set of AUs attributed to basic emotions.

4 Methodology

Our methodology consisted of, first, building a database of videos of deaf individuals and Libras interpreters. The facial expressions present in the videos were annotated using FACS (Section 4.1).

Table 1. Classification of Libras Facial Expressions associated with the Facial Action Coding System

	Facial Expression in Libras	FACS	Muscular / Action Description
Basic Expressions	*Upper face*		
	Joined eyebrows	AU1	Inner Brow Raiser
	Raised eyebrows	AU1+AU2	Inner Brow Raiser and Outer Brow Raiser
	Frown	AU4	Brow Lowerer
	Wide open eyes	AU5	Upper Lid Raiser
	Nose wrinkle	AU9	Nose Wrinkler
	Slightly closed eyes	AU41	Lid droop
		AU42	Slit
	Closed eyes	AU45	Eyes Closed
	Left / Right eye closed	AU46	Wink
	Direct the eyes	AU61	eyes to the left
		AU62	eyes to the right
		AU63	eyes up
		AU64	eyes down
	Lower face		
	Crooked mouth up	AU12	Lip corner puller
	Crooked mouth down	AU15+AU17	Lip corner depressor and Chin raiser
	Projected lips	AU18+AU23	Lip Puckerer and Lip Tightener
	Tongue in lisp position	AU19	Tongue show
	Swinging alveolar tongue		
	Sibilant tongue		
	Tip of the tongue touching the lips		
	Contracted lips	AU28	Lip suck
	Open mouth	AU25	Lips apart
		AU26	Jaw Drop
	Inflated cheeks	AU33	Cheek blow
	semi-open mouth (blowing)	AU34	Cheek puff
	Contracted cheeks	AU35	Cheek suck
Compound Expressions	**Grammatical Facial Expressions of Sentence**		
	WH-Question	AU4+AU53	Brief and upward movement of the head and frown
		AU4+AU18+AU23+AU53	Tilt back, frown and projected lips.
	Y/N Question	AU1+AU2+AU53	Brief and upward movement of the head and raised eyebrows
		AU4+AU18+AU23	Tilt to the side, frown and projected lips
		AU4+AU18+AU51+AU52	Balancing sideways, frown and projected lips
	Negative	AU15+AU17	Crooked mouth down
		AU4+AU15+AU17+AU54	Quick nod, frown and crooked mouth down
		AU51+AU52	Head balancing sideways
	Affirmative	AU53+AU54	Balance back and forth of the head
	Affective Facial Expressions		
	Basic Emotions	AU6+AU12	Happiness
		AU1+AU4+AU15	Sadness
		AU4+AU5+AU7+AU23	Anger
		AU1+AU2+AU5B+AU26	Surprise
		AU1+AU2+AU5+AU20+AU26	Fear
		AU1+AU4+AU5+AU7	
		AU9+AU15+AU16	Disgust
		R12A+R14A	Contempt

Second, we implemented a feature extraction process, responsible for creating the inputs to the evaluated networks (Section 4.2).

Finally, we proposed and evaluated two different deep neural network architectures (Sections 4.3 and 4.4) .

4.1 HM-LIBRAS Database

Our first prototype, the Head Movement in Libras (HM-Libras) database was built using parts of videos from the Internet of deaf individuals and sign language interpreters [52]. We downloaded videos distributed under the Creative Commons license, using different combinations of search keywords: Libras, questions, grammar, answer. Specifically, we target phrases with grammatical facial expressions for sentence. They were not chosen at random, but with the ad-

vice of a Libras expert, in such a way that these sentences represent a range of communication elements of the language. The HM-Libras database is composed of 80 FACS labeled videos being: 20 videos with statements, 20 videos with WH-questions, 20 videos with Yes/No questions and 20 videos with negation sentences.

We collected videos where the person starts facing straight to the camera to facilitate face detection. These videos are not always professionally curated and often suffer from perceptual artifacts, varying in illumination, and background. The set of videos has the presence of three women and seven men. In addition, HM-Libras includes a dataset matrix composed of facial points detected using Dlib [21]. The dataset is made available to all interested researchers upon request to the authors.

In summary, HM-Libras was created with the concept of studying head movement in Libras that occurs in the performance of certain types of sentences, where each frame was annotated by a single FACS coder.

4.2 Feature Extraction

We extracted 68 landmarks localized on the face placed alongside the ears, chin, eyes, nose and mouth (Fig. 3A) using DLib [21]. We resized the face images to 96×96 after cropping the face areas. In sign language, one has also to take into consideration the possible partial or total occlusion of the face as a result of the position and the movement of the hands. When that occurs we decide to remove the frames, to keep some continuity. The lost of face tracking happened in 3% of the entire database.

Following these steps, we choose to segment the face into two sets, a lower portion of the face comprehending the chin, mouth, and nose, as well as the upper portion of the face comprehending the forehead, eyebrows, and eyes.

This region related approach is adopted to increase the system activity sensitivity.

Given that the displacement of landmarks points is a measurable way to describe facial expression, we argue that the use of geometrical features could improve performance on models designed to learn AU classification. We add the geometric characteristics using the landmarks positions and by calculating some distances. As one can notice, AUs are measured by a change in face configuration. Thereby calculating the distance between the middle point in the lid tightener in the eyes can indicate if the eyes are open or closed (Fig. 3B). Likewise, for the mouth, we calculate the distance between the midpoints of the upper and lower lip. Each of these measurements was converted into a single gray pixel. In other words, we compose vectors with the face points p_i, $i = 1, \ldots, 68$ and the distance measures $d_2(p_j, p_k)$ with $(j, k) \in \{(3, 13), (17, 21), (21, 22), (22, 26), (38, 40),$ $(43, 47), (48, 54), (51, 57), (62, 66)\}$, later these values are scaled to the range $0-1$ and then encoded as gray levels. Finally, they were concatenated in the images respectively to the region of the face that belongs (see Fig. 4A).

4.3 CNN Based FACS classification

Based on Keras implementation [4], a CNN model was built following existing approaches [40, 45]. In Pramerdorfer *et. al.* (2016)[45] the architectures of a shallow CNN outperform modern deep CNNs. We use the information that a CNN with five hidden layers is already able to learn high-level discriminatory features to design our network. The model consists of a CNN where the image is passed through a stack of three convolutional layers. We use filters with a small kernel field: 2×2 for all convolutional layers, which can be seen as a linear transformation of the input channels followed by non-linearity. The convolution stride is fixed to one; the spatial padding of convolutional layer input is such that the spatial resolution is preserved after convolution, i.e., the filling is one pixel for 2×2 convolutional layers. Spatial pooling is carried out by two Max-Pooling layers, which follow the two first convolutional layers. Max-pooling is performed over a 3×3 window, with stride two. The last of convolutional layers is followed by three fully-connected layers: the first has 4096 channels, the second has 1024 channels and, the third can perform a 30-way AU classification or 50-way AU classification depending on the architecture. In the final layer, we use the softmax layer and thus contains 30 or 50 labels, one for each class whether it is for the upper part of the face or for the lower part of the face, respectively. The activation functions are all set to ReLu (Rectified Linear Functions). The configuration model and other details are shown in Fig. 4B.

4.4 CNN+LSTM for AUs classification

Since AUs are an observable event throughout time, learning the recognition of facial expressions can be improved by the knowledge of previous states. Naturally, we extended our system to address temporal context by designing a combination of both CNN and LSTM to fuse static features with temporal cues, inspired by [26, 39, 7].

More specifically, we propose a standard CNN with three convolutional layers alternated by two max-pooling layers. The convolutional layers are composed with a kernel of size three and stride one. The first two convolutional layers have 32 filters and the last one has 64 filters. The max-pooling layers have a

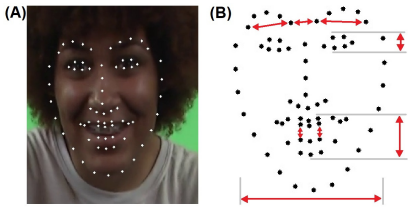

Fig. 3. A prototypical face mask is presented in (A) with white points $p_i = (x_i, y_i)$, $i = 1, \ldots, 68$. In (B) we mark with red line the chosen distances d_2 for measuring configuration of eyebrows, eyes and mouth.

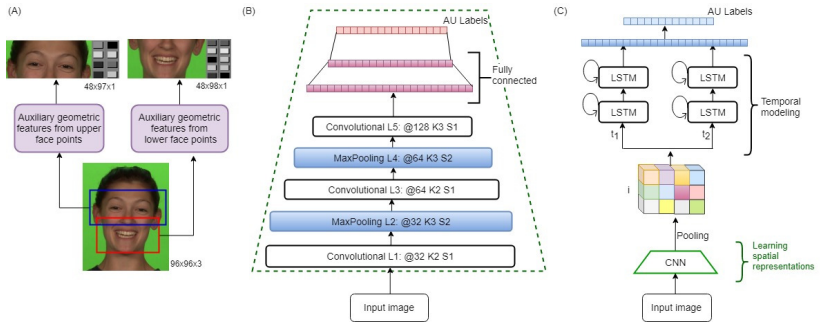

Fig. 4. Input image, CNN architecture and CNN+LSTM architecture. In the image (A), we present the input for our networks. The face image is cropped and later combined with extracted points from the face and their distances. In image (B), we present the architecture of our CNN, consisting of one input layer, three convolution layers, two max pooling layers, and two full connection layer. In image (C), we present the structure of the proposed hybrid network, where the input layer is feed into a convolutional neural network which is followed by a pooling layer that connects to an LSTM. Detailed descriptions are given in the text.

stride of size two. All activation functions are set as ReLU. The last layer is the region pooling layer. We model the correlations between spatial and temporal cues by adding a fusion layer. This fusion layer is a concatenation of feature maps, made to get regional features. So, from the CNN we have a pooling layer with 50 filters for the upper part of the face and 30 filters for the lower part of the face. These feature maps, are fed into stacks of LSTMs to fuse temporal dependency. We combine two frames of images as sequence into the LSTM. Then several stacks of LSTMs are used to capture facial actions temporal dependence. Finally, the outputs of LSTMs are aggregated into a dense layer to perform multi-label learning. In Fig.4C we present a scheme of CNN+LSTM network architecture.

5 Experiments

We evaluated the proposed architectures, performing experiments with the databases: Extended Cohn-Kanade dataset (CK+), DISFA (Denver Intensity of Spontaneous Facial Expressions), and the HM-Libras database.

CK+ Dataset [33] has the first release called CK which includes 486 sequences from 97 subjects posing the six basic emotions [20]. Each sequence starts with neutral and ends in apex of emotion and is AU coded. The second release is called CK+ and includes both posed and non-posed expressions [33]. Validated emotion labels have also been added to the metadata. In addition, CK+ provides baseline results for facial tracking, AU and emotion recognition. Is important to remark that the AU annotations were given at video and not frame wise.

DISFA Dataset [38] is a spontaneous database composed by videos of 27 subjects that vary in age from 18 to 50 years. The subjects are filmed while reacting to an emotional four-minute video stimuli. Also, it comprehends the manually labeled frame-based annotations of 5-level intensity of twelve FACS, labeled by two FACS coders. The lack of available data for comparing posed and spontaneous expression encouraged the same research group, to construct the Extended DISFA Dataset (DISFA+) [37], which contains the videos and AU annotations of posed and spontaneous facial expressions of 9 participants in the same format as DISFA.

Note that, DISFA and CK+ are standard datasets for AU detection where the AU correspond with the Libras' affective facial expressions class. Also, to the best of our knowledge, HM-Libras is the first Libras database with AU annotations. The combination of such datasets helped in diversifying training samples necessary in the sign language application, despite containing different sets of AUs.

For our first experiment, we choose to separate a percentage of HM-Libras database for testing, and the rest we combine with CK+ and DISFA databases to form the training set. Our train set was composed by 69624 frames and the test set was composed by 7736 frames. To demonstrate the effectiveness of the proposed model to the Libras application, extensive experiments have been conducted on our already described networks using a subject independent and cross databases approach.

Metrics. The performance of AU detection was evaluated on F1 frame-basic metric. F1 score is the harmonic mean of precision and recall, and it is widely used in AU detection [22, 66, 6]. Also, to further explore more details of our model, we also computed the average accuracy, precision, and recall.

Comparative methods. For a thorough comparison, we selected two popular deep network architectures that were designed and trained on ImageNet, and have been successfully applied to multiple vision problems: AlexNet [24] and VGG-16 [54]. To adapt both networks for our classification model, we modify the input and the output layer. The input layer was adjusted to accommodate images with a size of $60 \times 96 \times 3$. Also, the output layer was arranged to exit 30 labels for the upper part of the face and 50 labels for the lower part of the face. As a baseline, these networks were trained only with the same images, without geometric face information.

Unlike the common practice in AU literature [46, 68, 67, 17, 27, 5], where only 12 AUs are considered for a single dataset, our research encompass 39 AUs that are descriptive in Libras. The increased number of AUs makes difficult to fairly compare our approach with the state-of-art.

Implementation details. We train every architecture for up to 300 epochs and a fixed mini-batch size of 500 samples. Both models were initialized with a learning rate of 0.01, optimizing the cross-entropy loss using stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay of 0.001. Simard *et. al.* (2003)[53] have shown that if the data is augmented in a reasonable way, the model can perform better. For training data augmentation we use horizontal

mirroring, randomly rotations and two types of shift and zoom transformations. These are applied indiscriminately in each epoch creating twice the amount of data.

All experiments are performed on a PC with one NVIDIA GTX 1070 GPU. It took roughly 160 hours to train each network until convergence that happened around 250 epochs.

Results and Discussion.

Comparison between existent architectures. Table 2 shows the F1 metrics reported on AlexNet, VGG-16, CNN, and CNN+LSTM. Also, "Avg." for the mean score of both face parts. According to the results, both of our networks CNN and CNN+LSTM outperformed AlexNet and VGG-16 when trained with a cross-dataset and subject independence. In addition to the improvement by considering a shallow system, the performance gain of CNN can also be assigned to the usage of geometric features. These observations provide an evidence that the learned representation was transferable even when being tested across subjects and datasets.

CNN+LSTM performs a spatiotemporal fusion which consistently outperformed AlexNet and VGG-16 in all metrics. Our hybrid network uses small time steps window as we want to avoid the suppression of properly detected but short temporal series of AU activation, yet, if the temporal length of AU duration is short, then the CNN+LSTM model could not observe such actions [34]. In general, adding temporal information helped predicting AUs, but a more extensive study in the time steps sizes could be beneficial.

It can be seen that our CNN+LSTM does not bring a lot of gain over our CNN. Surprisingly, for the upper part of the face, we obtain 0.9018 of accuracy while in the lower part of the face 0.85 of accuracy in the CNN. The average accuracy for our AU classification using the CNN is 0.88. The value is comparable to other published results [58, 6, 15, 65].

Another way to compare our models' effectiveness is to use an off the shelf AU regressor. OpenFace [2] is an analysis platform capable of face detection and recognize a subset of AUs, specifically: 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28, and 45. Setting an experiment where the 80 categories are produced in post prediction and on our test set, the report results are 0.208 avg accuracy and 0.2115 for f1-score outcomes considerably lower than the results obtained by our models.

Moreover, models for AU classification handled typically only eight to twenty AU labels, while our experiment considers 80 categories (including compound expressions), which is much more challenging and realistic compared to many existing methods. Generally, our method explicitly inherits the advantage of information gathered from multiple local regions from complex AU acting as a deep feature ensemble in both architectures, and hence it naturally improves the recognition of basic AUs.

Comparison with manual transcription We compare our CNN AU detection framework with the human transcription of Libras' grammatical syntactic functions and affective facial expressions in Table 3. The human ratings are given

by the labeling agreement scores that were obtained by comparing the annotation between two coders. At the same time, the automatic AU detection ratings are collected by comparing the transcription between the CNN framework output and the two coders. The agreement measure chosen was Fleiss' kappa [12], which ranges from -1 to 1. The negative values indicate randomness in labeling or poor agreement; while the values in $0 - 0.2$ indicate slight agreement; between $0.21 - 0.4$, fair agreement; between $0.41 - 0.6$, moderate agreement; between $0.61 - 0.8$, substantial agreement; and, lastly, between $0.81 - 1$, perfect agreement. Kappas coefficients were calculated for five videos from the Sign Language Facial Action (SILFA) corpus [50]. SILFA contains videos of deaf interpreters signing in Libras and is transcribed with facial expressions classes and syntactic functions as defined in Table 1 where the labeling were made by two linguistics Libras experts.

In our approach, we consider the facial occlusion in sign language by annotation, i.e., by the usage of FACS visibility codes. Thereby, when the algorithm cannot detect the face, the output is the visibility code AU74, which means unscorable. However, if the occlusion is due to the hand being in front of the face, and consequentially, occluding the facial expression, the framework output is AU73, which translates to the entire face not visible. The fourth and fifth columns of Table 3 presents the occlusion agreement rates as almost perfect confidence for the human annotation and slight/fair confidence for our framework.

Our CNN AU detection framework using geometric and region of interest features outputs obtained fair/moderate confidence when compared to humans. Moreover, when averaging our prediction with the manual annotations, the performance can be further improved. This implies that learning Libras' facial expressions as a function of basic and complex AUs may be a more accurate and systematic way than learning facial expressions from the whole face. We also compare our model using CNN AU detection with [13] that has a model for recognition of grammatical facial expressions of sentence in Libras with shallow structure but uses only landmarks to predicted automatically from the output of a Multi-layer Perceptron and achieved F-scores over 80% for most of their experiments. Our results outperform [13], demonstrating the potential of our ensemble detection model if the AU prediction stage is improved.

Table 2. Performance comparison of proposed methods with state-of-the-art networks

Architecture	**Description**	**Accuracy**	**F1**	**Precision**	**Recall**
AlexNet	upper face	0.7322	0.7219	0.8565	0.6295
	lower face	0.6723	0.6639	0.8027	0.5719
	Avg.	0.7022	0.6929	0.8296	0.6007
VGG-16	upper face	0.6199	0.6199	0.6199	0.6199
	lower face	0.4800	0.4800	0.4800	0.4800
	Avg.	0.5499	0.5499	0.5499	0.5499
CNN	upper face	0.9018	0.8900	0.8972	0.8194
	lower face	0.8585	0.8522	0.8892	0.8091
	Avg.	**0.8805**	**0.8711**	**0.8932**	**0.8142**
CNN+LSTM	upper face	0.8828	0.8182	0.8714	0.7685
	lower face	0.8541	0.7047	0.8174	0.6697
	Avg.	0.8684	0.7614	0.8444	0.7191

6 Conclusions

The first contribution present in this paper is a novel model for recognition of grammatical and affective facial expressions in Brazilian Sign Language. Based on the literature, we construct a CNN and a hybrid CNN LSTM, which consisted of a feature extraction process where we segmented the face into the upper and lower part, creating two resembling networks that were trained in multiple databases. When compared with facial expression recognition works, we found similar results, although our model has capabilities of classification on more AUs labels. Secondly, was the construction of a database with Libras signers' fully annotated with FACS, the HM-Libras database. Also, a detailed survey of existing facial expressions and their syntactic functions in Libras were compiled. To facilitate and support further studies, we establish an association between the Facial Action Coding System and the listed Libras facial expressions. The action unit codification made it possible to observe that the number of facial expressions portrayed in Libras is superior to the prototypical emotion expressions evaluated in the literature.

When discussing our networks' accuracy performance, the interdisciplinary nature, and the amplitude of our study regarding AU recognition should be taken into account. Our built CNN presents an average accuracy of 0.88 by performing facial action unit classification in face images in terms of 80 AU codes, suggesting that our model gives some insight into AUs who are not usually included in other studies.

Given this comprehensiveness presented, our model can be generalized to other applications. Also, we can infer that the more significant number of compound AU influenced positively in recognition of basic AUs by analyzing our results. Though it is quite acceptable, the performance of the presented method can be improved in several respects: (1) our proposed method cannot encode the full range of Libras facial behavior; (2) the different characteristics actions between the upper part and the lower part of the face is not contemplate by our network architecture. Further efforts will be required if these limitations are to be addressed. Besides, it will be interesting to test the proposed method with a substantially extensive database.

Table 3. Agreement coefficient for comparison with manual annotation and automatic Libras' AU detection framework

Sentence Type	**Transcription Type**		**Visibility Code**	
	Human Annotation	**Our AU detection framework**	**Human Annotation**	**Our AU detection framework**
WH-Question	0.82	0.33	0.81	0.20
Yes/No Question	0.75	0.30	0.80	0.20
Negation	0.86	0.45	0.82	0.12
Afirmation	0.77	0.29	0.84	0.22
Affective	0.60	0.24	0.78	0.26

Acknowledgement

This paper and the research behind was financially supported by the Coordination for the Improvement of Higher Education Personnel (CAPES).

References

1. Antonakos, E., Roussos, A., Zafeiriou, S.: A survey on mouth modeling and analysis for sign language recognition. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. vol. 1, pp. 1–7. IEEE (2015)
2. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 59–66. IEEE (2018)
3. Benitez-Quiroz, C.F., Srinivasan, R., Feng, Q., Wang, Y., Martinez, A.M.: Emotionet challenge: Recognition of facial expressions of emotion in the wild (2017)
4. Chollet, F., et al.: Keras: The python deep learning library. Astrophysics Source Code Library (2018)
5. Chu, W.S., De la Torre, F., Cohn, J.F.: Modeling spatial and temporal cues for multi-label facial action unit detection. arXiv preprint arXiv:1608.00911 (2016)
6. Chu, W.S., De la Torre, F., Cohn, J.F.: Learning spatial and temporal cues for multi-label facial action unit detection. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). pp. 25–32. IEEE (2017)
7. Chu, W.S., De la Torre, F., Cohn, J.F.: Learning facial action units with spatiotemporal cues and multi-label sampling. Image and Vision Computing **81**, 1–14 (2019)
8. Ekman, P.: Facial expression and emotion. American psychologist **48**(4), 384 (1993)
9. Ekman, P., Friesen, W.V.: Manual for the facial action coding system. Consulting Psychologists Press (1978)
10. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern recognition **36**(1), 259–275 (2003)
11. Felipe, T.A.: The verbalvisual discourse in brazilian sign language–libras. Bakhtiniana: Revista de Estudos do Discurso **8**(2) (2013)
12. Fleiss, J.L., Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and psychological measurement **33**(3), 613–619 (1973)
13. Freitas, F.A., Peres, S.M., Lima, C.A., Barbosa, F.V.: Grammatical facial expression recognition in sign language discourse: a study at the syntax level. Information Systems Frontiers **19**(6), 1243–1259 (2017)
14. Freitas, F.A., Peres, S.M., de Moraes Lima, C.A., Barbosa, F.V.: Grammatical facial expressions recognition with machine learning. In: FLAIRS Conference (2014)
15. Gudi, A., Tasli, H.E., Den Uyl, T.M., Maroulis, A.: Deep learning based facs action unit occurrence and intensity estimation. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. vol. 6, pp. 1–5. IEEE (2015)
16. Hamm, J., Kohler, C.G., Gur, R.C., Verma, R.: Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. Journal of neuroscience methods **200**(2), 237–256 (2011)

17. Han, S., Meng, Z., Li, Z., O'Reilly, J., Cai, J., Wang, X., Tong, Y.: Optimizing filter size in convolutional neural networks for facial action unit recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5070–5078 (2018)
18. Hao, L., Wang, S., Peng, G., Ji, Q.: Facial action unit recognition augmented by their dependencies. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 187–194. IEEE (2018)
19. Hosie, J., Gray, C., Russell, P., Scott, C., Hunter, N.: The matching of facial expressions by deaf and hearing children and their production and comprehension of emotion labels. Motivation and Emotion **22**(4), 293–313 (1998)
20. Kanade, T., Tian, Y., Cohn, J.F.: Comprehensive database for facial expression analysis. In: fg. p. 46. IEEE (2000)
21. King, D.E.: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research **10**, 1755–1758 (2009)
22. Koelstra, S., Pantic, M., Patras, I.: A dynamic texture-based approach to recognition of facial actions and their temporal models. IEEE transactions on pattern analysis and machine intelligence **32**(11), 1940–1954 (2010)
23. Kolod, E.: How does learning sign language affect perception. Intel Science Talent Search pp. 1–20 (2004)
24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
25. Li, S., Deng, W.: Deep facial expression recognition: A survey. arXiv preprint arXiv:1804.08348 (2018)
26. Li, W., Abtahi, F., Zhu, Z.: Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1841–1850 (2017)
27. Li, W., Abtahi, F., Zhu, Z., Yin, L.: Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. arXiv preprint arXiv:1702.02925 (2017)
28. Liddell, S.K.: American sign language syntax, vol. 52. Mouton De Gruyter (1980)
29. Linh Tran, D., Walecki, R., Eleftheriadis, S., Schuller, B., Pantic, M., et al.: Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3190–3199 (2017)
30. Littlewort, G.C., Bartlett, M.S., Lee, K.: Automatic coding of facial expressions displayed during posed and genuine pain. Image and Vision Computing **27**(12), 1797–1803 (2009)
31. Liu, J., Liu, B., Zhang, S., Yang, F., Yang, P., Metaxas, D.N., Neidle, C.: Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions. Image and Vision Computing **32**(10), 671–681 (2014)
32. Liu, M., Shan, S., Wang, R., Chen, X.: Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1749–1756 (2014)
33. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. pp. 94–101. IEEE (2010)
34. Ma, C., Chen, L., Yong, J.: Au r-cnn: Encoding expert prior knowledge into r-cnn for action unit detection. Neurocomputing **355**, 35–47 (2019)

35. Majumder, A., Behera, L., Subramanian, V.K.: Automatic facial expression recognition system using deep network-based data fusion. IEEE transactions on cybernetics **48**(1), 103–114 (2018)
36. Martinez, B., Valstar, M.F., Jiang, B., Pantic, M.: Automatic analysis of facial actions: A survey. IEEE transactions on affective computing (2017)
37. Mavadati, M., Sanger, P., Mahoor, M.H.: Extended disfa dataset: Investigating posed and spontaneous facial expressions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–8 (2016)
38. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: Disfa: A spontaneous facial action intensity database. IEEE Transactions on Affective Computing **4**(2), 151–160 (2013)
39. Mei, C., Jiang, F., Shen, R., Hu, Q.: Region and temporal dependency fusion for multi-label action unit detection. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 848–853. IEEE (2018)
40. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. pp. 1–10. IEEE (2016)
41. Ong, S.C., Ranganath, S.: Automatic sign language analysis: A survey and the future beyond lexical meaning. IEEE Transactions on Pattern Analysis & Machine Intelligence (6), 873–891 (2005)
42. Pantic, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: The state of the art. IEEE Transactions on pattern analysis and machine intelligence **22**(12), 1424–1445 (2000)
43. Peterson, C.C., Siegal, M.: Deafness, conversation and theory of mind. Journal of child Psychology and Psychiatry **36**(3), 459–474 (1995)
44. Pigou, L., Dieleman, S., Kindermans, P.J., Schrauwen, B.: Sign language recognition using convolutional neural networks. In: European Conference on Computer Vision. pp. 572–578. Springer (2014)
45. Pramerdorfer, C., Kampel, M.: Facial expression recognition using convolutional neural networks: state of the art. arXiv preprint arXiv:1612.02903 (2016)
46. Romero, A., León, J., Arbeláez, P.: Multi-view dynamic facial action unit detection. Image and Vision Computing (2018)
47. Sanchez, E., Tzimiropoulos, G., Valstar, M.: Joint action unit localisation and intensity estimation through heatmap regression. arXiv preprint arXiv:1805.03487 (2018)
48. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE transactions on pattern analysis and machine intelligence **37**(6), 1113–1133 (2015)
49. Savran, A., Sankur, B., Bilge, M.T.: Regression-based intensity estimation of facial action units. Image and Vision Computing **30**(10), 774–784 (2012)
50. Silva, E., Costa, P., Kumada, K., De Martino, J.M.: Silfa: Sign language facial action database for the development of assistive technologies for the deaf. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG). pp. 382–386 (2020)
51. Silva, E.P., Costa, P.D.P.: Recognition of non-manual expressions in brazilian sign language. In: 12th IEEE International Conference on Automatic Face and Gesture Recognition. Doctoral Consortium. IEEE (2017)
52. da Silva, E.P., Costa, P.D.P.: Qlibras: A novel database for grammatical facial expressions in brazilian sign language. In: X Encontro de Alunos e Docentes do DCA/FEEC/UNICAMP (EADCA) (2017)

53. Simard, P.Y., Steinkraus, D., Platt, J.C.: Best practices for convolutional neural networks applied to visual document analysis. In: null. p. 958. IEEE (2003)
54. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
55. Von Agris, U., Knorr, M., Kraiss, K.F.: The significance of facial features for automatic sign language recognition. In: Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on. pp. 1–6. IEEE (2008)
56. Vural, E., Cetin, M., Ercil, A., Littlewort, G., Bartlett, M., Movellan, J.: Drowsy driver detection through facial movement analysis. In: International Workshop on Human-Computer Interaction. pp. 6–18. Springer (2007)
57. Walecki, R., Rudovic, O., Pavlovic, V., Schuller, B., Pantic, M.: Deep structured learning for facial action unit intensity estimation. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. pp. 5709–5718. IEEE (2017)
58. Wang, S., Hao, L., Ji, Q.: Facial action unit recognition and intensity estimation enhanced through label dependencies. IEEE Transactions on Image Processing (2018)
59. Wu, B.F., Lin, C.H.: Adaptive feature mapping for customizing deep learning based facial expression recognition model. IEEE Access **6**, 12451–12461 (2018)
60. Yabunaka, K., Mori, Y., Toyonaga, M.: Facial expression sequence recognition for a japanese sign language training system. In: 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS). pp. 1348–1353. IEEE (2018)
61. Yauri Vidalón, J.E., De Martino, J.M.: Brazilian Sign Language Recognition Using Kinect, pp. 391–402. Springer International Publishing, Cham (2016)
62. Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H., Presti, P.: American sign language recognition with the kinect. In: Proceedings of the 13th international conference on multimodal interfaces. pp. 279–286. ACM (2011)
63. Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., Dobaie, A.M.: Facial expression recognition via learning deep sparse autoencoders. Neurocomputing **273**, 643–649 (2018)
64. Zhang, K., Huang, Y., Du, Y., Wang, L.: Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Transactions on Image Processing **26**(9), 4193–4203 (2017)
65. Zhang, Y., Dong, W., Hu, B.G., Ji, Q.: Classifier learning with prior probabilities for facial action unit recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5108–5116 (2018)
66. Zhao, K., Chu, W.S., Martinez, A.M.: Learning facial action units from web images with scalable weakly supervised clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2090–2099 (2018)
67. Zhao, K., Chu, W.S., De la Torre, F., Cohn, J.F., Zhang, H.: Joint patch and multi-label learning for facial action unit detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2207–2216 (2015)
68. Zhao, K., Chu, W.S., Zhang, H.: Deep region and multi-label learning for facial action unit detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3391–3399 (2016)
69. Zhi, R., Liu, M., Zhang, D.: A comprehensive survey on automatic facial action unit analysis. The Visual Computer pp. 1–27 (2019)