Word Level Sign Language Recognition via Handcrafted Features

Daniel Sánchez-Ruiz , J. Arturo Olvera-López , and Ivan Olmos-Pineda ,

***Abstract*—The ability to be understood and convey feelings, requests or ideas through words (spoken or written) is one of the most undervalue by all the humans who have the privilege to do it. Deaf community faces this challenge every single day and, even though, sign languages exist as way to battle against this issue, not all in deaf community knows who to use them; in fact, hearing community knows in a smaller proportion how to interpret them. By this reason sign language recognition area becomes relevant as an effort to solve this issue and create new communication channels.**

This work aims a methodology for word level sign language recognition, as principal highlights a small set of handcrafted features are defined, between them non-manual features are explored deeply. Data augmentation and dimensionality reduction were performed to obtain a concise feature space. Two recognition models were used (Bidirectional Long Term Memory and Transformer) in LIBRAS dataset, and the best result was an accuracy of 94.33%, which was obtained with the bidirectional long term memory network.

***Index Terms*—Sign Language Recognition, Word Level Sign Language Recognition, Computer Vision, Pattern Recognition**

I. Introduction

As of the year 2021, approximately 430 million people worldwide have been diagnosed with hearing loss, which represents one in every ten people; it is estimated that this condition could increase to 700 million by 2050 [1]. As most of the communication technologies have been developed to support spoken or written language, sign language processing would definitely help to overcome communication barriers for deaf community.

Sign Languages (SLs) are the main way of communication in deaf communities, which are composed of movements from distinct parts of the body such as: fingers, hands, arms, head, pose or even facial expressions; these movements are divided in manual (first two features previously described) and non-manual features (latter four features) [2]–[5]. There are five main parameters in sign language, which are hand-shape, palm orientation, movement, location, and expression/non-manual signals. To obtain an accurate sign word, all of these five parameters must be performed correctly [2], [6].

There exist two main levels of SLR: isolated SLR (also known as word-level), which classifies recordings of individual signs into glosses (one sign per data input) and continuous SLR, which recognizes whole utterances (multiple signs per data input) in a sentence level [2]–[5].

Daniel Sánchez-Ruiz, J. Arturo Olvera-López and Ivan Olmos-Pineda are with Computer Science Faculty, Benemérita Universidad Autónoma de Puebla e-mail:daniel.sanchezru@viep.buap.mx.

Vast technological solutions have been developed in recent years in order to try to solve this problem, many of them employ wearable devices such as bracelets, gloves, or armbands connected to a smartphone. However, these solutions can be costly (since devices are required) and intrusive, making the communication process uncomfortable. Computer vision solutions can be much cheaper and less intrusive but they can be challenging since they depend on big datasets and problems such as: occlusions, background segmentation or data noise filtering need to be considered and addressed [7].

Furthermore, even in computer vision, methods can be divided according to the input data. From one side, in sequences of RGB or RGB-D (Depth) images or videos; usually these methods have better execution in terms of accuracy but are computationally more demanding. On the other side, input data as a sequence of body poses, represented by locations of skeletal joints and facial landmarks [2]–[5]. Methods based on this representation achieve lower accuracy, but classification models are lightweight and more suitable for real time processing, e.g. in mobile devices. Making SLR able to run on these devices dramatically increases their potential in everyday use.

Feature extraction step is determinant for the whole SLR process and it can be very complex since hand movements are very unique in shape variation, textures, and motion [4], [7]. Since it is not easy to define the features, some approaches employ the entire frame/image, as it is performed in deep learning feature extraction methods (e.g in convolutional neural networks), which have obtained relevant results. Nonetheless, this type of approaches also use too many hyperparameters, consider irrelevant regions/data and require considerable computational resources.

In this work we focus on word-level SLR based on manual and non-manual features. In particular, it is analyzed the use of gaze estimation and head orientation, which are non-manual features which have not been study thoughtfully in related works. Besides that, augmentation data, dimensionality reduction and continuous SLR input data were applied; Bidirectional Long Short Term Memory (BiLSTM) and Transformers were used as recognition methods.

The rest of this paper is organized as follows: In section 2, related work such as: sign language recognition types, characteristics and methods are approached. In section 3, general methodology is presented, first in a broadly manner and then in detail. Section 4 describes the design and the results obtained in the experimental section. Finally, in section 5 the conclusions of the work are addressed.

II. Related Work

The SLR system recognizes SL and transforms those signs into meaningful words or expressions for the hearing community. The SLR systems are strongly tied with human gesture recognition problem or human action recognition problem since the SL word is a collection of ordered gestures.

Researchers have proposed various SLRS over the last decade, using traditional machine learning and advanced deep learning methodologies. This section discusses current advances in SL recognition employing traditional machine learning and deep learning techniques and their drawbacks as well as advantages.

Traditional SLR systems classify a sequence of frames or images that reflect a specific sign word or sign gesture by extracting spatial and temporal information. Traditional techniques such as image segmentation, hand detection, contour detection, hand shape detection, and hand tracking are utilized as optional steps, As it is stated in Koller [7], most of the related work have employed a Convolutional Neural Network (CNN) to perform the recognition step. BLSTM have been chose widely as recognition method. Finally recent works have started to study attention and transformers approaches.

Espejel-Cabrera et al. [8] proposed a method of chromatic segmentation based on Mexican Sign Language in the HSV space. The proposed system uses a Neural Network (NN) to automatically detect the skin color in the images. It is stated that extracted features obtain a good performance without making use of techniques for feature selection. Various classifiers are employed but Support Vector Machine (SVM) outperforms the other classifiers. The principal limitation is skin color algorithms lacks robustness in all type of data and context.

In Marzouk et al. [9] SLR technique has been developed. The technique initially pre-processes the input frames by a weighted average filtering approach. Next, a CapsNet feature extractor produce a collection of feature vectors. To identify and classify sign language, deep convolutional auto encoder model is exploited in the study. At the final stage, the atom search optimization algorithm is utilized as a hyperparameter optimizer which in turn increases the efficacy of the model.

A Indian isolated SLR using Long Short Term Memory (LSTM) and Gated recurrent unit (GRU), which focus on different hand gestures is presented in Kothadiya et al. [10]. It is conveyed that increasing the number of layers in the LSTM and GRU, and applying LSTM followed by GRU, helps the model achieve higher accuracy in the recognition phase.

A new perspective that balances global and local temporal gesture information, namely Multilevel Temporal Relation Graph (MLTRG), is presented in Guou et al. [11]. This was performed in order to alleviate recognition blur caused by similar gesture movements. In particular, MLTRG is constructed by using the visual information from different time spans, and then a Graph Convolutional Network (GCN) Layer is used for feature fusion and propagation between different levels. Through this process, the method can effectively analyze the correlation between global and local movements. Also, the method can alleviate the recognition ambiguity caused by various gestures in continuous sign language.

Hu et al. [12] explore the multilingual sign language recognition topic. They proposed a unified framework, which consists of a shared visual encoder, and an independent sequential module for each language together with a shared sequential module. The shared visual encoder and shared sequential module benefit from large training data of different languages and are able to promote each independent module for its corresponding language task. Besides, a max-probability decoding scheme is proposed to align the videos and sign glosses for further visual encoder refinement.

Das et al. [13] proposed a model that use a Histogram Difference based keyframe extraction method and a combination of a CNN and handcrafted features for SLR. The work investigated the importance of the local handcrafted features for identifying SL words and the importance of using features from the convolution layer instead of the dense layer. Some limitations are the dataset, which has a uniform background, which simplifies the process of extracting key points from the hand region using the SIFT algorithm.

In Rodríguez-Moreno et al. [14] a SLR approach is presented; hand landmarks obtained through MediaPipe were used to create a set of signals. Common Spatial Patterns (CSP) algorithm is used to transform these signals and after extract features from them (variance, maximum, minimum and IQR values), classification is carried out. An advantage that authors mentioned is the small set of hyperparameters that are employed in the CSP algorithm in comparison with deep learning approaches.

Caliwag et al. [15] proposed a method, where a *movement-in-a-video* detection scheme was applied to extract unique spatial and temporal features from each gesture. The extracted features were subsequently used with a pre-trained CNN to classify sign language gestures. The proposed method identifies sign language with short, medium, and long gestures in the Argentinian and Chinese sign language datasets. Although in the experimental setup the method was only tuned for short gestures, the authors claim that it can be extended to medium gestures.

Li et al. [16] developed an end-to-end continuously dynamic gesture recognition system based on multi-mode fusion. In order to improve the accuracy of continuously dynamic gesture recognition, a fusion information of 10-dimensional Inertial Measurement Unit signal (including 3-dimensional accelerometer, 3-dimensional gyroscope, and quaternion) and 8-channel surface electromyography signal as gesture features were used. A unified end-to-end deep learning network is designed, without requiring the pre-segment gesture information.

In most of the related work a deep learning approach is used for feature extraction and classification tasks, although these approaches have reached state of the art results, they also generated a considerable set of features and use a high number of hyper parameters in the training phase.

Also, the majority employed key points related with body pose and although this information consider regions concern with non-manual features in SLR, it is well documented [7] that to the best of our knowledge, there is a lack of use of all possible non-manual features.

For these reasons in this work we focus in extract a small

set of handcraft features related with manual and non-manual characteristics in order to alleviate the training phase and to study the relevance of non-manual features that have not been study thoughtfully. Despite the fact continuous sign language data (sentence level) is used, the work only addressed word level recognition; by this reason alignment and semantic/grammar tasks are out of the scope of this work.

III. Methods

In this section, it is described the proposed method to perform SLR. At first data acquisition is addressed, then a region of interest (ROI) identification and tracking is performed, with these regions a feature extraction step follows. Using the extracted characteristics a data augmentation method is applied, after that, a dimensionality reduction step delivers the feature vectors that are employed by the recognition methods in the last step. Fig. 1 conveys a graphical description of the methodology previously described.

A. Dataset Obtained

Corpus LIBRAS (Brazilian Sign Language) dataset [17] was used for the experimental stage, in particular Florianópolis' data. All the videos have a resolution of 640x414 pixels, with a refresh rate of 30 frames per second. Signers of different ages and physical attributes were considered in the acquisition process.

Several topics were covered in the videos, such as dialogs, spoken poetry, interviews, and basic vocabulary. All the records consists of two signers, who were recorded from four different points of view: one from a lateral view, one from a top view showing both signers; and two from frontal views of each signer.

It is worth mentioning that not all the videos were annotated because the project was not though for computer vision tasks, the recordings were made with the purpose of diffusion and cultural preservation of the language. The videos that were annotated, used the open-source ELAN annotation tool [18]. Each annotation file contains the annotations for right-hand signs and left-hand signs in individuals tracks.

The main reasons to chose LIBRAS dataset are: it contains (in comparison to other datasets) a considerable vocabulary (glosses), it considers a conversation format for most of the videos, where two persons are interacting, it contains a considerable amount of annotations and it was recorder in an environment more challenging (occlusions by objects, different recording angles, signers physical appearance).

B. Region Of Interest Detection

SLs use body movements to convey ideas following specific grammar rules, specifically hands configurations and positions, which are known as manual features. Beside this features, it also exists non-manual features, which are used to communicate importance, sarcasm, doubt among other emotions about an idea; all of this through features such as: gaze direction, head position and tilt, lip movements or facial expressions [2]–[5].

Manual and non-manual features hold spatiotemporal characteristics, hence it is necessary to encapsulate both type of information into a feature space. Recent related work in the SLR area have employed deep learning techniques for feature extraction and recognition steps [2]–[5], this advances have reached state of the art results, the main disadvantage in most of this approaches is the huge feature space they generate and need, which represents a difficulty for real time and mobile solutions.

By this reason handcrafted spatiotemporal features are used in this work. Hands, arms and head are the regions where this work is focused. Body pose estimation through key points is one of the principal manners to obtain features related with arms, hands and facial expressions [7]. Aligning with this, MediaPipe [19] was used to estimate body key points, the advantage of this framework is that it was developed for mobile devices, for this reason it does not need a Graphic Processing Unit (GPU).

Although MediaPipe provides hands' information and even though it is very precise in most of the times, it occurs that sometimes this region can not be estimated properly, because SLR presents a unique context where occlusions happen very often. By this reason other hand pose estimators who have obtained good results, such as InterHand2.6M [20] were discarded. Instead and taking this into consideration it was developed a hands recognition method, in order to achieve this, YOLOv5 framework [21] was trained with a set of images extracted from the dataset.

YOLO is one of the most reliable frameworks for object detection [22] and it has been used in several applications such as: face detection [23], apple flower detection [24], ship detection [25], detection and tracking of objects in surveillance systems [26], among others. Through all its versions, which are related to upgrades and updates, the accuracy and speed inference have improved. In particular in v5 the largest contribution is to translate the original framework to the PyTorch framework. The original framework was written primarily in C and offers fine grained control over the operations encoded into the network. In many ways the control of the lower level language has many advantages, but it could make it slower.

From the corpus some videos were selected following a systematic sampling technique [27]. The systematic sampling is defined in the Eq. (1), where k is the size of the increment for the selection of each one of the elements in the subsampling, this value is calculated as N/n, where N is the total size of the sample and n is the size of the subsampling; lastly i is a random number selected in range $[1-k]$.

$$M = (i, i+k, i+2k, ..., i+(n-1)k) \tag{1}$$

From the subsampling videos an image extraction step was performed in order to generate the input for YOLOv5 training. Computer Vision Annotation Tool (CVAT) [28] was used to define bounding boxes and annotate the hand class in all images. The inference model it was generated is stored and loaded subsequently to identify hands region.

Finally, as it is reported in Koller [7], there is a deficit in the study of non-manual features in comparison with manual features. Hand region related with manual features is the

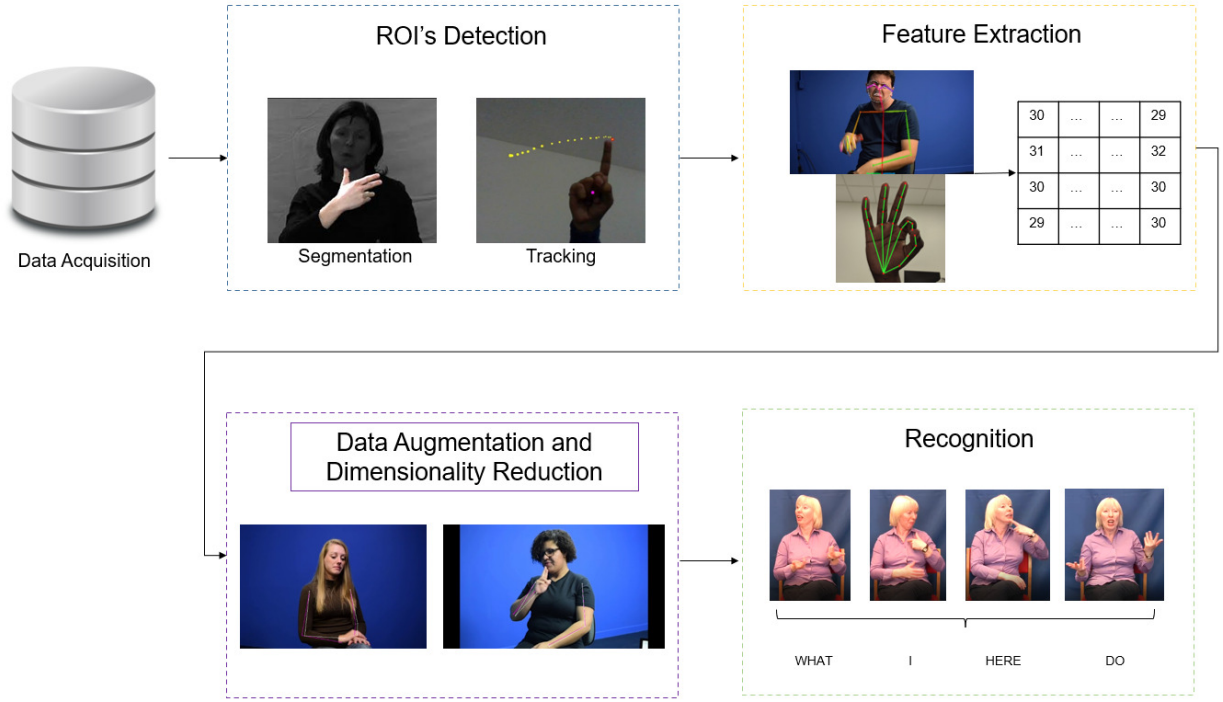


Fig. 1. General diagram of the proposed methodology.

region it conveys more information in SLs, however non-manual features can not be omitted since they also provided relevant information; this work addressed non-manual features considering data related to head pose and eye gaze estimation, OpenFace framework [29] is employed to do that.

C. Feature Extraction

Regarding deep learning approaches in related work, which commonly employ feature vector with one hundred characteristics or more, a small vector is proposed in this work considering manual and non-manual features.

With the inference model generated by YOLOv5 six features are extracted (three for each hand), all of them in relation with bounding box centroid: x and y coordinate and approximate speed, the latter feature is calculated based on the basic speed equation (Eq. (2)) following the algorithm for speed estimation in [30], where $distance\_meter$ is defined in Eq. (3), $distance\_pixel$ in Eq. (4), MPP (Meters Per Pixel) in Eq. (5) and $time$ is the elapsed time between two adjacent frames.

$$speed = distance\_meter/time \tag{2}$$

$$distance\_meter = distance\_pixel * MPP \tag{3}$$

$$distance\_pixel = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{4}$$

$$MPP = distance\_camera\_to\_signer * frame\_width \tag{5}$$

Commonly related work based on pose features only employ information concern to spatial position of every key point [2]–[5], in this work the use of spatial information is also used, but besides that, information related with euclidean distance between key points pairs is considered. Fig. 2 and Fig. 3

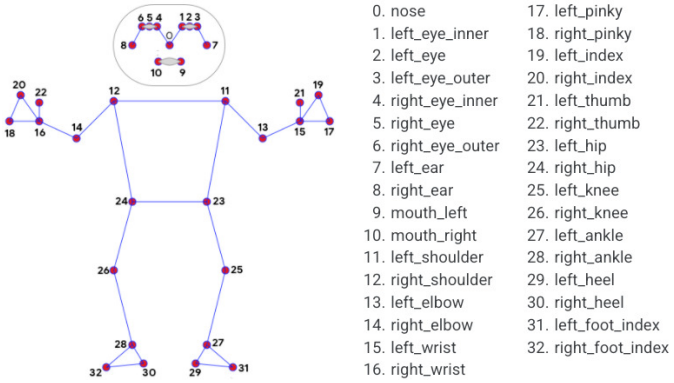


Fig. 2. Key points related to body pose, image taken from [19].

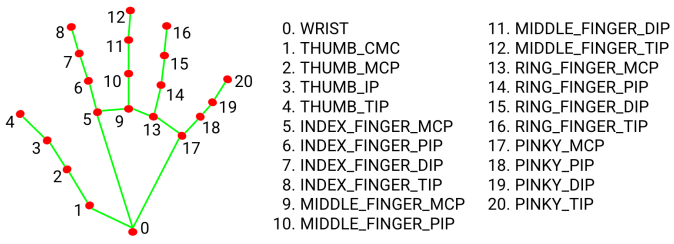


Fig. 3. Key points related to hands pose, image taken from [19].

shows the key points that MediaPipe provides for body and hands regions, Table I and Table II enumerated which one were selected, in the latter the pair of points displayed are the same for both hands.

For facial expressions, which are also part of non-manual features, euclidean distance is also calculated between a set of key points pairs ((0-17),(61-291),(0-94),(52-159),(282-386)),

TABLE I
KEY POINTS, WHERE ONLY SPATIAL ((x, y) COORDINATES) ARE USED (IN RELATION TO FIG. 2).

Region	Key points
Left Shoulder	11
Left Elbow	13
Left Wrist	15
Right Shoulder	12
Right Elbow	14
Right Wrist	16

TABLE II
PAIR OF POINTS, WHERE DISTANCE BETWEEN THEM IS CALCULATED (IN RELATION TO FIG. 3).

Region	Pair key points
Wrist - Middle Finger MCP	(0-9)
Wrist - Thumb MCP	(0-2)
Wrist - Pinky MCP	(0-17)
Pinky MCP - Pinky Tip	(17-20)
Ring Finger MCP - Ring Finger Tip	(13-16)
Middle Finger MCP - Middle Finger Tip	(9-12)
Index Finger MCP - Index Finger Tip	(5-8)
Thumb Finger MCP - Thumb Finger Tip	(2-4)

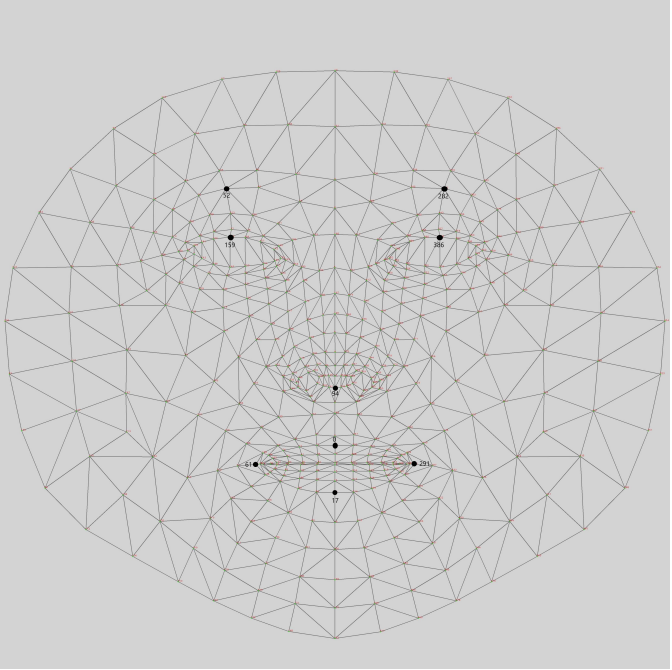


Fig. 4. Key points related to facial expressions, image taken from [19].

Fig. 4 exhibits these points.

Finally, as it was stated previously, through OpenFace are extracted non-manual features, in particular, for head pose spatial information concern about rotation angle in x, y and z axis is extracted; and for eye gaze estimation direction angle in x and y is extracted. A total of 44 manual and non-manual features are extracted to compose the feature vector.

D. Data Augmentation and Dimensionality Reduction

The use of data augmentation techniques for SLR has been studied in related work [2]–[5]. This have been done with the purpose to obtain a more complete model in training. It is very common that in some datasets exist labels that are gesticulated once, so it is impossible identify them correctly in the test phase without the use of some data augmentation technique.

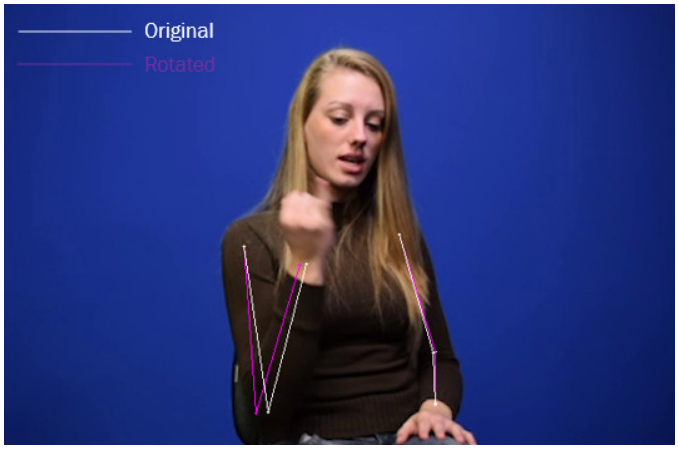


Fig. 5. Example of joints rotation.

Therefore in this work it was implemented a data augmentation technique defined by Boháček et al. [31], where the arm's spatial information extracted in the previous step was used to perform slightly rotations, which simulates the subtle variations between the signers gesticulation of each sign without changing its own semantic meaning.

In our approach, a sequential join rotation is carried out, where coordinates of both arms are passed successively, next the landmark is slightly rotated with respect to the current one. A probability of 30% is defined as the possibility of each joint to be rotated, the angle is a random angle up to ± 4 degrees.

The Eq. (6) and Eq. (7) are used for the rotation, where the center of rotation is the selected joint in the iteration (x_{sj}, y_{sj}), the x, y coordinates correspond to the adjacent joints and θ is the random selected angle. Fig. 5 conveys an example of the result of this operation.

$$f_{rotate}(x) = (x_{sj} - x)cos\theta - (y_{sj} - y)sin\theta \quad (6)$$

$$f_{rotate}(y) = (x_{sj} - x)sin\theta + (y_{sj} - y)cos\theta \quad (7)$$

Although the proposed feature vector is small, in order to find only the most discriminative features and with the intention to avoid overfitting and decrease the model complexity, principal component analysis (PCA) was used as dimensionality reduction technique. After performed this procedure, the final vector is established.

E. Recognition Methods

Computer vision seeks to understand digital images through various tasks, one of them is image recognition, which allows machines to identify objects, people, entities, and other variables (e.g. glosses in SLR) in images. To perform this activity, it deals with recognizing patterns and regularities in the image data, to later classify them into categories by learning and interpreting image pixel patterns. Related works have used diverse recognition techniques, from classical ones such as

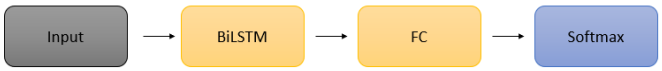


Fig. 6. Proposed BiLSTM architecture.

Hidden Markov Models (HMM), Recurrent Networks, Long Short Term Memory (LSTM) networks to more recent such as Transformers or CapNets [2]–[5], all of these methods have obtained good results in sequential problems.

A BiLSTM and a Transformer are utilized as the recognition methods in this work. BiLSTM networks have obtained acceptable results in sequences modeling problems such as: human activity recognition [32], emotion classification [33], facial expression recognition [34] or violence detection [35]. As it was mentioned previously, Koller stipulated in [7] that this networks have been used in SLR achieving relevant results.

Koller also mentioned Transformers have started to be considered as recognition methods for SLR and that the first results have been promising. Transformers are still considered a new technique, and even though the technique was developed originally for natural language processing tasks, recently have been adapted to computer vision tasks obtaining good results in problems such as: image classification, object detection, image segmentation, image super resolution or image denoising [36].

For the first method it was defined a baseline architecture, which is shown in Fig. 6 and it is composed by the following layers: starting with a basic BiLSTM network receives the input, the output goes through a Full Connected neural network, finally, a softmax layer returns the probabilities for each label in the vocabulary of the dataset.

The Transformer for recognition is a slight modification of the original work from Vaswani et al. [37] proposed by Boháček et al. [31]. In the decoder layer of the transformer the input is one query, which is decoded into the class representing the sign. The class query passes through a Multi-Head Projection module. This module is a special case of the Multi-Head Attention module, when there is only one element in the processed sequence.

For this case, the softmax in the attention module always results in 1 and thus the attention has no influence on the value vector. Hence, only the projection of the input vector into the value space has any meaning and it is not learned the key and query spaces in this module. This is the main difference in comparison to the original work.

IV. Experiments and Results

This section address the design experiment and the obtained results. Corpus LIBRAS dataset was used for all the experiments. The results obtained with YOLOv5 for the hands recognition and for SLR are described in detailed.

Google Colab [38], Python and PyTorch [39] were used as the platform, programming language and framework to perform the experimental stage. Data was divided in the training and test sets with a relation of 70%-30%, respectively.

TABLE III
Hands detection results by using YOLOv5 trained model for detection.

Metric	Result
mAP@0.5	96.22%
mAP@0.5-0.95	62.22%

A. YOLOv5 Results

From the dataset a subsampling had to be generated for YOLOv5 training, this videos were chosen following the systematic sampling technique previously described. A value of $k = 23$ and $i = 10$ were defined to finished with 50 videos, after that, from the subsampling set a sliding window of 15 seconds was employed to extract images, at the end 614 images were obtained.

All images were rescaled to 416x416, a batch with a size of 16 is established and the training was realized for 400 epochs. Mean average precision (mAP) is the metric to measure the performance of the detector. mAP with a threshold confidence value of 0.5 (mAP@0.5) and in the interval 0.5 to 0.95 (mAP@0.95) with a size step of 0.05.

Tab. (III) shows the obtained results, as it is expected with mAP@0.5-0.95 the result decreases, however they were good enough to employ trained model for the necessary inferences in the feature extraction step.

B. SLR Results

50 videos were used and a sliding window of 3 frames is defined. For PCA, Minka [40] method to find automatically the number of components is used, the value for percentage variance between the components to preserve is 95%. After performing the PCA process, from 44 features only 22 are maintained, among them those related to non-manual are facial expressions and eye gaze.

All the values in the final feature vector are normalized by the z-score method, which is the transformation of features by substracting the mean and dividing by standard deviation, Eq. (8) conveys this procedure.

$$x_{normalize} = (X - mean)/Std \tag{8}$$

Eaf annotations files, which are associated to each video, they contain the label for each gloss and their time intervals; in the feature vector each frame is verified if it is related to a label or not. As the videos are for continuous SLR (multiple glosses per video), transitions or rest states exists between glosses, so a blank_label is defined for this type of data.

Taking this in consideration, the label for each instance is established. Fig. 7 shows two relevant facts, first, as it was described, several labels present few instances and two, blank_label has more instances than any other label by a considerable margin, which it means the data is imbalanced.

With the purpose to investigate if decreasing blank_label instances helps to obtain better results, One Side, Repeated Edited, Tomek Lynks and All KNN under sampling methods are used through imbalanced-learn python module [41]. All of this methods are used to decrease the majority class without impacting the remaining classes labels.

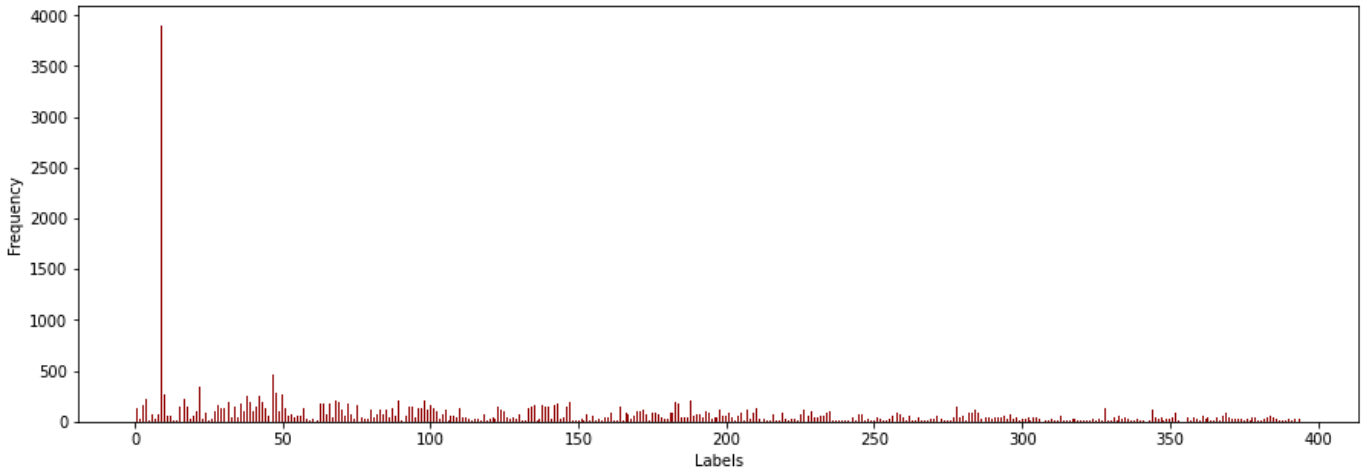


Fig. 7. Imbalance blank_label instances in respect to the rest.

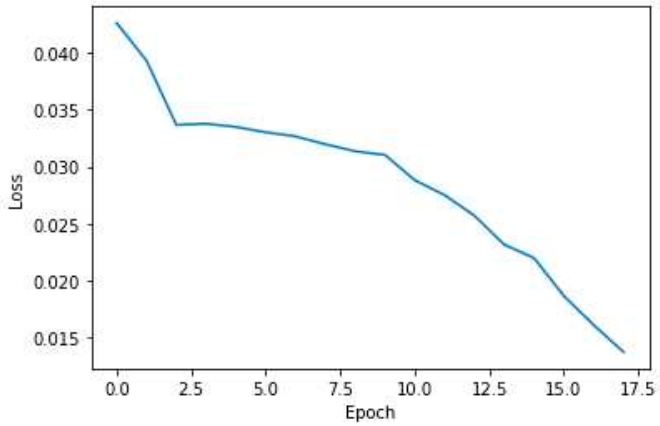


Fig. 8. Loss function behavior for BiLSTM training to determine epochs value.

The training and evaluation of BiLSTM are carried out for each video, this represent that a feature vector was obtained for each video. The metric employed was accuracy using a cross validation with $k = 3$. The parameters for the training were: 2 hidden layers, 128 cells in each hidden layer, a learning rate (LR) = 0.003 and a number of 18 epochs. This values were found through empirical experimentation; in particular for the epoch value and with the aim to avoid overfitting the model, evaluation in loss function behavior through training process (depicted in Fig. 8) was considered.

Tab. (IV) shows the average accuracy obtained for each video and the average standard deviation. Fig. (9) depicts the accuracy for each video in the cross validation process by each under sampling method. As the Tab. (IV) states trough standard deviation the behavior is vastly uniform in each fold for every video. It can be appreciated from Fig. (9) the video number 30 always achieved the lower accuracy; in a thoughtfully examination it was observed the illumination and the camera point of view generate shadows in face region that did not help in feature extraction stage. It is interesting to note the second best result was obtained without the use of any under sampling technique, which shows blank_label is not affecting the generation of a robust model inference.

TABLE IV
BILSTM RESULTS FOR SLR.

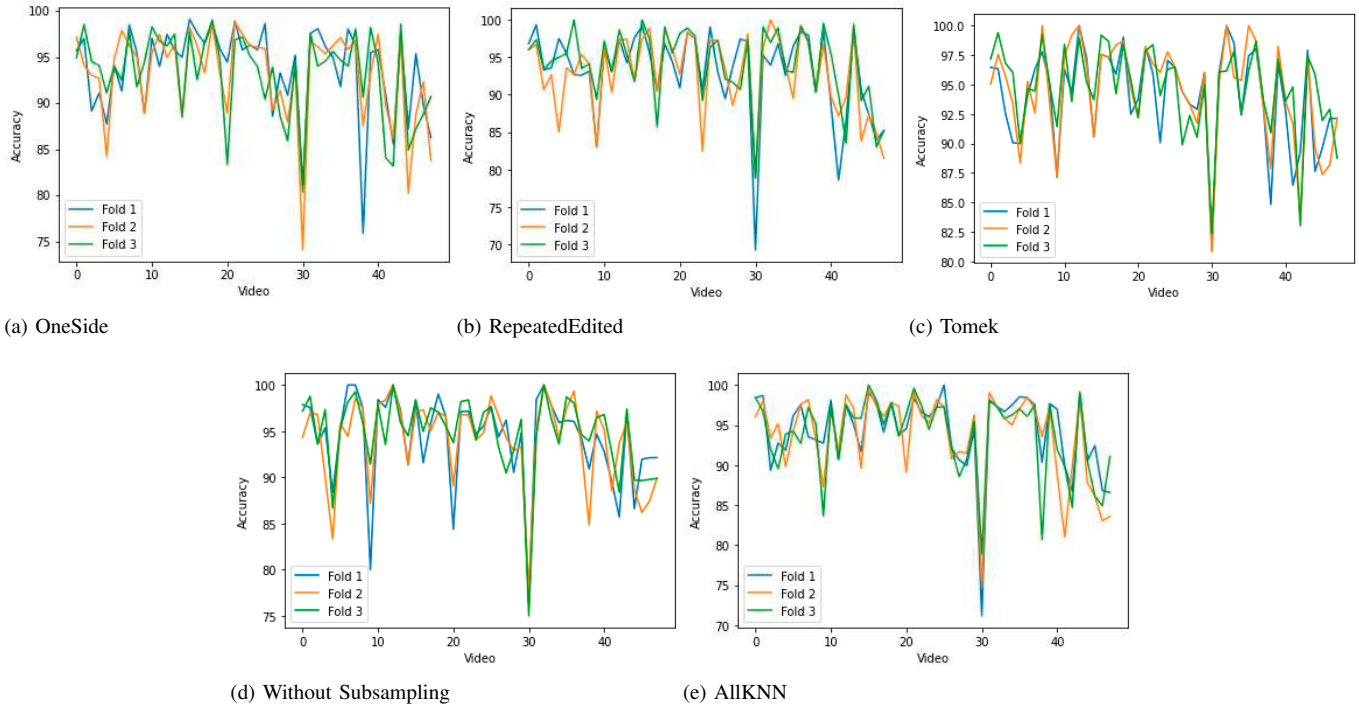
Under sampling tecnique	Accuracy	Std
Tomek	94.33%	± 3.81
Without subsampling	94.31%	± 4.38
AllKNN	93.85%	± 4.79
Repeated Edited	93.32%	± 5.01
One Side	93.29%	± 4.41

In the Transformer recognition method, the input vector had to be preprocessed, this due the fact each instance needs to contain all the data of all frames for each label. This means all adjacent instances who have the same label are collapsed in one instance. After perform the process for each video a new feature vector is generated for all data.

Under sampling techniques were discarded since when they were used, a relevant gain was not obtained. The Transformer model used as parameters: 6 encoder layers, 6 decoder layers, 22 hidden dimensions, feed-forward dimension = 2048 and 11 heads. 150 epochs and LR = 0.001 were defined for the training process and as BiLSTM model a softmax layer returns the probability for each label.

Fig. (10) depicts the training process in a accuracy/loss graph over the epochs, in the same manner epochs value was obtained through empirical experimentation. For this model a LR scheduler was implemented, a tolerance of 5 epochs were set, if the accuracy does not change, after the tolerance is met a new learning rate is defined ($new\_lr = lr * 0.1$), Fig. (11) shows this process. LR schedulers seek to adjust (generally reduce) the LR value during training as the epochs increase, going from a general to a specific optimization procedure. Finally the obtained results were 95.65% for the training set and 91.18% for the test set.

To the best of the authors' knowledge only Amaral et al. [42] and Passos et al. [43] have used Corpus Libras for SLR task. Tab. (V) shows a comparison between the best obtained results and the ones reported in Amaral's and Passos' work.



(a) OneSide (b) RepeatedEdited (c) Tomek

(d) Without Subsampling (e) AllKNN

Fig. 9. BiLSTM results for each under sampling unbalance technique.

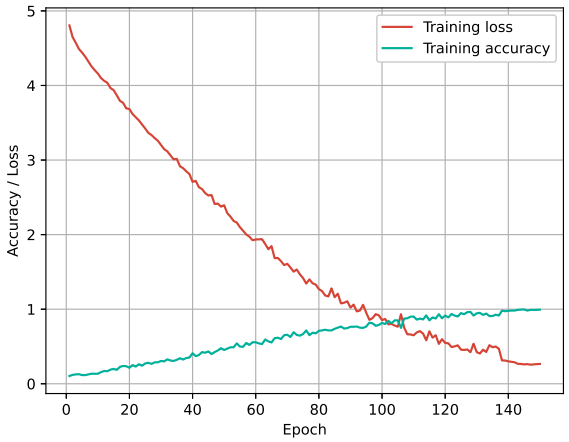


Fig. 10. Transformer accuracy/lost training.

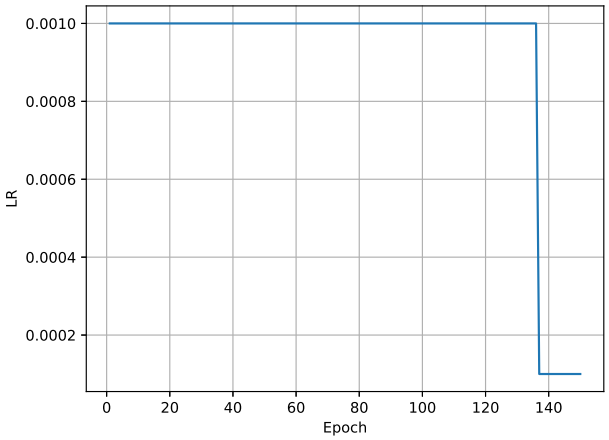


Fig. 11. LR automatic adjustment over epochs in training.

Both recognition methods had a better performance apart from that, something to take into account is that in Amaral et al. [42] only 10 different labels were used (100 for each label), and in Passos et al. [43] only 24 different labels were used meanwhile the proposed work employed almost 400 labels, some of them with a reduced number of instances, which shows the extracted and preserved features are discriminative enough even with a limited number of instances.

TABLE V
COMPARISON OF RESULTS WITH RELATED WORK.

Author	Recognition Method	Accuracy
Proposed work	BiLSTM	94.33%
Proposed work	Transformer	91.18%
Amaral et al. [42]	LSTM	88.4%
Passos et al. [43]	SVM	88.12%

V. CONCLUSIONS

SLR is a research area which can impact several people as new technological advances are developed. This work proposed a SLR methodology for word level by using a small set of handcrafted features. Also in a deeper manner in comparison with related work non-manual features are studied. After a ROI and tracking phase, spatio temporal features from body pose, facial expressions, hands region, head pose

and eye gaze estimation were extracted. A data augmentation and dimensionality reduction step is performed and finally in the recognition phase BiLSTM and Transformer models were used.

The experiments proves the performance of the methodology it is competitive in comparison with related work, the best result was an accuracy of 94.33%. In fact, the conditions (vocabulary size and instances per label) under the proposed work obtained the results were more challenging. Non-manual features were preserved after dimensionality reduction, which shows that this type of features contain relevant information. Also it was demonstrated that is not necessary a feature space of considerable size in order to generate recognition models for SLR.

As future work new data augmentations can be explored, also new recognition methods can be studied, in particular recent ones such as CapsNets or other Attention based methods [2]–[5]. Ephentesis moves could be of interest due to the fact continuous SLR data is used, its use as a preprocess step to identify the beginning and the end of each sign or as a new feature could help to discriminate more precisely blank_label instances from glosses instances.

Finally new datasets might be occupied to validate in a more robust manner the obtained results, LSA64 [44] and WLASL [45] datasets have been used by a vast related work [2]–[5].

ACKNOWLEDGMENTS

The first author thanks the support by the CONACyT PhD Scholarship 482941.

REFERENCES

[1] W. H. Organization, "Deafness and hearing lossv." https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

[2] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *Expert Systems with Applications*, vol. 164, p. 113794, 2021.

[3] A. Wadhawan and P. Kumar, "Sign language recognition systems: A decade systematic literature review," *Archives of Computational Methods in Engineering*, vol. 28, no. 3, pp. 785–813, 2021.

[4] R. Elakkiya, "Machine learning based sign language recognition: a review and its research frontier," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 7, pp. 7205–7224, 2021.

[5] E.-S. M. El-Alfy and H. Luqman, "A comprehensive survey and taxonomy of sign language research," *Engineering Applications of Artificial Intelligence*, vol. 114, p. 105198, 2022.

[6] S. Subburaj and S. Murugavalli, "Survey on sign language recognition in context of vision-based and deep learning," *Measurement: Sensors*, vol. 23, p. 100385, 2022.

[7] O. Koller, "Quantitative survey of the state of the art in sign language recognition," *arXiv preprint arXiv:2008.09918*, 2020.

[8] J. Espejel-Cabrera, J. Cervantes, F. García-Lamont, J. S. R. Castilla, and L. D. Jalili, "Mexican sign language segmentation using color based neuronal networks to detect the individual skin color," *Expert Systems with Applications*, vol. 183, p. 115295, 2021.

[9] R. Marzouk, F. Alrowais, F. N. Al-Wesabi, and A. M. Hilal, "Atom search optimization with deep learning enabled arabic sign language recognition for speaking and hearing disability persons," in *Healthcare*, vol. 10, p. 1606, MDPI, 2022.

[10] D. Kothadiya, C. Bhatt, K. Sapariya, K. Patel, A.-B. Gil-González, and J. M. Corchado, "Deepsign: Sign language detection and recognition using deep learning," *Electronics*, vol. 11, no. 11, p. 1780, 2022.

[11] J. Guo, W. Xue, L. Guo, T. Yuan, and S. Chen, "Multi-level temporal relation graph for continuous sign language recognition," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 408–419, Springer, 2022.

[12] H. Hu, J. Pu, W. Zhou, and H. Li, "Collaborative multilingual continuous sign language recognition: A unified framework," *IEEE Transactions on Multimedia*, pp. 1–12, 2022.

[13] S. Das, S. K. Biswas, and B. Purkayastha, "Automated indian sign language recognition system by fusing deep and handcrafted feature," *Multimedia Tools and Applications*, vol. 82, no. 11, pp. 16905–16927, 2023.

[14] I. Rodríguez-Moreno, J. M. Martínez-Otzeta, I. Goienetxea, and B. Sierra, "Sign language recognition by means of common spatial patterns: An analysis," *Plos one*, vol. 17, no. 10, p. e0276941, 2022.

[15] A. C. Caliwag, H.-J. Hwang, S.-H. Kim, and W. Lim, "Movement-in-a-video detection scheme for sign language gesture recognition using neural network," *Applied Sciences*, vol. 12, no. 20, p. 10542, 2022.

[16] J. Li, J. Meng, H. Gong, and Z. Fan, "Research on continuous dynamic gesture recognition of chinese sign language based on multi-mode fusion," *IEEE Access*, vol. 10, pp. 106946–106957, 2022.

[17] R. M. d. Quadros, "Documentação da língua brasileira de sinais," *Brasília: IPHAN - Ministerio da Cultura*, vol. 1, pp. 157–174, 2016.

[18] H. Sloetjes and P. Wittenburg, "Annotation by category-elan and iso dcr," in *6th international Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

[19] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for perceiving and processing reality," in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019.

[20] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee, "Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pp. 548–564, Springer, 2020.

[21] M. Horvat and G. Gledec, "A comparative study of yolov5 models performance for image localization and classification," in *Central European Conference on Information and Intelligent Systems*, pp. 349–356, Faculty of Organization and Informatics Varazdin, 2022.

[22] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *Digital Signal Processing*, vol. 126, p. 103514, 2022.

[23] N. K. Benamara, E. Zigh, T. B. Stambouli, and M. Keche, "Towards a robust thermal-visible heterogeneous face recognition approach based on a cycle generative adversarial network," *Int. J. Interact. Multimed. Artif. Intell*, vol. 7, no. 4, pp. 132–145, 2022.

[24] D. Wu, S. Lv, M. Jiang, and H. Song, "Using channel pruning-based yolo v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments," *Computers and Electronics in Agriculture*, vol. 178, p. 105742, 2020.

[25] J. Jiang, X. Fu, R. Qin, X. Wang, and Z. Ma, "High-speed lightweight ship detection algorithm based on yolo-v4 for three-channels rgb sar image," *Remote Sensing*, vol. 13, no. 10, p. 1909, 2021.

[26] M. Adimoolam, S. Mohan, G. Srivastava, *et al.*, "A novel technique to detect and track multiple objects in dynamic video surveillance systems," vol. 7, no. 4, pp. 112–120, 2022.

[27] R. K. Som, *Practical sampling techniques*. CRC press, 1995.

[28] B. Sekachev, N. Manovich, and A. Zhavoronkov, "Computer vision annotation tool," Oct. 2019. GitHub: https://github.com/opencv/cvat.

[29] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "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.

[30] A. Rosebrock, *Deep learning for computer vision with python: Starter bundle*. PyImageSearch, 2017.

[31] M. Boháček and M. Hrúz, "Sign pose-based transformer for word-level sign language recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 182–191, 2022.

[32] K. K. Verma and B. M. Singh, "Deep multi-model fusion for human activity recognition using evolutionary algorithms," vol. 7, no. 2, pp. 44–58, 2021.

[33] J. Yang, X. Huang, H. Wu, and X. Yang, "Eeg-based emotion classification based on bidirectional long short-term memory network," *Procedia Computer Science*, vol. 174, pp. 491–504, 2020.

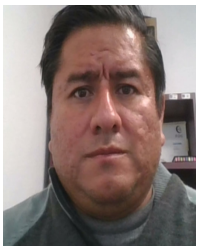
[34] L. Chen, Y. Ouyang, Y. Zeng, and Y. Li, "Dynamic facial expression recognition model based on bilstm-attention," in *2020 15th International Conference on Computer Science & Education (ICCSE)*, pp. 828–832, IEEE, 2020.

[35] R. Halder and R. Chatterjee, "Cnn-bilstm model for violence detection in smart surveillance," *SN Computer science*, vol. 1, no. 4, p. 201, 2020.

[36] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.
[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, p. 5999, 2017.
[38] E. Bisong, *Building machine learning and deep learning models on Google cloud platform: A comprehensive guide for beginners*. Apress, 2019.
[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8024–8035, 2019.
[40] T. Minka, "Automatic choice of dimensionality for pca," *Advances in neural information processing systems*, vol. 13, pp. 598–604, 2000.
[41] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.
[42] L. Amaral, V. Ferraz, T. Vieira, and T. Vieira, "Skelibras: A large 2d skeleton dataset of dynamic brazilian signs," in *Iberoamerican Congress on Pattern Recognition*, pp. 184–193, Springer, 2021.
[43] W. L. Passos, G. M. Araujo, J. N. Gois, and A. A. de Lima, "A gait energy image-based system for brazilian sign language recognition," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 11, pp. 4761–4771, 2021.
[44] F. Ronchetti, F. Quiroga, C. A. Estrebou, L. C. Lanzarini, and A. Rosete, "Lsa64: an argentinian sign language dataset," in *XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016).*, pp. 794–803, 2016.
[45] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1459–1469, 2020.

Daniel Sánchez-Ruiz received the bachelor degree in Computer Engineering from Autonomous University of Puebla. Also in the same institution completed the Master in Computer Science degree with specialization in distributed systems. He is mostly interested in problems related to the areas of computer vision, patterns recognition and digital image processing and analysis.

J. Arturo Olvera-López received the bachelor degree in Computer Science from Autonomous University of Puebla. He completed a Master and PhD degree in Computer Science at National Institute of Astrophysic, Optic and Electronic (INAOE). He is interested in problems related to the areas of pattern recognition, data mining, machine learning, data pre-processing, data reduction, digital image/signal processing & analysis and biometrics.



Ivan Olmos-Pineda received the bachelor degree in Computer Science from Autonomous University of Puebla. He completed a Master degree with specialization in computer networks at ITESM and a PhD degree in Computer Science at National Institute of Astrophysic, Optic and Electronic (INAOE). He is interested in problems related to the areas of pattern recognition, data mining, machine learning, data pre-processing, data reduction, digital image/signal processing & analysis and biometrics.