Spatial-Temporal Graph Convolutional Networks for Sign Language Recognition

Cleison Correia de Amorim
Centro de Informática
Universidade Federal de Pernambuco
50.740-560, Recife, PE, Brazil
cca5@cin.ufpe.br

David Macêdo
Centro de Informática
Universidade Federal de Pernambuco
50.740-560, Recife, PE, Brazil
dlm@cin.ufpe.br

Cleber Zanchettin
Centro de Informática
Universidade Federal de Pernambuco
50.740-560, Recife, PE, Brazil
cz@cin.ufpe.br

***Abstract*—The recognition of sign language is a challenging task with an important role in society to facilitate the communication of deaf persons. We propose a new approach of Spatial-Temporal Graph Convolutional Network to sign language recognition based on the human skeletal movements. The method uses graphs to capture the signs dynamics in two dimensions, spatial and temporal, considering the complex aspects of the language. Additionally, we present a new dataset of human skeletons for sign language based on ASLLVD to contribute to future related studies.**

I. Introduction

Sign language is a visual communication skill that enables individuals with different types of hearing impairment to communicate in society. It is the language used by most deaf people in their daily lives and, moreover, it is the symbol of identification between the members of that community and the main force that unites them. The sign language has a very close relationship with the culture of the country or even regions, and for this reason, each nation has its language [1].

According to the World Health Organization, the number of deaf people is about 466 million, and the organization estimates that by 2050 this number exceeds 900 million, which is equivalent to a forecast of 1 in 10 individuals around the world [2]. These data, in turn, highlight the breadth and importance of the sign language in the communication of people in different nations.

Despite this, there is still a small number of hearing people able to communicate through sign language. This ends up characterizing an invisible barrier that interferes with the communication between deaf and hearing persons, making a more effective integration among these people impossible [3]. In this context, it is essential to develop tools that can fill this gap by promoting integration among the population.

Research related to the sign recognition have been developed since the 1990s, and it is possible to verify significant results [4], [5]. The main challenges are primarily related to considering the dynamic aspects of the language, such as movements, articulations between body parts and non-manual expressions, rather than merely recognizing static signs or isolated hand positions. Besides, the sign language has thousands of signs, which sometimes differ only by subtle changes in movement, shape, or position of the hands and involve significant overlaps of fingers and occlusions. When combined with differences in signing style by distinct individuals and the variations arising from its non-universality and regionalisms, this field of research can become challenging for current artificial intelligence algorithms [6].

In this work, we present a new approach to performing the recognition of human actions based on spatial-temporal graphs called Spatial-Temporal Graph Convolutional Networks (ST-GCN) [7]. Using graph representations of the human skeleton we focus on body movement and the interactions between its parts, disregarding the interference of the environment around them. Besides, we address the movements under the spatial and temporal dimensions, and this allows to capture the dynamic aspects of the actions exercised over time. These characteristics make it a very relevant approach to dealing with the challenges and peculiarities of sign language recognition.

To develop the proposed method, however, it is first necessary to build a dataset of human skeletons associated with sign language that would serve as the primary information to feed the ST-GCN and enable the creation of its graphs. We used as a base the videos of the American Sign Language Lexicon Video Dataset (ASLLVD). It consists of a large set of the American Sign Language (ASL), on which we perform a segmentation and assignment of estimated skeleton labels.

The main contributions of the paper are: 1) the proposition of a new technique to sign recognition based on human movement, which considers different aspects of its dynamics and contributes to overcoming some of the main field challenges; and 2) the creation of a new dataset of human skeletons for sign language, currently non-existent, which aims to support the development of studies in this area.

Section II presents the related works and the details of ST-GCN. In Section III, the creation of the new database is discussed. Section IV addresses the adjustments in ST-GCN and Section V the conduction of the experiments. Finally, Sections VI and VII contains the results and final remarks respectively.

II. Related work

The recognition of sign languages obtained significant progress in recent years, which are motivated mainly by the advent of modern sensors, new machine learning techniques,

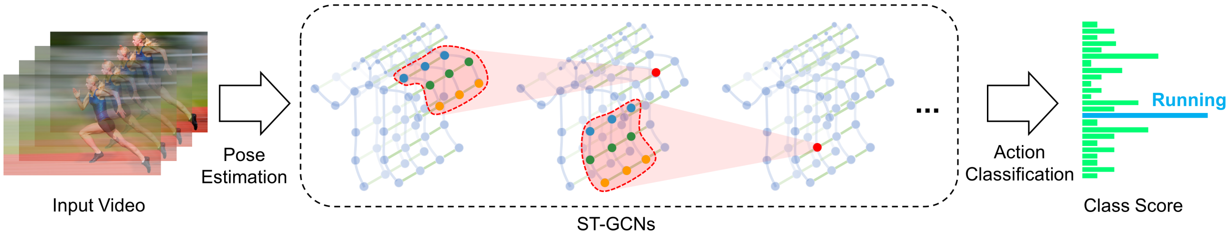


Figure 1. Overview of the ST-GCN approach [7, p. 3].

and more powerful hardware [5], [8]. Besides, approaches considered intrusive and requiring the use of sensors such as gloves, accelerometers, and markers coupled to the body of the interlocutor have been gradually abandoned, replaced by new approaches using conventional cameras and computer vision techniques.

Due to this movement, it is also notable the increase in the adoption of techniques for feature extraction such as SIFT [9], HOG [10], HOF [11] and STIP [11] to preprocess the images obtained by cameras and provide richer information to machine learning algorithms [4], [12].

Convolutional Neural Networks (CNN), as in many computer vision applications, obtained remarkable results in this field with accuracy reached 90% [12], [13], [14], [15] depending on the dataset. There are still some variations as 3D CNNs [16], the combination with other models such as Inception [17] or the Regions of Interest applications [18]. Recurrent Neural Networks [6] and Temporal Residual Networks [19] also obtained interesting results in the same purpose.

Despite the above advances, a large portion of these studies addressing static signs or single-letter images, from the dactylology[1] [12], [14], [16], [17], [18]. The problem is the negative effects on the intrinsic dynamics of the language, such as its movements, non-manual expressions, and articulations between parts of the body [20]. In this sense, it is extremely relevant that new studies observe such important characteristics, as in [6] and [19].

With this purpose, we present an approach based on skeletal body movement to perform sign recognition. This technique is known as Spatial-Temporal Graph Convolutional Network (ST-GCN)[2] and was introduced in [7]. The approach aims for methods capable of autonomously capturing the patterns contained in the spatial configuration of the body joints as well as their temporal dynamics. The authors suggest that previous methods for action recognition were limited by not explicitly exploring such spatial relations between the joints, which are crucial for the understanding of human actions. These methods simply used the joint coordinates in individual time steps to form feature vectors, applying a temporal analysis on them [7], [21], [22].

The ST-GCN uses as a base of its formulation a sequence of skeleton graphs representing the human body obtained from a series of action frames of the individuals. Figure 2 allows visualizing this structure, where each node corresponds to the point of articulation. The intra-body vertices are defined based on the body's natural connections. The inter-frame vertices, in turn, connect the same joints between consecutive frames to denote their trajectory over time [7].

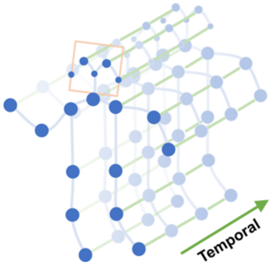


Figure 2. Sequence of skeleton graphs, denoting human movement in space and time, used by the ST-GCN [7, p. 1].

Figure 1 gives an overview of this technique. First, the estimation of individuals' skeletons in the input videos, as well as the construction of space-time graphs based on them. Then, multiple ST-GCN convolution layers are applied, gradually generating higher and higher level feature maps for the presented graphs. Finally, they are submitted to a classifier to identify the corresponding action.

To understand the operation of the ST-GCN, it is necessary first to introduce its sampling and partitioning strategies. When we are dealing with convolutions over 2D images, it is easy to imagine the existence of a rigid grid (or rectangle) around a central point that represents the sampling area of the convolutional filter, which delimits the neighborhood. In the case of graphs, however, it is necessary to look beyond this definition and consider the neighborhood of the center point as points that are directly connected by a vertex. Figure 4 shows this definition for a single frame. Note that for the red center points, the dashed edges represent the sampling area of the convolutional filter. Note also that although there are other points physically close to the central points (such as the points of the feet, knees, and waist), the method does not consider this points unless there is a vertex connecting them to the red points. This sequence of steps is the **sampling strategy** of ST-GCN.

[1] Dactylology - also known as the digital or manual alphabet. It consists of spelling words by the Deaf. It is generally used to introduce a word that does not yet have an equivalent sign [20], [1].

[2] Available at https://github.com/yysijie/st-gcn.

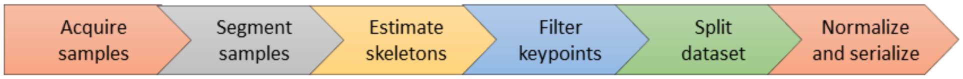


Figure 3. Preprocessing steps for creating the ASLLVD-Skeleton dataset.

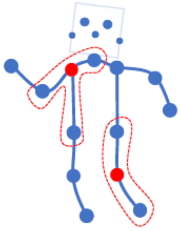


Figure 4. Sampling strategy in a convolution layer for a single frame [7, p. 5].

Figure 4 shows how the convolutional filter considers only points immediately connected to the central points. In other words, the delimitation of the filter area consider the neighbors with distance $D = 1$. The authors defined this distance in ST-GCN [7].

The **partitioning strategy** is based on the location of the joints and the characteristics of the movement of the human body, as shown in Figure 5. According to the authors, concentric or eccentric categorize the body parts movements, and the points in the sampling region are partitioned into three subsets:

- The root node (or center point, marked green in Figure 5);
- The centripetal group (blue dots in the Figure 5), which are the neighborhood nodes that are closest to the center of gravity of the skeleton (black cross in the Figure 5);
- The centrifugal group (yellow dots in the Figure 5), which are the nodes farther from the center of gravity.

The center of gravity is taken to be the mean coordinate of all joints of the skeleton in one frame. During convolution, each point of the body is labeled according to one of the above partitions named Spatial Configuration Partitioning [7]. It is through this method that the authors also establish the weights of the model, making each of the partitions receive a different weight to be learned.

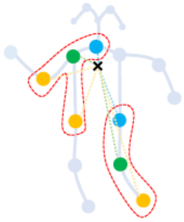


Figure 5. Spatial Configuration Partitioning strategy [7, p. 5].

In order to learn the temporal dimension, the ST-GCN extends the concept of graph convolution shown above to the scheme presented in the Figure 6. Considering this dimension as a sequence of skeletons graphs stacked consecutively, as in Figure 2. With this, we have in hands a set of graphs that are neighbors to each other. Let us now assume that each articulation of the body in a graph must be connected using a vertex to itself in the graph of the previous neighbor frame and also in the next neighbor frame. Given that, if we return to the definition of sampling introduced above, we verify that the convolutional filter contemplates those points belonging to the neighboring graphs, that now fit the requirements of being directly connected at a distance $D = 1$. In this way, ST-GCN considers the spatial and temporal dimensions and applies convolutions on them.

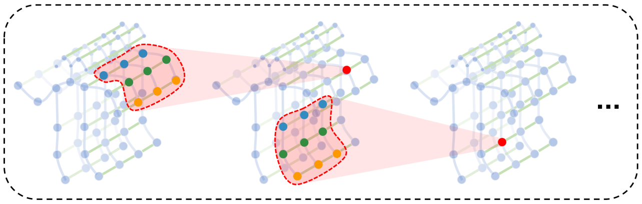


Figure 6. Convolution on the spatial and temporal dimensions, which considers the points directly connected in the current and neighboring graphs [7, p. 3].

Figure 7 (left) shows the model architecture and its convolutional layers, according to [7]. There are a total of nine sequentially positioned ST-GCN layers, which perform the extraction of features from the graphs. A normalization layer precedes them and followed by a global pooling and a softmax classification layer. On the right side of the image we also show the details of an ST-GCN convolutional unit.

To estimate the skeletons of individuals in videos, the authors [7] used a library called OpenPose. It is an open-source tool that uses deep learning algorithms to detect and estimate up to 130 human body points, as shown in Figure 8, and presented in [24], [25], [26].

III. New dataset of human skeletons for sign language

We introduce a new dataset of human skeletons for sign language based on the American Sign Language Lexicon Video Dataset (ASLLVD). The ASLLVD is a broad public dataset[3] containing video sequences of thousands of American Sign Language (ASL) signs, as well as their annotations, respective start, and end frame markings, and class labels for each sample [27], [28], [29].

According to the authors, each sign in ASLLVD is articulated by native individuals in ASL, and video sequences are collected using a four-camera system that simultaneously

[3] Available at http://csr.bu.edu/asl/asllvd/annotate/index.html.

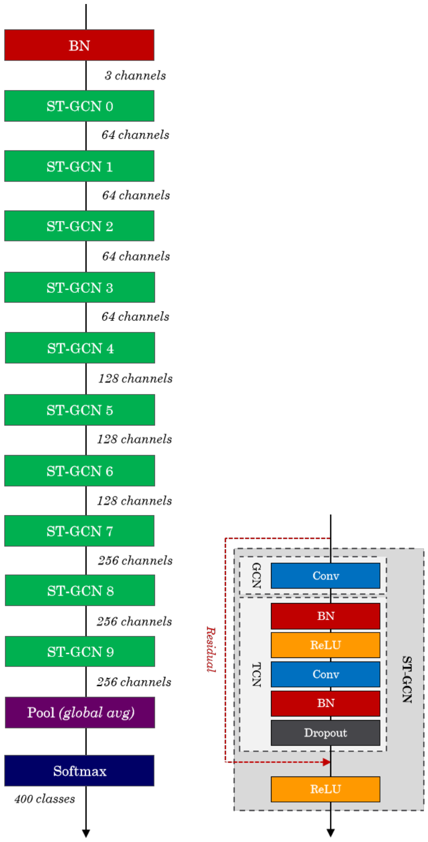


Figure 7. ST-GCN Model architecture (left) and details of one convolutional unit (right). Image adapted from [7].

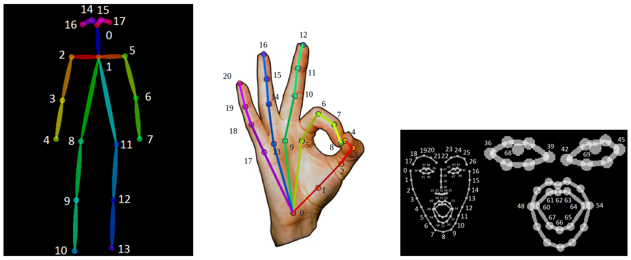


Figure 8. Representation of the 130 points estimated by OpenPose, 18 correspond to the body (left); 21 correspond to each of the hands (center); and 70 points correspond to the face (right) [23].

captures two frontal views, one side view and one enlarged view of the face of these individuals. Figure 9 exemplifies capturing three of these views for the “MERRY-GO-ROUND” sign.

The number and type of signs included in ASLLVD are similar in scale and scope to the set of lexical entries in existing English-to-ASL dictionaries. There is at least one example of video per sign for almost all those contained in the Gallaudet Dictionary of American Sign Language [27], [30].

There is a total of 2,745 signs, represented in approximately 10 thousand samples. Each sign contains between 1 and 18 samples articulated by different individuals (the average number of samples per sign is 4).

Figure 9. Example of the sign “MERRY-GO-ROUND” from three different perspectives, in the ASLLVD [27, p. 2].

To make the ASLLVD samples compatible with the input of the ST-GCN model, it was first necessary to apply a series of preprocessing steps, as shown in Figure 3. These steps gave rise to a new dataset containing the estimate of the skeletons for all the signs contained therein.

The new skeleton dataset was named **ASLLVD-Skeleton** and is public available[4] to contribute to the development of future studies on the recognition of sign languages.

The first step consists of **obtaining the videos** based on the associated metadata file used to guide this process. We consider only the videos captured by the frontal camera, once they simultaneously contemplate movements of the trunk, hands, and face of the individuals.

The next step is to **segment the videos** to generate a video sample for each sign. Every file in the original dataset corresponds to multiple signs per individual. Due to this, it is necessary to change the files organization in such a way that each sign is arranged individually with its respective label. We also considered the metadata file which contains the labels and the start and end frame marks to perform the segmentation for each sign. The output of this step consists of small videos with a few seconds, as shown in Figure 10.

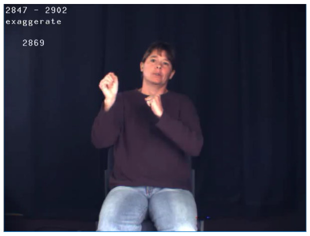


Figure 10. Representation of the “EXAGGERATE” sign, segmented from ASLLVD dataset.

The third step consists of **estimating the skeletons** of the individuals present in the segmented videos. In other words, the coordinates of the individuals’ joints are estimated for all frames, composing the skeletons that can be used to generate the graphs of the ST-GCN method. As in [7], we used the OpenPose library in this process, and a total of 130 key points were estimated (according to figure 8). Figure 11 illustrates the reconstruction in a 2D image of the estimated coordinates for the “EXAGGERATE” sign.

Figure 12 shows the content of an example file obtained at the end of this step. Each frame contains a section called

[4] Available at http://www.cin.ufpe.br/~cca5/asllvd-skeleton

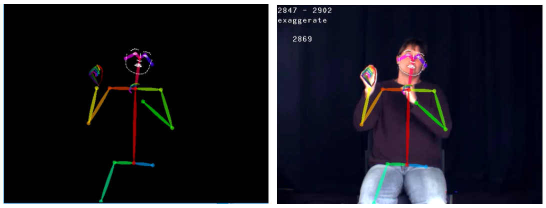
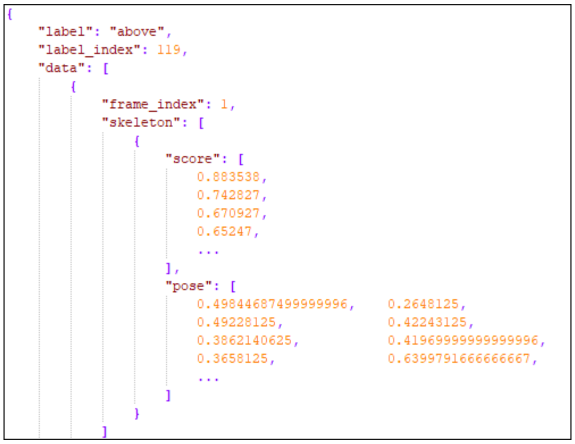


Figure 11. Reconstruction of the skeleton from the coordinates estimated by OpenPose for the sign “EXAGGERATE” (left); and the overlaping this skeleton in the original video (right).

“pose” with the coordinates of the X and Y axes estimated for the body joints, and a section “score” with the degree of confidence for each of these joints.



```
{
    "label": "above",
    "label_index": 119,
    "data": [
        {
            "frame_index": 1,
            "skeleton": [
                {
                    "score": [
                        0.883538,
                        0.742827,
                        0.670927,
                        0.65247,
                        ...
                    ],
                    "pose": [
                        0.49844687499999996,    0.2648125,
                        0.49228125,             0.42243125,
                        0.3862140625,           0.41969999999999996,
                        0.3658125,              0.6399791666666667,
                        ...
                    ]
                }
            ]
```

Figure 12. Example containing the estimated coordinates for the skeletons in a sign.

The fourth step involves **filtering the key points**. We use only 27 of the 130 estimated key points, which 5 refer to the shoulders and arms, and 11 refer to each hand, as illustrated in Figure 13. The output of this step consists of the same files as shown in Figure 12, but with a smaller number of coordinates per frame.

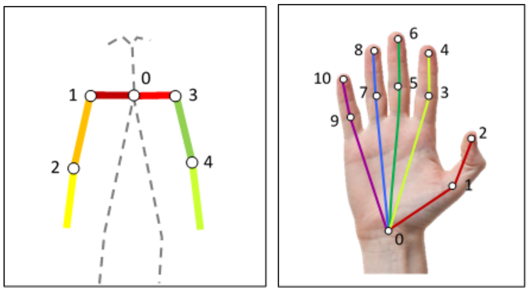


Figure 13. Representation of the 27 used key points, which 5 refer to the shoulders and arms (left) and 11 refer to each hand (right).

The fifth step concerns the **division of the dataset** into smaller subsets for training and test. For this procedure, a cross-validation dataset tool called “*train_test_split*”, available by the Scikit-Learn [31] library was used. In this division, we assign a proportion of 80% of the samples for training (corresponding to 7,798 samples) and 20% for tests (corresponding to 1,950 samples). This division is a commonly adopted proportion, and we understand that the number of samples in these groups is sufficient to validate the performance of most machine learning models.

Finally, the sixth step is to **normalize and serialize** the samples to make compatible with the ST-GCN format. Normalization aims to make the length of all samples uniform by applying the repetition of their frames sequentially to the complete filling of an established fixed number of frames. The number of fixed frames adopted is 63 (using a rate of 30 FPS, corresponds to a video with an approximate duration of 2 seconds). Serialization, in turn, consists of preloading the normalized samples from the subsets to translate them into physical Python [32] files, which contain their in-memory representations which are the format used by the ST-GCN. We adopted this format to optimize the data loading process. For each subset, we generate two physical files, the samples, and the labels.

The source code that performs the processing of these steps is also available in the ASLLVD-Skeleton download area.

IV. ST-GCN for Sign Language Recognition

Since the graph representation approach adopted by the ST-GCN is very flexible, it is not necessary to make modifications in the architecture of the model. Instead, only punctual adaptations to consider the new coordinates of the sign language domain.

The first one is to modify the algorithm to make it capable of using new customized layouts of graphs beyond those originally defined in [7]. Thus, a new type of layout called “custom” was defined in the configuration file of the model. In addition, a new parameter called “custom_layout” has been created within the “graph_args” attribute to allow the number of nodes, the central node, and the edges of new graphs to be informed.

Once this configuration is established, it is possible to map the topology of the graph used in this paper, which is composed by the 27 joints and vertices shown in Figure 13. Finally, small adjustments were required in the dimensions of the matrices used by the data feeders and also in the class that define the graph domain of the model, so that they can now support layouts with dynamic dimensions, as above.

The source code containing these adaptations is public available[5]. This code consists of a *forked* repository created from the one originally model developed by the authors in [7].

V. Experiments

We used as reference the experiments proposed in [4], which evaluate the performance in ASLLVD dataset of models as the Block-Based Histogram of Optical Flow (BHOF) method (see section II) and also popular techniques such as Motion Energy Image (MEI) [27], Motion History Image (MHI) [33],

[5] Available at http://www.cin.ufpe.br/~cca5/st-gcn-sl

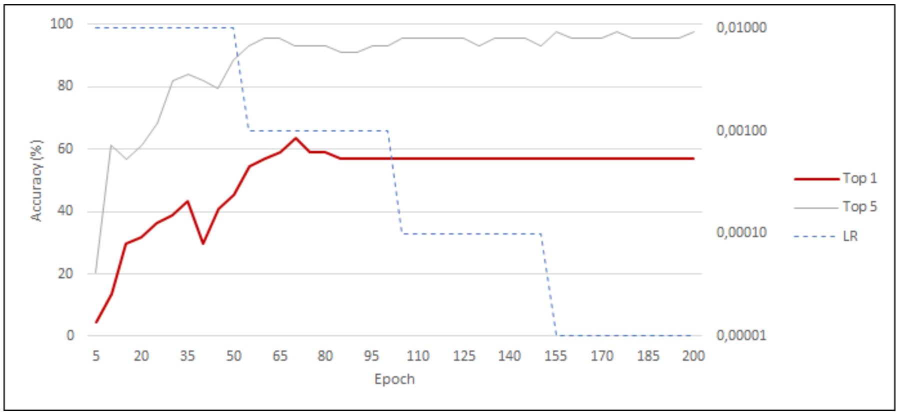


Figure 14. Accuracy obtained by the presented approach in recognition of the 20 signs selected from the ASLLVD.

Principal Component Analysis (PCA) [34] and Histogram of Optical Flow (HOF) [11].

The authors used a subset containing 20 signs selected from the ASLLVD, as presented in Table I. To reproduce this configuration, we selected the estimated skeletons for these signs from the ASLLVD-Skeleton dataset. We also identified that among the selected signs there were different articulations for the signs AGAIN, BASEBALL, CAN, CHAT, CHEAP, CHEAT, CONFLICT, DEPRESS, DOCTOR and DRESS. To solve this, we chose to keep only that articulation that contained the most significant number of samples for their respective signs.

Since the number of samples resulting from this process was small, totaling 131 items, it was necessary to apply a new division making 77% of this subset for training and 33% for tests. With this strategy, we improve the balance in the number of samples for an adequate evaluation of the model. Finally, as we have a new sample division, it was also necessary to normalize and serialize the samples to be compatible with the ST-GCN. The resulting subset is also public available[6] and named **ASLLVD-Skeleton-20**.

Table I
SELECTED SIGNS FOR THE EXPERIMENTS IN [4].

Dataset	Selected signs
ASLLVD	adopt, again, all, awkward, baseball, behavior, can, chat, cheap, cheat, church, coat, conflict, court, deposit, depressed, doctor, don't want, dress, enough

Due to the characteristics of this small dataset, the size of the batch was set to 8 samples after previous experiments. In the same way, as in the original implementation of the ST-GCN training algorithm, we used as optimizer the Stochastic Gradient Descent (SGD) with Nesterov Momentum.

For the learning rate, a decay strategy was adopted, which consists of initializing it with a higher value and gradually reducing it in the later epochs of the learning process to allow for more and more refined adjustments of the weights, as in [7]. Thus, in the experiment with the 20 selected signs, the total number of epochs was 200, adopting an initial rate of 0.01, which was decreased to the values of 0.001, 0.0001 and 0.00001 after the end of the epochs 50, 100 and 150, respectively.

In addition, an experiment with the complete ASLLVD dataset was also conducted to establish a reference value. In this experiment, only the batch size was changed. The batch size for this scenario with 2,745 signs was 24, and the learning rate had similar behavior to the experiment above.

The obtained results in both scenarios are presented in the following section, and the pre-trained models are available along with the adaptations made in the ST-GCN, according to section IV.

VI. Results

The first experiment was performed using the approach presented in [4], which considers the selection of 20 specific signs of the ASLLVD, whose performance is represented in Figure 14. The red line presents the accuracy of the model (*top-1*) and its evolution throughout training epochs. The gray line represents the *top-5* accuracy, which corresponds to the accuracy based on the 5 most likely responses presented by the model. Finally, the blue dashed line represents the evolution of the learning rate used in the respective epochs and its decay behavior.

It can be observed from the image that the model was able to achieve an accuracy of 56.82% from the epoch 80 in sign recognition. The *top-5* accuracy, in turn, was able to reach 95.45%. This performance was superior to the results presented by traditional techniques such as MEI, MHI, and PCA, but was not able to overcome that obtained by the HOF and BHOF techniques [4]. Table II presents the comparison of these results.

To establish a reference with the complete ASLLVD dataset, a second experiment was performed using its 2,745 signs. In this scenario, an accuracy (*top-1*) of 20.85% and a *top-5* accuracy of approximately 40.15% was obtained. Of course, this is a much more challenging task than the one proposed in [4], and these results reflect this complexity.

[6] Available at http://www.cin.ufpe.br/~cca5/asllvd-skeleton-20

Table II
SIGN RECOGNITION ACCURACY USING DIFFERENT APPROACHES AS PROPOSED IN [4].

	Accuracy (%)
MHI	10.00
MEI	25.00
PCA	45.00
ST-GCN SL	**56.82**
HOF	70.00
BHOF	85.00

From the table, we can see that the approach presented in this paper, based on graphs of the coordinates of human articulations, has not yet been able to provide such remarkable results as that based on the description of the individual movement of the hands through histograms adopted by BHOF. Indeed, the application of consecutive steps for optical flow extraction, color map creation, block segmentation and generation of histograms from them were able to ensure that more enhanced features about the hand movements were extracted favoring its sign recognition performance. This technique is derived from HOF and differs only by the approach of focusing on the hands of individuals while calculating the optical flow histogram.

Methods such as MEI and MHI, however, present more primitive approaches, which mainly detect the movements and their intensity from the difference between the consecutive frames of actions. They are not able to differentiate individuals or to focus on specific parts of their body, causing movements of any nature considered equivalent. The PCA, in turn, adds the ability to reduce the dimensionality of the components based on the identification of those with greater variance and that, consequently, are more relevant for the detection of movement in the frames.

VII. FINAL REMARKS

The results obtained with the presented approach did not reach such expressive performance as those obtained by some of the comparative techniques. However, its contributions are very important in guiding the next steps to be taken by future studies in this field.

Based on what is observed in this paper, it is relevant to seek approaches capable of enriching the information about the movements of the estimated coordinates, especially those of the hands and fingers, which, although very subtle, play a central role in the articulation and meaning of the signs.

This may be related, for example, to the definition of a new partitioning strategy that would allow more emphasis on the subtle traces of hands and fingers to the detriment of the other parts of the body. In addition, the definition of specific weights for these parts would enable the model to learn more about its dynamics; today the model does not distinguish the type of joint being learned. Finally, to include the depth information in the coordinates can provide the model with more details about the trajectory of the movement of these parts, enabling it to observe them through the three-dimensional plane.

REFERENCES

[1] M. C. d. C. Pereira, D. Choi, M. I. Vieira, P. Gaspar, and R. Nakasato, *Libras - Conhecimento Além Dos Sinais*, 1st ed. São Paulo: Pearson, 2011.

[2] W. H. Organization. (2018, mar) Deafness and hearing loss. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[3] S. M. Peres, F. C. Flores, D. Veronez, and C. J. M. Olguin, "Libras signals recognition: a study with learning vector quantization and bit signature," in *2006 Ninth Brazilian Symposium on Neural Networks (SBRN'06)*, Oct 2006, pp. 119–124.

[4] K. M. Lim, A. W. Tan, and S. C. Tan, "Block-based histogram of optical flow for isolated sign language recognition," *Journal of Visual Communication and Image Representation*, vol. 40, pp. 538 – 545, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320316301468

[5] L. Zheng, B. Liang, and A. Jiang, "Recent advances of deep learning for sign language recognition," in *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Nov 2017, pp. 1–7.

[6] D. Konstantinidis, K. Dimitropoulos, and P. Daras, "Sign language recognition based on hand and body skeletal data," in *2018 - 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, June 2018, pp. 1–4.

[7] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *CoRR*, vol. abs/1801.07455, 2018. [Online]. Available: http://arxiv.org/abs/1801.07455

[8] M. F. Tolba and A. S. Elons, "Recent developments in sign language recognition systems," in *2013 8th International Conference on Computer Engineering Systems (ICCES)*, Nov 2013, pp. xxxvi–xlii.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, June 2005, pp. 886–893 vol. 1.

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[12] S. S. Shanta, S. T. Anwar, and M. R. Kabir, "Bangla sign language detection using sift and cnn," in *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, July 2018, pp. 1–6.

[13] Y. Ji, S. Kim, and K. Lee, "Sign language learning system with image sampling and convolutional neural network," in *2017 First IEEE International Conference on Robotic Computing (IRC)*, April 2017, pp. 371–375.

[14] M. Taskiran, M. Kıllıoğlu, and N. Kahraman, "A real-time system for recognition of american sign language by using deep learning," 07 2018.

[15] G. A. Rao, K. Syamala, P. V. V. Kishore, and A. S. C. S. Sastry, "Deep convolutional neural networks for sign language recognition," in *2018 Conference on Signal Processing And Communication Engineering Systems (SPACES)*, Jan 2018, pp. 194–197.

[16] M. ElBadawy, A. S. Elons, H. A. Shedeed, and M. F. Tolba, "Arabic sign language recognition with 3d convolutional neural networks," in *2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS)*, Dec 2017, pp. 66–71.

[17] A. Das, S. Gawde, K. Suratwala, and D. Kalbande, "Sign language recognition using deep learning on custom processed static gesture images," in *2018 International Conference on Smart City and Emerging Technology (ICSCET)*, Jan 2018, pp. 1–6.

[18] T. D. Sajanraj and M. Beena, "Indian sign language numeral recognition using region of interest convolutional neural network," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, April 2018, pp. 636–640.

[19] L. Pigou, M. V. Herreweghe, and J. Dambre, "Gesture and sign language recognition with temporal residual networks," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017, pp. 3086–3093.

[20] R. M. d. Quadros and L. B. Karnopp, *Língua de sinais brasileira: estudos linguísticos*. Porto Alegre: Artmed, 2004, vol. 1.

[21] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 1290–1297.

[22] B. Fernando, E. Gavves, M. J. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 5378–5387.

[23] CMU Perceptual Computing Lab. (2018) OpenPose demo - output. [Online]. Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose/blob/master/doc/output.md

[24] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[25] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR*, 2017.

[26] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[27] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali, "The american sign language lexicon video dataset," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 06 2008, pp. 1–8.

[28] C. Neidle, A. Thangali, and S. Sclaroff, "Challenges in development of the american sign language lexicon video dataset (ASLLVD) corpus," in *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012, Istanbul, Turkey*, 05 2012. [Online]. Available: https://open.bu.edu/handle/2144/31899

[29] C. Vogler and C. Neidle, "A new web interface to facilitate access to corpora: development of the ASLLRP data access interface," in *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012, Istanbul, Turkey*, 05 2012. [Online]. Available: https://open.bu.edu/handle/2144/31886

[30] C. Valli and G. University, *The Gallaudet Dictionary of American Sign Language*, ser. The Gallaudet Dictionary of American Sign Language. Gallaudet University Press, 2005, no. v. 1.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[32] G. Rossum, "Python reference manual," CWI (Centre for Mathematics and Computer Science), Amsterdam, The Netherlands, The Netherlands, Tech. Rep., 1995.

[33] R. V. Babu and K. Ramakrishnan, "Recognition of human actions using motion history information extracted from the compressed video," *Image and Vision Computing*, vol. 22, no. 8, pp. 597 – 607, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885603002312

[34] P. Dreuw, D. Stein, T. Deselaers, D. Rybach, M. Zahedi, J. Bungeroth, and H. Ney, "Spoken language processing techniques for sign language recognition and translation," *Technology and Disability*, vol. 20, 04 2012.