ICANN19
28th International Conference on Artificial Neural Networks

Spatial-Temporal Graph Convolutional Networks for Sign Language Recognition

Cleison Correia de Amorim[1], David Macêdo[1], Cleber Zanchettin[1]
[1] Universidade Federal de Pernambuco
Centro de Informática
cca5@cin.ufpe.br, dlm@cin.ufpe.br, cz@cin.ufpe.br

INTRODUCTION

The recognition of sign languages obtained significant progress in recent years, which is motivated mainly by the advent of advanced sensors, new machine learning techniques, and more powerful hardware [1]. 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 and replaced by new approaches using conventional cameras and computer vision techniques.

Convolutional Neural Networks (CNN), as in many computer vision applications, obtained remarkable results in this field with accuracy reached 90% depending on the dataset.

However, a large portion of these studies addresses static signs or single-letter images, from the dactylology. The problem is the negative effect on the intrinsic dynamics of the language, such as its movements, non-manual expressions, and articulations between parts of the body. In this sense, it is extremely relevant that new studies observe such important characteristics.

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) and was introduced in [2]. 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.

Additionally, we present a new dataset of human skeletons for sign language based on ASLLVD to contribute to future related studies.

ST-GCN

The **Spatial-Temporal Graph Convolutional Network** 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 1 shows this structure, where each node corresponds to an articulation point. 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 [2].

To learn the temporal dimension, the ST-GCN extends the concept of graph convolution, considering this dimension as a sequence of skeletons graphs stacked consecutively, as in Figure 1.

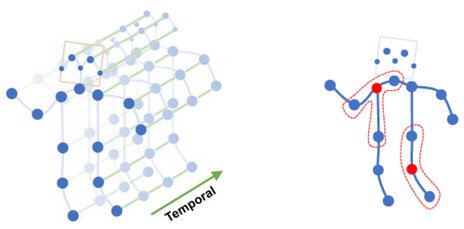

Figure 1: Sequence of skeleton graphs, denoting human movement in space and time (left). Sampling strategy in a convolution layer for a single frame (right) [2].

Figure 2 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 levels of feature maps for the presented graphs. Finally, they are submitted to a classifier to identify the corresponding action.

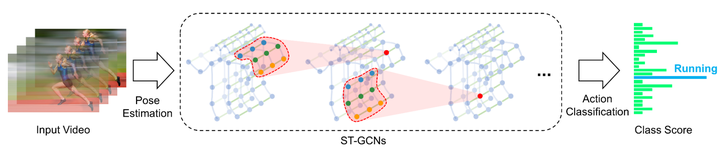

Figure 2: Overview of the ST-GCN approach [2, p. 3].

ASLLVD-SKELETON

We introduce a new dataset of human skeletons for sign language based on the American Sign Language Lexicon Video Dataset (ASLLVD) [3, 4].

Figure 4 shows a series of preprocessing steps to make the ASLLVD samples compatible with the ST-GCN model input. These steps, in turn, gave rise to a new dataset consisting of the skeletal estimates for all the signs contained therein, which was named ASLLVD-Skeleton. Figure 3 shows a series of preprocessing steps to make the ASLLVD samples compatible with the ST-GCN model input.

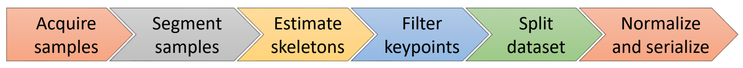

Figure 3: Preprocessing steps for creating the ASLLVD-Skeleton dataset.

- **Acquire samples**: obtain videos from ASLLVD dataset. We consider only the videos captured by the frontal camera;
- **Segment samples**: generate a video sample for each sign. The output of this step consists of small labeled videos with a few seconds;
- **Estimate skeletons**: 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. We used OpenPose [5, 6, 7] in this process;
- **Filter keypoints**: 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 4;
- **Split dataset**: divide the dataset into smaller subsets for training (80%) and test (20%);
- **Normalize and serialize**: make the samples compatible with the ST-GCN input.

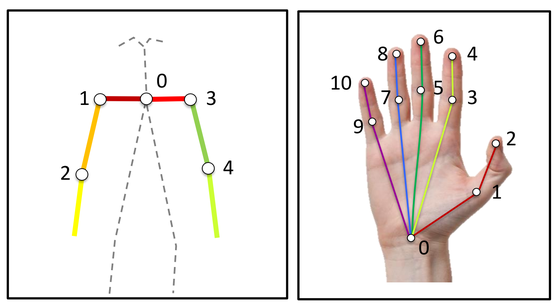

Figure 4: Representation of the 27 used key points, which 5 refer to the shoulders and arms (left) and 11 refer to each hand (right).

EXPERIMENTS

We used as reference the experiments proposed in [8], which evaluate the performance of models as the Block-Based Histogram of Optical Flow (BHOF) and popular techniques such as Motion Energy Image (MEI), Motion History Image (MHI), Principal Component Analysis (PCA), and Histogram of Optical Flow (HOF) in the ASLLVD dataset.

The authors used a subset containing 20 signs selected from the ASLLVD, as presented in Table 1. To reproduce this configuration, we selected the estimated skeletons for these signs from the ASLLVD-Skeleton dataset.

Table 1: Selected signs for the experiments in [8].

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 and based on preliminary experiments, the size of the used batch is of 8 samples. 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 [20]. 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.

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

RESULTS

Figure 5 presents the first experiment using the approach presented in [8]. 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 five 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.

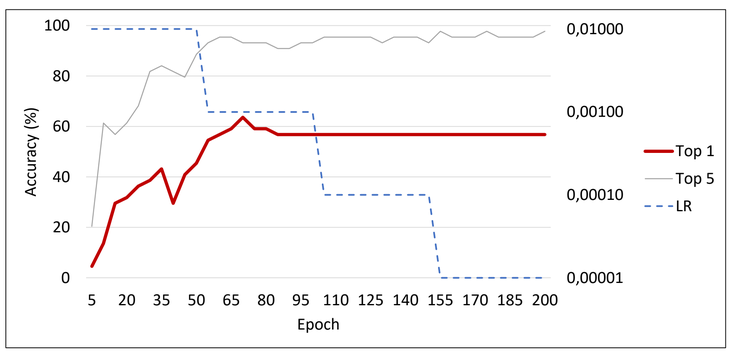

Figure 5: Accuracy obtained by the presented approach in recognition of the 20 signs selected from the ASLLVD.

We can observe 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 [8]. Table 2 presents the comparison of these results.

Table 2: Sign recognition accuracy using different approaches as proposed in [8].

	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.

The ASLLVD-Skeleton database, in turn, presented high relevance and was able to comply with its purpose of making feasible and supporting the development of this work. Through its approach and format, it allowed the benefits of adopting a robust and assertive technique to read human skeleton without restrictions or performance penalty during experiments. Considering that such techniques usually have a high computational cost, especially when combined with complex deep learning models, this is a factor that commonly restricts or impedes the progress of researches that seeks to evolve in this direction, based on the coordinates of the body. Thus, this paper contributes and extends to the academic community the benefits found here, encouraging and enabling other advances in sign language recognition.

REFERENCES

[1] Zheng, L., Liang, B., Jiang, A.: Recent advances of deep learning for sign language recognition. In: 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 1–7, November 2017.
[2] Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. CoRR abs/1801.07455 (2018).
[3] Athitsos, V., et al.: The American sign language lexicon video dataset. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–8, June 2008.
[4] Neidle, C., Thangali, A., Sclaroff, S.: 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, May 2012.
[5] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: CVPR (2017).
[6] Simon, T., Joo, H., Matthews, I.A., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping (2017).
[7] Wei, S., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4732, June 2016.
[8] Lim, K.M., Tan, A.W., Tan, S.C.: Block-based histogram of optical flow for isolated sign language recognition. J. Vis. Commun. Image Represent. 40, 538–545 (2016).

Manuscript ID: 368