ISSN 1870-4069

Sign Language Recognition through Manual and Non-Manual Features

Daniel Sánchez-Ruiz, José Arturo Olvera-López,
Iván Olmos-Pineda

Benemérita Universidad de Puebla,
Facultad de Ciencias de la Computación,
Mexico

daniel.sanchezruiz@viep.com.mx, jose.olvera@correo.buap.mx,
iolmos@cs.buap.mx

Abstract. Deaf community uses sign language as its main form of communication; however, most of the speaking community does not know how to understand that language, therefore the sign language recognition through technological developments has been an area of great interest for years. In this work, a proposal for this problem is presented, where regions of interest detection, manual and non-manual features extraction are carried out and for the recognition some BiLSTM networks with different architectures are used. The results obtained are an 73.99% accuracy, which are promising for the upcoming experiments. Finally, various actions are presented with the aim of improving the results as future work.

Keywords: Sign language recognition, computer vision, pattern recognition.

1 Introduction

People are considered to have a hearing loss when they are not able to hear under a hearing threshold of 25dB or less in both ears. Around 430 million people worldwide have disabling hearing loss, and it is estimated that by 2050 over 700 million people will suffer this kind of disability [1].

Hearing loss is one of the most common chronic impairments that appear with age as degeneration of sensory cells. It results from different congenital or acquired causes (e.g.: genetic causes, complications at birth, infectious diseases, exposure to excessive noise, among others).

Sign languages are classified as natural languages [2], which are used by the deaf community as their principal way of communication. Sign languages' visual, spatial nature and their variability, present a considerable research problem to be solved through technological developments.

Numerous areas are involved, such as linguistics, medicine, machine learning, computer vision, natural language processing, and computer graphics. Sign Language Recognition (SLR) is the scientific area responsible for capturing and translating sign speech using computer vision and artificial intelligence techniques [3]. Considering the

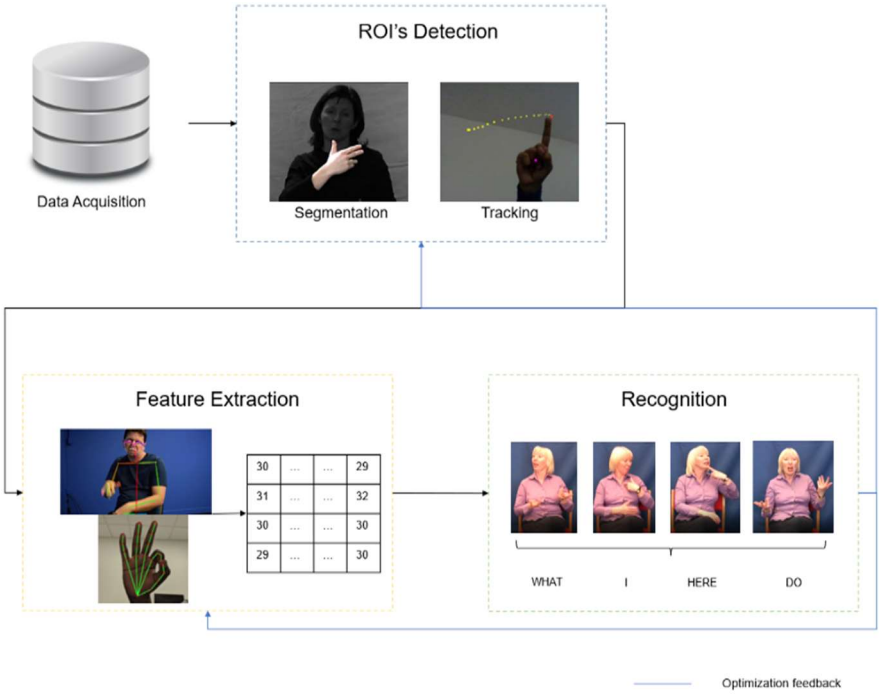


Fig. 1. Diagram of the proposed methodology for sign language recognition.

importance of sign languages for the communication of millions of people across the world and the rapid technological developments, this work proposes a methodology for sign language recognition employing several features.

The main proposal regarding the related work is to extract non-manual features based on the estimation of gaze and head pose along other well studied features; these two descriptors have not been studied thoughtfully, as can be seen in some works [4]. Another difference with recent works is the region of interest detections unlike the use of deep learning techniques as feature extractor, where is common to discard it, this is done to avoid adding unnecessary noise in the recognition phase.

This paper is organized as follows: in section 2 related works are analyzed and presented; section 3 is focused on the proposed methodology; section 4 describes the experiments that were designed and the obtained results and finally, in section 5 conclusions and future work are listed.

2 Related Work

SLR has been a research area very active since 90's [5], but recently important advances have been reported. At the beginning most of the studies focused in the used of gloves or haptic sensors to segment and track hands [6–8]. Nonetheless, deaf community felt very intrusive these types of methods because it can create practical difficulties in daily life and often limit their movements. By all these reasons, recent works are mainly

Fig. 2. Frames of example from the LIBRAS dataset.

focused on solutions based on computer vision, where the only necessary equipment are cameras.

As in any pattern recognition system, features extraction is an important stage. In SLR systems using cameras as input capture devices depend on the computer vision and image processing techniques. Common tasks performed are hand shape estimation, gesture segmentation, contour and boundary modeling, or color and motion cue identification.

Sign Languages have two types of features: manual and non-manual. Manual features consist in spatial and temporal descriptors base in the hand region; shape, position and motion are the most employed. As its name suggests non-manual features are all the cues related with the rest of the body.

They contain relevant information, which helps to recognize sign gestures with better accuracy. Non-manual features convey semantic or lexical properties, but also syntactical and grammatical functions, e.g.: negation, clausal type, question status, topics, or emphasis.

Several non-manual features have been studied, the principal are facial expressions and body pose. The SLR research can be classified in two principal types of investigation: isolated (ISLR) and continuous sign language recognition (CSLR). ISLR involves the recognition of a letter or a word at a time.

Some ISLR approaches employ Leap Motion (LM) sensors to recognize isolated words [9, 10]. In the former approach, they use fingertip information and their correlation, then a Support Vector Machine (SVM) [28] is used as recognition method to get an accuracy of 91.28% for ten ASL digits, whereas in the latter approach, they used extracts of 3D information and again a SVM. This system shows the best accuracy of 96.50% for ten ASL digits.

Kumar et al. [11] recognized 50 isolated signs using Kinect and LM sensors, an accuracy of 40.23% for all sign gestures are the results obtained. The principal disadvantage is that only manual features are considered. Ibrahim et al. [12] recognized 30 isolated Arabic signs and obtained an accuracy of 97%. However, to be implemented in real-time continuous sign sentences, more experiments in bigger vocabulary needs to be addressed.

CSLR concerns in recognizing one or more complete sentences. CSLR is more challenging than ISRL; problems of occlusion, alignment, or sign gestures identification in respect of transition movements are some of the difficulties that need to be considered.

It is difficult to recognize the transition movements because they are very subtle between all the different signs in data for CSLR, for this reason it is a topic of relevant interest in the research community. Common approaches to solve or mitigate this problem is eliminate epentheses movements (transition movements) by explicit modeling, implicit modeling or simply ignoring the transition movements.

Kong and Ranganath [13] presented a probabilistic approach based on the design and recognition of sign sub-segments and produced an 81.6% accuracy, the drawback is that movement epentheses are labeled manually. Li et al. [14] presented a scalable approach ignoring transition movements, the proposal gives an accuracy of around 87%. The inconveniences of the proposed approach are the use of a small vocabulary and its computer's execution time, which is considerable.

Elakkiya and Selvamani [15] proposed an automatic sign language classification, where they break down signs into subunits without any prior knowledge about the gestures. A Bayesian parallel hidden Markov model is used, its function is to combine manual and non-manual subunit features, but besides that it also handles the problem of movement ambiguities. An 82.1% of accuracy with signer independence was obtained as a result.

3 Methodology

The stages of the proposed methodology are depicted on the diagram in Figure 1. The first step consists of obtaining the input data to be used, the second and third steps are related to region of interest (ROI) detection and feature extraction, respectively; finally, the recognition task occurs. In the following subsections each one of these steps are described in detail.

3.1 Data Acquisition

LIBRAS (Brazilian Sign Language) dataset [16] was used for the experimental part, in particular Florianópolis' data, which contains 639 records, all the videos have a resolution of 640x414 pixels, with a refresh rate of 30 frames per second; besides that, an EAF file for the annotation of the signs is incorporated to each record. The topics that are covered in the dataset are dates, fruits, numbers, literature, interviews among others.

In every record two persons are present in the room having a conversation, which makes this dataset of continuous type. Four videos were recorded, one with an aerial angle, one with a lateral view of both persons and the last two with a direct view of each person.

For the problem the latter are the most suitable to use, however, as this dataset was not thought for SLR systems, in some records only one of the persons is gesticulating signs while the other is only watching, for this reason some videos are discarded. Figure 2 shows a couple of frames from one of the videos as an example.

3.2 ROI's detection

Based on sign language grammar [5], the regions of interest that were defined are the hands in order to extract manual descriptors and the body and the face in order to extract

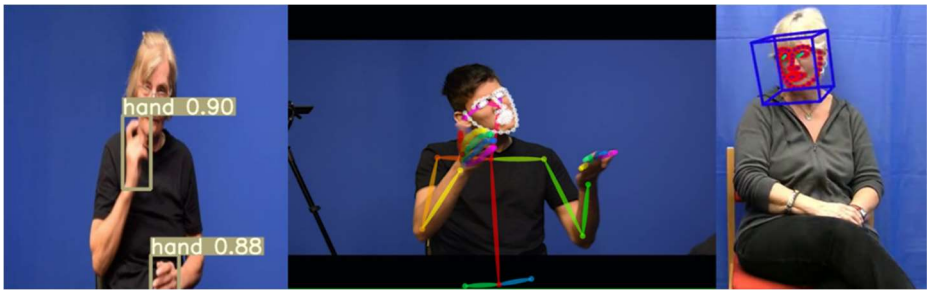


Fig. 3. Examples of ROI's detection through YOLOv5, OpenPose and OpenFace.

non-manual descriptors. For the task of detecting the hands' ROI, YOLOv5 system [17] was used, which is the state of the art in the object detection task.

However, since the features that are present in the data are very specific (deformations and occlusions), a training process from scratch with a manual annotated subset of LIBRAS images is performed to obtain a custom model.

For the body posture estimation OpenPose was used [18], which in addition to estimating the key points referring to the body joints, also brings the possibility of estimating key points related to the regions of the hand and face. Additionally, OpenFace [19] was also used for obtaining more characteristics related to the face and head. In the Figure 3 some examples of the obtained results of this stage are shown.

3.3 Feature Extraction

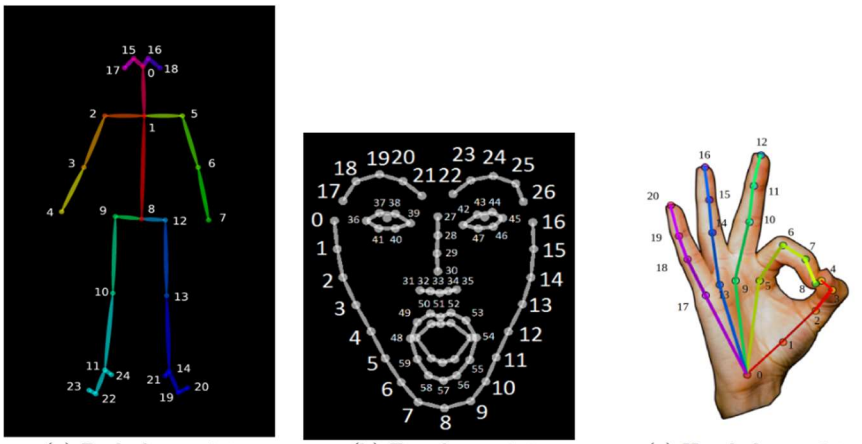
Aloysius and Geetha [20] stated that with approaches based on deep learning, such as the use of convolutional neural networks as feature extractors, it is no longer necessary to do ROIs detection and feature extraction locally.

Although the results have improved considerably, in most of these works in the input images irrelevant information is not previously discarded (context or even parts of the body such as legs that are not necessary). Furthermore, deep learning-based approaches work best with large amounts of data, which is not the case in most of the existing datasets.

For these reasons, the detection of regions of interest and the local extraction of descriptors based on manual and non-manual characteristics were proposed. As Koller [4] describes, several works have studied the relevance in the use of descriptors based on manual features (hands).

However, non-manual features (body and head) are also important in sign languages, and as Koller showed, they have been less explored, those related to the position of the head, or the direction of gaze have not been explored to the best of the knowledge of the authors. Taking this into account, the following features are extracted.

- Coordinates (x,y) for each hand. This relative to the centroid of the envelope frame detected with YOLOv5.
- Approximate speed for each hand. Tracking the change of the centroids' position every two distinct intervals of time (every 3 frames).



(a) Body key points (b) Face key points (c) Hands key points

Fig. 4. Key points obtained through OpenPose.

- Euclidean distance between selected key points related with facial expressions and hands. The OpenPose points considered are (0-9), (0-2), (17-20), (13-16), (9-12), (5-8), (2-4), (51,57), (48-54), (33-51), (19-37) and (24-44); they can be visualized in Figure 4a and Figure 4c.
- Key points coordinates (x, y) related to the arms. The points considered are (3, 4, 6, 7) and can also be seen in Figure 4b.
- Angular coordinates (x, y) of the gaze direction. Coordinates in radians averaged for both eyes and obtained with OpenFace.
- Rotation in radians around the X, Y, Z axes. Values obtained through OpenFace, which provides the head posture.

3.4 Sign Language Recognition

Since the problem to solve is a sequential problem, the recognition method to be used was a BiLSTM network, which have proven to be useful in several works [4, 20, 3]. Three architectures are proposed, the first one serves as a base to explain the last two and is shown in Figure 5, it is composed of an LSTM layer that is bidirectional, a fully connected layer and a softmax layer.

The second architecture has the BiLSTM layer, followed by a dropout layer, a ReLU layer, a fully connected layer and the softmax layer. Finally, the last architecture has the BiLSTM layer, a dropout layer, a fully connected layer that reduces the dimension of the features, followed by another fully connected layer and the softmax layer.

The classes that are going to be recognized are the written meaning of what is gestured, nonetheless, these annotated classes are not provided.

To deal with this issue, the instances generated within the defined window size (3 frames) are annotated with the corresponding class occupying the data provided in the EAF files associated to each record. Those instances that are not associated to any class

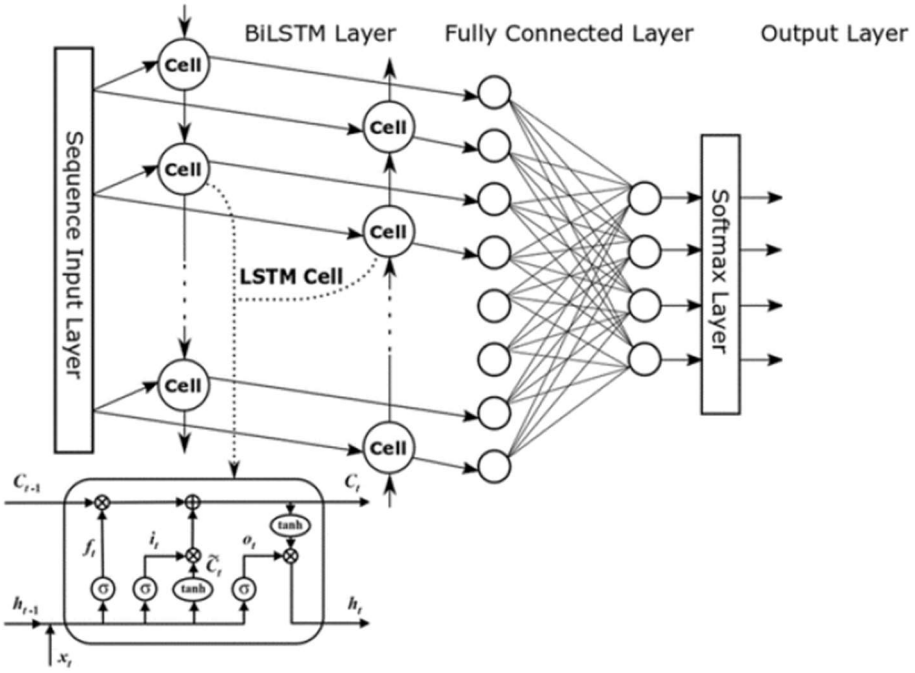


Fig. 5. Base architecture for the recognition task.

are labeled as *blank_transition*, these instances are concerned with transition movements or rest states.

4 Experimental Design and Results

For the experiments carried out, LIBRAS dataset is used. At the current stage of the investigation only 8 videos were used for the tests. In the training process, it was decided to use Google Colab. Available Colab hardware resource is a Tesla T4 graphics processing unit (GPU) that features 16GB RAM, 2,560 NVIDIA CUDA cores, and single-precision performance of 8.1 TFLOPS.

For training and evaluation processes, the data was divided into two sets: 70% for training and 30% for testing. Python language was used in the implementation of the proposal; PyTorch [27] was used for the BiLSTM networks.

The parameters that were defined for all the BiLSTM networks in training process are the number of epochs, which was set to 10 and which was defined empirically through experiments. Also, the batch size is set to 50; the number of cells in each hidden layer is 128; and the learning rate is 0.003.

The experiments were performed by each video using each one of the BiLSTM architectures that were presented in the previous chapter, at the end average accuracy (AvgAcc) and standard deviation (SD) were calculated; Table 1 shows the obtained results. At first glance it can be seen the accuracy is low, to improve the results and

Table 1. Obtained results by each BiLSTM architecture before (BPCA) and after (APCA) PCA process.

Architecture	AvgAcc (BPCA)	SD (BPCA)	AvgAcc (APCA	SD (APCA)
Base	64.79%	9.27%	73.99%	8.24%
Base+ReLU	63.51%	9.73%	72.09%	8.90%
Base+2FC	62.27%	9.46%	71.41%	8.99%

Table 2. Comparison of the proposed method with related work.

Author	Signing data	Accuracy
Amaral et al. [24]	Isolated	88.40%
Passos et al. [23]	Isolated	85.40%
Proposed Work	Continuous	73.99%

analyze the relevance of the features, Principal Components Analysis (PCA) [21] is applied before the recognition process.

PCA was implemented using scikit-learn framework, which gives the option to define the number of components or to define the percentage of the desire variability to preserve in the features. The latter was chosen, this approach used Minka's method [22] to automatically find the number of components, which in this case resulted to be 20. After this, the experiments were performed again and the obtained results showed an improvement, they are shown in the Table 1.

The best result was 73.99%, in order to compare it with related work, it was taken into consideration the review of Wadhawan and Kumar [25], where an extensive analysis by different sign languages of distinct countries was conducted.

This is done because LIBRAS dataset has not been occupied in another sign language recognition works to the best of the knowledge of the authors. Table 2 depicts the comparison of the best obtained result with other author's work, who used distinct datasets employing the same sign language.

Although they have better accuracies for word level recognition, the best result obtained is acceptable and it has the advantage that it was obtained by continuous signing data unlike the works, which have been described as a more complexed and challenging task in section 2.

Interestingly one of the features that it was preserved after the PCA step is the head pose, showing that this feature contributes relevant information. Another relevant finding that it was made is that some classes have between one or three instances and other have more than one hundred, so to improve the results it must be analyzed if data augmentation or imbalance data techniques might help to obtain a more robust recognition model.

Also, as it can be appreciated SD is still high, this needs to be addressed as future work; different difficulty in the sequences or instability in the training process could be the explanation of this behavior. Finally, the difference between the three BiLSTM architectures were so close, this behavior must be analyzed in depth in order to find out why the best result was obtained with the base network.

5 Conclusions and Future Work

In the present work, a proposal for sign language recognition using manual and non-manual features was conveyed. The descriptors were extracted locally to avoid add unnecessary noise in the recognition process, in addition, the relevance of descriptors such as head posture and gaze direction, which have not been used, was analyzed.

The results obtained from the designed experiments are promising, especially if is considered that the dataset used was not acquired and designed for the purpose of sign language recognition. As future work there are several possible actions to be carried out, the first could be to increase the data through data augmentation techniques to obtain a recognition model that has a greater number of instances on those classes that currently have few; in the same topic, the use of imbalance data techniques (subsampling) can also be a path to follow.

Besides that, the implementation of other recognition techniques that are suitable for the problem and that have shown good results, such as Connectionist Temporal Classification (CTC) [29] or Transformers [26], will be carried out. Finally, the design and execution of more experiments considering more data and benchmark datasets will be done to validate the results.

References

1. World health organization: Deafness and hearing loss (2022) www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss
2. Hockett, C. F.: The origin of speech. Scientific American, vol. 203, pp. 88–97 (1960)
3. Elakkiya, R.: 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 (2020) doi: 10.1007/s12652-020-02396-y
4. Koller, O.: Quantitative survey of the state of the art in sign language recognition (2020) doi: 10.48550/ARXIV.2008.09918
5. Ebrahim-Al-Ahdal, M., Nooritawati, M. T.: Review in sign language recognition systems. In: IEEE Symposium on Computers and Informatics (2012) doi: 10.1109/isci.2012.6222666
6. Fels, S. S., Hinton, G. E.: Glove-talk: a neural network interface between a data-glove and a speech synthesizer. IEEE Transactions on Neural Networks, vol. 4, no. 1, pp. 2–8 (1993) doi: 10.1109/72.182690
7. Grobel, K., Assan, M.: Isolated sign language recognition using hidden Markov models. In: IEEE International Conference on Systems, Man, and Cybernetics, Computational Cybernetics and Simulation, vol. 1, pp. 162–167 (1997) doi: 10.1109/ICSMC.1997.625742
8. Mehdi, S. A., Khan, Y. N.: Sign language recognition using sensor gloves. In: Proceedings of the 9th International Conference on Neural Information Processing, vol. 5, pp. 2204–2206 (2002) doi: 10.1109/ICONIP.2002.1201884
9. Marin, G., Dominio, F., Zanuttigh, P.: Hand gesture recognition with leap motion and kinect devices. In: IEEE International conference on image processing, pp. 1565–1569 (2014) doi: 10.1109/ICIP.2014.7025313
10. Marin, G., Dominio, F., Zanuttigh, P.: Hand gesture recognition with jointly calibrated leap motion and depth sensor. Multimedia Tools and Applications, vol. 75, pp. 14991–15015 (2016) doi: 10.1007/s11042-015-2451-6

11. Kumar, P., Gauba, H., Roy, P. P., Dogra, D. P.: A multimodal framework for sensor based sign language recognition. Neurocomputing, vol. 259, pp. 21–38 (2017) doi: 10.1016/j.neucom.2016.08.132
12. Ibrahim, N. B., Selim, M. M., Zayed, H. H.: An Automatic Arabic Sign Language Recognition System (ArSLRS). Journal of King Saud University - Computer and Information Sciences, vol. 30, no. 4, pp. 470-477 (2018) doi: 10.1016/j.jksuci.2017.09.007
13. Kong, W. W., Ranganath, S.: Towards subject independent continuous sign language recognition: A segment and merge approach. Pattern Recognition, vol. 47, no. 3, pp. 1294–1308 (2014) doi: 10.1016/j.patcog.2013.09.014
14. Li, K., Zhou, Z., Lee, C. H.: Sign transition modeling and a scalable solution to continuous sign language recognition for real-world applications. ACM Transactions on Accessible Computing, vol. 8, no. 2, pp. 1–23 (2016) doi: 10.1145/2850421
15. Elakkiya, R., Selvamani, K.: Extricating manual and non-manual features for subunit level medical sign modelling in automatic sign language classification and recognition. Journal of Medical Systems, vol. 41, no. 11 (2017) doi: 10.1007/s10916-017-0819-z
16. Quadros, R. M., Schmitt, D., Lohn, J., de Arantes Leite, T.: Corpus de libras (2020)
17. Jocher, G., Stoken, A., Borovec, J., Changyu, L., Hogan, A.: Ultralytics/yolov5: v4. 0-nn. silu () activations, weights and biases logging, pytorch hub integration. Zenodo (2021)
18. Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., Sheikh, Y.: OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172–186 (2021) doi: 10.1109/tpami.2019.2929257
19. Baltrusaitis, T., Zadeh, A., Lim, Y. C., Morency, L. P.: OpenFace 2.0: Facial behavior analysis toolkit. In: 13th IEEE International Conference on Automatic Face and Gesture Recognition (2018) doi: 10.1109/fg.2018.00019
20. Aloysius, N., Geetha, M.: Understanding vision-based continuous sign language recognition. Multimedia Tools and Applications, vol. 79, no. 31–32, pp. 22177–22209 (2020) doi: 10.1007/s11042-020-08961-z
21. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and Intelligent Laboratory Systems, vol. 2, no. 1–3, pp. 37–52 (1987) doi: 10.1016/01697439(87)80084-9
22. Minka, T.: Automatic choice of dimensionality for PCA. Advances in Neural Information Processing Systems 13 (2000)
23. Passos, W. L., Araujo, G. M., Gois, J. N., de Lima, A. A.: A gait energy image-based system for brazilian sign language recognition. In: IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 68, no. 11, pp. 4761–4771 (2021) doi: 10.1109/tcsi.2021.3091001
24. Amaral, L., Ferraz, V., Vieira, T., Vieira, T.: Skelibras: A large 2d skeleton dataset of dynamic brazilian signs. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Springer International Publishing, pp. 184–193 (2021) doi: 10.1007/978-3-030-93420-0_18
25. Wadhawan, A., Kumar, P.: Sign language recognition systems: A decade systematic literature review. Archives of Computational Methods in Engineering, vol. 28, no. 3, pp. 785–813 (2019) doi: 10.1007/s11831-019-09384-2
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Conference on Neural Information Processing Systems (2017)
27. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An imperative style, high-performance deep learning library. In: Proceedings of the 33rd

International Conference on Neural Information Processing Systems, no. 721, pp. 8026-8037 (2019) doi: 10.48550/arXiv.1912.01703
28. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning, vol. 20, no. 3, pp. 273–297 (1995) doi: 10.1007/bf00994018
29. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification. In: Proceedings of the 23rd International Conference on Machine Learning (2006) doi: 10.1145/1143844.1143891