Towards Portuguese Sign Language Identification Using Deep Learning

Ismael Costa, Domingos Martinho, and Ricardo Vardasca(✉)

ISLA Santarém, Largo Cândido dos Reis, 2000-241 Santarém, Portugal
{domingos.martinho,ricardo.vardasca}@islasantarem.pt

Abstract. In Portugal there are above 80,000 people with hearing impairment with the need to communicate through the sign language. Equal opportunities and social inclusion are the major concerns of the current society. It is aim of this research to create and evaluate a Deep Learning model that using a dataset with images of characters in Portuguese sign language can identify the gesture of a user, recognizing it. For model training, 5826 representative samples of the characters 'C', 'I', 'L', 'U' and 'Y' in Portuguese sign language. The Deep Learning model is based on a convolutional neural network. The model evaluated using the sample allowed for an accuracy of 98.5%, which is considered as a satisfactory result. However, there are two gaps: the existence of datasets with the totality of the alphabet in the Portuguese sign language and with the various representations of movement that each word has at the layout of letters. Using the proposed model with more complete datasets would allow to develop more inclusive user interfaces and equal opportunities for users with auditory difficulties.

Keywords: Deep learning · Inclusion user interfaces · Portuguese sign language

1 Introduction

One of the greatest current challenges is the digital transformation, which must be inclusive and must promote the integration of all individuals without any discrimination. In Portugal, there are about 83,000 people with reference to hearing loss (0.8% of the resident population) [1]. Hearing loss has a negative social and economic impact on people, families, and communities. Individuals with hearing difficulties or who have been deaf throughout life are distinct from congenital deaf people, as they have learned a language and can overcome their difficulty through video subtitles. Congenital deaf people have enormous difficulty in understand the written form, so any method-based text-only is inappropriate.

To help congenital deaf people, sign language was created and developed, the first proposal for a gesture dictionary corresponding to alphabet letters was proposed in XVII century for the Spanish language [2]. After which several dictionaries were created and developed. for different languages, where word creation is in addition to being letter-based, these are completed with hand movements and facial expressions.

T. Guarda et al. (Eds.): ARTIIS 2021, CCIS 1485, pp. 1–11, 2021.
https://doi.org/10.1007/978-3-030-90241-4_6

Among the existing sign languages, the most used dictionary worldwide is the American Sign Language (ASL), which is also the one that is more easily found in an image database that can be used for digital processing of sign language. The static example of the correspondence of the gesture to the letter in the ASL dictionary is shown in Fig. 1.

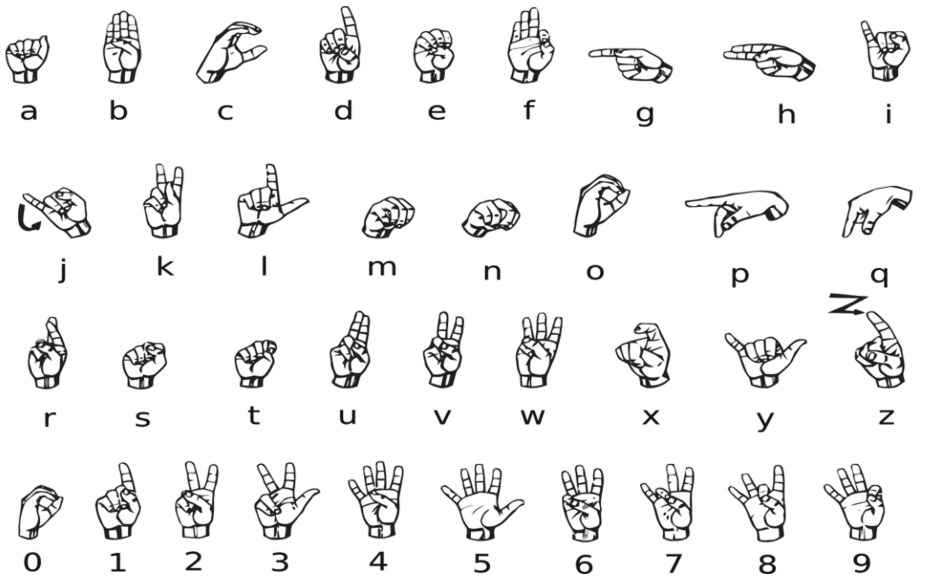

Fig. 1. The American sign language alphabet.

In the 19th century, by order of King João VI, the Institute for the Deaf and Blind (Instituto de Surdos-Mudos e Cegos) was founded in Casa Pia and the Swedish specialist Pär Aron Borg was invited to coordinate it. This was fundamental for the teaching of deaf people in Portugal and allowed them to communicate through an alphabet and a sign language of Swedish origin, a teaching method that was adopted in Portugal and had Borg as its creator [3]. Since 1997, the Portuguese Sign Language (PSL) has been enshrined in the Constitution of the Portuguese Republic, being one of the first countries to legislate a sign language [4]. The alphabet of the PSL is shown in Fig. 2.

It is aim of this research to create and evaluate a Deep Learning model that using a dataset with images of characters in Portuguese sign language can identify the gesture of a user, recognizing it.

The background related to this work is present in the next chapter. In Sect. 3 is presented the methodology used in development of the proposed solution, the results are described in Sect. 4, followed by the last chapter with the presentation of discussion, conclusion, and further work.

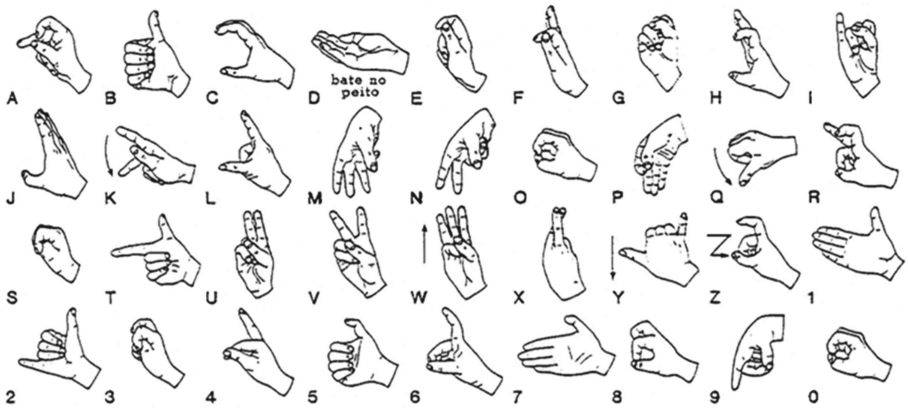

Fig. 2. The Portuguese sign language alphabet.

2 Background

The sign language gestures can be segmented in images using a global thresholding algorithm (in a binary image), infrared illumination based (camera provides a distance), software based (color separation), HSV (hue, saturation, lightness) brightness factor matching and Binary Large OBject method [5].

The feature extraction can be achieved through Edge Detection using HSV brightness matching, palm point detection, center of region detection, dimensionality reduction, scale-invariant feature transform technique and leap motion [5].

The classification techniques available for sign classification are Hidden Markov Models (HMM) (statistical model designed using Bayes network), Machine learning algorithms (Artificial Neural Network, k-Nearest Neighbors, Support Vector Machine) and deep learning techniques such as Convolutional Neural Network (CNN) [5].

Before the application of deep learning in sign language classification and detection it has also been performed through machine learning, which is less powerful than deep learning but faster.

The Portuguese sign language (PSL) signs were isolated and classified with machine learning methods using data gathered with a Microsoft Kinect and data gloves, the obtained accuracy results were 87.3% with Random Trees, 96.6% with Boost Cascade, 80.4% with Artificial Neural Networks, 98.2% with K-Nearest Neighbors, 96.8% with Naive Bayes and 100% with Support Vector Machine [6].

A dataset consisting of 2524 images with 70 images per category. Each of the 36 categories represented a different character of American Sign Language (ASL). All images were augmented to create a dataset of 14781 images, 75% were used for training and remaining 25% for test. The model of training was based in the VGG16 CNN architecture with 4 epochs, the average training accuracy was of 95.54%, being the digit 0 that who presented the worst accuracy. The validation accuracy after the 4 epochs was of 94.68 in predicting the given image gesture [7].

Previous approaches for classifying American sign language were: using real time video input with HMM and Euclidean distance achieved 90% accuracy [8], static images

input for Histogram of Orientation Gradient (HOG), Histogram of Boundary Description (HBD) and the Histogram of Edge Frequency (HOEF) with SVM reached 98.1% accuracy [9], real time video input for PCA and fuzzy logic obtained 91% accuracy [10] and real time video input for HMM and ANN with error back propagation algorithm achieved 91% accuracy [11].

With a dataset of 41258 training and 2728 testing samples. Each sample provides a RGB image (320 × 320 pixels), Depth map (320 × 320 pixels), Segmentation masks (320 × 320 pixels) for the classes: background, person, three classes for each finger and one for each palm, 21 Key points for each hand with their uv coordinates in the image frame. The pre-trained model based in a SqueezeNet model processed the images by dividing every pixel by 255 and then resizing the image to 244 × 244. The results of the squeezing process are concatenated and separated in two groups of 4 convulsions each being one single and other of 3 × 3 filters, being finally concatenated for the output. From the tests a maximum training accuracy of 87.47% is attained. The validation accuracy attained is 83.29% [12].

A model of CNN based in a convolutional layer of 16 filters (2 × 2) reducing spatial dimensions to 32 × 32 with added max pooling filters (increased to 5 × 5), a dropout rate of 20% and a SoftMax classifier for producing the output. All 1000 images of ASL of 26 characters and 10 digits were resized to 50 × 50 pixels and the total number of epochs used to train the network is 50 with a batch size of 500. For chars an accuracy of 90.04% was achieved in 4.31 s, characters 'A', 'C' and 'D' presented highest accuracy (100%), and 'Z' the lowest (67.78%). The digits obtained 93.44% accuracy in 3.93 s in real-time recognition, the best (100%) was reached by the digit '5' and the worst (83.33%) by the '8' digit [13].

Using the Modified National Institute of Standards and Technology database (MNIST) database of American Sign Language with 60000 images spread by 24-character classes, 'J' and 'Z' were excluded because they require motion. Principal Component Analysis (PCA) was used for sign identification with a custom CNN model with 11 layers: 4 Convolutional layers, 3 Pooling (Max) layers, 2 Dense Connected, 1 Flatten and 1 Dropout layer. A Rectified Linear Unit (ReLu) activation function was used to avoid negativities. The training dataset consists of 27455 images with 784 (28 × 28) features, on average each letter class has 1000 images, the validation set consists of 7172 images with 784 (28 × 28) features. A dropout layer with a given probability of 20% is included to avoid model overfitting as it drops out 20% of the hidden and visible units from these densely connected layers. The final training of model from scratch produces a considerably high level of 99% accurateness on the training set. The output is taken from the SoftMax Layer with 24 class classification. The model was trained to minimize loss by usage of cross entropy ADAM, for 10 epochs on a batch size of 128. The model was trained with a learning rate of 0.001 with 0 decay. The validation accuracy of the model is greater than 93% and only 6 epochs were required for a stable result. For further research it is suggested that the images should be segmented using the OpenCV library [14].

Two modified pre-trained AlexNet and VGG16 based CNN models have been proposed for gesture recognition and classification using a ASL dataset with 36 classes, characters and digits, 70 images per class (2520 images). The methodology consisted of

pre-processing (image resize, data augmentation, dataset split into training and test), feature extraction with a pre-trained CNN architecture and classification with SVM until the accuracy achieves the desired value. Using a 70% training and 30% test sets, the AlexNet showed troubles in achieving 100% accuracy with digits '0' and '6' and with 'E', 'M', 'O', 'U', 'W' and 'Z' characters. The VGG16 model also not classified 100% accurately the digits '0', '2', '6' and '9' and the chars 'I', 'K', 'M', 'S', 'T', 'W' and 'X'. the VGG16 model (99.82%) performed badly when compared with the AlexNet model (99.76%). In terms of training time the AlexNet model took 5.83 min when VGG16 model took 88 min [15].

A research on ASL recognition comparing the deep learning approaches AlexNet and model E on a dataset of 3000 × 29 100 × 100 pixel images of hand gesture language, found that the model E presented a training accuracy of 19.38% and validation accuracy of 30.94%, whereas the AlexNet model showed a training accuracy of 2.50% and validation accuracy of 3.28% with 50 epochs. Another important aspect was that the AlexNet requires 1.72 h while model E takes 0.71 h in training the model [16].

A total of 61.614 images were collected for 28 classes, comprised of 26 alphabets, including 'J' and 'Z', as well as two classes for space and delete. All the images were scaled to 224 × 224 pixels, and then, normalized to be fed to the VGG_Net architecture. A 70% training and 30% test sets were chosen. An accuracy of 98.53% was obtained for the training set, and 98.84% for the validation set in real-time. In terms of accuracy, the highest (99.95%) was obtained for the character 'L' and the lowest (97.31%) was obtained was for the character 'M' [17].

3 Methodology

From the related research, it can be observed that there is no freely available images dataset of the PSL, which has an impact on the development of research in the area and implementation of machine learning and deep learning solutions.

The development of a solution that allows the interpretation of sign language must be performed at several steps, initially each character and digit must be identified individually and statically, then the dynamic identification of characters that presupposes motion, and finally the recognition of the construction of words combined with a grammar and semantics.

Since there is no Portuguese sign language dataset, there was a need to adapt an existing dataset including the characters that are common to another sign language dictionary, in this case the American, which is the one that offered more freely accessible datasets.

As dataset for this research, an adaptation of the original dataset published on the Kaggle platform provided by "tecperson" [18] was used. The sign language dataset is presented in this work following the comma separated value (CSV) format, labeled with the written language characters and the monochromatic hue values of the pixels that make up the image. The standardized format of the classic MNIST dataset contains 28 × 28 pixel images with grayscale values 0–255.

From the original 24-character American Sign Language reference dataset only the characters common to PSL were considered (Fig. 3). For the learning and assessment

process, 5 labels (0–4) corresponding to the common characters between LGA and LGP were used: 0 – 'C', 1 – 'I', 2 – 'L', 3 – 'U' and 4 – 'Y'. The dataset considered for the 3 sign language symbols has 5826 samples.

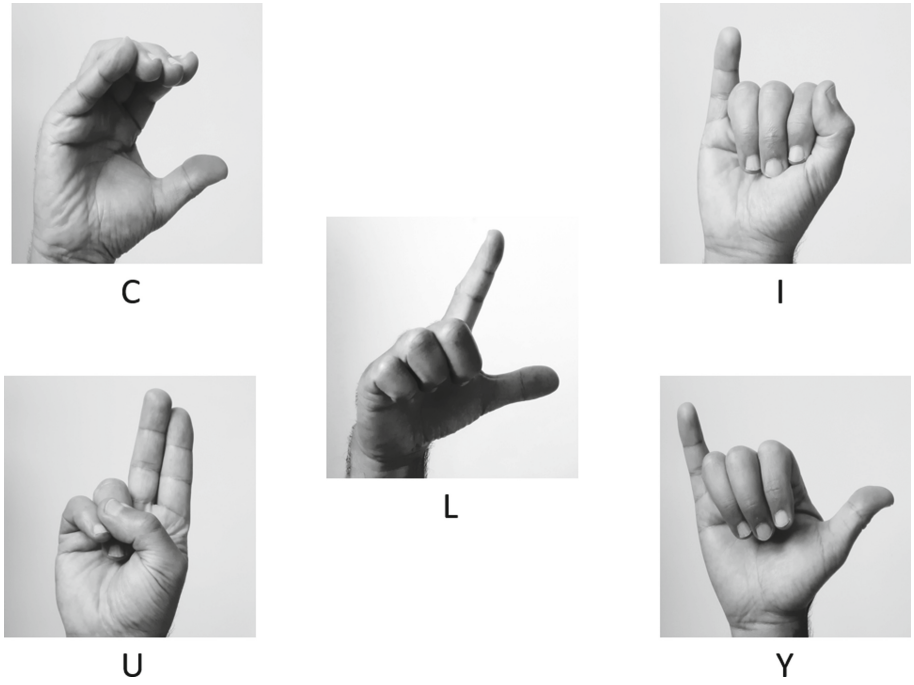

Fig. 3. Examples of images of selected characters for the dataset of this research.

3.1 Deep Learning Model

The algorithm of the deep learning classification model based on a CNN was carried out based on 2 different configurations and 2 evaluation methods. The code proposed in this work was adapted from the original code developed by Wan [19].

For the training operation of the algorithm, convulsion layers (Conv2d) were used, applying a 2D convolution over an input signal composed of several input layers. Subsequently, grouping layers (MaxPool2d) were applied by applying a maximum 2D grouping to an input signal composed of several input layers. Repeated new convulsive layers and then applied a linear transformation to the input data.

To optimize the model, in order to allow it to achieve the state of maximum precision, given the resource constraints, such as time, computing capacity, memory, etc., the stochastic gradient descent (SGD) optimization technique was used, which is made available through the "torch.optim" package.

The Cross-entropy loss technique was used in this work as a learning optimizer, this technique is very common in classification tasks both in Machine Learning and in Deep Learning [20].

The model will achieve a very satisfactory performance when the loss value (Cross entropy loss) is less than 0.02. When this value is less than 0.00, the algorithm will be in its perfect learning state.

The infrastructure of the learning model used in the implementation of this research is shown in Fig. 4. In that the features are the amount of grey value intensity, C – Convolutions, S – Subsampling.

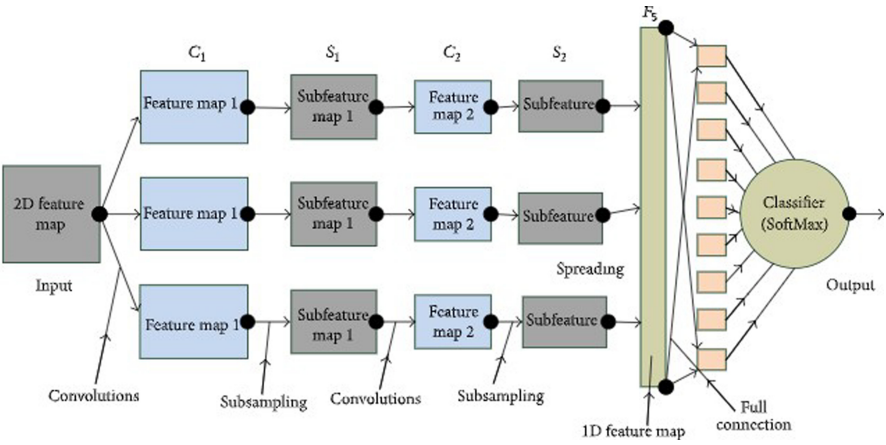

Fig. 4. The infrastructure of the classification model used in this research.

For the development and testing of the model, the programming language Python v.3.8 and the artificial vision library OpenCV v.4.4 were used.

A dataset composed of LGP images representative of the characters 'C', 'I', 'L', 'U' and Y" with a total of 5,826 samples was used to train the model. For this process, 12 iterations (epocs) were considered because this number of iterations presents a loss value = 0.0035 (loss) calculated by the average of 10 executions. As this value is less than 0.02, the performance of the model is considered very satisfactory.

For the model evaluation process, a set of 1,405 different samples from the samples used in the training process was considered. The performance evaluation of the model was carried out in two phases. All training samples (5,826 samples) were considered, and an evaluation performed using a pre-trained model available in the PyTorch library, a second evaluation performed based on the ONNX Runtime inference accelerator method [21].

In the last phase, the process of evaluating the model in real time is carried out by collecting the user's video image by the device's camera. In this real-time evaluation, the user must present a PSL gesture corresponding to one of the selected characters and check in the user interface whether the model is able to identify the correct character through the interpreted image. During the process of translating the PSL character read in real time, the model will search for a list of possible Portuguese characters. When the character is found in the list, it is displayed in the user interface.

4 Results

The verification of the number of executions necessary in training the model is shown in Fig. 5, 10 runs are satisfactory.

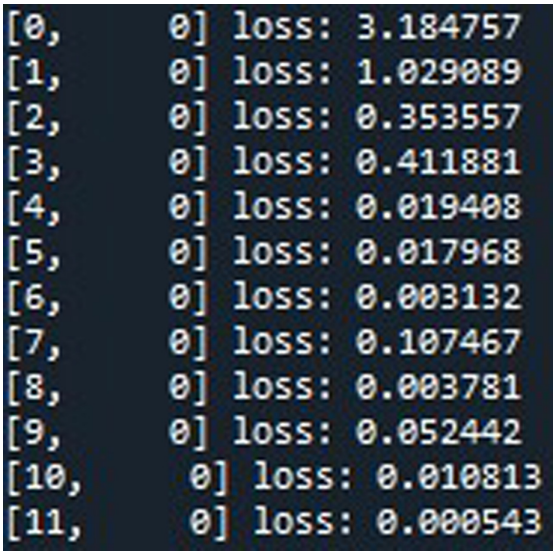

Fig. 5. Results of loss values (loss) at each iteration (epoch).

The accuracy of the training and validation of the model using the training and validation sets is shown in Fig. 6, the implementation using the PyTorch library presents a value more for training but lower for validation, however in both cases the accuracy is higher than 88.5%.

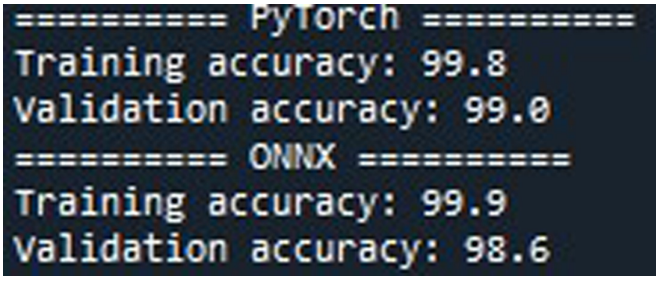

Fig. 6. Model evaluations with dataset using PyTorch and ONNX methods.

An example of the execution of a PSL gesture with recognition of the respective characters is shown in Figs. 7 and 8, the gesture corresponding to the characters ‘L’ and ‘C’ was correctly identified.

Fig. 7. Real-time PSL ‘L’ character recognition model evaluation user interface.

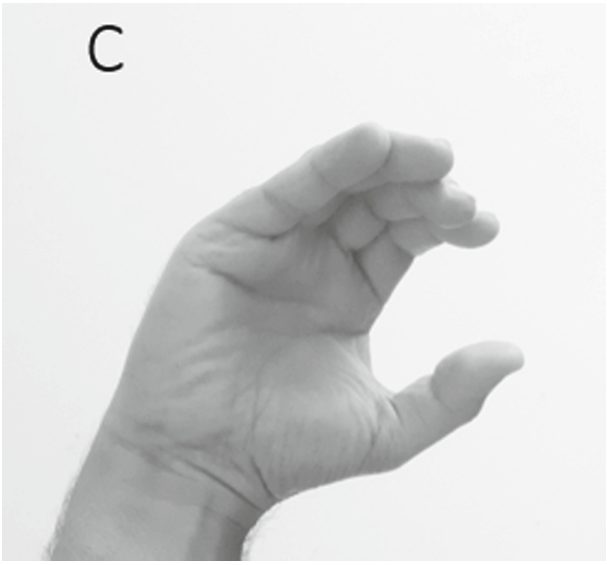

Fig. 8. Real-time PSL ‘C’ character recognition model evaluation user interface.

5 Discussion and Conclusion

The major barrier for the development of this research was the absence of a freely available PSL characters image dataset, to overcome this limitation a ASL characters image dataset [18] was used, selecting only the images that have correspondence between PSL and ASL.

This manuscript only covers the first phase in PSL recognition, it is the first step towards a more comprehensive investigation.

The PSL character classification results presented by this investigation using the developed deep learning model surpass most results presented by machine learning classifiers with data collected by smart gloves and Microsoft Kinetic system [6].

It is important to verify that the results of classification in training and validation presented by this simple and fast implementation surpass the results of other more complex implementations applied to ASL [11–13].

No entanto fica aquém dos resultados de outras implementações de mais complexas aplicadas à ASL usando VGG_Net e AlexNet [15–17], mas requer menor capacidade computacional.

For the best knowledge of the authors of this investigation, this is the first attempt to apply a low-complexity deep learning model to the identification of PSL characters.

It is concluded that the model demonstrated a high efficiency even considering the learning using a reduced number of samples. The proposed solution based on image collection using a commonly used webcam will be viable given the easy accessibility that this medium will have for most users.

For future work, it is proposed the development of a dataset that includes all PSL characters and digits, even those with motion, which will allow an evaluation of the model's effectiveness in translating the PSL more complete. The proposed deep learning model should also be tested for dynamic PSL words identification, even if for that purpose a grammar and lexicon are required to be developed.

This work was a first step towards the creation of fully integrative user interface for deaf people, which not only should be capable of recognize their communication gestures, as should provide them responses in the same language through the usage of an animated avatar also using a deep learning method, or even in a more advanced development be an important aid for PSL learning for subjects in need of it.

References

1. Gonçalves, C.: Enquadramento familiar das pessoas com deficiência: Uma análise exploratória dos resultados dos Censos 2001. Revista de Estudos Demográficos **33**, 69–94 (2003)
2. Bonet, J.P.: Reduction de las letras, y arte para enseñar a ablar los mudos. Por Francisco Abarca de Angulo (1930)
3. Olsson, C.G.: Omsorg och kontroll. En handikapphistorisk studie 1750–1930. Umeå: Umeå universitet (2010)
4. Correia, M.D.F.S., Coelho, O., Magalhães, A., Benvenuto, A.: Learning/teaching philosophy in sign language as a cultural issue. J. Educ Cult. Soci. **4**(1), 9–19 (2013)

5. Safeel, M., Sukumar, T., Shashank, K.S., Arman, M.D., Shashidhar, R., Puneeth, S.B.: Sign language recognition techniques-a review. In: 2020 IEEE International Conference for Innovation in Technology (INOCON), pp. 1–9 (2020)
6. Escudeiro, P., et al.: Virtual sign–a real time bidirectional translator of portuguese sign language. Procedia Comput. Sci. **67**, 252–262 (2015)
7. Masood, S., Thuwal, H.C., Srivastava, A.: American sign language character recognition using convolution neural network. In: Satapathy, S.C., Bhateja, V., Das, S. (eds.) Smart Computing and Informatics. SIST, vol. 78, pp. 403–412. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-5547-8_42
8. Nandy, A., Prasad, J.S., Mondal, S., Chakraborty, P., Nandi, G.C.: Recognition of isolated indian sign language gesture in real time. : Information Processing and Management, pp. 102–107 (2010). https://doi.org/10.1007/978-3-642-12214-9_18
9. Lilha, H., Shivmurthy, D.: Analysis of pixel level features in recognition of real life dual-handed sign language data set. In: Recent Trends in Information Systems (ReTIS), pp. 246–251 (2011)
10. Kishore, P.V.V., Kumar, P.R.: A video based indian sign language recognition system (INSLR) using wavelet transform and fuzzy logic. Int. J. Eng. Technol. **4**(5), 537 (2012)
11. Kishore, P.V.V., Prasad, M.V.D., Kumar, D.A., Sastry, A.S.C.S.: Optical flow hand tracking and active contour hand shape features for continuous sign language recognition with artificial neural networks. In: IEEE 6th International Conference on Advanced Computing (IACC), pp. 346–351 (2016)
12. Kasukurthi, N., Rokad, B., Bidani, S., Dennisan, D.: American sign language alphabet recognition using deep learning. CoRR (2019). https://arxiv.org/abs/1905.05487. Accessed 15 July 2021
13. Tolentino, L.K.S., Juan, R.O.S., Thio-ac, A.C., Pamahoy, M.A.B., Forteza, J.R.R., Garcia, X.J.O.: Static sign language recognition using deep learning. Int. J. Mach. Learn. Comput. **9**(6), 821–827 (2019)
14. Sabeenian, R.S., Sai Bharathwaj, S., Mohamed Aadhil, M.: Sign language recognition using deep learning and computer vision. J. Adv. Res. Dyn. Contr. Syst. **12**(05-Special Issue), 964–968 (2020)
15. Barbhuiya, A.A., Karsh, R.K., Jain, R.: CNN based feature extraction and classification for sign language. Multimedia Tools Appl. **80**(2), 3051–3069 (2020)
16. Pratama, Y., Marbun, E., Parapat, Y., Manullang, A.: Deep convolutional neural network for hand sign language recognition using model E. Bull. Electr. Eng. Inf. **9**(5), 1873–1881 (2020)
17. Kadhim, R.A., Khamees, M.: A real-time american sign language recognition system using convolutional neural network for real datasets. TEM J. **9**(3), 937 (2020)
18. Tecperson: Kaggle datasets: sign language MNIST. drop-in replacement for MNIST for hand gesture recognition tasks. https://www.kaggle.com/datamunge/sign-language-mnist. Accessed 15 July 2021
19. Wan, A.: How to build a neural network to translate sign language into english. https://github.com/alvinwan/sign-language-translator. Accessed 15 July 2021
20. Wang, B.: Loss functions in machine learning. https://medium.com/swlh/cross-entropy-loss-in-pytorch-c010faf97bab. Accessed 15 July 2021
21. Verucchi, M., et al.: A systematic assessment of embedded neural networks for object detection. In: 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), 1, pp. 937–944 (2020)